12-9-2021

Error Propagation and Algorithmic Design of Contour Integral Eigensolvers with Applications to Fiber Optics

Benjamin Quanah Parker
Portland State University

Recommended Citation

Parker, Benjamin Quanah, "Error Propagation and Algorithmic Design of Contour Integral Eigensolvers with Applications to Fiber Optics" (2021). *Dissertations and Theses.* Paper 5839.
https://doi.org/10.15760/etd.7710

Error Propagation and Algorithmic Design of Contour Integral Eigensolvers With Applications to Fiber Optics

by

Benjamin Quanah Parker

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

Doctor of Philosophy
in
Mathematical Sciences

Dissertation Committee:
Jay Gopalakrishnan, Chair
Jeffrey Ovall, Co-Chair
Bin Jiang
Erik Sánchez
Jacob Grosek

Portland State University
2021

Abstract

In this work, the finite element method and the FEAST eigensolver are used to explore applications in fiber optics. The present interest is in computing eigenfunctions u and propagation constants β satisfing the Helmholtz equation $\Delta u + k^2 n^2 u = \beta^2 u$. Here, k is the freespace wavenumber and n is a spatially varying coefficient function representing the refractive index of the underlying medium. Such a problem arises when attempting to compute confinement losses in optical fibers that guide laser light. In practice, this requires the computation of functions u referred to as guided modes and leaky modes. For guided modes, the location of the corresponding propagation constants in the complex plane is known in the optics literature, making the FEAST algorithm an ideal candidate to tackle this problem. In practice, one solves the Helmholtz equation by prescribing zero Dirichlet boundary conditions on a bounded, circular domain representing the fiber cross-section. In this work, we compute numerical solutions to this problem using the FEAST algorithm with the Discontinuous Petrov-Galerkin method for the underlying discretization. To compute leaky modes, we must find complex-valued propagation constants β with corresponding outgoing functions u. To compute such quantities, we employ a Perfectly Matched Layer (PML) to force exponential decay of the solutions, rendering the problem computationally tractable on a bounded domain. A frequency-dependent approach is taken for the PML: This transforms the weak formulation of the Helmholtz problem into a polynomial eigenvalue problem, thus motivating our the adaptation of the FEAST algorithm to solve such eigenvalue problems. We verify the results of our algorithm by finding leaky modes and propagation constants of a step-index fiber, where one can compare against known analytic solutions. Our algorithm is then applied to the task of computing confinement losses of a microstructure fiber, where lossy modes are expected. Confinement losses are computed from the imaginary parts

of the corresponding propagation constants, and quantify the loss of power as light travels through an optical fiber. Our numerical results show that in practice, there is a large preasymptotic regime prior to convergence of confinement losses, suggesting that confinement losses reported in the literature could be matched in this regime. In addition, we show that our computed confinement losses remain stable as the strength of the decay in the PML region and the PML region's size are varied. Our results also show that computed confinement losses are extremely sensitive to minor perturbations in fiber geometry.

Dedication

This work is dedicated to my mom, aunts, uncles, grandparents, great grandparents, ancestors, siblings, cousins, friends, and every graduate student in the struggle. You all give me the strength to keep going. Miigwech, hišəbaʔ, and hisḱʷuʔ.

Acknowledgments

As I conclude my degree program, I know that I would not have gotten to where I am without the support network I had coming to PSU, and the bigger one I developed during my time as a student. I want to take this time to acknowledge as many people as I can who I met, or with whom I worked, laughed, shared a drink, or just took part in grinding out our degrees over the years. Graduate school is a challenging place to be, and even harder in the midst of a global pandemic and times of profound loss. We have remember to be kind to ourselves, and extend that to one another.

In this spirit of this, I want to start by acknowledging I would not be where I am now without the help of Felix Osa. When the pandemic hit and the lockdowns took effect, we looked for as many ways as we could to survive while we waited in uncertainty as the world around us rapidly changed. We made our living room into a blanket fort, got a new pet, painted, gardened, cooked, binged new shows on whatever streaming service we could get our hands on, and really sought out whatever creative outlets we could find to take the edge off. In doing so, we helped one another, and I formed a really important, lasting friendship.

I also want to acknowledge Brittany Ellis, David Brown, Maricella Best-McKay, Tuyen Tran, David Hong, Tenchita Alzaga, Nathan Lawrence, Sam Reynolds, Chris Aagard, Robyn Reid, Jeff Kim, River Corbin, Zarek McGee, Sophie Kristensen, Jamie

Faue, Joseph Motto, Dow Drake, Tatha Goswami, Pieter Vandenberge, Julia Dancis, Bianca Lee, Tianyin Liu, Emily Leickly, James Delaney III, Lukas Kogler, and Rajesh Venkatachalapathy. During my time at Portland State University (PSU) and abroad, I got to know you all better through shared classes, shared offices, cramming for course examinations, hanging out as all of you prepared for comprehensive examinations or wrote your dissertations, house parties, driving to Cascade RAIN, singing our hearts out at karaoke, preparing for conferences and exploring Vienna, or just working through our personal struggles together (and afar) as we continued through the pandemic.

Another acknowledgement goes out to the native students and staff here at PSU, as well as my family in Portland, Washington, and North Dakota. This includes my cousins Sara and Justin Dauphinaus, my cousin Aja DeCoteau, my cousin in-law Raffaele Timarchi, my uncle Larry Dauphinais, my mom Jennifer DeCoteau, my grandparents, aunts, uncles, as well as Cante Nakanishi, Cory Cornelius, Brook Thompson, Matt Morseman, Naomi Rodriguez, Yolonda Salguiero, Tim Anderson, Robert Franklin, Alana Lamotte-St. John, and Trevino Brings Plenty. Whether it was running AISES before and during the pandemic, having a place to relax and take my mind off of the stresses of graduate school, having someone to talk to, or just hang out for a laugh, I could not have done what I have without you.

I would also like to acknowledge the support I have received from faculty members, collaborators, and my dissertation committee, including my advisors Jay Gopalakrishnan and Jeffrey Ovall, Bin Jiang, Erik Sánchez, and Jacob Grosek. I am appreciative of the support I have had from you all in courses, research, preparing for conferences, job hunting, and being able to travel across the world. In addition, I would

like to acknowledge the helpful conversations I have had with our collaborators Luka Grubišić and Jacob Grosek, as well as Professor Joachim Schöberl for hosting myself and Dow during the summer of 2019 for the NGSolve user meeting. I would also like to acknowledge Brittany Erickson for the help she provided me during her numerical analysis classes and the casual conversations we had about graduate school.

Finally, I would like to acknowledge the support received for the research I conducted over the course of my degree program. This work would not have been possible without support from the AFRL Cooperative Agreement #18RDCOR018, the NSF's Major Research Instrumentation grant DMS-1624776, AFOSR grant FA9550-19-1-0237, and NSF grant DMS-1912779.

This work and my survival in graduate school during a global pandemic would not have been possible without the help of everyone here, and even more people outside of PSU. Thank you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivating Problem

In this work, we are interested in finding the propagation constants β and outgoing solutions (or modes) u satisfying the following partial differential equation.

$$\Delta u + k^2 n^2 u = \beta^2 u \tag{1.1}$$

This particular form of the Helmholtz equation arises in applications of fiber optics [43], and as we see in the next section, is derived from Maxwell's equations under the assumption that such a medium is guiding laser light. We omit the treatment of boundary conditions for the moment, as this problem is typically solved on all of $\mathbb{R}^2$ [43] before applying numerical methods and solving this problem computationally.

One such medium we will explore is that of a step-index fiber. A step-index fiber is composed of a circular core region doped with a rare earth metal such as ytterbium (Yb) or Thulium (Tm) [44, 49]. This core region is surrounded by an outer, annular cladding region that is several times thicker than the core region. Such optical fibers can be several kilometers long [43], and as we will see later in this work, have

diameters on the length scale of hundreds of micrometers (μm).

Figure 1.1: Step-Index Fiber Cross-Section

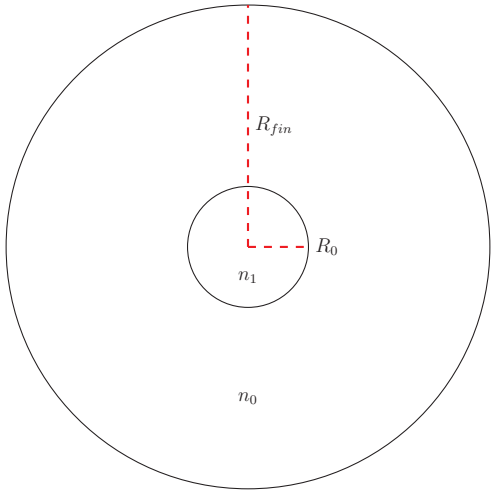


Figure 1.1: A cross-section of a step-index fiber.

For problem (1.1), the index of refraction is given by

$$n(x,y) = \begin{cases} n_1, & x^2+y^2 \leq R_0^2 \\ n_0, & x^2+y^2 > R_0^2 \end{cases}. \tag{1.2}$$

with $n_1 > n_0 > 0$ identically constant and satisfying, for our purposes, $n_1^2 - n_0^2 \ll 1$. In addition, $k = 2\pi/\lambda$ is the freespace wave number, λ is the operating wavelength in meters, $n = n(x,y)$ is the index of refraction of the optical fiber along any cross-section orthogonal to the direction of propagation, and β is the propagation constant of light through the optical fiber [43]. The radii R_0 and R_{fin} are the core and cladding radii, sometimes denoted r_{core} and r_{clad}, respectively. For step-index fibers, there are a finite number of guided modes u and propagation constants $\beta \in \mathbb{C}$ satisfying (1.1). For guided modes in an ideal step-index fiber we consider, the desired real-valued propagation constants β satisfy [43]

$$kn_0 < \beta < kn_1.$$

The number of propagation constants in this interval depends on several parameters, including the indices of refraction n_0 and n_1 of the cladding and core regions, the core radius, and the wavenumber k; indeed, a well-known estimate for the number of corresponding guided modes of a step-index fiber we consider is given by $\frac{1}{2}V^2$, with $V = R_0 k\sqrt{n_1^2 - n_0^2}$ [43].

Our goal in this work is to harness tools such as the finite element method and the FEAST eigensolver to find the propagation constants β and the corresponding eigenfunctions u. For hermitian problems, we have obtained results using FEAST with the Discontinuous Petrov Galerkin (DPG) method, including an application to finding guided modes of an ytterbium-doped optical fiber [24]. Later in this dissertation, we provide numerical results from applying a nonlinear eigensolver used to handle the task of computing modes for a hollow-core microstructure fiber considered in works such as [29, 50]. A rough sketch of the geometry of such a fiber is given in Figure 1.2 [29, 50], and further details of this geometry can be found in our other work [27].

The geometric parameters for the microstructure fiber are as follows [29]:

- R_{core} - the radius of the core region (μm)
- R_{to} - the outer radius of the hollow capillary tubes (μm)
- R_{ti} - the inner radius of the hollow capillary tubes (μm)
- t - the thickness the hollow capillary tubes (μm)

Figure 1.2: Microstructure Fiber Cross-Section

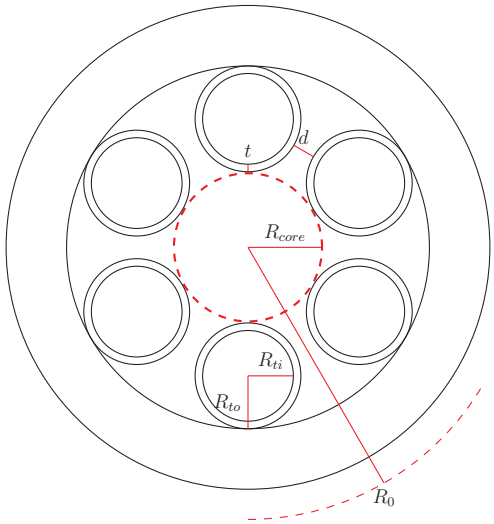


Figure 1.2: A cross-section of a microstructure fiber with six hollow capillaries.

- d - the azimuthal separation between two capillary tubes (μm)
- R_0 - the radial distance after which the outgoing medium is homogeneous

To get an intuition for the solutions we wish to compute, we will take a look at deriving the scalar equation (1.1) from Maxwell's Equations. This derivation will form the basis of comparison for the algorithm we develop based on the work of [46].

1.2 Derivation from Maxwell's Equations

We begin with a high-level overview of the problem we wish to solve, starting from Maxwell's Equations. We derive the problem we wish to solve under the assumption of working within an optical medium through which laser light is propagated, specifically an optical fiber with a cylindrical (outer) geometry. Such a derivation can be found in various works [42, 43, 52]. We briefly explore the details of such a derivation in this section to motivate the problem we wish to solve. In full, Maxwell's equations are given by [52]

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} - \mu_0 \frac{\partial \mathbf{M}}{\partial t} \tag{1.3a}$$

$$\nabla \times \mathbf{H} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} + \mathbf{j} \tag{1.3b}$$

$$\nabla \cdot (\varepsilon_0 \mathbf{E}) = -\nabla \cdot \mathbf{P} + \rho \tag{1.3c}$$

$$\nabla \cdot (\mu_0 \mathbf{H}) = -\nabla \cdot (\mu_0 \mathbf{M}) \tag{1.3d}$$

Where $\mathbf{E} = \begin{bmatrix} E_x & E_y & E_z \end{bmatrix}^T$ and $\mathbf{H} = \begin{bmatrix} H_x & H_y & H_z \end{bmatrix}^T$ are the electric and magnetic fields, $\mathbf{P}$ and $\mathbf{M}$ are the polarization and magnetic densities, ρ is the free space charge density, $\mathbf{j}$ is the current density, ε_0 is the vacuum permittivity, and μ_0 is the vacuum permeability [52]. For optical fibers, we can make an immediate simplification of Maxwell's equations: We assume a negligable magnetic density $\mathbf{M}$, as well as no free charge density, or current density, so Maxwell's equations reduce to

$$\begin{aligned} \nabla \times \mathbf{E} &= -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \\ \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} \\ \nabla \cdot \mathbf{D} &= 0 \\ \nabla \cdot \mathbf{B} &= 0 \end{aligned}$$

where $\mathbf{D} := \varepsilon_0 \mathbf{E} + \mathbf{P}$ is the electric charge displacement and $\mathbf{B} := \mu_0 \mathbf{H}$ is the magnetic

induction vector [42, 52]. Thus, Maxwell's equations can be written as

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t}$$

A further simplification allows us to write $\mathbf{P} = \varepsilon_0 \chi \mathbf{E}$, for which we can then replace $\mathbf{D}$ by $\varepsilon_0 \mathbf{E} + \varepsilon_0 \chi \mathbf{E} = \varepsilon_0 \varepsilon \mathbf{E}$, where $\varepsilon := 1 + \chi$ and where χ is the scalar electric susceptibility [52]. Then we have

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \tag{1.4a}$$

$$\nabla \times \mathbf{H} = \varepsilon_0 \varepsilon \frac{\partial \mathbf{E}}{\partial t} \tag{1.4b}$$

We then proceed by eliminating $\mathbf{H}$. Taking the curl of the Equation (1.4a) yields

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu_0 \frac{\partial (\nabla \times \mathbf{H})}{\partial t},$$

and then substution of the expression for $\nabla \times \mathbf{H}$ yields

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu_0 \varepsilon_0 \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

Next, we apply the identity

$$\nabla \times (\nabla \times \mathbf{E}) = \nabla (\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}.$$

By neglecting the term $\nabla(\nabla \cdot \mathbf{E})$ (under the assumption of a homogenous medium),

we obtain

$$-\nabla^2\mathbf{E} - \mu_0\varepsilon_0\varepsilon\frac{\partial^2\mathbf{E}}{\partial t^2} = 0.$$

where Δ^2 is the vector laplacian, i.e. the laplacian applied to each component of $\mathbf{E}$. Denoting $c_0 := (\varepsilon_0\mu_0)^{-1/2}$ to be the vacuum speed of light, we have that

$$-\nabla^2\mathbf{E} = \frac{\varepsilon}{c_0^2}\frac{\partial^2\mathbf{E}}{\partial t^2}. \tag{1.5}$$

While this certainly holds in a homogeneous medium, we will be dealing with applications in which ε (and hence the refractive index of our medium) is non-homogenous. Furthermore, we will assume that that that the solutions to the wave equation above are time-harmonic and propagate in the z-direction of our coordinate system. Hence, we assume an Ansatz of the form [42, 43]:

$$\mathbf{E} = \mathbf{E}_0(x, y)e^{i(\beta z - \omega t)} \tag{1.6}$$

for the electric field $\mathbb{R}$, where we write $\mathbf{E} = \begin{bmatrix} E_x & E_y & E_z \end{bmatrix}^T$ as before; a similar expression is taken for the magnetic field $\mathbf{H}$ [42]. Here, we have that the components E_x, E_y, E_z depend spatially on the coordinates x and y only. Furthermore, the Ansatz (1.6) assumes that our field propagates along the z-axis of our coordinate system, and that our solution is time-harmonic. In doing so, we have that for each component E_i and H_i of the electric and magnetic fields ($i = x, y, z$), the following holds from expanding out (1.4) (and similarly in [42]):

$$\frac{\partial E_z}{\partial y} - i\beta E_y = i\omega\mu_0 H_x \tag{1.7a}$$
$$-\frac{\partial E_z}{\partial x} + i\beta E_x = i\omega\mu_0 H_y \tag{1.7b}$$
$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = i\omega\mu_0 H_z \tag{1.7c}$$

$$\frac{\partial H_z}{\partial y} - i\beta H_y = -i\omega\varepsilon_0\varepsilon E_x \tag{1.7d}$$
$$-\frac{\partial H_z}{\partial x} + i\beta H_x = -i\omega\varepsilon_0\varepsilon E_y \tag{1.7e}$$
$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = -i\omega\varepsilon_0\varepsilon E_z \tag{1.7f}$$

Next, our goal is to write the longitudinal components E_z and H_z in terms of the corresponding components E_x, E_y, H_x, H_y. Our goal is to show that the longitudinal component E_z of the electric field $\mathbf{E}$ satisfies the Helmholtz problem we stated in the previous section. To this end, we just need a careful combination of the equations in (1.7). We begin by showing that the E_z component of the electric field $\mathbf{E}$ satisfies a Helmholtz equation used to solve for guided and leaky modes. To this end, we first begin by finding expressions for H_x, H_y, and then substitute them into (1.7f) to obtain the desired result. First, observe that Equations (1.7e) and (1.7a) can be written as

$$i\omega\mu_0 H_x = \frac{\partial E_z}{\partial y} - i\beta E_y \tag{1.8a}$$
$$i\beta H_x = \frac{\partial H_z}{\partial x} - i\omega\varepsilon_0\varepsilon E_y \tag{1.8b}$$

Multiplying Equation (1.8a) by $-i\omega\varepsilon_0\varepsilon$, Equation (1.8b) by $i\beta$, and then adding followed by division of $\kappa^2 := \omega^2\mu_0\varepsilon_0\varepsilon - \beta^2$ [42] yields

$$H_x = \frac{i}{\kappa^2}\left(\beta\frac{\partial H_z}{\partial x} - \omega\varepsilon_0\varepsilon\frac{\partial E_z}{\partial y}\right). \tag{1.9}$$

Similarly, we consider Equation (1.7b) and (1.7d).

$$i\omega\mu_0 H_y = -\frac{\partial E_z}{\partial x} + i\beta E_x \tag{1.10a}$$

$$i\beta H_y = \frac{\partial H_z}{\partial y} + i\omega\varepsilon_0\varepsilon E_x \tag{1.10b}$$

Multiplying Equation (1.10a) by $-i\omega\varepsilon_0\varepsilon$, Equation (1.10b) by $i\beta$, and then adding followed by division of κ^2 yields

$$H_y = \frac{i}{\kappa^2}\left(\beta\frac{\partial H_z}{\partial y} + \omega\varepsilon_0\varepsilon\frac{\partial E_z}{\partial x}\right). \tag{1.11}$$

Substitution and of Equations (1.9) and (1.11) into the left-hand-side of Equation (1.7f) yields

$$\begin{aligned}\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= \frac{i}{\kappa^2}\left(\beta\frac{\partial^2 H_z}{\partial x\partial y} + \omega\varepsilon_0\varepsilon\frac{\partial^2 E_z}{\partial x^2}\right) - \frac{i}{\kappa^2}\left(\beta\frac{\partial^2 H_z}{\partial y\partial x} - \omega\varepsilon_0\varepsilon\frac{\partial^2 E_z}{\partial y^2}\right)\\ &= \frac{i\omega\varepsilon_0\varepsilon}{\kappa^2}\left(\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2}\right).\end{aligned}$$

Equating this result with the right-hand-side of Equation (1.7f) yields

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} = \kappa^2 E_z,$$

where we have taken the initiative to divide out boths sides of Equation (1.7f) by the common factor of $i\omega\varepsilon_0\varepsilon$. Writing κ^2 as

$$\begin{aligned}\kappa^2 &= \omega^2\mu_0\epsilon_0\varepsilon - \beta^2 \\ &= \frac{\omega^2}{c_0^2}\varepsilon - \beta^2 \\ &= k^2 n^2 - \beta^2\end{aligned}$$

where $k = \omega/c_0 = 2\pi/\lambda$ and $n = \sqrt{\varepsilon}$ [52][1]. Hence, the E_z component of the electric field $\mathbf{E}$ satisfies

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + k^2 n^2 E_z = \beta^2 E_z.$$

A similar derivation shows that the H_z component of the magnetic field $\mathbf{H}$ satisfies the same differential equation. In the next section, we will tackle semi-analytic solutions of this problem for functions representing the E_z component of the electric field $\mathbf{E}$ for guided modes. A similar analysis will hold for leaky modes.

[1]It should be noted that in other works, such as Marcuse's *Light Transmission Optics*, the refractive index n of a dielectric material is defined to be $n = \sqrt{\varepsilon_i/\varepsilon_0}$, where ε_i is the permittivity of medium with which we are working and ε_0 is the vacuum permittivity (see, for example, [42, Chapter 1, §6]).

1.2.1 Guided Modes of a Step-Index Optical Fiber

In the previous section, we stepped through a derivation of the wave equation satisfied by the desired electric field solution to Maxwell's equations. In addition, we assumed a time-harmonic solution, a simplification allowing one instead solve a PDE in xy-coordinates. Since the present interest is to find solutions to (1.1) for step-index fibers, our next step is to find guided mode solutions and propagation constants β satisfying (1.1). In polar coordinates, we have

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} - \beta^2 u + k^2 n^2 u = 0, \tag{1.12}$$

Note that in Cartesian coordinates, the above would appear as the familiar differential equation $\Delta u + k^2 n^2 u = \beta^2 u$. Using separation of variables by taking a solution of the form $u(r, \theta) = R(r)T(\theta)$, we obtain

$$R''T + \frac{1}{r}R'T + \frac{1}{r^2}RT'' + \kappa^2 u = 0,$$

with $\kappa^2 = k^2 n^2 - \beta^2$. We denote $\kappa_1 := k^2 n_1^2 - \beta^2$ and $\kappa_0^2 := k^2 n_0^2 - \beta^2$. Collecting functions of like variables on either side, we have

$$r^2 \frac{R''}{R} + r\frac{R'}{R} + r^2\kappa^2 = -\frac{T''}{T}.$$

Each side is a function of just the variable r or t, so both sides must agree on the same constant, which we denote by ℓ^2 as in [52]. Then we have immediately that

$$T'' + \ell^2 T = 0,$$

which has two linearly independent solutions $T(\theta) = e^{\pm i\ell\theta}$ for ℓ a non-negative integer. This forces our azimuthal solution T to be 2π periodic, which in turn yields the desired continuity we need in our solutions. Next, we split the differential equation for $R(r)$ into one part corresponding to the core region and another part corresponding to the cladding region, respectively. Before splitting, recall that

$$r^2\frac{R''}{R} + r\frac{R'}{R} + r^2\kappa^2 = \ell^2.$$

Upon multiplying by R and moving all terms to the left-hand-side, we have

$$r^2R'' + rR' + r^2\left(\kappa_1^2 - \frac{\ell^2}{r^2}\right)R = 0, r \leq R_0$$

$$r^2R'' + rR' - r^2\left(\kappa_0^2 + \frac{\ell^2}{r^2}\right)R = 0, r > R_0$$

The above differential equations have general solutions consisting of Bessel functions $J_\ell(\kappa_1 r), Y_\ell(\kappa_1 r)$ and modified Bessel functions $I_\ell(\kappa_0 r), K_\ell(\kappa_0 r)$, respectively [52]. For guided modes, we seek solutions that are bounded in the core region and decay exponentially in the cladding. For the purposes of computing a semianalytic solution, we have tacitly assumed that the cladding region extends infinitely outward from the center of the fiber cross-section as in [43]. In anticipation of the task of numericall solving such a problem, we discuss the enforcement of boundary conditions later in this dissertation. In the present discussion, the radial solution over the core and the cladding regions simplifies to

$$R(r) = \begin{cases} AJ_\ell(\kappa_1 r), & r \leq R_0 \\ BK_\ell(\kappa_0 r), & r > R_0 \end{cases}.$$

Since we desire a unique solution, we need to close this problem with conditions at the

interface of the core and cladding regions. Specifically, the solution u and its normal derivative should be continuous at the interface of the core and cladding regions [52]. In practice, this amounts to enforcing

$$\begin{aligned} AJ_\ell(\kappa_1 R_0) &= BK_\ell(\kappa_0 R_0) \\ A\kappa_1 J'_\ell(\kappa_1 R_0) &= B\kappa_0 K'_\ell(\kappa_0 R_0) \end{aligned}$$

or

$$\begin{aligned} AJ_\ell(\kappa_1 R_0) - BK_\ell(\kappa_0 R_0) &= 0 \\ A\kappa_1 J'_\ell(\kappa_1 R_0) - B\kappa_0 K'_\ell(\kappa_0 R_0) &= 0. \end{aligned},$$

This system of equations in the unknowns A and B has nontrivial solutions when the determinant of the matrix corresponding to the above linear system is zero, i.e. when

$$\kappa_0 J_\ell(\kappa_1 R_0) K'_\ell(\kappa_0 R_0) - \kappa_1 K_\ell(\kappa_0 R_0) J'_\ell(\kappa_1 R_0) = 0.$$

Next, we use the Bessel function identities [1, 52]

$$J'_\ell(x) = (-\ell/x) J_\ell(x) + J_{\ell-1}(x)$$

$$K'_\ell(x) = (-\ell/x) K_\ell(x) - K_{\ell-1}(x),$$

so at $r = R_0$, we have

$$\begin{aligned} &\kappa_0 R_0 J_\ell(\kappa_1 R_0)((-\ell/(\kappa_0 R_0)) K_\ell(\kappa_0 R_0) - K_{\ell-1}(\kappa_0 R_0)) - \\ &\kappa_1 R_0 K_\ell(\kappa_0 R_0)((-\ell/(\kappa_1 R_0)) J_\ell(\kappa_1 R_0) + J_{\ell-1}(\kappa_1 R_0)) = 0. \end{aligned}$$

Simplifying yields

$$-\kappa_1 K_\ell(\kappa_0 R_0) J_{\ell-1}(\kappa_1 R_0) - \kappa_0 J_\ell(\kappa_1 R_0) K_{\ell-1}(\kappa_0 R_0) = 0. \tag{1.13}$$

Upon moving $\kappa_0 J_\ell(\kappa_1 R_0) K_{\ell-1}(\kappa_0 R_0)$ to the right-hand-side of (1.13) and dividing both sides by $\kappa_1 J_\ell(\kappa_1 R_0) K_{\ell-1}(\kappa_0 R_0)$, we obtain

$$-\frac{K_\ell(\kappa_0 R_0) J_{\ell-1}(\kappa_1 R_0)}{J_\ell(\kappa_1 R_0) K_{\ell-1}(\kappa_0 R_0)} = \frac{\kappa_0}{\kappa_1}, \tag{1.14}$$

or what Reider [52] refers to as the mode condition for cylindrical step-index waveguides. The above equation is used to determine the propagation constants β contained within the constants κ_0 and κ_1 [52]. In practice, either Equation (1.13) or Equation (1.14) can be solved numerically to determine the corresponding propagation consant β; one such package for doing so in the `Python` programming language is the `cxroots` package [48].

The next step is to determine the radial solution R for the guided modes. Under the assumption that the mode condition (1.14) is satisfied, we pick $A = K_\ell(\kappa_0 R_0)$ and $B = J_\ell(\kappa_1 R_0)$. Then the solution to 1.12 is given by

$$u(r,\theta) = \begin{cases} K_\ell(\kappa_0 R_0) J_\ell(\kappa_1 r), & r \leq R_0 \\ J_\ell(\kappa_1 R_0) K_\ell(\kappa_0 r), & r > R_0 \end{cases}. \tag{1.15}$$

The solution (1.15) to (1.12) will be used to verify the our theoretical error estimates in Chapter 4, where we compute guided modes and (scaled) propagation constants

using the FEAST algorithm. Similarly, a leaky mode solution will be used for the verification of our polynomial eigensolver in Chapter 6.

1.2.2 Leaky Modes of a Step-Index Fiber

In addition to computing guided modes, we wish to compute leaky modes u with corresponding propagation constants $\beta \in \mathbb{C}$ solbingg problem (1.1). Leaky modes, which are also known as resonances or quasi-normal modes [25, 43, 46], are outgoing solutions that together with a corresponding propagation constant β satisfy (1.1). In this section, we will go one step further to make the problem computationall tractable, beginning with a non-dimensionalization of problem (1.1).

Since computing leaky modes requires the computation of complex-valued propagation constants, we will opt to use a perfectly matched layer (PML) to transform the modes we seek to compute into a non-selfadjoint problem from which complex propagation constants can be computed. In addition, the PML forces exponential decay of the modes we compute, and this helps make the problem suitable for solving on a finite-sized computational domain. We address this topic in further detail in Chapters 2 and 5. Here, we focus on rescaling the partial differential equation (1.1) and define certain quantities used later in this work.

Rescaling the model problem

One of the first challenges to address is that of scaling. The dimensions of optical fibers are measured in micrometers (μm), where 1 μm $= 10^{-6}$ meters (m). In contrast, propagation constants β are several orders of magnitude larger. Consequently, we will choose a characteristic length scale $L > 0$ by which we nondimensionalize, in the hope that the quantities analogous to the propagation constant we compute are potentially

of the order $O(1)$. Let $x = (x_1, x_2)^T \in \mathbb{R}^2$ denote our coordinates with dimensions in the physical space of the fiber. Then define $\hat{x} := x/L$, i.e.

$$\hat{x}_i = x_i/L, i = 1, 2 \tag{1.16}$$

Then for $\hat{u}(\hat{x}) = u(L\hat{x}_1, L\hat{x}_2)$, we have $\Delta = L^{-2}\hat{\Delta}$, where $\hat{\Delta} := \partial/\partial\hat{x}_1^2 + \partial/\partial\hat{x}_2^2$ is the nondimensional analogue of the Laplacian Δ. After transforming to dimensionless coordinates, multiplying through by L^2, and subtracting $L^2\beta^2\hat{u}$ from both sides, we have

$$\hat{\Delta}\hat{u} + (L^2k^2\hat{n}(\hat{x})^2 - L^2\beta^2)\hat{u} = 0 \tag{1.17}$$

where $\hat{n}(\hat{x}) = n(L\hat{x})$ is the index of refraction in dimensionless coordinates. Next, we define the quantities

$$Z^2 = L^2(k^2n_0^2 - \beta^2), \tag{1.18a}$$

$$X^2 = L^2(k^2n_1^2 - \beta^2), \tag{1.18b}$$

$$V(\hat{x}_1, \hat{x}_2) = L^2k^2(n_0^2 - \hat{n}^2). \tag{1.18c}$$

In Chapters 5 and 6, we use this nondimensionalization to find leaky modes of step-index and microstructure fibers. The quantity Z, for our polynomial eigensolver, is the eigenvalue we seek to compute, as it contains the propagation β. The quantity X arises when we split our nondimensionalized problem into the respective core and cladding regions. Putting everything together by subtracting $L^2(k^2n_0^2 - \beta^2)\hat{u}$ from

both sides of (1.17) and letting $\hat{r} := ||\hat{x}||$, we obtain

$$-\hat{\Delta}\hat{u} + V^2\hat{u} = Z^2\hat{u}, \quad \hat{x} \in \mathbb{R}^2 \tag{1.19a}$$

$$\hat{u} \text{ is outgoing as } \quad \hat{r} \to \infty \tag{1.19b}$$

Consequently, the radial distance R_0 from the center of the physical fiber is also rescaled, namely $\hat{R}_0 := R_0/L$. In addition, define $\hat{r} = r/L$, where $r = \sqrt{x_1^2 + x_2^2}$. Note also that for $\hat{r} > \hat{R}_0$, (1.19) reduces to

$$-\hat{\Delta}\hat{u} = Z^2\hat{u}, \quad \hat{x} \in \mathbb{R}^2 \tag{1.20a}$$

$$\hat{u} \text{ is outgoing as } \quad \hat{r} \to \infty \tag{1.20b}$$

Under circumstances in which Z is real-valued, we would normally enforce the Sommerfeld Radiation condition

$$\lim_{\hat{r}\to\infty} \sqrt{\hat{r}} \left(\frac{\partial \hat{u}}{\partial \hat{r}} - iZ\hat{u} \right) = 0 \tag{1.21}$$

to seek outgoing solutions [27]. In practice, however, the non-dimensional quantities Z we wish to compute have positive real part and negative imaginary part, so we cannot directly enforce the Sommerfeld condition. Since the general solution in the outgoing region $\hat{r} > \hat{R}_0$, however, can be expressed using Hankel functions of the first kind, namely [27]

$$\hat{u}(\hat{r},\theta) = \sum_{\ell=-\infty}^{\infty} c_\ell H_\ell^{(1)}(Z\hat{r})e^{i\ell\theta}, \qquad \hat{r} > \hat{R}_0 \tag{1.22}$$

we prescribe that $\hat{u}$ is outgoing and thus require that $\hat{u}$ has the form (1.22) provided we use the analytic continuation of the Hankel functions from the real line into the complex plane.

Returning to the scenario of a step-index fiber, we choose $L = R_0$ for our characteristic length scale. Hence, we have that the function V satisfies

$$V(\hat{r}) = \begin{cases} -V_1^2, & \hat{r} \leq 1 \\ 0, & \hat{r} > 1, \end{cases} \tag{1.23}$$

and where $V_1^2 := L^2(k^2 n_1^2 - n_0^2) = X^2 - Z^2$. To compute analytic solutions to (1.19) corresponding to this particular construction of V, we repeat a similar process of ensuring continuity and smoothness where the material properties of the fiber change [27]. Separating the non-dimensional PDE into the corresponding nondimensional core and cladding regions, our goal is to find functons $\hat{u}$ and corresponding $Z \in \mathbb{C}$ satisfying

$$\hat{\Delta}\hat{u} + X^2\hat{u} = 0, \quad \hat{r} \leq 1 \tag{1.24}$$

$$\hat{\Delta}\hat{u} + Z^2\hat{u} = 0, \quad \hat{r} > 1 \tag{1.25}$$

$$[\![\hat{u}]\!] = [\![\partial\hat{u}/\partial\hat{r}]\!] = 0, \quad \hat{r} = 1. \tag{1.26}$$

Here, $X^2 := V_1^2 + Z^2$, and $[\![\cdot]\!]$ denotes the jump of a function defined at the interface

of the core and cladding regions $\hat{r} = 1$. In this case, we are enforcing continuity of the solution $\hat{u}$ and its normal derivative. Assuming a separation of variables solution as done for (1.12), but now for

$$\frac{\partial^2 \hat{u}}{\partial \hat{r}^2} + \frac{1}{\hat{r}} \frac{\partial \hat{u}}{\partial \hat{r}} + \frac{1}{\hat{r}^2} \frac{\partial^2 \hat{u}}{\partial \theta^2} + X^2 \hat{u} = 0 \tag{1.27a}$$

in the core region $\hat{r} \leq 1$ and

$$\frac{\partial^2 \hat{u}}{\partial \hat{r}^2} + \frac{1}{\hat{r}} \frac{\partial \hat{u}}{\partial \hat{r}} + \frac{1}{\hat{r}^2} \frac{\partial^2 \hat{u}}{\partial \theta^2} + Z^2 \hat{u} = 0 \tag{1.27b}$$

in the cladding region $\hat{r} > 1$. Using separation of variables as in the case of guided modes, our solution $\hat{u}(\hat{r}, \theta)$ can be expressed as

$$\hat{u}(\hat{r}, \theta) = \begin{cases} A J_\ell(X\hat{r}) e^{i\ell\theta}, & \hat{r} \leq 1 \\ B H_\ell^{(1)}(Z\hat{r}) e^{i\ell\theta}, & \hat{r} > 1 \end{cases} \tag{1.28}$$

Note that, in contrast to the analytic solution for guided modes of a step-index fiber, we have chosen the solution in the region $\hat{r} > 1$ to be the Hankel function of the first kind. In matrix-vector notation, we can express the interface conditions at $\hat{r} = 1$ by

$$T \begin{bmatrix} A \\ B \end{bmatrix}, \quad T = \begin{bmatrix} J_\ell(X) & -H_\ell^{(1)}(Z) \\ X J_\ell'(X) & -Z(H_\ell^{(1)})'(Z) \end{bmatrix} \tag{1.29}$$

We can find nontrivial solutions to (1.29) provided that we can find a root Z satisfying

$f(Z) = 0$, where f is defined by

$$\begin{aligned} f(Z) &= \det(T) \\ &= -ZJ_\ell(X)(H_\ell^{(1)})'(Z) + XJ_\ell'(X)H_\ell^{(1)}(Z) \\ &= -ZJ_\ell\left(\sqrt{V_1^2+Z^2}\right)(H_\ell^{(1)})'(Z) + \left(\sqrt{V_1^2+Z^2}\right)J_\ell'\left(\sqrt{V_1^2+Z^2}\right)H_\ell^{(1)}(Z) \end{aligned} \tag{1.30}$$

Using the Bessel function identities [1]

$$J_\ell'(z) = \frac{\ell}{z}J_\ell(z) - J_{\ell+1}(z) \tag{1.31a}$$

$$(H_\ell^{(1)})'(z) = \frac{\ell}{z}H_\ell^{(1)}(z) - H_{\ell+1}^{(1)}(z) \tag{1.31b}$$

we can simplify the second line of (1.30) to obtain an expression that involves no derivatives of Bessel or Hankel functions:

$$\begin{aligned} f(Z) &= -ZJ_\ell(X)(H_\ell^{(1)})'(Z) + XJ_\ell'(X)H_\ell^{(1)}(Z) \\ &= -ZJ_\ell(X)\left[\frac{\ell}{Z}H_\ell^{(1)}(Z) - H_{\ell+1}^{(1)}(Z)\right] + XH_\ell^{(1)}(Z)\left[\frac{\ell}{X}J_\ell(X) - J_{\ell+1}(X)\right] \\ &= -\ell J_\ell(X)H_\ell^{(1)}(Z) + ZJ_\ell(X)H_{\ell+1}^{(1)}(Z) + \ell J_\ell(X)H_\ell^{(1)}(Z) - XJ_{\ell+1}(X)H_\ell^{(1)}(Z) \\ &= ZJ_\ell(X)H_{\ell+1}^{(1)}(Z) - XJ_{\ell+1}(X)H_\ell^{(1)}(Z) \\ &= ZJ_\ell\left(\sqrt{V_1^2+Z^2}\right)H_{\ell+1}^{(1)}(Z) - \left(\sqrt{V_1^2+Z^2}\right)J_{\ell+1}\left(\sqrt{V_1^2+Z^2}\right)H_\ell^{(1)}(Z) \end{aligned} \tag{1.32}$$

In practice, the value of Z solving $f(Z) = 0$ for a fixed positive integer ℓ is computed numerically using an appropriate root solver, and multiple values of Z may correspond to a given value of ℓ. Assuming such a root exists, we set $A = H_\ell^{(1)}(Z)$ and $B = J_\ell(X)$, where $X = \sqrt{V_1^2 + Z^2}$. This yields the analytic solution

$$\hat{u}(\hat{r}, \theta) = \begin{cases} H_\ell^{(1)}(Z_\ell) J_\ell(X_\ell \hat{r}) e^{i\ell\theta}, & \hat{r} \leq 1 \\ J_\ell(X_\ell) H_\ell^{(1)}(Z_\ell \hat{r}) e^{i\ell\theta}, & \hat{r} > 1 \end{cases} \tag{1.33}$$

In non-dimensional coordinates, the solution we wish to compute is given by

$$u(r, \theta) = \begin{cases} H_\ell^{(1)}(R_0\kappa_0) J_\ell(\kappa_1 r) e^{i\ell\theta}, & r \leq R_0 \\ J_\ell(R_0\kappa_1) H_\ell^{(1)}(\kappa_0 r) e^{i\ell\theta}, & r > R_0. \end{cases} \tag{1.34}$$

In Chapter 6, we use (1.33) to verify our numerical results when computing leaky modes of a step-index fiber.

1.2.3 Confinement Losses in Step-Index and Microstructure Fibers

Up to this point, we have shown that the Helmholtz problem has analytic solutions, and serves as a means of verifying the correctness of our work later in this dissertation. Solving this problem, however, serves as just one step in a larger endeavor, which is to push the limits of our numerical discretizations to compute confinement losses in optical fibers. Many works are interested in understanding the different mechanisms by which optical fibers exhibit confinement loss, or loss of power, when guiding laser light. Loss of power can occur due to physical stresses such as bending [57], or due to the structures of the fibers themselves [29, 38, 50]. The model we have derived

thus far applies to step-index and microstructure fibers with no perturbations to the ideal geometries we have presented. Optical fibers, however, can have physical deviations from their intended geometries, and this leads to confinement losses in the power of laser light propagating through such fibers. Works such as [42, 43] explore perturbations to ideal geometries of various waveguides as a first step to showing how losses can occur. In this work, computing confinement losses requires that we are able to compute as many digits of precision of the imaginary part of the propagation constant β as possible. In the literature, confinement losses are computed, for example, via the formula [8, 11, 15, 39, 56, 60]

$$\text{CL} = \frac{20}{\log 10} k \Im(n_{eff}) = \frac{20}{\log 10} \Im(\beta), \tag{1.35}$$

where $\Im(\beta)$ is the imaginary part of β, $n_{eff} = \beta/k$ is the effective index [60], and the units of confinement loss (CL) are in decibels per meter (dB/m). It should be noted that in other works [29], the confinement loss is reported up to a sign change in the imaginary part of β.

1.3 The Thesis at a Glance

The remainder of the thesis will be organized as follows. In Chapter 2, we discuss the foundational numerical methods for solving our problem of interest. We divide this into three sections: The classifical finite element method (FEM), the Discontinuous Petrov Galerkin (DPG) method, and Perfectly Matched Layers (PML).

In Chapter 3, we introduce the FEAST eigensolver, a state-of-the art tool we use and extend in this work. Starting from the context of subspace iteration, we use this

to talk about the FEAST algorithm, its applications, and the relevant theory needed to solidify the utility of the algorithm. We also take this time to discuss advances in the FEAST algorithm, as well as brief comparison to other contour-integral methods.

In Chapter 4, we go into further detail on the application of the DPG method to solving linear, self-adjoint eigenvalue problems. In this chapter, we discuss some of the theoretical results as they relate to error propagation and error estimates for eigenspaces and eigenvalues.

Chapter 5 discusses the results of another paper which has recently appeared in the journal Wave Motion. This chapter also introduces the nonlinear eigenvalue problem, including how it arises in the context of the motivating problem we wish to solve. Our algorithm and theoretical results are presented that show the correctness of our approach.

In Chapter 6, we discuss numerical verification of the algorithm discussed in chapter 5, as well as numerical results for the microstructure fiber problem in which we are interested. This section covers several experimental results to test our algorithm when finding leaky modes of microstructure fibers, as well as the stability of our results with changing parameters that directly affect our application of PML.

We conclude in Chapter 7 with a brief discussion of our results, as well as goals for future research.

Chapter 2

Numerical Methods for Partial Differential Equations

2.1 Introduction

In this chapter, we briefly discuss two methods used for solving elliptic partial differential equations employed in this thesis: The classical finite element method, and the discontinuous Petrov Galerkin method. In each section, we give a brief overview of relevant theory for source problems, as well as a brief look at eigenvalue problems within the classical finite element method. The focus on source problems is needed since the eigenvalue problems we seek to solve in fiber optics are done using the FEAST eigensolver [51], for which the work of computing eigenvalues and eigenvectors boils down to performing several linear system solves. In addition, we will also look at how Perfectly matched layers (PML) are used to make computational problems tractable for numerical computation.

2.2 Classical FEM

2.2.1 An Example Problem

Let $\Omega \subset \mathbb{R}^2$ be a bounded, polygonal domain. Suppose I wish to solve the following problem: Find a function $u : \Omega \to \mathbb{R}$ satisfying

$$-\Delta u = f, x \in \Omega \tag{2.1}$$

$$u\Big|_{\partial\Omega} = 0 \tag{2.2}$$

where f is square integrable on Ω, i.e. $f \in L^2(\Omega)$. To solve this problem in a variational setting, we use the tools of integration by parts to cast this problem into its weak form. Take any $v \in H_0^1(\Omega)$, the space of functions v for which v is square integrable, its first (weak) derivatives are square integrable, and whose boundary trace is zero. We multiply $v \in H_0^1(\Omega)$ to both sides of the PDE in (2.1) and integrate by parts over Ω to obtain

$$\begin{aligned}\int_\Omega -\Delta u v dx &= \int_\Omega \nabla u \cdot \nabla v dx - \int_{\partial\Omega} v \nabla u \cdot n dA \\ &= \int_\Omega \nabla u \cdot \nabla v dx\end{aligned}$$

and

$$\int_\Omega f v dx$$

on the right-hand-side. Defining $a(u,v) = \int_\Omega u'v'dx$ and $l(v) = \int_\Omega fvdx$ for $u, v \in H_0^1(\Omega)$, we seek to find a weak solution $u \in \mathcal{V} = H_0^1(\Omega)$ satisfying

$$a(u,v) = l(v) \quad \text{for all} \quad v \in \mathcal{V}. \tag{2.3}$$

To show that such a variational problem has a unique solution u, we turn to tools from functional analysis, as well as the Poincaré inequality from partial differential equations. Indeed, to solve 2.3, we need to show that $a(\cdot,\cdot)$ forms an inner product on $\mathcal{V}$. It certainly follows that a is linear in one of its arguments when the other is fixed, as this follows from the linearity of the gradient operator ∇ and the integral. Furthermore, for any $u \in \mathcal{V}$, we have that $|\nabla u|^2 \geq 0$, and hence $a(u,u) \geq 0$. To show that $a(u,u) = 0$ precisely when $u = 0$, we need the Poincaré inequality, which states that for $u \in \mathcal{V}$

$$||u||_{L^2(\Omega)} \leq C|u|_{H^1(\Omega)} = C(a(u,u))^{1/2}$$

for some constant $C > 0$ [4, 45].[1] To this end, suppose that $u \in \mathcal{V}$ is arbitrary and that $a(u,u) = 0$. By the Poincaré inequality, we have that

$$0 = a(u,u) \geq C^{-2}||u||^2_{L^2(\Omega)},$$

which holds precisely when $u = 0$. Hence, a is an inner product for $H^1_0(\Omega)$, and we denote the norm induced by the inner product $a(\cdot,\cdot)$ by $||u||_a = \sqrt{a(u,u)}$. It remains to verify that $l(v)$ is a bounded linear form for all $v \in \mathcal{V}$. This follows from an application of the Cauchy-Schwarz inequality [4, 45] and the Poincaré inequality.

$$|l(v)| = \left|\int_\Omega fv dx\right| \leq \left(\int_\Omega |f|^2 dx\right)^{1/2}\left(\int_\Omega |v|^2 dx\right)^{1/2} \leq C_P||f||_{L^2(\Omega)}|v|_{H^1(\Omega)}.$$

Then by the Riesz representation theorem [4, 10, 45] we know there exists a unique $u \in \mathcal{V}$ such that (2.3) holds.

[1] Sometimes, this is stated with the constant C on the left-hand-side, i.e. $C||u||_{L^2(\Omega)} \leq |u|_{H^1(\Omega)}$. See, for example, [24] with c_P used for the constant in the Poincaré inequality instead of C.

The Finite-Dimensional Setting

For convenience, let $\mathcal{V} = H_0^1(\Omega)$. In the finite-dimensional setting, we wish to find an approximation to $u \in \mathcal{V}$ in some closed, finite-dimensional space $\mathcal{V}_h \subset \mathcal{V}$. In practice, we typically choose the Lagrange finite element space, i.e. we set

$$\mathcal{V}_h = \left\{ v \in \mathcal{V} : v\Big|_K \in P_p(K) \forall K \in \mathcal{T}_h \right\}.$$

Then the goal is to find $u_h \in \mathcal{V}_h$ satisfying

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in \mathcal{V}_h \tag{2.4}$$

The analogous problem (2.3) was shown to be well-posed, and we can go a step further here by showing that the error $u - u_h$ satisfies a best-approximation-error. We begin by observing that (2.3) holds for all $v \in \mathcal{V}$, so it certainly holds for all $v_h \in \mathcal{V}_h$. Subtracting (2.4) from (2.3), we see that

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in \mathcal{V}_h \tag{2.5}$$

a condition known as Galerkin orthogonality [4][§2.5, Proposition 2.5.9]. Next, observe that for any $w \in \mathcal{V}_h$, we have that

$$\begin{aligned} ||u - u_h||_a^2 &= a(u - u_h, u - u_h) \\ &= a(u - u_h, u - w + w - u_h) \\ &= a(u - u_h, u - w) + a(u - u_h, w - u_h) \end{aligned}$$

$$
\begin{aligned}
&= a(u - u_h, u - w) \\
&\leq |u - u_h|_{H^1(\Omega)} |u - w|_{H^1(\Omega)} \\
&= ||u - u_h||_a ||u - w||_a.
\end{aligned}
$$

Dividing through by $||u - u_h||_a$, we have that

$$
||u - u_h||_a \leq ||u - w||_a \quad \forall w \in \mathcal{V}_h,
$$

hence

$$
||u - u_h||_a \leq \inf_{w \in \mathcal{V}_h} ||u - w||_a.
$$

On the other hand, we certainly have that $\inf_{w \in \mathcal{V}_h} ||u - w||_a \leq ||u - u_h||_a$, and hence

$$
||u - u_h||_a = \inf_{w \in \mathcal{V}_h} ||u - w||_a.
$$

Since $\mathcal{V}_h$ is a closed subspace of $\mathcal{V}$, the element realizing the infimum is contained in $\mathcal{V}_h$, and so we have the error in the norm $|| \cdot ||_a$ is optimal, i.e.

$$
||u - u_h||_a = \min_{w \in \mathcal{V}_h} ||u - w||_a. \tag{2.6}
$$

Now suppose we wish to compute an approximate weak solution to (2.1) in the setting

of (2.4). Suppose that $\dim(\mathcal{V}_h) = n$. If $\beta = \{\varphi_j\}_{j=1}^n$ is the basis of $\mathcal{V}_h$, then we can substitute the representation of u_h in β into (2.4). Letting $v_h = \varphi_i$ for $1 \leq i \leq n$, we have

$$a\left(\sum_{j=1}^{n} c_j \varphi_j, \varphi_i\right) = l(\varphi_i) \tag{2.7}$$

or

$$\sum_{j=1}^{n} c_j \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i dx = \int_\Omega f\varphi_i dx. \tag{2.8}$$

Defining $A \in \mathbb{R}^{n\times n}$ by $A_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i dx$ and $\vec{f_i} = \int_\Omega f\varphi_i dx$, we have

$$A\vec{u} = \vec{f} \quad \forall v \in \mathcal{V}_h \tag{2.9}$$

where $\vec{c} = \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix}$. At this stage, a wealth of numerical methods can be used to solve this problem to obtain the coefficients in $\vec{c}$ [64, 65].

2.3 Eigenvalue Problems in the Variational Setting

The types of eigenvalue problems in which we are interested in the finite element method can be summarized starting from the departure point of the following model problem, analogous to the problems (1.1) and (2.1): Find a function u and scalar λ satisfying

$$-\Delta u = \lambda u, \quad x \in \Omega \tag{2.10}$$

$$u\Big|_{\partial\Omega} = 0 \tag{2.11}$$

This is a common problem to probe in the broad umbrella of partial differential equations. Indeed, when Ω, say, is the unit square $(0,1)^2 \subset \mathbb{R}^2$, the eigenfunctions and eigevalues are well-known and readily computed by techniques such as separation of variables: For the integer indices $m, n \geq 1$, we have [45]

$$\lambda_{m,n} = \pi^2(m^2 + n^2), \quad u_{m,n}(x) = \sin(m\pi x_1)\sin(n\pi x_2).$$

This is often a problem that is often used for stress-testing numerical algorithms for computing eigenvalues and eigenvectors, especially in the finite-element community (see, for example, [23]). In practice, we tackle problems like (2.10) using the same techniques as in section 2.2.1. As before, we convert a problem such as (2.10) to weak form by taking a $v \in \mathcal{V}$, multiplying it to the PDE in (2.10), and integrating by parts to obtain

$$\int_\Omega \nabla u \cdot \nabla v dx = \lambda \int_\Omega uv dx. \tag{2.12}$$

The goal, in this case, is to find (λ, u) satisfing (2.12) for all $v \in \mathcal{V}$. Defining $b(u,v) = \int_\Omega uv dx$ for $u, v \in \mathcal{V}$, we seek to find u satisfying $a(u,v) = \lambda b(u,v)$ for all $v \in \mathcal{V}$. In the finite-dimensional setting of $\mathcal{V}_h$, we seek to find u_h and λ_h satisfying

$$a(u_h, v_h)dx = \lambda_h b(u_h, v_h) \quad \forall v_h \in \mathcal{V}_h \tag{2.13}$$

As in the case for the source problem (2.4), we substitute the representation of u_h in the basis of $\mathcal{V}_h$ and let $v_h = \varphi_i$ for $1 \leq i \leq n$ to now obtain the generalized eigenproblem

$$A\vec{c} = \lambda_h B\vec{c} \tag{2.14}$$

where A and $\vec{c}$ are defined as they were for the source problem (2.4), and $B_{ij} = \int_\Omega \varphi_j \varphi_i dx$ for $1 \leq i, j, \leq n$. The matrix B is often referred to as the mass matrix [41]. As we see later in this work, the matrices A and B will be used as ingredients to the FEAST algorithm, which in practice requires solutions to source problems of the form $(z + \Delta)u = f$ for some $z \in \mathbb{C}$ and functions u and f. Further details of such a problem can be found in [23, 24].

2.4 The Discontinuous Petrov Galerkin Method

2.4.1 Definitions

Before proceeding, we require some definitions in what follows. First, let X, Y be Hilbert spaces with inner products $(\cdot, \cdot)_X$ and $(\cdot, \cdot)_Y$. In the case that X and Y are spaces over the field of complex numbers, we say $b : X \times Y \to \mathbb{C}$ is a sesquilinear form if, for $u, w \in X$, $v, y \in Y$, and scalars $\alpha, \beta \in \mathbb{C}$, b satisfies the following [35]:

$$b(\alpha u + \beta w, v) = \alpha b(u, v) + \beta b(w, v) \tag{2.15a}$$

$$b(u, \alpha v + \beta y) = \bar{\alpha} b(u, v) + \bar{\beta} b(u, y), \tag{2.15b}$$

In the case that X and Y are Hilbert spaces of the real numbers $\mathbb{R}$, then the second propery (2.15b) holds without conjugation of α and β, and b is called a bilinear form. Furthermore, we denote the spaces of conjugate linear functionals from X to $\mathbb{C}$ and those from Y to $\mathbb{C}$ by X^* and Y^*, respectively [35].

2.4.2 Abstract Framework

In this section, we summarize three components needed to show the well-posedness of a source problem using the DPG method. The goal here is to show that under the right conditions, that one has a solution to the variational problem of finding a solution $u \in X$ satisfying

$$b(u, v) = l(v) \tag{2.16}$$

for all $v \in Y$, where X, Y are Hilbert spaces. Similar to the previous section, $b : X \times Y \to \mathbb{C}$ is a sesquilinear form, and $l : Y^* \to \mathbb{C}$ is a conjugate linear functional. The purpose of this cursory review is to connect abstract results for the DPG method to the solution of a source problem. This is particularly important in Chapter 4, as the application of the FEAST algorithm requires the solution of several linear systems at a time. In practice, this means solving a variational formulation of a source problem, which we require to be well-posed.

Relevant Theory

To begin, suppose as in the previous section we wish to find $u \in X$ satisfying (2.16). To do so, we need to ensure certain properties of the form b, which we summarize here. Further discussions of these topics can be found in several works [21, 24]. We use the remainder of this chapter to cover the relevant theory from [21].

Theorem 1 (Theorem 1, [21])**.** Suppose X is a Banach space and Y is a reflexive Banach space. The following three statements are equivalent:

a) For any $l \in Y^*$, there is a unique $x \in X$ satisfying

$$b(x, y) = l(y) \quad \forall y \in Y \tag{2.17}$$

b) $\{y \in Y : b(z, y) = 0 \forall z \in Z\} = \{0\}$ and there is a $C_1 > 0$ such that

$$\inf_{0 \neq z \in X} \sup_{0 \neq y \in Y} \frac{|b(z, y)|}{||z||_X ||y||_Y} \geq C_1 \tag{2.18}$$

c) $\{z \in X : b(z, y) = 0 \forall y \in Y\} = \{0\}$ and there is a $C_2 > 0$ such that

$$\inf_{0 \neq y \in Y} \sup_{0 \neq z \in X} \frac{|b(z, y)|}{||y||_Y ||z||_X} \geq C_2 \tag{2.19}$$

Theorem 2 (Theorem 2, [21])**.** Suppose X and Y are Hilbert spaces, $X_h \subset X$ and $Y_h \subset Y$ are finite dimensional subspace, $\dim(X_h) = \dim(Y_h)$, and suppose one of (a), (b), or (c) of Theorem 1 hold. If, in addition, there exists a constant $C_3 > 0$ such that

$$\inf_{0\neq z_h\in X_h}\sup_{0\neq y_h\in Y_h}\frac{|b(z,y)|}{||z_h||_X||y_h||_Y}\geq C_3 \tag{2.20}$$

then there is a unique $x_h \in X_h$ satisfying

$$b(x_h, y_h) = l(y_h) \quad \forall y_h \in Y_h, \tag{2.21}$$

and

$$||x - x_h||_X \leq \frac{C_2}{C_3}\inf_{z_h\in X_h}||x - z_h||_X \tag{2.22}$$

where $C_2 > 0$ is any constant for which the inequality $|b(x, y)| \leq C_2||x||_X||y||_Y$ holds for all $x \in X$ and $y \in Y$.

The important takeaway is that we have precisely two theorems that state the ingredients needed for well-posedness of the infinite and finite-dimensional problems we wish to solve. The analogous discrete inf-sup condition is also important, and in fact necessary for when we make the problem of interest computationally tractable by finding an approximate solution in the finite-dimensional setting. An important difference between the classical finite element method and what we are setting out to do here is that the arguments of the form b are not required to be in the same spaces as is the case for the classical finite element method.

Next, we introduce some necessary machinery to make the construction leading to the defintion of the Discontinuous Petrov-Galerkin Method. To this end, we call the

finite-dimensional spaces $X_h \subset X$ and $Y_h \subset Y$ trial and test spaces, respectively [21]. The appeal of the DPG method is that we can pick an ideal combination of these spaces for solving the problem (2.17), which often is motivated by the task of finding solutions to partial differential equations. To find an optimal choice of the spaces X_h and Y_h, we define the optimal test space Y_h^{opt} corresponding to the continuous sesquilinear form b (see [21]) by

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

where $T : X \to Y$ is the trial-to-test operator defined by [21]

$$(Tz, y)_Y = b(z, y) \quad \forall y \in Y, z \in X.$$

Indeed, for $z \in X$ fixed, $l(y) = b(z, y)$ is a bounded, conjugate linear functional from Y^* to $\mathbb{C}$. By the Riesz representation theorem applied to the Hilbert space Y with the inner product $(\cdot, \cdot)_Y$ [10], there is a unique $w \in Y$ such that $(w, y)_Y = l(y)$ for all $y \in Y$. In this case, we define the operator $T : X \to Y$ to be $Tz = w$ for the corresponding fixed $z \in X$.

Next, given a choice of trial space X_h and ideal test space $Y_h^{opt} = T(X_h)$, we define the ideal Petrov-Galerkin (ideal PG) method as finding an $x_h \in X_h$ [21]

$$b(x_h, y_h) = l(y_h) \quad \forall y_h \in Y_h^{opt} \tag{2.23}$$

In addition, one can show that the solution $x_h \in X_h$ to (2.23) satisfies a best approximation error, much in the same manner as problems in classifical finite elements. To

do so, we need to develop the machinery for this. First, as in [21], denote the Riesz map $R_Y : Y \to Y^*$ by $(R_Y y)(v) = (y, v)_Y$ for all $y, v \in Y$, and $B : X \to Y^*$ the operator generated by b, i.e. $Bz(y) = b(z, y)$ for $x \in X$ and $y \in Y$. Then it follows that $T = R_Y^{-1} \circ B$, which we verify from the definition of T and the Riesz map R_Y:

$$\begin{aligned} R_Y(Tz)(v) &= (Tz, v)_Y \\ &= b(z, v) \\ &= Bz(v), \end{aligned}$$

hence $R_Y \circ T = B \Longleftrightarrow T = R_Y^{-1} \circ B$ since the map R_Y is an isometric isomorphism [21]. In anticipation of the next result, define the energy norm $|||z|||_X := ||Tz||_X$ [21]. Indeed, we have that $||| \cdot |||_X$ and $|| \cdot ||_X$ are equivalent, provided we make some additional assumptions. First, observe that

$$\sup_{0 \neq z \in X} \frac{|b(z, v)|}{||v||_Y} = \sup_{0 \neq z \in X} \frac{|(Tz, v)_Y|}{||v||_Y} = ||Tz||_Y.$$

Then under the assumption that condition (2.19) holds and there exists a $C_2 > 0$ for which

$$\sup_{0 \neq z \in X} \frac{|b(z, y)|}{||z||_X} \leq C_2 ||y||_Y \quad \forall y \in Y$$

holds (see Assumption 7 in [21]), we have that for $z \in X$,

$$|||z|||_X = ||Tz||_Y = \sup_{0 \neq y \in Y} \frac{|b(z, y)|}{||y||_Y} \geq C_1 ||z||_X$$

Likewise, we have that

$$|||z|||_X = ||Tz||_Y = \sup_{0 \neq y \in X} \frac{|b(z,y)|}{||y||_Y}, \leq C_2 ||z||_X$$

hence $|||\cdot|||_X$ and $||\cdot||_X$ are equivalent. We are now ready to present a result connecting the unique solution of the ideal Petrov-Galerkin method (2.23) to the best-approximation-error estimate in $|||\cdot|||_X$.

Theorem 3. (Residual Minimization [21]) Suppose that condition (2.19) holds, there is a $C_2 > 0$ such that

$$\sup_{0 \neq z \in X} \frac{|b(z,y)|}{||z||_X} \leq C_2 ||y||_Y \quad \forall y \in Y,$$

and that $x \in X$ solves (2.17). Then the following are equivalent statements:

i) $x_h \in X_h$ is the unique solution of the ideal PG method (2.23).

ii) x_h is the best approximation to x from X_h in the following sense:

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

iii) x_h minimize residual in the following sense:

$$x_h = \text{argmin}_{z_h \in X_h} ||l - Bz_h||_{Y^*}$$

Proof. Suppose that x_h solves (2.17). Since this holds for for any $y \in Y$, it certainly holds for all $y_h \in Y_h^{opt}$. Subtracting (2.17) from (2.23) with $y = y_h \in Y_h^{opt}$, we have that $b(x - x_h, y_h) = 0$ for all $y_h \in Y_h^{opt}$. Expanding this expression leads to

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

$$
\begin{aligned}
&= b(x - x_h, Tz_h)_Y \\
&= b(x - x_h, Tz_h - Tx_h)_Y \\
&= (T(x - x_h), Tz_h - Tx_h)_Y \\
&= (T(x - x_h), Tz_h - Tx + Tx - Tx_h)_Y \\
&= (T(x - x_h), T(z_h - x))_Y + (T(x - x_h), T(x - x_h))_Y \\
&= -(T(x - x_h), T(x - z_h))_Y + (T(x - x_h), T(x - x_h))_Y
\end{aligned}
$$

and hence $|||x - x_h|||_X^2 = (T(x - x_h), T(x - x_h))_Y = (T(x - x_h), T(x - z_h))_Y$. Then then have that

$$
\begin{aligned}
|||x - x_h|||_X^2 &= (T(x - x_h), T(x - z_h))_Y \\
&\leq ||T(x - x_h)||_Y ||T(x - z_h)||_Y \\
&= |||x - x_h|||_X |||x - z_h|||_X
\end{aligned}
$$

Assuming that $|||x - x_h|||_X \neq 0$, we divide through, obtaining

$$
|||x - x_h|||_X \leq |||x - z_h|||_X \quad \forall z_h \in X_h,
$$

and hence

$$
|||x - x_h|||_X \leq \inf_{z_h \in X_h} |||x - z_h|||_X.
$$

Furthermore, the infimum $\inf_{z_h \in X_h} |||x - z_h|||_X$ is certainly less than $|||x - z_h|||_X$ for an arbitrary $z_h \in X_h$, hence *ii)* follows.

To show that *ii)* is equivalent to iii), observe as in [21] that

$$\begin{aligned}
|||x-x_h|||_X = \inf_{z_h\in X_h} |||x-z_h|||_X &\Longleftrightarrow ||T(x-x_h)||_Y = \inf_{z_h\in X_h} ||T(x-z_h)||_Y \\
&\Longleftrightarrow ||R_Y^{-1}B(x-x_h)||_Y = \inf_{z_h\in X_h} ||R_Y^{-1}B(x-z_h)||_Y \\
&\Longleftrightarrow ||B(x-x_h)||_{Y^*} = \inf_{z_h\in X_h} ||B(x-z_h)||_{Y^*} \\
&\Longleftrightarrow ||l-Bx_h||_{Y^*} = \inf_{z_h\in X_h} ||l-Bz_h||_{Y^*}
\end{aligned}$$

where we recall that $l = Bx$, hence the result follows. □

Let us return to the problem (2.17), for which we assume theorem 1 holds. For an x_h solving (2.23), define $\tilde{l} : Y^* \to \mathbb{C}$ by $\tilde{l}(y) = l(y) - b(x_h, y)$ (i.e. $\tilde{l} = l - Bx_h$), which is bounded and continuous since l and b are themselves bounded and continuous. Hence, there is a unique $\varepsilon \in Y$ solving

$$(\varepsilon, y)_Y = \tilde{l}(y) \quad \forall y \in Y \tag{2.24}$$

by the Riesz representation theorem [10]. That ε takes its specific form comes from directly from (2.24), which immediately implies

$$R_Y(\varepsilon) = l - Bx_h \Longleftrightarrow \varepsilon = R_Y^{-1}(l - Bx_h) \in Y.$$

The error representation function ε appears in several works as an important component of the formulation of the DPG method and error control (see, for example, [6,24]).

Its presence is important also in the equivalence between problem (2.23) and a mixed formulation that explicitly incorporates ε.

Theorem 4. (Equivalence of the ideal Petrov-Galerkin method and a mixed formulation [21])

The following are equivalent statements.

i) $x_h \in X_h$ solves the ideal Petrov-Galerkin method (2.23).

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

$$(\varepsilon, y)_Y + b(x_h, y) = l(y) \quad \forall y \in Y, \tag{2.25a}$$

$$b(z_h, \varepsilon) = 0 \quad \forall z_h \in X_h. \tag{2.25b}$$

Proof. We follow [21]. Suppose that *i)* holds. Previous computations show that

$$(\varepsilon, y)_Y + b(x_h, y) = l(y) \quad \forall y \in Y,$$

so it remains to verify (2.25b). Indeed, we have by the definition of T that

$$\begin{aligned} b(z_h, \varepsilon) &= (Tz_h, \varepsilon)_Y \\ &= (Tz_h, R_Y^{-1}(l - Bx_h))_Y \\ &= (Tz_h, T(x - x_h))_Y \end{aligned}$$

Note, however, that $(Tz_h, T(x - x_h))_Y = \overline{(T(x - x_h), Tz_h)_Y} = \overline{b(x - x_h, Tz_h)}$, which by *i)* and application of theorem 1 is equal to zero.

Now suppose that *ii)* holds. To show that this implies *i)*, it suffices to show that

$(\varepsilon, y)_Y = 0$ for all $y \in Y_h$. Indeed, we have that

$$\begin{aligned} 0 &= b(z_h, \varepsilon) \\ &= (Tz_h, \varepsilon)_Y \\ &= (y, \varepsilon)_Y \end{aligned}$$

where $Tz_h = y \in Y_h$, and hence $b(x_h, y) = l(y)$ for all $y \in Y_h$, establishing *i)*. □

2.4.3 An Ideal DPG Method

To form an ideal Discontinuous Petrov-Galerkin Method, we go back to the definition of the space Y in (2.17). To do so, we take $\Omega \subset \mathbb{R}^n$ open, and partition Ω into disjoint open subsets K called elements; in practice, these elements usually are intervals in $\mathbb{R}$, triangles in $\mathbb{R}^2$, and tetrahedra in $\mathbb{R}^3$, although other choices are certainly possible in $\mathbb{R}^n$ for $n \geq 2$ [4]. Taken as a collection which we denote Ω_h, the elements $K \in \Omega_h$ satisfy $\cup_{K \in \Omega_h} \overline{K} = \bar{\Omega}$. To highlight the discontinuous qualifier of DPG, we let $Y(K)$ for an arbitrary element $K \in \Omega_h$ denote a Hilbert space with corresponding inner product $(\cdot, \cdot)_{Y(K)}$. Then an ideal DPG method is an ideal PG method (2.23) which uses a Hilbert space Y of the form [21]

$$Y = \prod_{K \in \Omega_h} Y(K) \tag{2.26}$$

with the corresponding inner product

$$(y, v)_Y = \sum_{K \in \Omega_h} (y|_K, v|_K)_{Y(K)} \quad \forall y, v \in Y.$$

In the definition above, $y|_K$ for $y \in Y$ refers to the restriction of y to its $Y(K)$-component. A common example of such a space Y is $H^1(\Omega_h) = \{v \in L^2(\Omega) : v|_K \in H^1(K) \forall K \in \Omega_h\}$ [21, 24]. Note that continuity of functions in $H^1(\Omega_h)$, for example, is not specified for elements $K_1, K_2 \in \Omega_h$ which share a common edge along their boundaries. This makes apparent the qualifier *discontinuous* in the DPG method: Since continuity along elements sharing an edge is not explicitly enforced, there is certainly no need to require or expect it in general. Another interesting consequence is that the definition of Y allows the computation of the action of the trial-to-test operator T to be done locally on each element, and independent of other elements [21].

2.4.4 A Practical DPG Method

In practice, to compute an optimal test space Y_h from X_h, we require an application of the trial-to-test operator T, which means applying an operator that normally acts on infinite-dimensional spaces. While some worked examples in [21] have closed-form expressions for T, we cannot expect this to be the case in general. Hence, a practical approach is needed: We will instead work with an approximation to the ideal test space by considering, instead of the Hilbert space Y, a finite-dimensional subspace $Y^r \subset Y$, where r is related to the dimension of Y^r [21]. To this end, we define a new trial-to-test operator $T^r : X_h \to Y^r$ by $(T^r w, y)_Y = b(w, y)$ for all $y \in Y^r$. Then a DPG method for solving (2.17) is given by the following [21]: Find $x_h \in X_h$ solving

$$b(x_h, y_h) = l(y_h) \quad \forall y_h \in Y_h^r, \tag{2.27}$$

where $Y_h^r = T^r(X_h)$.

As in the previous section, we can also define a norm that incorporates the trial-to-test operator T^r for the practical DPG method. Indeed, define the norm $|||x|||_r = ||T^r x||_Y$ for $x \in X_h$ [21]. Next, let $R_{Y^r} : Y^r \to (Y^r)^*$ denote the Reisz map defined by $R_{Y^r}(y)(v) = (y, v)_Y$ for all $y, v \in Y^r$, and let $B : X \to Y^*$ be defined by $Bx(y) = b(x, y)$ as before [21]. Analogous to how we defined T in relation to B and the Riesz map R_Y earlier, we have that that $T^r = R_{Y^r}^{-1} \circ B$, which we verify from the definition of T^r and the Riesz map R_{Y^r}:

$$\begin{aligned} R_{Y^r}(T^r z)(v) &= (T^r z, v)_Y \\ &= b(z, v) \\ &= Bz(v), \end{aligned}$$

hence $R_{Y^r} \circ T^r = B$, from which $T^r = R_{Y^r}^{-1} \circ B$ follows since R_{Y^r} is an isometric isomorphism beteen Y^r and $(Y^r)^*$ [21]. With these tools in hand, we have an analogous best approximation error result and residual minimization result for the practical DPG method, with $|||\cdot|||_X$ and $||\cdot||_{Y^*}$ replaced by $|||\cdot|||_r$ and $||\cdot||_{(Y^r)^*}$ when $x_h \in X_h$ uniquely solves (2.27) [21, Theorem 37].

Next, we wish to state a result to show that the variational formulation (2.27) is equivalent to a mixed finite element method that explicitly incorporates an error estimation term. We give its definition below.

Definition 1. (The Practical DPG Method Error Estimator [21]) Let x_h solve (2.17). The quantity $\varepsilon^r = R_{Y^r}^{-1}(l - Bx_h)$ is the error estimator of an $x_h \in X_h$, and is the unique element of Y^r satisfying $(\varepsilon^r, y)_Y = l(y) - b(x_h, y)$.

Let us return to the problem (2.17), for which we assume theorem 1 holds. For an x_h

solving (2.27), define the linear functional $\tilde{l} : (Y^r)^* \to \mathbb{C}$ by $\tilde{l}(y) = l(y) - b(x_h, y)$ (i.e. $\tilde{l} = l - Bx_h$), which is bounded and continuous since l and b are themselves bounded and continuous. Hence, there is a unique $\varepsilon^r \in Y^r$ solving

$$(\varepsilon^r, y)_Y = \tilde{l}(y) \quad \forall y \in Y^r. \tag{2.28}$$

by the Riesz representation theorem [10]. That ε^r takes its specific form comes from directly from (2.28), which immediately implies

$$R_{Y^r}(\varepsilon^r) = l - Bx_h \Longleftrightarrow \varepsilon^r = R_{Y^r}^{-1}(l - Bx_h) \in Y^r.$$

With this result in hand, we can now state the equivalence of (2.27) to a mixed method incorporating the error estimator ε^r.

Theorem 5. (Equivalence of the practical DPG method and a mixed formulation [21])

The following are equivalent statements.

i) $x_h \in X_h$ solves the DPG method (2.27).

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

$$(\varepsilon^r, y)_Y + b(x_h, y) = l(y) \quad \forall y \in Y^r, \tag{2.29a}$$

$$b(z_h, \varepsilon^r) = 0 \quad \forall z_h \in X_h. \tag{2.29b}$$

Proof. Suppose that *i)* holds. Previous computations show that

$$(\varepsilon^r, y)_Y + b(x_h, y) = l(y) \quad \forall y \in Y_h^r,$$

so it remains to verify (2.29b). Indeed, we have by the definition of T^r that

$$\begin{aligned} b(z_h, \varepsilon^r) &= (T^r z_h, \varepsilon^r)_Y \\ &= (T^r z_h, R_{Y^r}^{-1}(l - Bx_h))_Y \\ &= (T^r z_h, T^r(x - x_h))_Y \end{aligned}$$

Note, however, that $(T^r z_h, T^r(x - x_h))_Y = \overline{(T^r(x - x_h), T^r z_h)_Y} = \overline{b(x - x_h, T^r z_h)}$, which by *i)* and application of theorem 1 is equal to zero.

Now suppose that *ii)* holds. To show that this implies *i)*, it suffices to show that $(\varepsilon^r, y)_Y = 0$ for all $y \in Y_h^r$. Indeed, we have that

$$\begin{aligned} 0 &= b(z_h, \varepsilon^r) \\ &= (T^r z_h, \varepsilon^r)_Y \\ &= (y, \varepsilon^r)_Y \end{aligned}$$

where $T^r z_h = y \in Y_h^r$, and hence $b(x_h, y) = l(y)$ for all $y \in Y_h^r$. □

2.5 Perfectly Matched Layers

Recall that one of our goals is to accurately compute leaky modes and their propagation constants for problem (1.1). For leaky modes, the desired propagation constants β are complex-valued, and the imaginary part of the propagation constants are used to compute confinement losses for step-index and microstructure fibers [27, 50]. Consequently, a Perfectly Matched Layer (PML) provides a first step in tackling such

a problem. One consequence of using PML is that the problem we seek to solve mathematically is no longer self-adjoint; in the case of a frequency-dependent PML seen in Chapter 5 and [27, 46], the eigenproblem we wish to solve is no longer linear. As a result, the use of PML in conjuction with pushing the limits of our numerical discretization motivates the task of accurately computing propagation constants and leaky modes for problem (1.1). Such an endeavor culminates in the development of the extension of the FEAST algorithm to solve polynomial eigenvalue problems. As we see in Chapters 5, this arises from a frequency-dependent formulation of a perfectly matched layer as in [46], where the complex coordinate transformation used in creating the PML is dependent upon the eigenvalue being computed.

The method of Perfectly Matched Layers was originally developed by Berenger for Maxwell's equations [2], and many works use perfectly matched layers for solving problems such as the computation of scattering resonances, fiber bending, thin membrane photonics, and confinement losses of microstructure fibers [25,46,50,57]. Collino and Monk's work [9] shows that PML can be thought of as a complex coordinate transformation for problems posed on unbounded domains. In their treatment of a time-harmonic scattering problem, for example, they seek to find outgoing solutions for the magnetic field $\hat{H}$ and eigenvalues k^2 satisfying $\Delta\hat{H} + k^2\hat{H} = 0$ in $\mathbb{R}^2 \setminus \overline{\Omega}$, where $\Omega \subset \mathbb{R}^2$ is a smooth, bounded domain [9]. To tackle such a problem, the radial distance ρ from the computational domain Ω is recast as $\tilde{\rho}$ using a complex coordinate transformation, and solutions as $\rho \to \infty$ are represented using Hankel functions. Under the right conditions on the imaginary part of $k\tilde{\rho}$, the outgoing solutions become exponentially decaying solutions as $\rho \to \infty$, allowing the truncation of the unbounded domain to one that is finite in size. At this stage, one could impose zero dirichlet boundary conditions on this finitely sized domain and apply numerical methods such

as finite elements to solve the truncated problem.

In Chapter 6, we will compare our implementation of the frequency-dependent PML to the built in PML that comes with the NGSolve software package, which we use in conjunction with the standard linear FEAST algorithm. In NGSolve, the PML is implemented via a complex mesh deformation $\tilde{x}(x) = x + i\alpha d(x)$, where d is some function measuring the distance from the center of our computational domain [59]. The decay strength $\alpha > 0$ is such that for larger α, the decay of the computed solution in the PML region is more rapid. Moving forward, we refer to this as the NGSolve auto PML.

Chapter 3

The FEAST Algorithm for Eigenproblems

3.1 Introduction

When solving eigenvalue problems, we typically are interested in finding a $\lambda \in \mathbb{C}$ and $x \in \mathbb{C}^n$ satisfying

$$Ax = \lambda Bx \tag{3.1}$$

for $A, B \in \mathbb{C}^{n \times n}$. Depending on the structure of the matrices A and B, one typically has a variety of tools at their disposal for solving such problems. In the case we wish to compute all eigenvalues and eigenvectors for (3.1), we can use algorithms such as the QZ method, or Krylov methods such as Arnoldi if we wish to compute eigenvalues on the extreme ends of the spectrum for the matrix pair (A, B) [64, 65]. If, however, we wish to compute a subset of eigenvalues and eigenvectors at an arbitrary location in the complex plane, the FEAST algorithm allows us to do so with savings in expended computational resources. This is important for our motivating problem in fiber optics, as the number of unknowns (and hence the matrix dimensions) needed to compute propagation constants to high precision is on the order of $O(10^6)$. To motivate our discussion of the FEAST algorithm, we briefly discuss two algorithms designed for

computing eigenvalues and their corresponding eigenspaces: The power iteration and subspace iteration. While there exists a wealth of information on various iterative techniques and factorizations for computing eigenvectors and eigenvalues, we focus on two algorithms here to keep the discussion brief before introducing the FEAST eigensolver.

3.2 Subspace Iteration

3.2.1 A Motivating Algorithm: Power Iteration

Before taking a deeper look at FEAST, we look first at the concept of subspace iteration through the lense of a motivating algorithm, the power iteration [54,64,65]. Under certain assumptions, this algorithm finds an approximation to the eigenvector whose eigenvalue λ of A is is the largest in magnitude of all eigenvalues of $A \in \mathbb{C}^{n\times n}$ (see algorithm 1).

Algorithm 1 Power iteration

Matrix $A \in \mathbb{C}^{n\times n}$, initial vector $v_0 \in \mathbb{C}^n$, stopping tolerance $\epsilon > 0$.

```
for i = 1, 2, ...
    y = A v_{i-1}
    v_i = y/||y||
    if ||v_i − v_{i-1}||/||v_i|| < ε
        stop
    endif
endfor
```

In Algorithm 1, we simply repeat the process of applying A to an initial guess, normalize the iterate y using a norm $||\cdot||$ such as the vector norm $||\cdot||_2$ or $||\cdot||_\infty$, and repeat. Assuming the right conditions for the algorithm to terminate, we recover the approximate eigenvector v_i for some $i \geq 1$ and corresponding (approximate) eigenvalue $\lambda_i = v_i^* A v_i/(v_i^* v_i)$. Viewed through the lense of subspace iteration, we are

computing the vector whose span is the subspace span($\{v_i\}$). The termination of the algorithm, however, is another matter for discussion, so let us assume for now that the eigenvalues of A are semisimple. This means for any given eigenvalue λ of A, the algebraic multiplicity $m \geq 1$ of λ is the same as the number m of linearly independent eigenvectors that span the eigenspace corresponding to λ. In addition, we assume that the n eigenvalues of A satisfy $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. In the case that $|\lambda_1| > |\lambda_2|$, we call λ_1 the dominant eigenvalue of A [65]. Then we can expand any initial guess, say $v_0 \in \mathbb{C}$, as a linear combination of the eigenvectors $x_1, x_2, \ldots, x_n \in \mathbb{C}^n$ of A [65]:

$$v_0 = c_1 x_1 + c_2 x_2 + \ldots + c_n x_n.$$

Upon applying A and normalizing by λ_1, we have

$$\begin{aligned} v_1 &= \frac{1}{\lambda_1} A v_0 \\ &= c_1 \frac{1}{\lambda_1} A x_1 + c_2 \frac{1}{\lambda_1} A x_2 + \ldots + c_n \frac{1}{\lambda_1} A x_n \\ &= c_1 x_1 + c_2 \frac{\lambda_2}{\lambda_1} x_2 + \ldots + c_n \frac{\lambda_n}{\lambda_1} x_n. \end{aligned}$$

Repeating this process j times for $j \geq 1$ yields

$$v_j = \frac{1}{\lambda_1} A v_{j-1} = c_1 x_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^j x_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^j x_n$$

In order to make our analysis easier, we assume further that λ_1 is the dominant eigenvalue of A, and we show that $v_j \to c_1 x_1$ as $j \to \infty$. To this end, we have that in any vector norm $||\cdot||$ and $j \geq 0$,

$$
\begin{aligned}
||v_j - c_1 x_1|| &= \left|\left| c_1 x_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^j x_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^j x_n - c_1 x_1 \right|\right| \\
&= \left|\left| c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^j x_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^j x_n \right|\right| \\
&\leq |c_2| \left|\frac{\lambda_2}{\lambda_1}\right|^j ||x_2|| + |c_3| \left|\frac{\lambda_3}{\lambda_1}\right|^j ||x_3|| + \ldots + |c_n| \left|\frac{\lambda_n}{\lambda_1}\right|^j ||x_n|| \\
&\leq C \left|\frac{\lambda_2}{\lambda_1}\right|^j
\end{aligned}
$$

where

$$
C = \sum_{j=2}^{n} |c_j| \cdot ||x_j||.
$$

Then as $j \to \infty$, $||v_j - c_1 x_1|| \to 0$. In many applications, however, we often wish to compute multiple eigenvalues, often from the extreme ends of the spectrum, such as the largest or smallest eigenvalues, or perhaps somewhere between the extremes. This last idea is motivated by problems that come directly from the optics literature [43], but there are further challenges to address. To do so, we turn to subspace iteration.

3.2.2 Subspace Iteration at a Glance

Again, we return to problem (3.1). Our interest is now to compute a subset of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ for $k < n$. In problems where the matrices A and B are large and sparse, we often assume $k \ll n$, especially as the dimension n grows arbitrarily large. For larger problems, it becomes infeasible to attempt to compute most or all of the eigenvalues corresponding to (3.1); indeed, error estimates for eigenvalues grow as a function of the eigenvalue we are trying to find [41, §6.2, Theorem 6.7].

Returning to the topic of subspace iteration, our goal is to compute eigenvalues $\Lambda := \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ for $k < n$ and corresponding eigenvectors $v_1, v_2, \ldots, v_k$ satisfying (3.1). In the simplest case, assume that $B = I$, the $n \times n$ identity matrix. Then the subspace iteration algorithm reads as follows (see, for example, [54]):

Algorithm 2 Subspace iteration

Matrix $A \in \mathbb{C}^{n \times n}$, initial vectors $v_1, v_2, \ldots, v_k \in \mathbb{C}^n$ stored as columns in $V_0 \in \mathbb{C}^{n \times k}$.

1 **for** $i = 1, 2, \ldots$
2 **compute** $Y = AV_{i-1}$
3 **orthonormalize** Y**, set** $V_i = Y$
4 **endfor**

In detail, the algorithm is as follows: Step 2 applies the matrix A to an initial guess to the vectors in V_0 spanning the desired eigenspace $E = \{v \in \mathbb{C}^n : Av = \lambda v \,\text{for}\, \lambda \in \Lambda\}$. Step 3 reorthogonalizes the resulting matrix Y through a technique such as the QR method [54, 64], and the iteration repeats. The criteria for termination have been purposely omitted, as this requires a more careful discussion for measuring the distance between subspaces. Several works address and apply such a concept using the gap metric between subspaces of a Banach space (see [23,24,35], for example). In the meantime, we will refer to such a metric as d in the same manner as Watkins [65], taking its existence for granted to state a meaningful result from the same author about the convergence of subspace iteration. We state this theorem without proof [65, Theorem 6.2.3].

Theorem 6 (Convergence of subspace iteration)**.** Let $A \in \mathbb{F}^{n \times n}$ be semisimple with linearly independent eigenvectors $v_1, \ldots, v_n \in \mathbb{F}^n$ and associated eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{F}$, satisfying $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. Suppose $\lambda_k > \lambda_{k+1}$ for some k. Let $\mathcal{U}_k = \text{span}\{v_1, \ldots, v_k\}$ and $\mathcal{V}_k = \text{span}\{v_{k+1}, \ldots, v_n\}$. Let $\mathcal{S}$ be any k-dimensional

subspace of $\mathbb{F}^n$ such that $\mathcal{S} \cap \mathcal{V}_k = \{0\}$. Then there is a constant C such that

$$d(A^j\mathcal{S}, \mathcal{U}_k) \leq C \left| \frac{\lambda_{k+1}}{\lambda_k} \right|^j, j = 0, 1, 2, \ldots \tag{3.2}$$

Thus, $A^j\mathcal{S} \to \mathcal{U}_k$ linearly with convergence ratio $|\lambda_{k+1}/\lambda_k|$.

Of note is the expression $A^j\mathcal{S}$ in the statement of theorem 3.2. For a subset $S \subset \mathbb{C}^n$, the set AS is given by

$$A\mathcal{S} = \{Ax : x \in \mathcal{S}\},$$

and likewise $A^j\mathcal{S} = \{A^j x : x \in \mathcal{S}\}$ for $j \geq 0$. In addition, we see that the convergence rate of the iterates V_i of subspace iteration to the desired eigenspace is dictated by the ratio between the first eigenvalue that we do not want to the smallest eigenvalue that we do want. Making this ratio as small as possible is an important task, and algorithms such as FEAST work to realize this possibility.

In the case that subspace iteration converges, we must also state precisely to what vectors the iterates V_i in Algorithm 2 converge. To begin, let $X_0 = [x_1, x_2, \ldots, x_k] \in \mathbb{C}^{n\times k}$ be the matrix whose columns are initial guesses to the vectors spanning the desired eigenspace, and let $Q = [q_1, q_2, \ldots, q_k] \in \mathbb{C}^{n\times k}$ be the Schur vectors associated to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Recall also that for $A \in \mathbb{C}^{n\times n}$, the Schur factorization of A is given by [64]

$$A = QTQ^*$$

for an upper triangular matrix T and unitary matrix U. In the case of a matrix A whose eigenvalues and eigenvectors we wish to find, the eigenvalues of A lie on the

diagonal of T, and for any eigenvector x of A, the corresponding eigenvectors q of $Tq = \lambda q$ are related to the eigenvectors x solving $Ax = \lambda x$ by $q = Q^*x$. Under similar conditions as those in theorem 6, Saad shows that the iterates V_i in Algorithm 2 converge column-by-column to the corresponding Schur vectors, each up to multiplication by a possibly different complex phase factor $e^{i\theta}$ (see [54][Theorem 5.1, §5.1]).

Returning to subspace iteration, orthonormalization can be an expensive process to perform every iteration. Saad, for example, proposes a modification to algorithm 2 that computes $Z = A^{iter}X$ after some nontrivial number of iterations *iter*, re-orthogonalizes, and then repeats this process. Care must be taken to choose *iter* appropriately to balance the work done with orthogonalization versus the possibility that the column vectors of Z become linearly dependent [54]. We can take algorithm 2 a step further by considering Saad's proposal on subspace iteration with projection, either by computing a (small) $k \times k$ Schur factorization or eigendecomposition to recover the eigenvalues and update the approximation to the eigenvectors [54].

The primary takeaway from both methods presented here is that we require a strict separation of the eigenvalues we wish to compute from those we wish to ignore. While several strategies for computing eigenvalues and eigenvectors exploit this, we often wish to take this idea one step further by transforming the spectrum we compute, often to the advantage of hastening convergence of the method. Furthermore, we may wish to compute some subset of the eigenvalues of our problem of interest that do not necessarily lie on the extreme ends of the spectrum. To this, we turn to the FEAST algorithm.

3.3 The FEAST Algorithm

Suppose one wishes to compute some subset of eigenvalues from the problem (3.1). Many problems of interest often involve $A, B \in \mathbb{C}^{n\times n}$ that are large and sparse, often arising from the discretization of a partial differential equation using, say, finite differences, finite volumes, or the finite element method. Since the dimenion n of A and B can be upwards of $n = O(10^6)$ to accurately compute numerical solutions to a partial differential equation, it becomes infeasible to attempt to compute every eigenvalue. Often, the goal is to compute a subset of the eigenvalues (and corresponding eigenvectors) of such a problem, up to the algebraic multiplicity of the eigenvalues and their corresponding geometric multiplicities. To address this, several contour-integral based methods have risen to the challenge, often addressing differences in structure of the matrices A and B in (3.1). One such algorithm we address in this section is the FEAST algorithm developed by Eric Polizzi and others [19, 31, 36, 37, 51, 62].

The idea behind FEAST is to transform the eigenvalues such that the corresponding eigenspace is the dominant eigenspace, all while ignoring unwanted eigenvalues. This algorithm has been applied to several types of problems, including self-adjoint, non-self-adjoint, and nonlinear eigenvalue problems [17,18,34,37,51,62]. For non-self-adjoint problems, FEAST computes the left and right eigenspaces for the generalized eigenproblem of finding eigenpairs (λ, x) solving $Ax = \lambda Bx$. In addition to finding desired eigenspaces, non-selfadjoint FEAST contains a number of heuristics for mainting bi-orthogonality of the approximate left and right eigenspaces [37].

The FEAST algorithm has been applied to a number of problems for finding eigenvectors and eigenvalues of dense matrices [51], as well as sparse matrices that discretize

partial differential equations using classical finite elements and the Discontinuous Petrov-Galerkin (DPG) Method [23,24]. The latter work has been extended to problems containing a reaction-diffusion term, an important generalization for finding the eigenvalues and eigenfunctions corresponding to the propagation constants and leaky modes of optical fibers. A detailed analysis of the convergence of FEAST using classical finite elements and the DPG Method can be found in [23,24]. Chapter 4 focuses on a recently published work [24] that highlights the use of the DPG method within the FEAST algorithm.

Now recall for subspace iteration that we require a strict separation between the eigenvalues we wish to compute and those we wish to ignore. The idea behind the FEAST algorithm is to map the desired eigenvalues to a value much larger than the eigenvalues we wish to ignore. In essence, this idea is used ensure rapid convergence of, say, subspace iteration to the desired eigenvalues and corresponding eigenspaces. The inspiration for this method comes from Cauchy's integral formula [62]:

$$f(\xi) = \frac{1}{2\pi i} \oint_{\Gamma} \frac{1}{z-\xi} dz \tag{3.3}$$

Here, we assume Γ is a simple, closed loop such as a circle of radius $\gamma > 0$, and $\xi \in \mathbb{C}$. Letting $U \subset \mathbb{C}$ be the bounded, open set contained in the interior of Γ and such that $\partial U = \Gamma$, we see that $f(\xi)$ approximates the characteristic function $\chi_U(\xi)$. In this case, $f(\xi) = 1$ for $\xi \in U$ and $f(\xi) = 0$ on $\mathbb{C} \setminus \overline{U}$, whereas $\chi_U(\xi) = 0$ for $\xi \in \mathbb{C} \setminus U$.[1] This motivates the definition of the spectral projector [35,51], the main computational tool of the FEAST algorithm:

[1] In the case that $\xi \in \Gamma$, $f(\xi)$ takes the Cauchy principal value 1/2 (see, for example, [55]).

$$S = \frac{1}{2\pi i} \int_{\Gamma} (z - A)^{-1} dz \tag{3.4}$$

As before, Γ is a simple, closed countour, forming the boundary of an open set $U \subset \mathbb{C}$ that contains eigenvalues of the matrix A, and $R(z) := (z - A)^{-1}$ is the *resolvent* operator for the matrix $A \in \mathbb{C}^{n \times n}$. It is further assumed that no eigenvalues of A lie on the contour Γ. In practice, however, S is an infinite-dimensional object, so we have to approximate S using a quadrature rule. Using Cauchy's integral formula to motivate this, we first approximate the contour integral (3.3) by a rational function $r_N(\xi)$ using a suitable N-point quadrature rule for $N \geq 1$ specified by weights and nodes w_j, z_j for $i = 0, 1, \ldots, N - 1$:

$$r_N(\xi) = \sum_{j=0}^{N-1} w_j \frac{1}{z_j - \xi} \tag{3.5}$$

Typical choices of quadrature rules include Gaussian or trapezoidal quadrature [24], though other choices are certainly possible. If Γ, for example, is a circle in the complex plane with radius $\gamma > 0$ and center $y \in \mathbb{C}$, then the corresponding wieghts w_j and nodes z_j for $j = 0, 1, \ldots, N - 1$ for the trapezoid rule are given by [23]

$$w_j = \frac{\gamma}{N} e^{i\theta_j} \tag{3.6a}$$

$$z_j = y + \gamma e^{i\theta_j} \tag{3.6b}$$

where $\theta_j = \frac{2\pi j}{N}$ for $j = 0, 1, \ldots, N - 1$. In practice, this means approximating S using a rational function of the matrix A, shown below:

$$r_N(A) = \sum_{j=0}^{N-1} w_j (z_j - A)^{-1} \tag{3.7}$$

In works such as [31, 51, 62], $S_N = r_N(A)$ is an approximation of S using Gaussian quadrature, and is also denoted $r_N(M)$ for $M = B^{-1}A$ when the matrix B in (3.1) is invertible. In this case, we have that the approximation to the corresponding integral

$$S = \frac{1}{2\pi i} \int_C (z - B^{-1}A)^{-1} dz = \frac{1}{2\pi i} \int_C (zB - A)^{-1} B dz \tag{3.8}$$

is given by

$$r_N(M) = \sum_{j=0}^{N-1} w_j (z_j B - A)^{-1} B. \tag{3.9}$$

Note that this form of r_N is chosen even when B is not invertible. Moving forward, we provide a summary of the FEAST algorithm below, which can be found in numerous works such as [31, 51, 62].

First, we have that step 2 of Algorithm 3 applies the approximation S_N of the spectral projector S to the current iterate Q_{i-1}. Then steps 3-4 solve a small eigenvalue problem for which the approximate eigenvalues are computed, and the corresponding eigenvectors $W \in \mathbb{C}^{k \times k}$ are used to update the next iterate Q_i whose column span is closer to that of the desired eigenspace. At each iteration, we check for convergence, and repeat the iterations if stopping criteria for the algorithm have not been met.

To ensure convergence of the FEAST method, we recall an important theorem from

Algorithm 3 The FEAST Algorithm

Matrix $A \in \mathbb{C}^{n\times n}$, initial guess $Q_0 \in \mathbb{C}^{n\times k}$.

for $i = 1, 2, \ldots$
    compute $Y = S_N Q_{i-1}$
    compute $A_Y = Y^* A Y$**,** $B_Y = Y^* B Y$
    compute $W, \Lambda_Y \in \mathbb{C}^{k\times k}$ **solving** $A_Y W = B_Y W \Lambda_Y$ **for which** $W^* B_Y W = I$
    set $Q_i = YW$**,** $\Lambda = \Lambda_Y$
    if Q_i **has converged**
        return Q_i, Λ
    endif
endfor

[62]. To begin, observe that for an eigenpair (λ, x) of (3.1), the corresponding eigenvalue of $r_N(B^{-1}A)$ (assuming for now that B is invertible) is mapped to $r_N(\lambda)$ [62]. Furthermore, denote the mapped eigenvalues by $\gamma_j = r_N(\lambda_j)$ for $i = 1, 2, \ldots, k$, and let $\mathcal{Q}_i = \text{span}(Q_i)$, the column span of the iterate $Q_i \in \mathbb{C}^{n\times k}$ in Algorithm 3. Furthermore, let eigenvalues λ_i be ordered such that the mapped eigenvalues satisfy $|\gamma_1| \geq |\gamma_2| \geq \ldots \geq |\gamma_k| > |\gamma_{k+1}| \geq \ldots \geq |\gamma_n|$. Then by [62, §4, Theorem 4.1], then there is a vector $s_j \in \mathcal{Q}_i$ satisfying $||x_j - s_j||_B \leq \alpha|\gamma_{k+1}/\gamma_j|^i$, where $||y||_B := \sqrt{yB^*y}$. In particular, if $P_{(i)}$ is the B-orthogonal projector onto $\mathcal{Q}_i$, then we have that $||(I - P_{(k)})x_j||_B \leq \alpha|\gamma_{k+1}/\gamma_j|^i$ [62]. Effectively, this result states that as the FEAST iterations proceed, we can find a vector in the column span of the iterates that is arbitrarily close to a desired eigenvector, and approaches such a vector at a rate $|\gamma_{k+1}/\gamma_j| \ll 1$.

Moving forward, we will see the efficacy of the the FEAST algorithm as it is applied to self-adjoint and nonlinear eigenvalue problems in subsequent chapters. While our exposition here primarily deals with matrices, we will see in the next chapter a means of tying together discretization errors in the finite element method to the numerical

accuracy of the FEAST algorithm when applied to solving eigenvalue problems for partial differential equations.

Chapter 4

DPG Discretization Errors in FEAST

4.1 Introduction

In this chapter, we visit a more abstract formulation of the problem (1.1). Such a problem has been tackled in [23] using classical finite elements, and error bounds bounds for eigenspace and eigenvalue approximations were derived using tools such as the gap metric for distances between subspaces of Banach spaces [23, 35]. In our recent work [24], we revisit the work of [23], but with an application of the DPG method to finding the guided modes and propagation constants of a step-index fiber. As mentioned in our recent work, the focus is to develop an understanding of the discretization error when approximating the resolvent operator $z \mapsto R(z) = (z - A)^{-1}$ using the DPG discretization, and where A is a differential operator. As we shall see, the error bounds obtained largely depend on the closeness of unwanted eigenvalues to the desired eigenvalues computed using the FEAST algorithm.

The problem of interest is to compute portions of the spectrum and corresponding eigenspaces of the unbounded operator $A = -\Delta - \nu$ in $L^2(\Omega)$, where $\Omega \subset \mathbb{R}^2$ is bounded with Lipschitz boundary, $\nu \in L^\infty(\Omega)$, and the domain of A is taken to be $H_0^1(\Omega)$. In practice, we will take ν to be piecewise constant, as it will represent the index of refraction in the applications of interest described later. The domain of Ω

for practical applications is circular, as the problem domains of interest will be the cross-sections of optical fibers.

As discussed in chapter 3, the FEAST eigensolver has been applied to several problems of interest in physics and engineering. As is done in [23, 34], the goal is to bring in more granularity into the error analysis when approximations to the resolvent arise from a spatial discretization such as the finite element method or the use of spectral methods. Given the context in which the problem we seek to solve arises, namely finding the (approximate) eigenspaces of an unbounded differential operator on an infinite-dimensional space, it is important to understand how the discretization directly affects the application of the FEAST algorithm. In the following sections, we will introduce the reader to abstract framework, applications, numerical results, and relevant theory.

4.2 The Abstract Framework

We summarize the abstract framework of the FEAST algorithm in this section as given in [23, 27]. We begin by letting A be a linear, closed, selfadjoint operator $A : \mathrm{dom}(A) \to \mathcal{H}$ where $\mathcal{H}$ is a complex Hilbert space and $\mathrm{dom}(A) \subseteq \mathcal{H}$. In addition, we denote the (real) spectrum of A by $\Sigma(A)$. Our goal is to approximate a subset $\Lambda \subset \Sigma(A)$ and the corresponding eigenspace E. Of note is that the set Λ consists of a finite set of eigenvalues with finite multiplicity.

The FEAST algorithm uses a rational function

$$r_N(\xi) = w_N + \sum_{k=0}^{N-1} \frac{w_k}{z_k - \xi}, \tag{4.1}$$

where the choices of $w_k, z_k \in \mathbb{C}$ are chosen by a quadrature rule such as gaussian or trapezoidal quadrature rules. The motivation for this choice is to compute approximations to the Dunford-Taylor contour integrals of the form

$$S = \frac{1}{2\pi i} \oint_\Gamma R(z)dz \tag{4.2}$$

where $R(z) = (z - A)^{-1}$ is the resolvent of the operator A at $z \in \mathbb{C}$. In addition, Γ is a positively-oriented, simple, closed contour that surround the elements of Λ and excludes $\Sigma(A) \setminus \Lambda$; this operator S is referred to as the spectral projector onto the eigenspace E.

To make our endeavour computationally tractable, we define a quadrature approximation S_N to S below.

$$S_N = r_N(A) = w_N + \sum_{k=0}^{N-1} w_k R(z_k). \tag{4.3}$$

Since S_N is still an infinite-dimensional object (due to the presence of A in $R(z_k) = (z_k - A)^{-1}$), we go one step further with an approximation to S_N by S_N^h, where $h > 0$ is a parameter related to the discretization of $(z_k - A)$ for each $k = 0, 1, \ldots, N-1$; in practice, h is the mesh size of a triangulation of some bounded, finite domain $\Omega \subset \mathbb{R}^2$. We define S_N^h by

$$S_N^h = w_N + \sum_{k=0}^{N-1} w_k R_h(z_k), \tag{4.4}$$

where $R_h : \mathcal{H} \to \mathcal{V}_h$ is a finite-rank approximation of $R(z)$ and $\mathcal{V}_h$ is a subspace of a complex Hilbert space $\mathcal{V}$ that is embedded (continuously) in $\mathcal{H}$. As we shall see later, there is no need to assume that the resolvent approximations yield a self-adjoint S_N^h.

To this end, we create an analogous FEAST algorithm using the approximations S_N^h to S as follows: Let $E_h^{(0)} \subset \mathcal{V}_h$ be an initial approximation to the desired eigenspace E. Compute the subspace iterates

$$E_h^{(\ell)} = S_N^h E_h^{(\ell-1)}, \quad \ell = 1, 2, \ldots \tag{4.5}$$

until a desired convergence criterion is met for some $\ell \geq 0$. In the case that A is selfadjoint and finite-dimensional, say $A \in \mathbb{C}^{n\times n}$ is hermitian, then the discussion of a discretization parameter h is moot, as S_N can be used directly. While this is feasible for many applications, our exploration stresses the importance of context. Since we have to use a discretization to solve a continuous problem, it is important to measure how that discretization affects the results we obtain, and more importantly, to measure how close the FEAST approximations to eigenvalues and eigenspaces are to the desired eigenvalues and eigenspaces to the continuous problem. Since we will be working with hermitian (differential) operators A, however, we will use similar assumptions on the separation of Λ from the rest of the spectrum of A we wish to ignore, namely $\Sigma(A) \setminus \Lambda$. To begin, fix $y, \delta, \gamma \in \mathbb{R}$ such that $\gamma, \delta > 0$. Define the

inside and outside sets [23, 24]

$$I_\gamma^y = \{x \in \mathbb{R}^2 : |x - y| \leq \gamma\} \tag{4.6a}$$

$$O_{\delta,\gamma}^y = \{x \in \mathbb{R}^2 : |x - y| \geq (1 + \delta)\gamma\} \tag{4.6b}$$

and the quantities [23, 24]

$$W = \sum_{k=0}^{N} |w_k| \tag{4.7a}$$

$$\hat{\kappa} = \frac{\sup_{x \in O_{\delta,\gamma}^y} |r_N(x)|}{\inf_{x \in I_\gamma^y} |r_N(x)|} \tag{4.7b}$$

We now introduce important assumptions given in [23, 24] in order to introduce relevant results from our work. The first assumption states the desired set of eigenvalues Λ is contained in the inside set I_γ^y, that our chosen quadrature rule is well-behaved and such that the nodes z_k are neither elements of or limit points of the spectrum, and that our error reduction factor $\hat{\kappa}$ is smaller than unity.

Assumption 1. There are $y \in \mathbb{R}$, $\delta > 0$, and $\gamma > 0$ such that

$$\Lambda \subset I_\gamma^y, \quad \Sigma(A) \setminus \Lambda \subset O_{\delta,\gamma}^y, \tag{4.8}$$

and that r_N is a rational function of the form (4.4) with the following properties:

$$z_k \notin \overline{\Sigma(A)}, \quad W < \infty, \quad \text{and} \quad \hat{\kappa} < 1.$$

Our next assumption assumes a continuous embeddeding of Hilbert spaces, and that

the space $\mathcal{V}$ in which we seek solutions is an invariant subspace of $R(z)$ for all $z \in \rho(A)$, the resolvent set of A.

Assumption 2. The Hilbert space $\mathcal{V}$ is such that $E \subseteq \mathcal{V} \subseteq \mathcal{H}$, there is a $C_{\mathcal{V}} > 0$ such that for all $u \in \mathcal{V}$, $||u||_{\mathcal{H}} \leq C_{\mathcal{V}}||u||_{\mathcal{V}}$, and $\mathcal{V}$ is an invariant subspace of $R(z)$ for all z in the resolvent set of A, i.e., $R(z)\mathcal{V} \subseteq \mathcal{V}$.

The next assumption assumes that the error in the discretization of the resolvent $R(z)$ by $R_h(z)$ goes to zero across all quadrature points z_k as we approach the continuous problem ($h \to 0$).

Assumption 3. The operators $R_h(z_k)$ and $R(z_k)$ are bounded in $\mathcal{V}$ and satisfy

$$\lim_{h\to 0} \max_{k=0,\ldots,N-1} ||R_h(z_k) - R(z_k)||_{\mathcal{V}} = 0. \tag{4.9}$$

Assumption 4. Assume that $\mathcal{V}_h$ is contained in $\mathrm{dom}(a)$, where $a(\cdot,\cdot)$ denotes the symmetric (possibly unbounded) sesquilinear form associated to the operator A and $\mathrm{dom}(a)$ is the domain of a (see, for example, [35] and [23, §5]).

4.2.1 Consequences of Important Assumptions

To describe some of the consequences of our assumptions, let the desired eigenvalues consist of the finite set $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, counting multiplicities so that $m = \dim(E)$, the dimension of the desired eigenspace E. The assumptions given in section 4.2 yield some important consequences that one can find in [23]; we go over two of them here. First, Assumption 1 implies that

$$\sup_{x \in O^y_{\delta,\gamma}} |r_N(x)| < \inf_{x \in I^y_\gamma} |r_N(x)|,$$

and so we can find a simple, closed contour $\Gamma \subset \mathbb{C}$, the interior of which contains the

mapped eigenvalues $\mu = r_N(\lambda)$ for $\lambda \in I_\gamma^y$ and excludes eigenvalues $\mu = r_N(\lambda)$ for $\lambda \in O_{\delta,\gamma}^y$.

The consequence of Assumption 3 is that S_N^h converges to S_N in norm as $h \to 0$. Indeed,

$$\begin{aligned}
||S_N^h - S_N||_{\mathcal{V}} &= \left|\left|\sum_{k=0}^{N-1} w_k(R_h(z_k) - R(z_k))\right|\right|_{\mathcal{V}} \\
&\leq \sum_{k=0}^{N-1} |w_k| \cdot ||R_h(z_k) - R(z_k)||_{\mathcal{V}} \\
&\leq \left(\sum_{k=0}^{N-1} |w_k|\right) \max_{k=0,1,\ldots,N-1} ||R_h(z_k) - R(z_k)||_{\mathcal{V}} \\
&= W \max_{k=0,1,\ldots,N-1} ||R_h(z_k) - R(z_k)||_{\mathcal{V}}.
\end{aligned}$$

By Assumption 1, $W < \infty$, and so by Assumption 3, $||S_N^h - S_N||_{\mathcal{V}} \to 0$ as $h \to 0$. Then, for sufficiently small h, we have that [24]

$$P_h = \frac{1}{2\pi i} \oint_\Gamma (z - S_N^h)^{-1} dz$$

is the spectral projector associated with the contour Γ. Defining E_h to be the range of P_h, we revisit the feast iterations (4.5). For this problem, we also assume that $E_h^{(0)} \subseteq \mathcal{V}_h$ is chosen with $\dim E_h^{(0)} = \dim(P_h E_h^{(0)})$; in practice, this is not a restrictive assumption, as we often start with more vectors than the dimension of the desired eigenspace, and some vectors are removed as the iterations proceed.

In practice, FEAST computes a set of eigenvalue approximations and an approximation to the desired eigenspace E. In order to measure how far apart the approximate

eigenspace is from the desired eigenspace E, we use the gap metric, which measures distances between subspaces M and L of a Banach space $\mathcal{V}$ [24, 35]. This metric is defined by

$$\text{gap}_{\mathcal{V}}(M, L) = \max\left[\sup_{m \in U_M^{\mathcal{V}}} \text{dist}_{\mathcal{V}}(m, L), \sup_{l \in U_L^{\mathcal{V}}} \text{dist}_{\mathcal{V}}(l, M)\right], \tag{4.10}$$

where $U_M^{\mathcal{V}}$ (respectively, $U_L^{\mathcal{V}}$) is the unit ball in M (respectively, L) in the norm $||\cdot||_{\mathcal{V}}$ and

$$\text{dist}_{\mathcal{V}}(m, S) = \inf_{s \in S} ||m - s||_{\mathcal{V}}$$

for a subspace $S \subset \mathcal{V}$.

Remark 7. Intuitively, the gap between subspaces can be thought of as a generalization of the angle between two vectors, say, in $\mathbb{R}^2$. In that setting, the cosine of the angle between two vectors $u, v \in \mathbb{R}^2$ satisfies the well-known formula $\cos\theta = (u \cdot v)/(||u||_2||v||_2)$. The angle θ can be interpreted as measuring how "far apart" the spaces $M = \text{span}(\{u\})$ and $L = \text{span}(\{v\})$ are from each other.

In addition, we need to measure the accuracy of our approximate eigenvalues computed using the FEAST algorithm. To do so, define the set of approximations Λ_h to Λ by [24]

$$\Lambda_h = \{\lambda_h \in \mathbb{R} : \exists 0 \neq u_h \in E_h \text{ satisfying } a(u_h, v_h) = \lambda_h(u_h, v_h) \text{ for all } v_h \in E_h\}.$$

We measure how far apart Λ_h and Λ are using the Hausdorff distance, defined by [24]

$$\operatorname{dist}(\Upsilon_1, \Upsilon_2) = \max\left[\sup_{\mu_1 \in \Upsilon_1} \operatorname{dist}(\mu_1, \Upsilon_2), \sup_{\mu_2 \in \Upsilon_2} \operatorname{dist}(\mu_2, \Upsilon_1)\right],$$

where $\Upsilon_1, \Upsilon_2 \subset \mathbb{C}$ and $\operatorname{dist}(\mu, \Upsilon) = \inf_{\nu \in \Upsilon} |\mu - \nu|$ for any $\Upsilon \subset \mathbb{C}$. Next, let C_E be any constant sich that for all $e_1, e_2 \in E$, we have that $|a(e_1, e_2)| \leq C_E ||e_1||_{\mathcal{H}} ||e_2||_{\mathcal{H}}$. We state the following result from [24], the proof of which can be found in [23]. The important takeaway of this result is that as the FEAST iterations progress, the approximation error of the FEAST iterates $E_h^{(\ell)}$ to the space E_h can be made arbitrarily small, and similarly for the approximation E_h to the desired eigenspace E.

Theorem 8. Suppose Assumptions 1–3 hold. Then there are constants $C_N, h_0 > 0$ such that, for all $h < h_0$,

$$\lim_{\ell \to \infty} \operatorname{gap}_{\mathcal{V}}(E_h^{(\ell)}, E_h) = 0, \tag{4.11}$$

$$\lim_{h \to 0} \operatorname{gap}_{\mathcal{V}}(E, E_h) = 0, \tag{4.12}$$

$$\operatorname{gap}_{\mathcal{V}}(E, E_h) \leq C_N W \max_{k=0,\ldots,N-1} \left|\left|\left[R(z_k) - R_h(z_k)\right]\big|_E\right|\right|_{\mathcal{V}}. \tag{4.13}$$

If, in addition, Assumption 4 holds and $||u||_{\mathcal{V}} = |||A|^{1/2} u||_{\mathcal{H}}$, then there are $C_1, h_1 > 0$ such that for all $h < h_1$,

$$\operatorname{dist}(\Lambda, \Lambda_h) \leq (\Lambda_{max})^2 \operatorname{gap}_{\mathcal{V}}(E, E_h)^2 + C_1 C_E \operatorname{gap}_{\mathcal{H}}(E, E_h)^2, \tag{4.14}$$

where $\Lambda_{max} = \sup_{e_h \in E_h} |||A|^{1/2} e_h||_{\mathcal{H}} / ||e_h||_{\mathcal{H}}$ satisfies

$$(\Lambda_{max})^2 \leq \left[1 - \operatorname{gap}_{\mathcal{V}}(E, E_h)\right]^{-2} C_E.$$

4.3 Applications of the DPG Discretization

4.3.1 The Dirichlet Operator

As a first step, we look at the eigenvalues of the negative Laplacian $-\Delta$ for a zero-dirichlet eigenvalue problem. To this end, we define

$$A = -\Delta, \quad \mathcal{V} = H_0^1(\Omega), \quad \mathcal{H} = L^2(\Omega), \quad \text{dom}(A) = \{\psi \in H_0^1(\Omega) : \Delta\psi \in L^2(\Omega)\} \tag{4.15}$$

for a bounded, polyhedral, Lipschitz domain $\Omega \in \mathbb{R}^n$ for $n \geq 2$. For Sobolev spaces X, we use the standard notation $||\cdot||_X$ for norms and $|\cdot|_X$ for seminorms. To see that assumption 2 holds in this context, we revisit the discussion in [23]. We first observe that the form $a(\cdot,\cdot)$ arises from integration by parts for the problem of finding $\lambda \in \mathbb{C}$ and a funtion u satisfying

$$-\Delta u = \lambda u, \quad u|_{\partial\Omega} = 0.$$

Multipling the PDE above by $\bar{v}$ for $v \in \mathcal{V}$ and integrating by parts yields

$$\int_\Omega \nabla u \cdot \nabla \bar{v} dx = \lambda \int_\Omega u\bar{v} dx,$$

and so the form to which $A = -\Delta$ is associated is given by

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla \bar{v} dx, \quad u, v \in \mathcal{V}.$$

By the Poincaré inequality, we have that the norm $||u||_{\mathcal{V}}$ is equivalent to $||\,|A|^{1/2}u||_{\mathcal{H}} = ||A^{1/2}u||_{\mathcal{H}} = ||\nabla u||_{L^2(\Omega)} = |u|_{H^1\Omega}$ (further details can be found, for example, in [23, §5]).

Next, we need to make sure that, in leading up to the application of the resolvent $u = R(z)^{-1}v$, that one can solve the operator equation $(z - A)u = v$. Multiplying by a test function $w \in H_0^1(\Omega)$ and integrating by parts yields the problem of finding a $u \in \mathcal{V}$ satisfying

$$b(u, w) = (v, w)_{\mathcal{H}} \quad \text{for all} w \in \mathcal{V}, \tag{4.16}$$

where

$$b(u, w) = z(u, w) - a(u, w) = z \int_\Omega u\bar{w}dx - \int_\Omega \nabla u \cdot \nabla \bar{w}dx$$

for $u, w \in \mathcal{V}$. We begin with an inf-sup and continuity estimate for the form b, with the assumption that z is in the resolvent set of A.

Lemma 9. (An inf-sup condition for (4.16) [24]) For all $v \in H_0^1(\Omega)$,

$$\sup_{y \in H_0^1(\Omega)} \frac{|b(v, y)|}{|y|_{H^1(\Omega)}} \geq \beta(z)^{-1}|v|_{H^1(\Omega)},$$

where $\beta(z) = \sup\{|\lambda|/|\lambda - z| : \lambda \in \Sigma(A)\}$.

Proof. Following the work of [24], we begin by letting $v \in H_0^1(\Omega)$, and let $w = \bar{z}R(\bar{z})v$, i.e. let w solve $(\bar{z} - A)w = \bar{z}v$. Multiplying both sides by $\bar{s}$ for $s \in H_0^1(\Omega)$, we have that

$$\int_\Omega (\bar{z} - A)w\bar{s}dx = \bar{z}\int_\Omega w\bar{s}dx - \int_\Omega Aw\bar{s}dx = \bar{z}\int_\Omega w\bar{s}dx - a(w, s) = b(w, s)$$

on the left-hand-side and

$$\bar{z}\int_{\Omega} v\bar{s}dx = \bar{z}(v,s)_{\mathcal{H}}$$

on the right-hand side. Taking the complex conjugate of the left and right sides, we have that

$$z(s,w)_{\mathcal{H}} - a(s,w) = b(s,w) = z(s,v)_{\mathcal{H}}$$

for all $s \in H_0^1(\Omega)$. Choosing $s = v$, we have that

$$\begin{aligned} b(v, v-w) &= b(v,v) - b(v,w) \\ &= b(v,v) - z(v,v)_{\mathcal{H}} \\ &= z(v,v)_{\mathcal{H}} - a(v,v) - z(v,v)_{\mathcal{H}} \\ &= -a(v,v) \\ &= -|v|_{H^1(\Omega)}. \end{aligned}$$

In addition, we have that

$$\begin{aligned} v - w &= v - \bar{z}R(\bar{z})v \\ &= (R(\bar{z})^{-1} - \bar{z})R(\bar{z})v \\ &= (\bar{z} - A - \bar{z})R(\bar{z})v \\ &= -AR(\bar{z})v. \end{aligned}$$

Using [35, p. 273, Equation (3.17)], we have that $||AR(z)||_{\mathcal{H}} = \beta(z)$ holds for any z in

the resolvent set of A. Then since $|s|_{H^1(\Omega)} = ||A^{1/2}s||_{\mathcal{H}}$ for all $s \in H_0^1(\Omega) = \text{dom}(a) = \text{dom}(A^{1/2})$, and since $A^{1/2}$ commutes with $AR(z)$ (see the second representation theorem in [35, Theorem V2.23]), it follows that

$$\begin{aligned} |v - w|_{H^1(\Omega)} &= |AR(\bar{z})v|_{H^1(\Omega)} \\ &= ||A^{1/2}AR(\bar{z})v||_{\mathcal{H}} \\ &= ||AR(\bar{z})A^{1/2}v||_{\mathcal{H}} \\ &\leq \beta(\bar{z})||A^{1/2}v||_{\mathcal{H}} \\ &= \beta(\bar{z})|v|_{H^1(\Omega)} \end{aligned}$$

where $\beta(\bar{z}) = \beta(z)$ because the spectrum of A is real. It then follows that

$$\sup_{0 \neq y \in H_0^1(\Omega)} \frac{|b(v,y)|}{|y|_{H^1(\Omega)}} \geq \frac{|b(v, v-w)|}{|v-w|_{H^1(\Omega)}} \geq \frac{|v|_{H^1(\Omega)}}{\beta(z)|v|_{H^1(\Omega)}} = \beta(z)^{-1},$$

completing the proof. □

4.3.2 The DPG Resolvent Discretization

For the discretization using the discontinuous Petrov-Galerkin method, we assume that Ω is partitioned by a simplicial conforming mesh Ω_h; triangular elements in this mesh are denoted by K, and the mesh size h is given by $h = \max_{K \in \Omega_h} \text{diam}(K)$. Further details on the DPG method can be found in [12–14, 21, 28, 66] along with our own brief exposition in Chapter 2. Next, define the spaces

$$H^1(\Omega_h) = \prod_{K \in \Omega_h} H^1(K), \quad Q = H(\text{div}, \Omega) / \prod_{K \in \Omega_h} H_0(\text{div}, K)$$

with corresponding norms

$$||v||_{H^1(\Omega_h)} = \left(\sum_{K \in \Omega_h} ||v||^2_{H^1(K)} \right)^{1/2},$$

$$||q||_Q = \inf \left\{ ||q - q_0||_{H(\mathrm{div}, \Omega_h)} : q_0 \in \prod_{K \in \Omega_h} H_0(\mathrm{div}, K) \right\}$$

for $v \in H^1(\Omega_h)$ and $q \in Q$. In the definition of the space Q, $H(\mathrm{div}, \Omega)$ is the space of all vector-valued functions $q \in L^2(\Omega)^d$ $(d \geq 2)$ such that $\mathrm{div}(q) \in L^2(\Omega)$. Furthermore, for each $K \in \Omega_h$, $H_0(\mathrm{div}, K)$ refers to the vector-valued functions $q \in H(\mathrm{div}, K)$ for which the normal trace of $q \cdot n$ is zero on ∂K [16]. On every $K \in \Omega_h$, the normal trace $q \cdot n$ is in $H^{-1/2}(\partial K)$ for each $K \in \Omega_h$, with $\langle q \cdot n, v \rangle_{\partial K}$ denoting the action of $q \cdot n|_{\partial K}$ on the trace $v|_{\partial K}$ of $v \in H^1(K)$. [24]. Next, define the form b_h by

$$b_h((u, q), v) = \sum_{K \in \Omega_h} \left(z \int_K u \bar{w} dx - \int_K \nabla u \cdot \nabla \bar{w} dx + \langle q \cdot n, \bar{v} \rangle_{\partial K} \right)$$

As in the case of classical FEM, we need to pose the solution to our problem using carefully chosen finite-dimensional subspaces. To this end, we let $\mathcal{V}_h$ denote the Lagrange finite element space of $\mathcal{V} = H^1_0(\Omega)$ whose members are continuous functions which, when restricted to an arbitrary element $K \in \Omega_h$, are in the space of polynomials $P_p(K)$ of degree at most p; here, we assume $p \geq 1$.

For our additional task of approximating the unknown q, we leverage the Raviart-Thomas finite element space $RT_h \subset H(\mathrm{div}, \Omega)$, defined by

$$RT_h = \{ q \in H(\mathrm{div}, \Omega) : q \in P_{p-1}(K)^n + x P_{p-1}(K) \}$$

Letting

$$Q_h = \{q_h \in Q : q_h|_K \in P_{p-1}(K)^n + xP_{p-1}(K) + H_0(\text{div}, K)\}$$

and $Y_h = \{\varepsilon_h \in H^1(\Omega_h) : \varepsilon_h|_K \in P_{p+n+1}(K)\}$, we are now read to state the weak formulation of the equation $u = R(z)f$ for any $f \in L^2(\Omega)$. Specifically, we seek a $(u_h, q_h) \in \mathcal{V}_h \times Q_h$ with $u_h = R_h(z)f$ and $\varepsilon_h \in Y_h$ satisfying

$$(\varepsilon_h, \eta_h)_{H^1(\Omega_h)} + b_h((u_h, q_h), \eta_h) = \int_\Omega f\,\bar{\eta}_h\,dx, \qquad \text{for all } \eta_h \in Y_h, \tag{4.17a}$$

$$b_h((w_h, r_h), \varepsilon_h) = 0, \qquad \text{for all } w_h \in \mathcal{V}_h, r_h \in Q_h. \tag{4.17b}$$

where

$$(\varepsilon_h, \eta_h)_{H^1(\Omega_h)} = \sum_{K\in\Omega_h} \int_K (\varepsilon_h\bar{\eta}_h + \nabla\varepsilon_h \cdot \nabla\bar{\eta}_h)\,dx.$$

The next result we present bounds the error between u and u_h. We assume for the next and following results, that z varies in some bounded subset $D \subset \mathbb{C}$ of the resolvent set of A. As in [24], we write $t_1 \lesssim t_2$ whenever there is a positive constant C satisfying $t_1 \leq Ct_2$ independent of the mesh parameter $h = \max_{K\in\Omega_h} \text{diam}K$, but dependent on other quantities such as the diameter of D and the shape regularity of the mesh Ω_h. In addition, we the quantity $\beta(z)$ in our estimates to indicate the deterioration of estimates for z close to the desired portion of the spectrum we wish to compute.

Lemma 10. (Resolvent error estimate [24]) For all $f \in L^2(\Omega)$,

$$||R(z)f - R_h(z)f||_\mathcal{V} \lesssim \beta(z)\left[\inf_{w_h\in L_h} ||u - w_h||_{H^1(\Omega)} + \inf_{q_h\in RT_h} ||q - q_h||_{H(\text{div},\Omega)}\right],$$

where $u = R(z)f$ and $q = \nabla u$.

Proof. We follow the proof from [24]. We need to verify the conditions for well-posedness of the DPG method using the results of [28, Theorem 2.1]. This result immediately yields our claim, provided that we verify the conditions and assumptions used within the theorem. The first two conditions require showing that the operator generated by the form $b_h(\cdot,\cdot)$ is a bijection. For our purposes, we will state a form of this verification that is dual to the assumptions given for the conditions of the theorem in [28]. The first condition to verify is

$$\{\eta \in H^1(\Omega_h) : b_h((w,r),\eta) = 0 \text{ for all } (w,r) \in H_0^1(\Omega) \times Q\} = \{0\}. \tag{4.18a}$$

The second condition is that there are $C_1, C_2 > 0$ such that

$$C_1\big[|w|^2_{H^1(\Omega)} + ||r||^2_Q\big]^{1/2} \leq \sup_{\eta \in H^1(\Omega_h)} \frac{|b_h((w,r),\eta)|}{||\eta||_{H^1(\Omega_h)}} \leq C_2\big[|w|^2_{H^1(\Omega)} + ||r||^2_Q\big]^{1/2} \tag{4.18b}$$

for all $w \in H_0^1(\Omega)$ and $r \in Q$. Finally, the third condition is the existence of a bounded linear operator $\mathit{\Pi}_h : H^1(\Omega_h) \to Y_h$ such that

$$b_h((w_h, r_h), \eta - \mathit{\Pi}_h\eta) = 0. \tag{4.18c}$$

Once these conditions are verified, [28, Theorem 2.1] implies

$$|u - u_h|_{H^1(\Omega)} \leq \frac{C_2||\mathit{\Pi}||}{C_1}\left[\inf_{w_h \in L_h} |u - w_h|_{H^1(\Omega)} + \inf_{q_h \in \mathrm{RT}_h} ||q - q_h||_{H(\div,\Omega)}\right] \tag{4.19}$$

with $u = R(z)f$ and $u_h = R_h(z)f$. We begin by verifying verifying conditions (4.18a) and (4.18b) on $b_h(\cdot,\cdot)$ using the properties of the form $b(\cdot,\cdot)$. Namely, note that in [7, Lemma 2.2, Theorem 2.3], we have that

$$||r||_Q = \sup_{v \in H^1(\Omega_h)} \frac{|\sum_{K \in \Omega_h} \langle r \cdot n, v \rangle_{\partial K}|}{||v||_{H^1(\Omega_h)}}.$$

This result and [7, Theorem 3.3] together imply that the inf-sup condition we proved in Lemma 9 implies an inf-sup condition for b_h, specifically that the lower equality of (4.18b) holds for

$$\frac{1}{C_1^2} = \beta(z)^2 + \left[\beta(z)(1 + |z|) + 1\right]^2.$$

Combining this with the continuity estimate of b_h with $C_2 = 1 + |z|$, we have that $C_2/C_1 = O(\beta(z))$. Finally, condition (4.18c) follows from the Fortin operator constructed in [28, Lemma 3.2] whose norm is a constant bounded independently of z. Hence the lemma follows from (4.19).

□

4.3.3 FEAST Iterations with the DPG Discretization

In this section, we provide additional results needed to compute an approximation to the desired subspace $E \subseteq \mathcal{V}$, and round out the approximation error of approximating E by E_h, which in turn is an application of Theorem 8. To do so, we require the following assumption on regularity as specified in [24].

Assumption 5. Suppose there are positive constants C_{reg} and s such that the solution $u^f \in \mathcal{V}$ of the Dirichlet problem $-\Delta u^f = f$ admits the regularity estimate

$$||u^f||_{H^{1+s}(\Omega)} \leq C_{reg}||f||_{\mathcal{H}} \quad \text{for any } f \in \mathcal{V}. \tag{4.20}$$

Also suppose that

$$||u^f||_{H^{1+s_E}(\Omega)} \leq C_{reg}||f||_{\mathcal{H}} \quad \text{for any } f \in E. \tag{4.21}$$

(Since $E \subseteq \mathcal{V}$, (4.20) implies (4.21) with s in place of s_E, but in many cases (4.21) holds with s_E larger than s, see for example [45]. This is the reason for additionally assuming (4.21).)

In the case that Ω is convex, Assumption 5 holds with $s = 1$ [24]. Otherwise, if Ω has a maximum interior angle π/α located at a corner for $1/2 < \alpha < 1$, then the same assumption holds for any positive $s < \alpha$ [30]. Proceeding forward, we introduce another lemma to quantify the error in applying the resolvent using the DPG discretization.

Lemma 11. (Resolvent discretization errors [24]) Suppose Assumption 5 holds. Then,

$$||R(z)f - R_h(z)f||_{\mathcal{V}} \lesssim \beta(z)^2 h^{\min(p,s,1)}||f||_{\mathcal{V}}, \qquad \text{for all } f \in \mathcal{V}, \tag{4.22}$$

$$||R(z)f - R_h(z)f||_{\mathcal{V}} \lesssim \beta(z)^2 h^{\min(p,s_E)}||f||_{\mathcal{V}}, \qquad \text{for all } f \in E. \tag{4.23}$$

Proof. Using Lemma 10, we can apply standard finite element error estimates for the Lagrange and Raviart-Thomas finite spaces [4, 16] with $u = R(z)f$ and $u_h = R_h(z)f$ to obtain

$$\begin{aligned} ||u - u_h||_{H^1(\Omega)} &\lesssim \beta(z)\left[\inf_{w_h \in L_h}||u - w_h||_{H^1(\Omega)} + \inf_{q_h \in RT_h}||q - q_h||_{H(\text{div},\Omega)}\right] \\ &\lesssim \beta(z)\left[h^r|u|_{H^{1+r}(\Omega)} + h^r|q|_{H^r(\Omega)} + h^r|\text{div}\, q|_{H^r(\Omega)}\right] \end{aligned} \tag{4.24}$$

for $r \leq p$ and where $q = \nabla u$. Since u satrisfies $b(u, v) = (f, v)_{\mathcal{H}}$ for all $v \in H_0^1(\Omega)$, we

have by Lemma 9 that

$$\beta(z)^{-1}|u|_{H^1(\Omega)} \leq \sup_{y\in H_0^1(\Omega)} \frac{|b(u,y)|}{|y|_{H_0^1(\Omega)}} = \sup_{y\in H_0^1(\Omega)} \frac{|(f,y)_{\mathcal{H}}|}{|y|_{H_0^1(\Omega)}} = ||f||_{H^{-1}(\Omega)}. \tag{4.25}$$

By the Poincaré inequality, we have

$$||u||_{\mathcal{H}} \leq C_{\mathcal{V}}|u|_{\mathcal{V}} \leq C_{\mathcal{V}}\beta(z)||f||_{H^{-1}(\Omega)} \leq C_{\mathcal{V}}\beta(z)||f||_{\mathcal{H}}. \tag{4.26}$$

Applying elliptic regularity to $\Delta u = f - zu$, for all $r \leq s$ and $r \leq 1$,

$$\begin{aligned}
|u|_{H^{1+r}(\Omega)} &\leq C_{reg}(||f||_{\mathcal{H}} + |z|||u||_{\mathcal{H}}) && \text{by (4.20)}, \\
&\lesssim \beta(z)||f||_{\mathcal{H}} && \text{by (4.26)}, && (4.27)\\
|q|_{H^r(\Omega)} &= |\text{grad}u|_{H^r(\Omega)} \lesssim \beta(z)||f||_{\mathcal{H}}, && \text{by (4.27)}, && (4.28)\\
|\text{div}\, q|_{H^r(\Omega)} &= |f - zu|_{H^r(\Omega)} \\
&\lesssim |f|_{H^r(\Omega)} + |z|\beta(z)||f||_{\mathcal{H}} && \text{by (4.27)}, \\
&\lesssim \beta(z)||f||_{\mathcal{V}} && \text{since } r \leq 1. && (4.29)
\end{aligned}$$

Bringing results (4.27), (4.28) and (4.29) together and applying them to the last expression in (4.24) (and using the assumption that $||f||_{\mathcal{H}} \leq C_{\mathcal{V}}||f||_{\mathcal{V}}$, the claim (4.22) follows.

Next, we seek to show that (4.23) also holds, but now with $f \in E$. Since f may have higher regularity, the results (4.27) and (4.28) hold now for $r \leq s_E$. Analogously, we have that $|\text{div}\, q|_{H^r(\Omega)} \leq |f|_{H^r(\Omega)}$, so it remains to bound $|f|_{H^r(\Omega)}$. To do

so, we note that $-\Delta f \in E$, and hence (4.21) implies $||f||_{H^{1+r}(\Omega)} \lesssim ||f||_{\mathcal{H}}$. Then it follows that

$$|\operatorname{div} q|_{H^r(\Omega)} \lesssim \beta(z)||f||_{\mathcal{V}} \qquad \text{for } r \leq s_E,$$

so for $f \in E$, the estimates (4.27), (4.28) and (4.29) hold for all $0 \leq r \leq s_E$. As before, we apply these results to (4.24), proving (4.23). □

Next, we quantify the error when approximating the desired eigenspace E by E_h. This will round out the analysis of error using the DPG method by quantifying the error in terms of h, p, and s_E from Assumption 5, and can be found in [24].

Theorem 12. Suppose Assumption 1 (on spectral separation) and Assumption 5 (on elliptic regularity) hold. Then, there are positive constants C_0 and h_0 such that for all $h < h_0$, the FEAST iterates $E_h^{(\ell)}$ obtained using the DPG approximation of the resolvent converge to E_h and

$$\operatorname{gap}_{\mathcal{V}}(E, E_h) \leq C_0\, h^{\min(p, s_E)}, \tag{4.30}$$

$$\operatorname{dist}(\Lambda, \Lambda_h) \leq C_0\, h^{2\min(p, s_E)}. \tag{4.31}$$

Here C_0 is independent of h, but may depend on $\beta(z_k)^2$, W, C_N, p, Λ, C_{reg}, and the shape regularity of the mesh.

Proof. We follow the same proof given in [24], beginning with the application of Theorem 8. As we have already noted, Assumption 2 holds for the model Dirichlet problem with the settings in (4.15). Estimate (4.22) of Lemma 11 verifies Assumption 3. Thus, since Assumptions 1–3 hold, we may now apply (4.11) of Theorem 8

to conclude that $\mathrm{gap}_{\mathcal{V}}(E_h^{(\ell)}, E_h) \to 0$. Moreover, the inequality (4.13) of Theorem 8, when combined with the rate estimate (4.23) of Lemma 11 at each z_k, proves (4.30).

Finally, to prove (4.31), noting that the $\mathcal{V}_h$ set to the Lagrange finite element space L_h satisfies Assumption 4, we appeal to (4.14) of Theorem 8 to

$$\mathrm{dist}(\Lambda, \Lambda_h) \lesssim \mathrm{gap}_{\mathcal{V}}(E, E_h)^2 + \mathrm{gap}_{\mathcal{H}}(E, E_h)^2. \tag{4.32}$$

To control the last term, first note that $||e||_{\mathcal{V}}^2 = a(e,e) \leq C_E ||e||_{\mathcal{H}}^2$ for all $e \in E$. Moreover, by Assumption 2, $\mathrm{dist}_{\mathcal{H}}(e, E_h) \leq C_{\mathcal{V}} \mathrm{dist}_{\mathcal{V}}(e, E_h)$. Hence

$$\delta_h^{\mathcal{H}} := \sup_{0 \neq e \in E} \frac{\mathrm{dist}_{\mathcal{H}}(e, E_h)}{||e||_{\mathcal{H}}} \lesssim \sup_{0 \neq e \in E} \frac{\mathrm{dist}_{\mathcal{V}}(e, E_h)}{||e||_{\mathcal{V}}} \leq \mathrm{gap}_{\mathcal{V}}(E, E_h). \tag{4.33}$$

Note that

$$\mathrm{gap}_{\mathcal{H}}(E, E_h) = \max \left[\delta_h^{\mathcal{H}}, \sup_{m \in U_{E_h}^{\mathcal{H}}} \mathrm{dist}_{\mathcal{H}}(m, E) \right].$$

Now, by the already proved estimate of (4.30), we know that $\mathrm{gap}_{\mathcal{V}}(E, E_h) \to 0$. Hence, when h is sufficiently small, $\mathrm{gap}_{\mathcal{V}}(E, E_h) < 1$, so $\dim(E_h) = \dim(E) = m$. Taking h even smaller if necessary, $\delta_h^{\mathcal{H}} < 1$ by (4.33), so by [35, Theorem I.6.34], there is a closed subspace $\tilde{E}_h \subseteq E_h$ such that $\mathrm{gap}_{\mathcal{H}}(E, \tilde{E}_h) = \delta_h^{\mathcal{H}} < 1$. But this means that $\dim(\tilde{E}_h) = \dim(E) = \dim(E_h)$, so $\tilde{E}_h = E_h$. Summarizing, for sufficiently small h, we have

$$\mathrm{gap}_{\mathcal{H}}(E, E_h) = \delta_h^{\mathcal{H}} \lesssim \mathrm{gap}_{\mathcal{V}}(E, E_h).$$

Returning to (4.32), we conclude that

$$\mathrm{dist}(\Lambda, \Lambda_h) \lesssim \mathrm{gap}_{\mathcal{V}}(E, E_h)^2,$$

and the proof is finished using (4.30). □

4.3.4 A Generalization to Additive Perturbations

To extend our analysis for solving problems such as 1.1, we generalize the results for the Dirichlet operator by considering a perturbation by a function $\nu : \Omega \to \mathbb{R}$ in $L^\infty(\Omega)$. In this case, we wish to state something meaningful about the form

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla \bar{v} dx - \int_\Omega \nu u v dx \tag{4.34}$$

for $u, v \in \mathrm{dom}(a) = \mathcal{V} = H_0^1(\Omega)$. In this case, the operator A is the unbounded operator on $\mathcal{H} = L^2(\Omega)$ generated by the form a, for example, via an appropriate representation theorem [58]. The following lemma shows that in this newer setting, we still have a well-posed problem. The function $d(z)$ in Lemma 13 is given by

$$d(z) = 1 + \frac{c_z}{c_P}(|z| + \mu)$$

where

$$c_z = \sup\left\{ \frac{|\lambda + \mu|^{1/2}}{|z - \lambda|} : \lambda \in \Sigma(A) \right\}$$

and c_P is the constant arising from the Poincaré inequality. We state the lemma and its proof below.

Lemma 13 (Generalization of Lemma 9 [24])**.** Suppose a as in (4.34), $b(u,v) = z(u,v)_{\mathcal{H}} - a(u,v)$, z is in the resolvent set of A, and $d(z)$ is as defined above. Then for all $v \in H_0^1(\Omega)$,

$$\sup_{y \in H_0^1(\Omega)} \frac{|b(v,y)|}{|y|_{H^1(\Omega)}} \geq d(z)^{-1} |v|_{H^1(\Omega)}.$$

Proof. Analogous to Lemma 9, we construct a $w \in H_0^1(\Omega)$. This time, for any $v \in H_0^1(\Omega)$, we let

$$w = R(\bar{z})(\bar{z}v + \nu v).$$

This choice of w solves $b(s, w) = z(s, v)_{\mathcal{H}} + (\nu s, v)_{\mathcal{H}}$ for all $s \in H_0^1(\Omega)$. Choosing $s = v$ as in Lemma 9, we have

$$\begin{aligned} b(v, v - w) &= b(v, v) - b(v, w) \\ &= b(v, v) - z(v, v)_{\mathcal{H}} - (\nu v, v)_{\mathcal{H}} \\ &= z(v, v)_{\mathcal{H}} + (\nu v, v)_{\mathcal{H}} - a(v, v) - z(v, v)_{\mathcal{H}} - (\nu v, v)_{\mathcal{H}} \\ &= -a(v, v) \\ &= -|v|_{H^1(\Omega)} \end{aligned} \tag{4.35}$$

Now for any $\mu > ||\nu||_{L^\infty(\Omega)}$, the form domain $\mathrm{dom}(a) = H_0^1(\Omega)$ is equal to $\mathrm{dom}((A + \mu)^{1/2})$ by [58, Proposition 10.5], which also states the result

$$a(u, v) = ((A + \mu)^{1/2}u, (A + \mu)^{1/2}v)_{\mathcal{H}} - \mu(u, v)_{\mathcal{H}} \qquad \forall u, v \in H_0^1(\Omega).$$

Then it follows that

$$\begin{aligned} |w|^2_{H^1(\Omega)} &= a(w, w) + (\nu w, w)_{\mathcal{H}} \\ &\leq a(w, w) + \mu||w||^2_{\mathcal{H}} \\ &= ||(A + \mu)^{1/2}w||^2_{\mathcal{H}}. \end{aligned} \tag{4.36}$$

Now suppose that z is in the resolvent set of the operator A. Then by the functional calculus results from [5, Theorem 6.4.1], we have that the the spectrum of the normal operator $(A+\mu)^{1/2}R(z)$ consists of elements of the form

$$\frac{(\lambda+\mu)^{1/2}}{z-\lambda}$$

for $\lambda \in \Sigma(A)$, hence $(A+\mu)^{1/2}R(z)$ is a bounded operator with norm

$$c_z = \sup_{\lambda \in \Sigma A} \left| \frac{(\lambda+\mu)^{1/2}}{z-\lambda} \right|.$$

Then (4.36) implies that

$$\begin{aligned} |w|_{H^1(\Omega)} &\leq ||(A+\mu)^{1/2}R(\bar{z})(\bar{z}v+\nu v)||_{\mathcal{H}} \\ &\leq c_z||(\bar{z}v+\nu v)||_{\mathcal{H}} \\ &\leq (|z|+\mu)c_z||v||_{\mathcal{H}} \\ &\leq (|z|+\mu)\frac{c_z}{c_P}|v|_{H^1(\Omega)} \end{aligned}$$

by the Poincaré inequality $c_P||v||_{\mathcal{H}} \leq |v|_{H^1(\Omega)}$. Then it follows that

$$\begin{aligned} |v-w|_{H^1(\Omega)} &\leq |v|_{H^1(\Omega)} + |w|_{H^1(\Omega)} \\ &= |v|_{H^1(\Omega)} + (|z|+\mu)\frac{c_z}{c_P}|v|_{H^1(\Omega)} \\ &= d(z)|v|_{H^1(\Omega)} \end{aligned} \tag{4.37}$$

with $d(z) = 1 + (|z| + \mu)\dfrac{c_z}{c_P}$. We then combine the results of (4.35) and (4.37) to obtain

$$\sup_{y \in H_0^1 \Omega} \frac{|b(v,y)|}{|y|_{H^1(\Omega)}} \geq \frac{|b(v, v-w)|}{|v-w|_{H^1(\Omega)}} \geq \frac{|v|^2_{H^1(\Omega)}}{d(z)|v|_{H^1(\Omega)}} = d(z)^{-1}|v|_{H^1(\Omega)},$$

completing the proof. □

4.4 Numerical Verification

For numerical verification, we look at two examples of solving $-\Delta u = \lambda u$ with zero dirichlet boundary conditions on two different domains: The unit square, and an L-shaped domain. We use these problems to confirm that the use of the DPG discrtetization produces the expected decrease in error based on the theory developed earlier. The software used for numerically solving our problem with the finite element method is NGSolve [59], a C++ library with a Python front-end. In conjunction with our in-house pythonic implementation of FEAST `pyeigfeast` [26], we use a circular contour of radius $\gamma > 0$, center $y \in \mathbb{C}$, and a shift $\phi > 0$ to prevent any quadrature points from coinciding with an eigenvalue we wish to compute. Analogous to 3.6, we define the weights and nodes for the (shifted) trapezoidal quadrature by

$$w_j = \frac{\gamma}{N} e^{i(\theta_j + \phi)} \tag{4.38a}$$

$$z_j = y + \gamma e^{i(\theta_j + \phi)} \tag{4.38b}$$

for $j = 0, 1, \ldots, N-1$, with $\theta_j = 2\pi j/N$ and $\phi = \pm\pi/N$. For our numerical studies, we use $N = 8$ equally spaced quadrature points about the circular contour Γ.

4.4.1 Discretization Errors on the Unit Square

For this verification, we let $\Omega = (0,1)^2$. Our goal is to compute approximations to the eigenvalues λ solving $-\Delta u = \lambda u$ with $u|_{\partial\Omega} = 0$. In this case, our circular contour Γ has a radius of $\gamma = 45$ and center $y = 20$. The exact eigenvalues we wish to compute are $\Lambda = \{2\pi^2, 5\pi^2\}$, for which the first eigenvalue $\lambda_1 = 2\pi^2$ has multiplicity one, and $\lambda_2 = \lambda_3 = 5\pi^2$ has multiplicity two. The corresponding eigenfunctions, which can be derived using separation of variables, take the form $u_{mn}(x,y) = \sin(m\pi x)\sin(n\pi y)$ for integers $m, n \geq 1$.

For the numerical study, we started with an initial mesh size of $h = 2^{-2}$, and performed five uniform mesh refinements until the mesh size decreased to $h = 2^{-7}$. For each mesh size h, we numerically compute the eigenvalues and eigenvectors for $p = 1, 2, 3$. To measure convergence of eigenfunctions, we approximated the following quantities

$$\delta_i^{(1)} = \min_{0 \neq e \in E} |e_{i,h} - e|_{H^1(\Omega)} = \text{dist}_{H_0^1(\Omega)}(e_{i,h}, E), \tag{4.39a}$$

$$\delta_i^{(2)} = \min_{0 \neq e_h \in E_h} |e_i - e_h|_{H^1(\Omega)} = \text{dist}_{H_0^1(\Omega)}(e_i, E_h). \tag{4.39b}$$

by

$$\delta_{i,h}^{(1)} = \text{dist}_{H_0^1(\Omega)}(e_{i,h}, I_h E), \tag{4.40a}$$

$$\delta_{i,h}^{(2)} = \text{dist}_{H_0^1(\Omega)}(I_h e_i, E_h). \tag{4.40b}$$

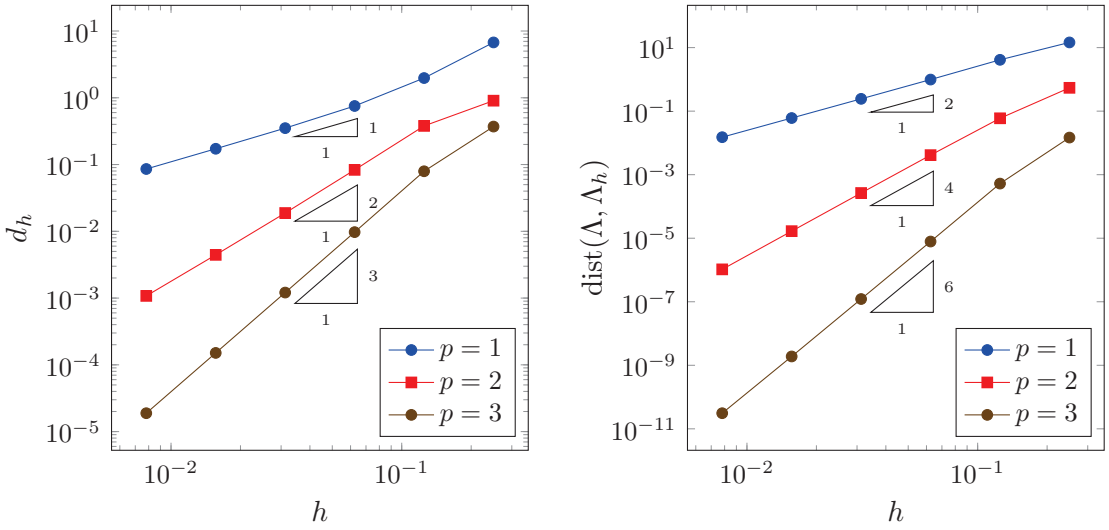
where I_h is a standard interpolant into $\mathcal{V}_h$. To keep matters brief, we plot the be-

haviour of their sum

$$d_h = \sum_{i=1}^{3} \sum_{j=1}^{2} \delta_{i,h}^{(j)} \tag{4.41}$$

for decresing mesh sizes h and polynomial degees $p = 1, 2, 3$ in figure 4.1 [24]. In addition, we also plot the Hausdorff distance between the approximate eigenvalues Λ_h computed by FEAST and the exact eigenvalues Λ. In agreement with the developed theory, we see that d_h goes to zero as $O(h^p)$, and that the error in the eigenvalue approximations decreases as $O(h^{2p})$.

Figure 4.1: Unit Square Convergence Results



(a) Convergence rates for eigenfunctions (b) Convergence rates for eigenvalues

Figure 4.1: Convergence results for the unit square.

4.4.2 Convergence Rates on an L-shaped Domain

We explore the same task of computing eigenvalues and eigenfunctions for same partial differential equation and boundary condition in the previous section, but we now

Table 4.1: L-shaped Domain Errors and Convergence Rates

	λ_1		λ_2		λ_3	
h	ERR	NOC	ERR	NOC	ERR	NOC
2^{-2}	6.29e-02	—	3.29e-02	—	5.95e-02	—
2^{-3}	2.41e-02	1.39	2.65e-03	3.63	4.05e-03	3.88
2^{-4}	9.48e-03	1.34	2.55e-04	3.38	2.59e-04	3.97
2^{-5}	3.75e-03	1.34	2.99e-05	3.09	1.63e-05	3.99
2^{-6}	1.49e-03	1.34	4.03e-06	2.89	1.02e-06	4.00

Table 4.1: Eigenvalue errors (ERR) and numerical order of convergence (NOC) for the smallest three eigenvalues on the L-shaped domain.

let $\Omega = (0,2)^2 \setminus [1,2]^2$. Furthermore, we use a circular contour with a radius of $\gamma = 15$ and center $y = 8$ with $N = 8$ equally spaced quadrature points. Our interest is in computing the first three eigenvalues contained within this contour along with their numerical order of convergence. The corresponding eigenvalues are given by [24] $\lambda_1 \approx 9.6397238$, $\lambda_2 \approx 15.197252$, and $\lambda_3 = 2\pi^2$. We show the results in table 4.1 [24]. The quantity $\text{ERR} = \text{ERR}(h) = |\lambda_{i,h} - \lambda_i|$, where $\lambda_{i,h}$ is the ith eigenvalue computed for the given mesh size $h > 0$. The numerical order of convergence is given by $\text{NOC} = \text{NOC}(h) = \log_2\left(\frac{\text{ERR}(2h)}{\text{ERR}(h)}\right)$.

For our convergence studies, we begin with a mesh size of $h = 2^{-2}$ and perform four uniform mesh refinements to a final mesh size of $h = 2^{-6}$. Our study was focused on observing convergence rates using a polynomial degree $p = 2$. As we can see, our convergence rates are affected by the interior corner in our domain: With an interior angle of $\alpha = \frac{3\pi}{2}$, so the quantity s_E can be chosen satisfying $s_E < \pi/\alpha = 2/3$. Hence, we see that the numerical order of convergence corresponding to the smallest eigenvalue is approximately 4/3, which agrees with the rate of $O(h^{2\min\{s_E,p\}})$ from (4.31) of theorem 12. Keep in mind that this estimate holds for the entire set of eigenvalues we are computing, but not necessarily for all individual eigenvalues. Indeed, we see

that the largest eigenvalue approximately converges at a rate of $O(h^{2p}) = O(h^4)$, as the corresponding eigenfunction is analytic.

Chapter 5

Polynomial Eigenvalue Problems

5.1 Introduction

Polynomial eigenvalue problems extend the work that we have completed for linear eigenproblems, and will be the focus of this section. These problems are of interest because they arrive in several different contexts, including systems of differential equations, engineering, and photonics [25, 32, 40, 63]. One such example comes from a paper by Nannen and Wess for computing scattering resonances. The non-linear eigenvalue problem they solve arises from the application of a frequency-dependent perfectly matched layer [46].

The means by which these problems are solved vary greatly. Algorithms such as Neumier's Residual Inverse Iteration tackled problems for semisimple eigenvalues, though Neumeier also demonstrated his algorithm for eigenvalues with multiplicity greater than one [47]. A broad survey of numerical methods for nonlinear eigenvalue problems is explored by Ruhe, including discussion of Neumaier's algorithm and the use of Newton-type methods. As Ruhe points out, the challenge when using these methods is providing a feasible initial guess to the eigenvectors and corresponding eigenvalues we seek. [53].

More recent non-linear eigensolvers have taken advantage of contour-integral based methods: This includes the methods of Beyn and Polizzi, the former relying on the computation of the first and second moments [3]

$$A_0 = \int_\Gamma P(z)^{-1} dz \tag{5.1a}$$

$$A_1 = \int_\Gamma z P(z)^{-1} dz \tag{5.1b}$$

to transform a nonlinear eigenvalue problem into a small, linear eigenproblem. Here, Γ is a simple, closed contour in $\mathbb{C}$ containing the desired eigenvalues. Beyn's method effectively probes a singular value decomposition. The requirement of Beyn's method, however, is that many quadrature points must be used in order to resolve the computed eigenvalues to the desired accuracy [3]. In Gavin and Polizzi's work [18], that the contour integral

$$Q = \int_\Gamma (X - T(z)^{-1} R(X, \Lambda))(zI - \Lambda)^{-1} dz$$

be used in the effort to solve nonlinear eigenvalue problems. Here $T : \mathbb{C} \to \mathbb{C}^{n \times n}$ is a matrix-valued function of z, Λ is a diagonal matrix whose diagonal entries are eigenvalue approximations, and R represents a block residual for the matrix X whose columns are approximations to the desired eigenvectors [18]. With some carefully chosen examples, we will see that their choice of contour integral has potential drawbacks depending on the choice of $T(z)$. First, however, we begin with a discussion on nonlinear eigenvalue problems.

5.2 Nonlinear eigenvalue problems

Let $T : \mathbb{C} \to \mathbb{C}^{n\times n}$ be a matrix-valued function. For nonlinear eigenvalue problems, we seek to find $\lambda \in \mathbb{C}$ and $x, \tilde{x} \in \mathbb{C}^n$ satisfying the following[1]:

$$T(\lambda)x = \mathbf{0} \tag{5.2a}$$

$$T(\lambda)^*\tilde{x} = \mathbf{0}. \tag{5.2b}$$

In contrast to linear eigenvalue problems, we have to exercise caution when computing eigenvalues and eigenvectors for nonlinear problems. In general, one has to be careful about defective eigenvalues. Such eigenvalues have algebraic multiplicity greater than the geometric multiplicity, or dimension, of the corresponding eigenspace [32]. Another challenge is that of computing eigenvectors. For nonlinear eigenvalue problems, eigenvectors corresponding to distinct eigenvalues can be linearly dependent. A further challenge is the presence of infinite eigenvalues, which themselves can be defective or share the same eigenspace as finite eigenvalues [32].

5.3 Polynomial Eigenproblems

A polynomial eigenproblem is a specific case of a nonlinear eigenproblem for which the function $T : \mathbb{C} \to \mathbb{C}^n$ is a polynomial in its argument, and for which the coeffiecients of the polynomial are matrices in $\mathbb{C}^{n\times n}$, i.e.

[1] In the case that T, for example, is of the form $T(\lambda) = \sum_{i=0}^{d} \lambda^i A_i$ for some integer $d \geq 1$ and $A_i \in \mathbb{C}^{n\times n}$, then $[T(\lambda)]^* = \sum_{i=0}^{d} \overline{\lambda}^i A_i^*$. Such functions T are considered in subsequent discussions on polynomial eigenvalue problems.

$$T(z) = A_0 + A_1 z + \ldots + A_d z^d, \tag{5.3}$$

where $d \geq 1$ is the degree of the polynomial. For our purposes, we assume that A_d is nonzero. In addition, we assume that T is regular, meaning that $\det T(z)$ is not identically zero for all $z \in \mathbb{C}$ for which T is well-defined [32]. As stated in the previous section, our goal is to find $\lambda \in \mathbb{C}$ and $x, \tilde{x} \in \mathbb{C}^n$ satisfying equations (5.2) where T now takes the form of (5.3). As with other nonlinear eigenvalue problems, we do have to grapple with possibly defective or infinite eigenvalues. In this case, we say $\lambda = \infty$ is an eigenvalue of T if 0 is an eigenvalue of the reversal $z^d T(z^{-1})$ [32,63]. In practice, we find the infinite eigenvalues of T by finding the 0 eigenvalue of A_d. Consider the following example of finding the eigenvalues and eigenvectors corresponding to

$$T(z) = A_0 + A_1 z + A_2 z^2, z \in \mathbb{C} \tag{5.4}$$

where

$$A_0 = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \tag{5.5a}$$

$$A_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{5.5b}$$

$$A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{5.5c}$$

We begin by computing the eigenvalues of T by solving $\det T(z) = 0$. Indeed, we

have that

$$\det T(z) = \begin{vmatrix} z^2 & z \\ 1 & 1 \end{vmatrix} = z^2 - z = z(z-1).$$

In this case, we have two finite eigenvalues given by $\lambda = 0, 1$. Since $\det A_2 = 0$, we know at once that $\lambda = \infty$ is also an eigenvalue. To find the eigenvectors, we will find the corresponding eigenvectors using the first companion linearization [32], which we discuss later.

5.4 Solving Polynomial Eigenproblems

To solve polynomial eigenvalue problems, we look the technique [20, 63] of linearizations. The topic of linearizations forms a study in its own right [32, 63]; we leave this as an exploration for the interested reader. For our purposes, we are interested in the use of the first companion linearization for the matrix pencil $\mathcal{A} - z\mathcal{B}$ [20, 63] given by

$$\mathcal{A} = \begin{bmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & I \\ A_0 & A_1 & \cdots & A_{d-2} & A_{d-1} \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} I & 0 & \cdots & \cdots & 0 \\ 0 & I & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & I & 0 \\ 0 & \cdots & \cdots & 0 & -A_d \end{bmatrix}. \tag{5.6}$$

Our recent work shows that the task of solving the nonlinear eigenvalue problem [24] where T is a matrix polynomial (5.3) is equivalent to solving for the eigenvalues and corresponding eigenvectors for the matrix pencil $\mathcal{A} - z\mathcal{B}$. To show this equivalence, we require the following definition [27, 32].

Definition 2 (Jordan Chains)**.** Let $T : \mathbb{C} \to \mathbb{C}^{n\times n}$ and let d^lT/dz^l denote the lth derivative of T with respect to z for $l > 0$. We call $x_0, x_1, \ldots, x_k \in \mathbb{C}$ a *right Jordan chain* and $\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_k \in \mathbb{C}$ a *left Jordan chain* of T if the following set of conditions is satisfied:

$$\sum_{l=0}^{j} \frac{1}{l!} T^{(l)}(\lambda) x_{j-l} = 0, \qquad \sum_{l=0}^{j} \frac{1}{l!} [T^{(l)}(\lambda)]^* \tilde{x}_{j-1} = 0, \qquad j = 0, 1, \ldots, k-1. \tag{5.7}$$

Suppose $k = 1$. Then the above conditions are reduced to

$$T(\lambda)x_0 = 0, \qquad \tilde{x}_0^* T(\lambda) = 0, \tag{5.8}$$

which are the precise conditions for left and right eigenvectors $x_0, \tilde{x}_0$ corresponding to an eigenvalue λ of $T(z)$. For $k > 1$, these chains of vectors are referred to as right (respectively left) generalized eigenvectors of T [32]. For our purposes, our interest is in computing a cluster of eigenvalues $\Lambda \subset \mathbb{C}$. We give the following definition for the left and right generalized eigenspaces associate to all $\lambda \in \Lambda$ [27].

Definition 3 (Algebraic eigenspaces of a cluster of eigenvalues Λ.)**.** Let $\Lambda \subset \mathbb{C}$ be a set of eigenvalues of the nonlinear, matrix-valued function $T : \mathbb{C} \to \mathbb{C}^{n\times n}$. The right and left eigenspaces of a set of nonlinear eigenvalues $\lambda \in \Lambda$ are, respectively, the span of all right and left nonlinear generalized eigenvectors associated to every $\lambda \in \Lambda$.

As noted in [27, 32], these definitions generalize the standard notions for generalized eigenspaces for linear eigenproblems when $T(z) = A - zB$ for $A, B \in \mathbb{C}^{n\times n}$. With these definitions in hand, we can now tackle the problem of finding a cluster of nonlinear eigenvalues Λ and the associated left and right eigenspaces.

5.5 Equivalence of Eigenproblems

To begin our journey into finding the eigenvalues and corresponding left and right eigenspaces for a nonlinear eigenproblem, we need to use the tools of linearizations discussed previously. As mentioned before, we will rely on the use of the first companion linearization (5.6) for devising our algorithm. In order to do so, our goal is to show equivalence between the nonlinear eigenproblem for T a matrix polynomial and the linear eigenproblem of finding eigenvalues λ and corresponding right and left eigenvectors $X, \tilde{X} \in \mathbb{C}^{nd}$ satisfying

$$\mathcal{A}X = \lambda \mathcal{B}X, \quad \tilde{X}^*\mathcal{A} = \lambda \tilde{X}^*\mathcal{B} \tag{5.9}$$

To begin to show this equivalence, let $Y \in \mathbb{C}^{nd \times m}$ be block partitioned as

$$Y = \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_{d-1} \end{bmatrix} \tag{5.10}$$

where $Y_i \in \mathbb{C}^{n \times m}$ for $i = 0, 1, \ldots, d-1$. In addition, we define the operators $F, L \in \mathbb{C}^{n \times nd}$ by

$$F = \begin{bmatrix} I & 0 & \cdots & 0 \end{bmatrix} \tag{5.11a}$$

$$L = \begin{bmatrix} 0 & 0 & \cdots & I \end{bmatrix} \tag{5.11b}$$

where I is the $n \times n$ identity matrix. These operators extract the first (respectively last) n rows of a matrix Y with nd rows, and specifically are used to extract the corrospinding right and left eigenvectors for the nonlinear eigenproblems we wish to solve.

Moving forward, Goldberg, Lancaster, and Rodman [20] show that a given $\lambda \in \mathbb{C}$ is a nonlinear eigenvalue with multiplicity k of (5.2) for T a matrix polynomial if and only if λ is a linear eigenvalue of multiplicity k for (5.9), which justifies interest in pursuing linearizations to solve these types of problems [61, 63].

For our purposes, we are interested in connecting the nonlinear (polynomial) eigenvalue problems with the linearization, as well as providing our own implementation of a FEAST algorithm that takes advantage of the problem structure. As in our own work [27], we provide the ingredients needed to solve this problem by constructing a FEAST algorithm for the linearizartion (5.6).

First, define $\mathcal{S}, \tilde{\mathcal{S}}$ and their corresponding quadrature approximations $\mathcal{S}_N, \tilde{\mathcal{S}}_N$ by [27]

$$\begin{aligned} \mathcal{S} &= \frac{1}{2\pi i} \oint_\Gamma (z\mathcal{B} - \mathcal{A})^{-1}\mathcal{B}\, dz, \quad \tilde{\mathcal{S}} = \frac{1}{2\pi i} \oint_\Gamma (z\mathcal{B} - \mathcal{A})^{-*}\mathcal{B}^*\, dz, \\ \mathcal{S}_N &= \sum_{k=0}^{N-1} w_k (z_k\mathcal{B} - \mathcal{A})^{-1}\mathcal{B}, \qquad \tilde{\mathcal{S}}_N = \sum_{k=0}^{N-1} \bar{w}_k (z_k\mathcal{B} - \mathcal{A})^{-*}\mathcal{B}^*. \end{aligned} \tag{5.12}$$

Where $\Gamma \subset \mathbb{C}$ is a simple, closed contour containing the desired cluster of eigenvalues Λ we wish to compute. Let $\mathcal{E}_0, \tilde{\mathcal{E}}_0 \subset \mathbb{C}^{nd}$ be initial guess to the desired right and left eigenspaces corresponding to the matrix pencil $\mathcal{A} - z\mathcal{B}$. The FEAST algorithm, in practice, generates sequences of right and left subspace approximations $\mathcal{E}_\ell, \tilde{\mathcal{E}}_\ell$ by [27]

$$\mathcal{E}_\ell = \mathcal{S}_N \mathcal{E}_{\ell-1}, \qquad \tilde{\mathcal{E}}_\ell = \tilde{\mathcal{S}}_N \tilde{\mathcal{E}}_{\ell-1} \qquad \text{for } \ell = 1, 2, \ldots. \tag{5.13}$$

The goal in applying this algorithm is to generate approximations to the desired right and left eigenspaces $\mathcal{E}$ and $\tilde{\mathcal{E}}$ of $\mathcal{A} - z\mathcal{B}$ and their corresponding eigenvalues Λ contained in the interior of the contour Γ. As a next step, we show the relationship between the eigenvalues and eigenvectors of the nonlinear eigenvalue problem we wish to solve and the linear eigenproblem we tackle with algorithm 4.

Theorem 14. (Relation between eigenspaces [27]) Let E and $\tilde{E}$ be the right and left algebraic eigenspaces of the nonlinear eigenvalues of $T(z)$ enclosed in Γ, respectively. Then

1. $E = F\mathcal{E}$,
2. $\tilde{E} = L\tilde{\mathcal{E}}$.

When the iterations of the FEAST algorithm converge for ℓ sufficiently large, then truncation by the operators F and L yields the desired right and left eigenspace approximartions for the nonlinear eigenproblem we set out to solve. Before giving a proof of Theorem 14, we provide some needed machinery in the form of a lemma from [27]. Before proceeding, let N denote a $k \times k$ nilpotent matrix

$$N = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & 1 \\ 0 & 0 & \dots & \dots & 0 \end{bmatrix},$$

let I denote the $k \times k$ identity matrix, and let $J = \lambda I + N$ be a $k \times k$ Jordan matrix. We are now ready to state and prove the lemma [27] needed for the proof of theorem 14.

Lemma 15. A sequence $v_0, v_1, \dots v_{k-1}$ in $\mathbb{C}^n$ is a nontrivial Jordan chain of a nonlinear eigenvalue λ of $T(z)$ in the sense of definition (2), if and only if $v_0 \neq 0$ and $V = [v_0, v_1, \dots v_{k-1}] \in \mathbb{C}^{n\times k}$ satisfies

$$\sum_{i=0}^{d} A_i V J^i = 0. \tag{5.14}$$

Proof. As in [27], define the quantity $s_{i\ell}$ for $i, \ell \geq 0$ by

$$s_{i\ell} = \binom{i}{\ell} \lambda^{i-\ell} A^i V N^\ell$$

Note that $s_{i\ell} = 0 \in \mathbb{C}^{n\times k}$ if $\ell > i$ or $\ell > k-1$, so for convenience, define $m_{ik} = \min\{i, k-1\}$. The sum (5.14) can be written by expanding the powers of J:

$$\begin{aligned}
\sum_{i=0}^{d} A_i V J^i &= \sum_{i=0}^{d} A_i V (\lambda I + N)^i \\
&= \sum_{i=0}^{d} A_i V \left[\binom{i}{0} \lambda^i I + \binom{i}{1} \lambda^{i-1} N + \ldots + \binom{i}{\ell} \lambda^{i-\ell} N^\ell + \ldots + \binom{i}{i} N^i \right] \\
&= \sum_{i=0}^{d} A_i V \sum_{\ell=0}^{m_{ik}} \binom{i}{\ell} \lambda^{i-\ell} N^\ell \\
&= \sum_{i=0}^{d} \sum_{\ell=0}^{m_{ik}} \binom{i}{\ell} \lambda^{i-\ell} A_i V N^\ell \\
&= \sum_{i=0}^{d} \sum_{\ell=0}^{m_{ik}} s_{i\ell}
\end{aligned}$$

$$= \left(\sum_{i=k}^{d} + \sum_{i=0}^{k-1} \right) \sum_{\ell=0}^{m_{ik}} s_{i\ell}$$

$$= \left(\sum_{i=k}^{d} \sum_{\ell=0}^{m_{ik}} + \sum_{i=0}^{k-1} \sum_{\ell=0}^{m_{ik}} \right) s_{i\ell}$$

$$= \left(\sum_{i=k}^{d} \sum_{\ell=0}^{k-1} + \sum_{i=0}^{k-1} \sum_{\ell=0}^{i} \right) s_{i\ell}$$

$$= \left(\sum_{\ell=0}^{k-1} \sum_{i=k}^{d} + \sum_{\ell=0}^{k-1} \sum_{i=\ell}^{k-1} \right) s_{i\ell}$$

$$= \sum_{\ell=0}^{k-1} \sum_{i=\ell}^{d} s_{i\ell}. \tag{5.15}$$

Our next step is to relate the last expression in (5.15) to definition (5.7). For clarity, we expand out the sums $\sum_{i=\ell}^{d} s_{i\ell}$, $i \geq 0$ up to an arbitrary index ℓ. We then have

$$\begin{aligned} \sum_{i=0}^{d} s_{i\ell} &= \sum_{i=0}^{d} \binom{i}{0} \lambda^{i-0} A^i V N^0 \\ &= \sum_{i=0}^{d} \lambda^i A^i V \\ &= P(\lambda) \begin{bmatrix} v_0 & v_1 & v_2 & \dots & v_{k-1} \end{bmatrix}, \end{aligned} \tag{5.16a}$$

$$\begin{aligned}\sum_{i=1}^{d} s_{i\ell} &= \sum_{i=1}^{d} \binom{i}{1} \lambda^{i-1} A^i V N^1 \\ &= \sum_{i=1}^{d} \frac{1}{1!} i \lambda^{i-1} A^i V N \\ &= \frac{1}{1!} P^{(1)}(\lambda) \begin{bmatrix} 0 & v_0 & v_1 & \dots & v_{k-2} \end{bmatrix}, \end{aligned} \tag{5.16b}$$

$$\begin{aligned}\sum_{i=2}^{d} s_{i\ell} &= \sum_{i=2}^{d} \binom{i}{2} \lambda^{i-2} A^i V N^2 \\ &= \sum_{i=2}^{d} \frac{1}{2!} i(i-1) \lambda^{i-2} A^i V N^2 \\ &= \frac{1}{2!} P^{(2)}(\lambda) \begin{bmatrix} 0 & 0 & v_0 & \dots & v_{k-3} \end{bmatrix}, \\ &\vdots \end{aligned} \tag{5.16c}$$

and so on, until for $i = \ell$, we have

$$\begin{aligned}\sum_{i=\ell}^{d} s_{i\ell} &= \sum_{i=\ell}^{d} \binom{i}{\ell} \lambda^{i-\ell} A^i V N^\ell \\ &= \sum_{i=\ell}^{d} \frac{1}{\ell!} i(i-1) \cdots (i-\ell+1) \lambda^{i-\ell} A^i V N^\ell \\ &= \frac{1}{\ell!} P^{(\ell)}(\lambda) \begin{bmatrix} 0 & \dots & 0_\ell & v_0 & \dots & v_{k-\ell-1} \end{bmatrix}, \end{aligned} \tag{5.16d}$$

where $0, \dots, 0_\ell$ denote ℓ columns of zero vectors. Then the sum (5.15) is really just the sum

$$\sum_{\ell=0}^{k-1} \sum_{i=\ell}^{d} s_{i\ell} = \sum_{\ell=0}^{k-1} \frac{1}{\ell!} P^{(\ell)}(\lambda) \begin{bmatrix} 0 & \dots & 0_\ell & v_0 & \dots & v_{k-\ell-1} \end{bmatrix}$$

$$= \sum_{\ell=0}^{k-1} \frac{1}{\ell!} \begin{bmatrix} 0 & \dots & 0_\ell & P^{(\ell)}(\lambda)v_0 & \dots & P^{(\ell)}(\lambda)v_{k-\ell-1} \end{bmatrix} \tag{5.17}$$

but $0 = \sum_{i=0}^{d} A_i V J^i = \sum_{\ell=0}^{k-1} \sum_{i=\ell}^{d} s_{i\ell}$. Consequently, the jth column of (5.17) is

$$\sum_{\ell=0}^{j} \frac{1}{\ell!} P^{(\ell)}(\lambda) v_{j-\ell}$$

for $j = 0, 1, \dots, k-1$, which is just a restatement of the conditions that the vectors $v_0, v_1, \dots, v_{k-1}$ form a (right) Jordan chain of a nonlinear eigenvalue λ as stated in definition 2. □

With lemma 15 proven, we move to the proof of Theorem 14, which we closely follow from [27].

Proof. Suppose that λ is a nonlinear eigenvalue enclosed by the contour Γ. By lemma 15, the vectors $v_0, v_1, \dots, v_{k-1} \in \mathbb{C}^n$ form a right Jordan chain of vectors corresponding to the nonlinear eigenvalue λ iff equation (5.14) holds. This is equivalent to the last n equations holding in the $nd \times nd$ system

$$\begin{bmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & I \\ A_0 & A_1 & \cdots & A_{d-2} & A_{d-1} \end{bmatrix} \begin{bmatrix} V \\ VJ \\ VJ^2 \\ \vdots \\ VJ^{d-1} \end{bmatrix} = \begin{bmatrix} I & 0 & \cdots & \cdots & 0 \\ 0 & I & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & I & 0 \\ 0 & \cdots & \cdots & 0 & -A_d \end{bmatrix} \begin{bmatrix} V \\ VJ \\ VJ^2 \\ \vdots \\ VJ^{d-1} \end{bmatrix} J, \tag{5.18}$$

where J and V are given here as in lemma 15. For convenience, let

$$\mathcal{V} = \begin{bmatrix} V \\ VJ \\ VJ^2 \\ \vdots \\ VJ^{d-1} \end{bmatrix} \in \mathbb{C}^{nd \times k},$$

and let $\mathcal{V}_i$ denote the $(i+1)$th column of $\mathcal{V}$ for $i = 0, 1, \ldots, k-1$, so we can write $\mathcal{V} = \begin{bmatrix} \mathcal{V}_0 & \mathcal{V}_1 & \ldots & \mathcal{V}_{k-1} \end{bmatrix}$. Then (5.18) can be stated succinctly as $\mathcal{A}\mathcal{V} = \mathcal{B}\mathcal{V}J$. Written in terms of the columbs of $\mathcal{V}$, we see that

$$\begin{aligned} \mathcal{A}\mathcal{V}_0 &= \lambda\mathcal{B}\mathcal{V}_0 \\ \mathcal{A}\mathcal{V}_1 &= \mathcal{B}(\mathcal{V}_0 + \lambda\mathcal{V}_1) = \mathcal{B}\mathcal{V}_0 + \lambda\mathcal{B}\mathcal{V}_1 \\ \mathcal{A}\mathcal{V}_2 &= \mathcal{B}(\mathcal{V}_1 + \lambda\mathcal{V}_2) = \mathcal{B}\mathcal{V}_1 + \lambda\mathcal{B}\mathcal{V}_2 \\ &\vdots \\ \mathcal{A}\mathcal{V}_{k-1} &= \mathcal{B}(\mathcal{V}_{k-2} + \lambda\mathcal{V}_{k-1}) = \mathcal{B}\mathcal{V}_{k-2} + \lambda\mathcal{B}\mathcal{V}_{k-1} \end{aligned}$$

which is just a statement that the columns $\mathcal{V}_i$ of $\mathcal{V}$ form a (right) Jordan chain for the matrix pencil $\mathcal{A} - \lambda\mathcal{B}$. In addition, we have that $F\mathcal{V}_i = v_i$, which tells us that $\mathcal{V}_i$ forms a Jordan chain for $\mathcal{A} - \lambda\mathcal{B}$ iff $v_0, v_1, \ldots, v_{k-1}$ form a Jordan chain for the nonlinear eigenvalue λ of T.

We repeat the same idea for the second statement in theorem 14, and now starting with the left Jordan chain $\tilde{v}_0, \tilde{v}_1, \ldots, \tilde{v}_{k-1} \in \mathbb{C}^n$; for convenience, we let $\tilde{V} = \begin{bmatrix} \tilde{v}_0 & \tilde{v}_1 & \ldots & \tilde{v}_{k-1} \end{bmatrix} \in \mathbb{C}^{n \times k}$. Using definition (5.7), this left Jordan chain satisfies

$$\sum_{\ell=0}^{j} \frac{1}{\ell!}[T^{(\ell)}(\lambda)]^* \tilde{v}_{j-\ell} = 0 \text{ for } j = 0, 1, \ldots, k-1.$$

Applying lemma 15 to $[T(z)]^*$, we have that $\tilde{V}$ satisfies

$$\sum_{i=0}^{d} A_i^* \tilde{V} \bar{J}^i = 0. \tag{5.19}$$

Next, let $W_{d-i} := -\sum_{j=0}^{i-1} A_{d-j}^* \tilde{V} \bar{J}^{i-1-j}$ for $i = 2, \ldots, d$. We wish to show that (5.19) holds iff

$$\begin{bmatrix} 0 & 0 & \ldots & 0 & A_0^* \\ I & 0 & \ldots & 0 & A_1^* \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & A_{d-2}^* \\ 0 & \ldots & 0 & I & A_{d-1}^* \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{d-2} \\ \tilde{V} \end{bmatrix} = \begin{bmatrix} I & 0 & \cdots & \cdots & 0 \\ 0 & I & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & I & 0 \\ 0 & \cdots & \cdots & 0 & -A_d^* \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{d-2} \\ \tilde{V} \end{bmatrix} \bar{J}, \tag{5.20}$$

also holds. We begin by using the definition of W_{d-i} to complete this step, expanding out the details of W_{d-i} for $i = 2, \ldots, d$ below.

$$\begin{aligned} W_{d-2} &= -\sum_{j=0}^{1} A_{d-j}^* \tilde{V} \bar{J}^{2-1-j} \\ &= -(A_d^* \tilde{V} \bar{J} + A_{d-1}^* \tilde{V}) \\ W_{d-3} &= -\sum_{j=0}^{2} A_{d-j}^* \tilde{V} \bar{J}^{3-1-j} \\ &= -(A_d^* \tilde{V} \bar{J}^2 + A_{d-1}^* \tilde{V} \bar{J}^1 + A_{d-2}^* \tilde{V}) \end{aligned}$$

$$
\begin{aligned}
&\vdots \\
W_{d-i} &= -\sum_{j=0}^{i-1} A_{d-j}^* \tilde{V} \bar{J}^{i-1-j} \\
&= -(A_d^* \tilde{V} \bar{J}^{i-1} + A_{d-1}^* \tilde{V} \bar{J}^{i-2} + \ldots + A_{d-(i-1)}^* \tilde{V}) \\
&\vdots \\
W_1 &= -\sum_{j=0}^{d-2} A_{d-j}^* \tilde{V} \bar{J}^{d-2-j} \\
&= -(A_d^* \tilde{V} \bar{J}^{d-2} + A_{d-1}^* \tilde{V} \bar{J}^{d-3} + \ldots + A_2^* \tilde{V}) \\
W_0 &= -\sum_{j=0}^{d-1} A_{d-j}^* \tilde{V} \bar{J}^{d-1-j} \\
&= -(A_d^* \tilde{V} \bar{J}^{d-1} + A_{d-1}^* \tilde{V} \bar{J}^{d-2} + \ldots + A_2^* \tilde{V} \bar{J} + A_1^* \tilde{V})
\end{aligned}
$$

By equating the first n rows of the left- and right-hand sides of (5.20), we obtain $A_0^* \tilde{V} = W_0 \bar{J}$, which is a restatement of (5.19). For the remaining $n(d-1)$ rows of (5.20), we have that

$$
W_{d-i-1} + A_{d-i}^* \tilde{V} = W_{d-i} \bar{J}, \quad i = 2, \ldots, d-1.
$$

This tells us that the remaining matrices W_{d-i} can be obtained through this recurrence with the definition of W_0 corresponding to $i = d$ as a base case, provided we take $W_{-1} \equiv 0$; another route to this conclusion follows from the definition of the matrices W_{d-i}, which satisfy $W_{d-i} \bar{J} - W_{d-i-1} = A_{d-i}^* \tilde{V}$. Consequently, (5.20) holds precisely when (5.19) holds.

Next, let

$$\tilde{\mathcal{V}} = \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{d-2} \\ \tilde{V} \end{bmatrix}.$$

Writing $\tilde{\mathcal{V}} = \begin{bmatrix} \tilde{\mathcal{V}}_0 & \tilde{\mathcal{V}}_1 & \dots & \tilde{\mathcal{V}}_{k-1} \end{bmatrix}$ where $\tilde{\mathcal{V}}_i$ denotes the ith column of $\tilde{\mathcal{V}}$, we have by (5.20) that $\mathcal{A}^*\tilde{\mathcal{V}}_0 = \bar{\lambda}\mathcal{B}^*\tilde{\mathcal{V}}_0$, as well as

$$\mathcal{A}^*\tilde{\mathcal{V}}_i = \mathcal{B}^*(\tilde{\mathcal{V}}_{i-1} + \bar{\lambda}\tilde{\mathcal{V}}_i), \quad i = 1, \dots, k-1,$$

which is equivalent to stating that $\tilde{\mathcal{V}}_0^*\mathcal{A} = J^T\tilde{\mathcal{V}}_0^*\mathcal{B}$ and

$$\tilde{\mathcal{V}}_i^*\mathcal{A} = (\tilde{\mathcal{V}}_{i-1}^* + \lambda\tilde{\mathcal{V}}_i^*)\mathcal{B}, \quad i = 1, \dots, k-1.$$

This tells us that the columns of $\tilde{\mathcal{V}}$ form a left Jordan chain for the matrix pencil $\mathcal{A} - z\mathcal{B}$. Since $\tilde{V} = L\tilde{\mathcal{V}}$, we have that $\tilde{V}$ is a left Jordan chain corresponding to the eigenvalue λ of T. □

5.6 A FEAST Algorithm for Polynomial Eigenproblems

In this section, we lay the foundation for an efficient implementation of the FEAST algorithm for polynomial eigenvalue problems. Given that linearizations are pursued in order to find solutions to such problems, one cost we have to think about is implementation. The practical implementation of FEAST requires that we compute quadrature approximations to the contour integrals used to form the Riesz projections onto the desired eigenspaces for the matrix pencil $\mathcal{A} - z\mathcal{B}$. In practice, this means having to

solve several linear systems of the form $(z\mathcal{B} - \mathcal{A})X = Y$ and $(z\mathcal{B} - \mathcal{A})^*\tilde{X} = W$ for $\tilde{X}, X, Y, W \in \mathbb{C}^{nd\times m}$. In doing so directly, we pay a price: With dimensions $nd \times nd$ for the matrices $\mathcal{A}$ and $\mathcal{B}$, along with the requirement of FEAST having to perform several factorizations of the pencil $z\mathcal{B} - \mathcal{A}$ at different points $z \in \mathbb{C}$, such a direct approach could be prohibitively costly. For the cubic eigenproblem we wish to solve with $n = O(10^6)$ and higher, the linearized problem is nine times larger than the original problem. To get around this constraint, we use the structure of the linearization so that the only factorizations explicitly required are those of $T(z)$ at various quadrature points z. We show how this is done with a theorem from our paper [27].

Theorem 16. (Resolvent application [27]) Suppose $T(z)$ is invertible at some $z \in \mathbb{C}$ and consider $X, \tilde{X}, Y, W \in \mathbb{C}^{nd}$ block partitioned as in (5.10). Then the following identities hold.

1. The block components of $X = (z\mathcal{B} - \mathcal{A})^{-1}Y$ are given by

$$X_0 = T(z)^{-1}\left(-Y_{d-1} - A_dY_{d-1} + \sum_{i=1}^{d} A_i \sum_{j=0}^{i-1} z^{i-1-j}Y_j\right) \tag{5.21a}$$

$$X_i = zX_{i-1} - Y_{i-1}, \qquad i = 1, 2, \ldots, d-1. \tag{5.21b}$$

2. The block components of $\tilde{X} = (z\mathcal{B} - \mathcal{A})^{-*}W$ are given by

$$\tilde{X}_{d-1} = -T(z)^{-*}\sum_{j=0}^{d-1} \bar{z}^j W_j, \quad \tilde{X}_{d-2} = -W_{d-1} - \bar{z}A_d^*\tilde{X}_{d-1} - A_{d-1}^*\tilde{X}_{d-1}, \tag{5.22a}$$

$$\tilde{X}_i = -W_{i+1} + \bar{z}\tilde{X}_{i+1} - A_{i+1}^*\tilde{X}_{d-1}, \qquad i = 0, 1, \ldots, d-3. \tag{5.22b}$$

Proof. We begin with (5.21). We obtain the block components of X by looking at the structure of $(z\mathcal{B} - \mathcal{A})X = Y$ in further detail.

$$\begin{bmatrix} zI & -I & 0 & \cdots & 0 \\ 0 & zI & -I & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & zI & -I \\ -A_0 & -A_1 & \cdots & -A_{d-2} & -(zA_d + A_{d-1}) \end{bmatrix} \begin{bmatrix} X_0 \\ X_1 \\ \vdots \\ X_{d-2} \\ X_{d-1} \end{bmatrix} = \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_{d-2} \\ Y_{d-1} \end{bmatrix} \tag{5.23}$$

The first $n(d-1)$ rows of (5.23) yield the recurrence

$$zX_{i-1} - X_i = Y_{i-1}, \quad i = 1, 2, \ldots, d, \tag{5.24}$$

which confirms (5.21b). The last n rows form the sum

$$-A_0X_0 - A_1X_1 - \ldots - A_{d-2}X_{d-2} - A_{d-1}X_{d-1} - zA_dX_{d-1} = Y_{d-1}. \tag{5.25}$$

To confirm (5.21a), we begin by recursively expanding out the definitions of the X_i terms whos recurrence we confirmed matched the desired identity in this theorem. Then

$$\begin{aligned} X_1 &= zX_0 - Y_0 \\ X_2 &= zX_1 - Y_1 \\ &= z^2X_0 - zY_0 - Y_1 \\ X_3 &= zX_2 - Y_2 \end{aligned}$$

$$
\begin{aligned}
&= z^3 X_0 - z^2 Y_0 - z Y_1 - Y_2 \\
&\vdots \\
X_i &= z X_{i-1} - Y_{i-1} \\
&= z^i X_0 - z^{i-1} Y_0 - z^{i-2} Y_1 - \ldots z Y_{i-2} - Y_{i-1} \\
&= z^i X_0 - \sum_{j=0}^{i-1} z^{i-1-j} Y_j
\end{aligned}
$$

for $i = 1, 2, \ldots, d-1$. Then it follows that

$$
A_i X_i = z^i A_i X_0 - A_i \sum_{j=0}^{i-1} z^{i-1-j} Y_j,
$$

and so the left-hand-side of (5.25) can be written as

$$
\begin{aligned}
-\sum_{i=0}^{d-1} z^i A_i X_0 + \sum_{i=0}^{d-1} A_i \sum_{j=0}^{i-1} z^{i-1-j} Y_j - z^d A_d X_0 + A_d \sum_{j=0}^{d-2} z^{d-1-j} Y_j &= -T(z) X_0 \\
&\quad + \sum_{i=0}^{d} A_i \sum_{j=0}^{i-1} z^{i-1-j} Y_j \\
&\quad - A_d Y_{d-1}
\end{aligned}
$$

equating this with the right-hand-side of (5.25) yields

$$
-T(z) X_0 + \sum_{i=0}^{d} A_i \sum_{j=0}^{i-1} z^{i-1-j} Y_j - A_d Y_{d-1} = Y_{d-1}.
$$

Upon moving all terms not involving $T(z)$ to the right-hand-side and applying $T(z)^{-1}$ to both sides, we obtain

$$X_0 = T(z)^{-1}\left(-Y_{d-1} - A_d Y_{d-1} + \sum_{i=0}^{d} A_i \sum_{j=0}^{i-1} z^{i-1-j} Y_j\right),$$

which establishes (5.21a). Next, we verify (5.22) by observing the structure of $(z\mathcal{B} - \mathcal{A})^* \tilde{X} = W$. In this case, we have that

$$\begin{bmatrix} \bar{z}I & 0 & \dots & 0 & -A_0^* \\ -I & \bar{z}I & \dots & 0 & -A_1^* \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \bar{z}I & -A_{d-2}^* \\ 0 & \dots & 0 & -I & -(\bar{z}A_d^* + A_{d-1}^*) \end{bmatrix} \begin{bmatrix} \tilde{X}_0 \\ \tilde{X}_1 \\ \vdots \\ \tilde{X}_{d-2} \\ \tilde{X}_{d-1} \end{bmatrix} = \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{d-2} \\ W_{d-1} \end{bmatrix} \tag{5.26}$$

We begin by establishing that the first n rows of the left- and right-hand-sides of (5.26) tell us $\bar{z}\tilde{X}_0 - A_0^*\tilde{X}_{d-1} = W_0$. The next $n(d-2)$ rows yield the reccurence

$$-\tilde{X}_i + \bar{z}\tilde{X}_{i+1} - A_{i+1}^*\tilde{X}_{d-1} = W_{i+1}$$

for $i = 0, 1, \dots, d-3$, immediately establishing (5.22b). To tackle the computation of X_{d-1}, we list the following three relations we have established through (5.26)

$$\bar{z}\tilde{X}_0 - A_0^* X_{d-1} = W_0 \tag{5.27a}$$

$$-\tilde{X}_i + \bar{z}\tilde{X}_{i+1} - A_{i+1}^*\tilde{X}_{d-1} = W_{i+1}, \quad i = 0, \dots, d-3 \tag{5.27b}$$

$$-\tilde{X}_{d-2} - \bar{z}A_d^*\tilde{X}_{d-1} - A_{d-1}^*\tilde{X}_{d-1} = W_{d-1} \tag{5.27c}$$

To proceed, multiply (5.27b) by $\bar{z}^{i+1}$ to obtain

$$-\bar{z}^{i+1}\tilde{X}_i + \bar{z}^{i+2}\tilde{X}_{i+1} - \bar{z}^{i+1}A_{i+1}^*\tilde{X}_{d-1} = \bar{z}^{i+1}W_{i+1}, \quad i = 0, \ldots, d-3 \tag{5.28}$$

Next, compute the sum from $i = 0$ to $d-3$ of 5.27b to obtain

$$\sum_{i=0}^{d-3}(-\bar{z}^{i+1}\tilde{X}_i + \bar{z}^{i+2}\tilde{X}_{i+1}) - \sum_{i=0}^{d-3}\bar{z}^{i+1}A_{i+1}^*\tilde{X}_{d-1} = -\bar{z}\tilde{X}_0 + \bar{z}^{d-1}\tilde{X}_{d-2} - \sum_{i=0}^{d-3}\bar{z}^{i+1}A_{i+1}^*\tilde{X}_{d-1}$$
$$= \sum_{i=0}^{d-3}\bar{z}^{i+1}W_{i+1} \tag{5.29}$$

Adding both side of (5.27a) to the last two expressions in (5.29) yields

$$\bar{z}^{d-1}\tilde{X}_{d-2} - \sum_{i=0}^{d-2}\bar{z}^i A_i^*\tilde{X}_{d-1} = \sum_{i=0}^{d-2}\bar{z}^i W_i \tag{5.30}$$

Now multiply (5.27c) by $\bar{z}^{d-1}$ and add this to (5.30) to obtain

$$-\sum_{i=0}^{d}\bar{z}^i A_i^*\tilde{X}_{d-1} = \sum_{i=0}^{d-1}\bar{z}^i W_i \tag{5.31}$$

This simplifies to

$$-T(z)^*\tilde{X}_{d-1} = \sum_{i=0}^{d-1}\bar{z}^i W_i,$$

which establishes the first part of (5.22a). The second part is established by (5.27c).

□

For polynomial eigenproblems, we give our algorithm for the nonlinear eigensolver we

created in our recent work [27]. Algorithm 4 is designed under the assumption that the number of eigenvalues m we wish to compute is far fewer than the dimension n of the matrices required for the algorithm. This is certainly the case for our applications, as the matrices we construct to solve the cubic eigenproblem arising from our optics applications are large and sparse with $n = O(10^6)$. In addition, we assume that the computed eigenvalues are semisimple so that we may leverage existing tools as needed, primarily to avoid stable computations of generalized eigenvectors (i.e. numerically computing a Jordan decomposition) [27].

Algorithm 4 begins with setup steps for factorizing T at the specified input quadrature nodes $z_k \in \mathbb{C}$, as well as allocating scratch space. We then compute $R = \mathcal{S}_N Y$ and $\tilde{R} = \tilde{\mathcal{S}}_N \tilde{Y}$ in steps 7-13 to obtain our updated left and right eigenvector approximations. At this stage, the scratch matrices $R, \tilde{R}$ are used to probe which eigenvector approximations may be in the null space of $\mathcal{B}$ in steps 16-22. Any vectors found to be in the null space are then removed, and we then assemble a small $m \times m$ (with m possibly smaller than initially specified) Ritz system in step 23. This small, dense eigenproblem is then solved in steps 24, from which we extract the eigenvalue approximations and update the eigenvector approximations in step 25. Users of the algorithm can specify how often (say, every three iterations) to see if any eigenvalue approximations lie outside of the contour Γ. If so, the corresponding eigenvector approximations are removed, and the number of eigenvalues m which we seek to approximate is reduced. In practice, users may need to address potential issues surrounding the computation of defective eigenvalues; we omit this potential step from the algorithm for now [27].

One other aspect of the algorithm to touch upon involves the matrix $\mathcal{B}$. It is not

necessary that $\mathcal{B}$ is invertible, and one of our steps in the algorithm checks the kernel of $\mathcal{B}$ to filter out eigenvalues at infinity. This is due to the fact that $\mathcal{K} = \ker(\mathcal{B})$ is contained in the zero eigenspace of $\mathcal{S}$ and $\mathcal{S}_N$. Steps 16-22 are specifically designed to filter out vectors in the nullspace of $\mathcal{B}$, as the dominant eigenvalues of operators $\mathcal{S}$ and $\mathcal{S}_N$ are nonzero. Letting $K = \ker(A_d)$, we see upon closer inspection that $\mathcal{K} = L^T K$. This is due to the fact that for any $X \in \mathbb{C}^{nd}$ that is also in $\mathcal{K}$, it follows that the first $n(d-1)$ rows of X must be zero, with the last n components of X forming an element of K. Likewise, we have that for any $x \in K$, the vector $X := L'x \in \mathcal{K}$. But any vector in K is, by definition, an eigenvector corresponding to the eigenvalue $\lambda = \infty$, so infinite eigenvalues are filtered out in our algorithm.

Algorithm 4 Polynomial FEAST Eigensolver for Problem (put ref here)

Input contour Γ, quadrature z_k, w_k, sparse coefficient matrices $A_0, \ldots, A_{d-1}, A_d \in \mathbb{C}^{n\times n}$, initial right and left eigenvector iterates given as columns of $Y, \tilde{Y} \in \mathbb{C}^{nd\times m}$, respectively, block partitioned as in (put equation reference here) into $Y_j, \tilde{Y}_j \in \mathbb{C}^{n\times m}$, and tolerance $\varepsilon > 0$.

setup
  Prepare $T(z_k)^{-1}$ by sparse factorization at each quadrature point z_k.
repeat
  Set all entries of workspace $\tilde{R}, R \in \mathbb{C}^{nd\times m}$ to 0.
  for each $z_k, \quad k = 0, \ldots, N-1$, do:
    Compute block components of $X \in \mathbb{C}^{nd\times m}$:
      $$X_0 \leftarrow T(z_k)^{-1} \sum_{i=1}^{d} \sum_{j=0}^{i-1} z_k^{i-1-j} A_i Y_j,$$
      for $i = 1, \ldots, d-1$ do: $X_i \leftarrow z_k X_{i-1} - Y_{i-1}$.
    Increment $R \mathrel{+}= w_k X$.
    Compute block components of $\tilde{X} \in \mathbb{C}^{nd\times m}$:
      $$\tilde{X}_{d-1} \leftarrow T(z_k)^{-*} \sum_{j=0}^{d-1} \bar{z}_k^j \tilde{Y}_j,$$
      $\tilde{X}_{d-2} \leftarrow -\tilde{Y}_{d-1} - \bar{z}_k A_d^* \tilde{X}_{d-1} - A_{d-1}^* \tilde{X}_{d-1}$,
      for $i = d-3, \ldots, 1, 0$, do: $\tilde{X}_i \leftarrow A_{i+1}^* \tilde{X}_{d-1} - \tilde{Y}_{i+1} + \bar{z}_k \tilde{X}_{i+1}$.
    Increment $\tilde{R} \mathrel{+}= \bar{w}_k \tilde{X}$.
  endfor
  $G \leftarrow \tilde{R}^* \mathcal{B} R$.
  Compute biorthogonal $V, \tilde{V} \in \mathbb{C}^{m\times m}$ such that $\tilde{V}^* G V = \mathrm{diag}(d_1, \ldots, d_m)$.
  $Y \leftarrow RV$, $\tilde{Y} \leftarrow \tilde{R}\tilde{V}$.
  for $\ell = 1, \ldots, m$ do:
    If $d_\ell \approx 0$: then remove ℓth columns of $\tilde{Y}$ and Y,
    else: rescale ℓth column of $\tilde{Y}$ and Y by $|d_\ell|^{-1/2}$.
  endfor
  Assemble small Ritz system: $A_Y \leftarrow \tilde{Y}^* \mathcal{A} Y, \quad B_Y \leftarrow \tilde{Y}^* \mathcal{B} Y$.
  Compute Ritz values $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ and $W, \tilde{W} \in \mathbb{C}^{m\times m}$ satisfying $\tilde{W}^* A_Y W = \Lambda, \quad \tilde{W}^* B_Y W = I$.
  $Y \leftarrow YW, \quad \tilde{Y} \leftarrow \tilde{Y}\tilde{W}$.
  Periodically check: if λ_ℓ falls outside G, remove ℓth columns of Y and $\tilde{Y}$.
until maximal difference of successive Λ iterates is less than ε.
output eigenvalue cluster $\{\lambda_\ell\}$, left and right eigenvectors in $L\tilde{Y}$ and FY.

Remark 17 (Other algorithms for generalizing FEAST). In our recent paper, specifically *Remark 3*, we explored why one might not pursue

$$S = \oint_C T(z)^{-1} dz$$

as a means of generalizing the FEAST algorithm to solve the nonlinear eigenproblem (5.2). In this instance, C is a simple, closed contour in $\mathbb{C}$. As was done in our paper [27], we look at an example for which

$$T(z) = (z^2 - 1)I, z \in \mathbb{C}$$

with $I \in \mathbb{C}^{n \times n}$ being the identity matrix. In this case, $\det T(z) = z^2 - 1$, hence the eigenvalues are given by $\lambda = \pm 1$. In this case, the corresponding eigenspace is just $\mathbb{C}^n$. For $z \in \mathbb{C}$ that are not eigenvalues of T, we have that

$$T(z)^{-1} = \frac{1}{z^2 - 1} I = \left(\frac{A}{z - 1} + \frac{B}{z + 1} \right) I.$$

Then for a simple closed contour C surrounding both eigenvalues $\lambda = \pm 1$, we have

$$S = \oint_C \left(\frac{1}{z^2 - 1} \right) I dz = \left(\oint_C \frac{1/2}{z - 1} - \frac{1/2}{z + 1} dz \right) I = 0 \cdot I = 0 \in \mathbb{C}^{2 \times 2}.$$

by Cauchy's Integral Formula [55]. In this case, we see that this particular formulation is not viable for devising an algorithm to find eigenvalues and eigenvectors.

Remark 18 (A remark on Polizzi's algorithm). In our paper, Remark 5 [27] talks about how Gavin and Polizzi's algorithm [18] finds solutions in a different space than what we were hoping to see, and shows how we perhaps need to treat the solutions of nonlinear eigenvalues on a case-by-case basis. We provide the details of the same

example, but with futher granularity.

Let $T(z) = A_0 + zA_1 + z^2A_2$, where

$$A_0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, A_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The eigenpairs of this problem are given by

$$\left(0, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right), \left(1, \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right), \left(\infty, \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right).$$

The last eigenpair comes from finding the zero eigenvalue and corresponding eigenvector(s) of the coefficient matrix of z^2, i.e. finding a vector $\mathbf{x} \in \mathbb{C}^2$ satisfying

$$A_2\mathbf{x} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x} = \mathbf{0}.$$

where $\mathbf{0} \in \mathbb{C}^2$ is the zero vector. Moving ahead, let us take a look at what happens when a contour surrounds the eigenvalues $\lambda = 0$ and $\lambda = 1$. Call this contour C, and let it be parameterized by $r(t) = \gamma e^{it} + 1/2$ with $0 \leq t < 2\pi$ and $\gamma > 1/2$. Next, we do some book-keeping for this problem. For a fixed $\vec{\mu} = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}^T$ with each μ_i in the interior of C and each μ_i not an eigenvalue of T, we have

$$
\begin{aligned}
(T(z) - T(\mu_i))(z - \mu_i)^{-1} &= \left(\begin{bmatrix} 1 & z \\ 1 & z^2 \end{bmatrix} - \begin{bmatrix} 1 & \mu_i \\ 1 & \mu_i^2 \end{bmatrix} \right) (z - \mu_i)^{-1} \\
&= \left(\begin{bmatrix} 0 & z - \mu_i \\ 0 & z^2 - \mu_i^2 \end{bmatrix} \right) (z - \mu_i)^{-1} \\
&= \begin{bmatrix} 0 & 1 \\ 0 & z + \mu_i \end{bmatrix}
\end{aligned}
$$

Next, we have that

$$
T(z)^{-1} = \frac{1}{z(z-1)} \begin{bmatrix} z^2 & -z \\ -1 & 1 \end{bmatrix}
$$

Before proceeding, here is why we do these computations individually for each μ_i: One can show that the contour integral method from Gavin and Polizzi's nonlinear FEAST paper, when applied to a block of vectors $X := \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_m \end{bmatrix} \in \mathbb{C}^{n \times m}$ with the corresponding eigenvalue approximations $\Lambda := \operatorname{diag}(\mu_1, \mu_2, \dots, \mu_m) \in \mathbb{C}^{m \times m}$, is equivalent to applying their contour integral method to each eigenvector individually as though one applied the contour integral to a single eigenpair, i.e.

$$
\oint_C (X - T(z)^{-1} R(X, \Lambda))(zI - \Lambda)^{-1} dz
$$

where $R(X, \Lambda)$ (or $T(X, \Lambda)$ in [18]) is the *block residual* defined by

$$
R(X, \Lambda) := \sum_{i=0}^{2} A_i X \Lambda^i,
$$

is equivalent to

$$\begin{bmatrix} S_2^{(\mu_1)}\mathbf{x}_1 & S_2^{(\mu_2)}\mathbf{x}_2 & \dots & S_2^{(\mu_m)}\mathbf{x}_m \end{bmatrix} = \begin{bmatrix} S_2^{(\mu_1)} & S_2^{(\mu_2)} & \dots & S_2^{(\mu_m)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_2 & \dots & \mathbf{0} \\ \vdots & & & \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{x}_m \end{bmatrix}$$

where

$$S_2^{(\mu_i)} := \oint_C (T(z) - T(\mu_i))(z - \mu_i)^{-1} dz, i = 1, 2, \dots, m$$

and $\mathbf{0}$ is the $n \times 1$ zero vector. Fleshing out the details, we have that

$$\begin{aligned} S_2^{(\vec{\mu})} &= \oint_C (X - T(z)^{-1} R(X, \Lambda))(zI - \Lambda)^{-1} dz \\ &= \oint_C T(z)^{-1}(T(z)X - R(X, \Lambda))(zI - \Lambda)^{-1} dz \\ &= \oint_C P(z)^{-1} \left(T(z)X - \sum_{i=0}^{2} A_i X \Lambda^i \right) (zI - \Lambda)^{-1} dz \end{aligned}$$

Before integration, notice that the jth column of $(T(z)X - \sum_{i=0}^{2} A_i X \Lambda^i)(zI - \Lambda)^{-1}$ is just

$$\begin{aligned} \left(T(z)\mathbf{x}_j - \sum_{i=0}^{2} \mu_j^i A_i \mathbf{x}_j \right) (z - \mu_j)^{-1} &= \left(T(z)\mathbf{x}_j - T(\mu_j)\mathbf{x}_j \right) (z - \mu_j)^{-1} \\ &= [(T(z) - T(\mu_j))(z - \mu_j)^{-1}]\mathbf{x}_j, \end{aligned}$$

hence the jth column of $S_2^{(\vec{\mu})}$ is just

$$\oint_C T(z)^{-1}(T(z)\mathbf{x}_j - T(\mu_j)\mathbf{x}_j)(z-\mu_j)^{-1}dz = S_2^{(\mu_j)}\mathbf{x}_j.$$

Proceeding forward, we have that

$$\begin{aligned} T(z)^{-1}(T(z)-T(\mu_i))(z-\mu_i)^{-1} &= \frac{1}{z(z-1)}\begin{bmatrix} z^2 & -z \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 0 & z+\mu_i \end{bmatrix} \\ &= \frac{1}{z(z-1)}\begin{bmatrix} 0 & -z\mu_i \\ 0 & z+\mu_i-1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & -\mu_i/(z-1) \\ 0 & 1/z+\mu_i/(z^2-z) \end{bmatrix} \end{aligned}$$

A brief partial fraction decomposition yields

$$\frac{1}{z^2-z} = -\frac{1}{z} + \frac{1}{z-1}$$

and hence

$$\frac{1}{z} + \frac{\mu_i}{z^2-z} = \frac{1}{z} - \frac{\mu_i}{z} + \frac{\mu_i}{z-1} = \frac{1-\mu_i}{z} + \frac{\mu_i}{z-1},$$

and so

$$T(z)^{-1}(T(z)-T(\mu_i))(z-\mu_i)^{-1} = \begin{bmatrix} 0 & -\mu_i/(z-1) \\ 0 & (1-\mu_i)/z + \mu_i/(z-1) \end{bmatrix}.$$

Then

$$S_2^{(\mu_i)} = \oint_C \begin{bmatrix} 0 & -\mu_i/(z-1) \\ 0 & (1-\mu_i)/z + \mu_i/(z-1) \end{bmatrix} dz, i = 1, 2.$$

The entry-by-entry computations can be done using Cauchy's integral formula. Since C surrounds both eigenvalues 0 and 1, and since the matrix entries have either (or both) eigenvalue(s) as a simple pole, we can split C into two smaller, simple, closed countours that just surround each eigenvalue, call them C_0 and C_1. Hence,

$$S_2^{(\mu_i)} = \oint_{C_0} \begin{bmatrix} 0 & -\mu_i/(z-1) \\ 0 & (1-\mu_i)/z + \mu_i/(z-1) \end{bmatrix} dz + \oint_{C_1} \begin{bmatrix} 0 & -\mu_i/(z-1) \\ 0 & (1-\mu_i)/z + \mu_i/(z-1) \end{bmatrix} dz.$$

This results in the following:

$$S_2^{(\mu_i)} = \begin{bmatrix} 0 & 0 \\ 0 & 1-\mu_i \end{bmatrix} + \begin{bmatrix} 0 & -\mu_i \\ 0 & \mu_i \end{bmatrix} = \begin{bmatrix} 0 & -\mu_i \\ 0 & 1 \end{bmatrix}$$

Bringing everything together, Gavin and Polizzi's block analogue of their contour integral can be written as

$$S_2^{(\vec{\mu})} = \begin{bmatrix} 0 & -\mu_1 & 0 & -\mu_2 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

This yields one of two possibilities, namely $\mu_1 = \mu_2$ or $\mu_1 \neq \mu_2$. Since μ_1 and μ_2 are taken to be arbitrary approximations to the corresponding eigenvalues 0 and 1, what if they end up being same (i.e. $\mu_i = \mu \in \mathbb{C}$ for $i = 1, 2$)? In that case, we

would have that the range of $S_2^{(\vec{\mu})}$ would then be one-dimensional (namely, the span of $\begin{bmatrix} -\mu & 1 \end{bmatrix}^T$), even though the desired eigenspace is two-dimensional. The rationale behind this scenario is that numerically computing the desired eigenspace might be challenging if the eigenvalues are approximated badly enough.

Now suppose that $\mu_1 \neq \mu_2$, a more likely scenario given that computations are done in finite-precision arithmetic. Our goal is to see if we can find vectors $\mathbf{x}_1, \mathbf{x_2}$ in

$$E = \text{span}\left(\left\{\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right\}\right) = \mathbb{R}^2$$

such that applying $S_2^{(\vec{\mu})}$ only lands us in a proper subspace of E. The idea is that, analogous to linear FEAST, we should be able to give $S_2^{(\vec{\mu})}$ two linearly independent vectors, for which the application of $S_2^{(\vec{\mu})}$ should yield a set of vectors whose span is E. Indeed, we will go one step further and provide $S_2^{(\vec{\mu})}$ two linearly independent vectors in E. In this case, we will choose $\mathbf{x}_1 := \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, $\mathbf{x}_2 := \begin{bmatrix} 0 & 1 \end{bmatrix}^T$, and we let

$$Y = \begin{bmatrix} \mathbf{x}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then

$$S_2^{(\vec{\mu})} Y = \begin{bmatrix} 0 & -\mu_2 \\ 0 & 1 \end{bmatrix},$$

which clearly has a one-dimensional column span, regardless of the value of μ_2. A similar choice of Y, namely

$$Y = \begin{bmatrix} \mathbf{x}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

shows that

$$S_2^{(\vec{\mu})} Y = \begin{bmatrix} -\mu_1 & 0 \\ 1 & 0 \end{bmatrix}.$$

Before we wrap up, we address what happens if we only seek a single eigenpair. By reusing the details from earlier, we redefine $S_2^{(\mu_i)}$ for $i = 1, 2$ as

$$S_2^{(\mu_i)} = \oint_{C_i} \begin{bmatrix} 0 & -\mu_i/(z-1) \\ 0 & (1-\mu_i)/z + \mu_i/(z-1) \end{bmatrix} dz, i = 1, 2.$$

where C_i, $i = 0, 1$ are simpled, closed contours surrounding the eigenvalues 0 and 1, respectively, and μ_i, $i = 1, 2$ are the respective eigenvalue approximations to 0 and 1. Then

$$S_2^{(\mu_1)} = \begin{bmatrix} 0 & 0 \\ 0 & 1-\mu_1 \end{bmatrix}, S_2^{(\mu_2)} = \begin{bmatrix} 0 & -\mu_2 \\ 0 & \mu_2 \end{bmatrix}.$$

While computing individual eigenpairs does not seem to pose an issue in this case, the idea appears to break down when computing a cluster of eigenpairs. Whether or not this is related to the eigenspance $\lambda = \infty$ being contained in the eigenspace corresponding to $\lambda = 0, 1$ remains to be determined.

5.7 Leaky Modes of Optical Fibers

Our application of interest is computing leaky modes of optical fibers. To recall, our motivation is to find solutions u and propagation constants $\beta \in \mathbb{C}$ satisfying (1.1) with an index of refraction given by (1.2). We will explore the results of these computations for step-index fibers and then for microstructure fibers.

To this end, we are interested in computing *leaky modes* (also known as resonances or quasi-normal modes) [25, 43, 46] which are outgoing solutions that together with a corresponding propagation constant β satisfy (1.1). Such solutions should satisfy our motivating problem on all of $\mathbb{R}^2$. To make this problem computationally tractable, we need to solve on a bounded domain, and the circular cross-section of an optical fiber serves as an ideal candidate.

Since leaky modes can grow arbitrarily large in magnitude as we move further out from the center of our computational domain, we will use a PML to force exponential decay of our solutions, doing so in a way that treats the PML as arising from a complex coordinate transformation [9, 46]. In Chapter 1, we detail the nondimensionalization of problem (1.1) in Section 1.2.2, so we move forward now with the application of a PML to our problem of interest.

5.7.1 Discretization Based on PML

To make our problem of interest computationally tractable, we need to ensure that we truncate the original domain (namely $\mathbb{R}^2$) to one of finite size. To help facilitate this, we need to first transform our problem from one that has solutions growing arbitrarily large in magnitude to one whose solutions decay exponentially. In practice,

this will allow us to truncate the domain of the problem to a finite size. At this point, we can apply either Dirichlet or Neumann boundary conditions, and we will choose the latter approach. This allows one to compute a solution numerically, and recompute by growing the size of the PML region if the initial computed solution has not decayed sufficiently in the PML [27].

To begin, we first need to define the transformation needed to create the PML for our problem. To this end, define $\tilde{x}$ by

$$\tilde{x} := \frac{\eta(\hat{r})}{\hat{r}}\hat{x}, \tag{5.32}$$

where $\hat{x}$ is defined by $\hat{r} = ||\hat{x}||$ for $||\cdot||$ the vector 2-norm. Furthermore, we have

$$\eta(\hat{r}) := \begin{cases} \hat{r}, & \hat{r} \leq \hat{R} \\ \dfrac{1+\alpha i}{Z}(\hat{r} - \hat{R}) + \hat{R}, & \hat{r} > \hat{R} \end{cases} \tag{5.33}$$

where $\alpha > 0$ is the strength of the decay in the PML, Z is defined as in Equation (1.18a), and where $\hat{R} = R/L$ is fixed and satisfies $\hat{R} > \hat{R}_0$. Note that for $\hat{r} < \hat{R}$, we have that $\tilde{x} = \hat{x}$, and so the problem on the interior domain $\Omega_{int} := \{\hat{x} \in \mathbb{R}^2 : ||\hat{x}|| \leq \hat{R}\}$ remains the same. Thus, we are choosing the PML to begin some fixed distance $\hat{R}$ into the outer region of our domain where the index of refraction is a fixed constant.

5.7.2 Problem Formulation

Next, we need to revisit the PDE in which we are interested in solving, and convert it to weak form. To begin, we let $\mathcal{V} = H^1(\Omega)$ be the Sobolev space of square integrable

functions whose first weak derivatives are also square integrable. Here, we take Ω to be the bounded computational domain $\Omega := \{\hat{x} \in \mathbb{R}^2 : ||\hat{x}|| < \hat{R}_{fin}\}$, where $\hat{R}_{fin} > \hat{R}$ is fixed. As an approximation to the problem we seek to solve on the unbounded domain $\mathbb{R}^2$, we seek to find an outgoing solution $\hat{u}$ and corresponding $Z \in \mathbb{C}$ satisfying

$$-\Delta\hat{u} + V\hat{u} = Z^2\hat{u}, \quad \hat{x} \in \Omega, \tag{5.34}$$

where we recall that Z is defined in equation (1.18a). Upon assuming zero Neumann boundary conditions, multiplying through by a test fuction $v \in \mathcal{V}$, and integrating by parts, we obtain the weak formulation of our same problem: Find a $\hat{u} \in \mathcal{V}$ and $Z \in \mathbb{C}$ satisfying

$$\int_\Omega \hat{\nabla}\hat{u} \cdot \hat{\nabla}\hat{v}d\hat{x} + \int_\Omega V\hat{u}\hat{v}d\hat{x} = Z^2 \int_\Omega \hat{u}\hat{v}d\hat{x} \quad \forall v \in \mathcal{V}, \tag{5.35}$$

where $\hat{\nabla} := (\partial/\partial\hat{x}_1, \partial/\partial\hat{x}_2)^T$. To transform this problem using PML, let $\tilde{u} := \hat{u} \circ \tilde{x}$ and define $\Omega_{pml} := \Omega \setminus \bar{\Omega}_{int}$. Next, we compute gradient $\hat{\nabla}\tilde{u}$ in terms of the Jacobian of the coordinate transformation $\tilde{x}$. To begin, note that in the region Ω_{int}, $\tilde{x} = \hat{x}$, and hence

$$\frac{\partial\tilde{x}_i}{\partial\hat{x}_j} = \delta_{ij},$$

where δ_{ij} is the kronecker delta. The reinforces the notion that the problem in Ω_{int} remains unchanged, and hence $\tilde{u} = \hat{u}$ in this region. In the PML region, however, we have by the chain rule that $\hat{\nabla}\tilde{u} := \hat{\nabla}(\hat{u} \circ \tilde{x})$ can be written componentwise as

$$\frac{\partial \tilde{u}}{\partial \hat{x}_1} = \frac{\partial \tilde{u}}{\partial \tilde{x}_1}\frac{\partial \tilde{x}_1}{\partial \hat{x}_1} + \frac{\partial \tilde{u}}{\partial \tilde{x}_2}\frac{\partial \tilde{x}_2}{\partial \hat{x}_1}$$

$$\frac{\partial \tilde{u}}{\partial \hat{x}_2} = \frac{\partial \tilde{u}}{\partial \tilde{x}_1}\frac{\partial \tilde{x}_1}{\partial \hat{x}_2} + \frac{\partial \tilde{u}}{\partial \tilde{x}_2}\frac{\partial \tilde{x}_2}{\partial \hat{x}_2},$$

And so it follows that $\hat{\nabla}\tilde{u} = J^T\hat{\nabla}\hat{u}$. Thus, we can restate the weak formulation of the PDE as finding a $\tilde{u} \in V$ and a $Z \in \mathbb{C}$ satisfying

$$\int_\Omega \hat{\nabla} J^{-T}\tilde{u} \cdot J^{-T}\hat{\nabla} v \det(J) d\hat{x} + \int_\Omega V\tilde{u}v \det(J) d\hat{x} = Z^2 \int_\Omega \tilde{u}v \det(J) d\hat{x} \tag{5.36}$$

for all $v \in V$, and where $J^{-T} := [J^{-1}]^T$. In the first integral, we can make the order of operations clearer by rewriting the dot product of the gradients as

$$J^{-T}\hat{\nabla}\tilde{u} \cdot J^{-T}\hat{\nabla} v = \hat{\nabla} v^T J^{-1} J^{-T}\hat{\nabla}\tilde{u} = (J^{-1}J^{-T}\hat{\nabla}\tilde{u}) \cdot \hat{\nabla} v$$

We can then absorb the factor $\det(J)$ into the term $J^{-1}J^{-T}$ to obtain

$$\int_\Omega a(\hat{x})\hat{\nabla}\tilde{u} \cdot \hat{\nabla} v d\hat{x} + \int_\Omega V\tilde{u}v \det(J) d\hat{x} = Z^2 \int_\Omega \tilde{u}v \det(J) d\hat{x} \tag{5.37}$$

where $a(\hat{x}) := \det(J)J^{-1}J^{-T}$. To make matters convenient, define the bilinear forms $a : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ and $b : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ by

$$a(u, v) = \int_\Omega a(\hat{x})\hat{\nabla} u \cdot \hat{\nabla} v d\hat{x} + \int_\Omega Vuv \det(J) d\hat{x} \tag{5.38a}$$

$$b(u,v) = \int_{\Omega} uv \det(J) d\hat{x} \tag{5.38b}$$

for $u, v \in \mathcal{V}$. We further split the bilinear forms a and b into forms on the interior subdomain Ω_{int} and PML subdomain Ω_{pml}. In this case, we define the bilinear forms corresponding to Ω_{int} for $u, v \in \mathcal{V}$ by

$$a_{int}(u,v) = \int_{\Omega_{int}} \hat{\nabla} u \cdot \hat{\nabla} v d\hat{x} + \int_{\Omega_{int}} V uv d\hat{x} \tag{5.39a}$$

$$b_{int}(u,v) = \int_{\Omega_{int}} uv d\hat{x} \tag{5.39b}$$

and the bilinear forms corresponding to Ω_{pml} for $u, v \in \mathcal{V}$ by

$$a_{pml}(u,v) = \int_{\Omega_{pml}} a(\hat{x}) \hat{\nabla} u \cdot \hat{\nabla} v d\hat{x} + \int_{\Omega_{pml}} V uv \det(J) d\hat{x} \tag{5.40a}$$

$$b_{pml}(u,v) = \int_{\Omega_{pml}} uv \det(J) d\hat{x} \tag{5.40b}$$

Explicit Computation of the Jacobian

Our next step is to compute an expression for the Jacobian J. To begin, we need a few shorthand notations for convenience. Define $\dot{\eta}$ by

$$\dot{\eta} := \frac{\partial \eta}{\partial \hat{r}} = \frac{\partial}{\partial \hat{r}} \left[\frac{1 + \alpha i}{Z} (\hat{r} - \hat{R}) + \hat{R} \right] = \frac{1 + \alpha i}{Z}.$$

In addition, we have that

$$\frac{\partial \hat{r}}{\partial \hat{x}_i} = \frac{\partial}{\partial \hat{x}_i}\left(\sqrt{\hat{x}_1^2 + \hat{x}_2^2}\right) = \frac{\hat{x}_i}{\sqrt{\hat{x}_1^2 + \hat{x}_2^2}} = \frac{\hat{x}_i}{\hat{r}}$$

for $i = 1, 2$. Then for $i, j = 1, 2$, we have

$$\begin{aligned}
\frac{\partial}{\partial \hat{x}_j}\left(\frac{\eta(\hat{r})}{\hat{r}} x_i\right) &= \frac{\eta(\hat{r})}{\hat{r}}\delta_{ij} + \frac{\partial}{\partial x_j}\left(\frac{\eta(\hat{r})}{\hat{r}}\right) x_i \\
&= \frac{\eta(\hat{r})}{\hat{r}}\delta_{ij} + \frac{\partial\eta/\partial\hat{x}_j \cdot \hat{r} - \eta \cdot \partial\hat{r}/\partial\hat{x}_j}{\hat{r}^2}\hat{x}_i \\
&= \frac{\eta(\hat{r})}{\hat{r}}\delta_{ij} + \frac{\dot{\eta} \cdot \hat{x}_j/\hat{r} \cdot \hat{r} - \eta \cdot \partial\hat{x}_j/\hat{r}}{\hat{r}^2}\hat{x}_i \\
&= \frac{\eta(\hat{r})}{\hat{r}}\delta_{ij} + \frac{\dot{\eta}\hat{r} - \eta}{\hat{r}^3}\hat{x}_i\hat{x}_j.
\end{aligned}$$

In terms of matrices and vectors, we can write this expression as

$$J = aI + b\hat{x}\hat{x}^T, \tag{5.41}$$

where $I \in \mathbb{R}^{2\times 2}$ is the identity matrix, $a = \eta(\hat{r})/\hat{r}$ and $b = (\dot{\eta}\hat{r} - \eta)/\hat{r}^3$; for convenience, we have temporarily reused the symbols a and b to simplify computations related to the Jacobian J. If we wish to reference the matrix in the first integrand of (5.37), we will explicitly state it as $a(\hat{x})$. Similarly, we will explicitly state *the bilinear form* a (respectively b) in reference to (5.38). Next, we compute the determinant of the Jacobian J:

$$
\begin{aligned}
\det(J) &= \begin{vmatrix} a + b\hat{x}_1^2 & b\hat{x}_1\hat{x}_2 \\ b\hat{x}_1\hat{x}_2 & a + b\hat{x}_2^2 \end{vmatrix} \\
&= (a + b\hat{x}_1^2)(a + b\hat{x}_2^2) - b^2\hat{x}_1^2\hat{x}_2^2 \\
&= a^2 + ab(\hat{x}_1^2 + \hat{x}_2^2)b^2\hat{x}_1^2\hat{x}_2^2 - b^2\hat{x}_1^2\hat{x}_2^2 \\
&= a^2 + ab\hat{r}^2 \\
&= \frac{\eta^2}{\hat{r}^2} + \frac{\eta(\dot{\eta}\hat{r} - \eta)}{\hat{r}^4}\hat{r}^2 \\
&= \frac{\eta^2}{\hat{r}^2} + \frac{\dot{\eta}\eta\hat{r} - \eta^2}{\hat{r}^2} \\
&= \frac{\dot{\eta}\eta}{\hat{r}}.
\end{aligned}
$$

This will be a useful simplification moving forward.

Inverting the Jacobian

For the inverse of the Jacobian $J = aI + b\hat{x}\hat{x}^T$, we will defer to the well-known Sherwood-Morrison-Woodbury formula [33] for the inverse of a sum of an invertible matrix $A \in \mathbb{C}^{2\times 2}$ and a rank one update uv^T for $u, v \in \mathbb{C}^2$:

$$
(A + uv^T)^{-1} = A^{-1}\left(I - \frac{1}{1 + v^T A^{-1} u} uv^T A^{-1}\right)
$$

In our case, we have that $A := aI$, $u := b\hat{x}$, and $v = \hat{x}$. Then we have that

$$\begin{aligned}
(aI + b\hat{x}\hat{x}^T)^{-1} &= a^{-1}I\left(I - \frac{1}{1 + (b/a)\hat{x}^T I\hat{x}}(b/a)\hat{x}\hat{x}^T\right) \\
&= a^{-1}\left(I - \frac{b/a}{1 + (b/a)\hat{r}^2}\hat{x}\hat{x}^T\right) \\
&= a^{-1}I - \frac{b}{a^2 + ab\hat{r}^2}\hat{x}\hat{x}^T \\
&= a^{-1}I - \left(\frac{b}{\dot{\eta}\eta/\hat{r}}\right)\hat{x}\hat{x}^T \\
&= \frac{\hat{r}}{\eta}I - \left(\frac{\dot{\eta}\hat{r} - \eta}{\hat{r}^2\dot{\eta}\eta}\right)\hat{x}\hat{x}^T.
\end{aligned}$$

Specially chosen Test Functions

To solve this eigenproblem, we use test functions of the form [46]

$$\tilde{v}(\hat{x}) = \begin{cases} v(\hat{x}), & \hat{x} \in \Omega_{int} \\ \dfrac{\eta(\hat{r})}{\hat{R}}v(\hat{x}), & \hat{x} \in \Omega_{pml} \end{cases} \tag{5.42}$$

For our specially chosen test function $\tilde{v}$, we have that

$$\begin{aligned}
\frac{\partial \tilde{v}}{\partial \hat{x}_i} &= \frac{\partial}{\partial \hat{x}_i}\left(\frac{\eta(\hat{r})}{\hat{R}}v\right) \\
&= \frac{\partial}{\partial \hat{x}_i}\left(\frac{\eta(\hat{r})}{\hat{R}}\right)v + \frac{\eta(\hat{r})}{\hat{R}}\frac{\partial v}{\partial \hat{x}_i} \\
&= \dot{\eta}\frac{\hat{x}_i}{\hat{r}\hat{R}}v + \frac{\eta(\hat{r})}{\hat{R}}\frac{\partial v}{\partial \hat{x}_i}
\end{aligned}$$

for $i = 1, 2$, and hence

$$\hat{\nabla}\tilde{v} = \frac{\dot{\eta}v}{\hat{r}\hat{R}}\hat{x} + \frac{\eta(\hat{r})}{\hat{R}}\hat{\nabla}v. \tag{5.43}$$

5.7.3 Simplification of the Weak Formulation in Ω_{pml}

Next, we focus on simplifying the integrals needed to convert (5.37) into the polynomial eigenproblem we seek to solve. For ease of computation, we compute $a(\hat{x})\hat{\nabla}\tilde{u}\cdot\hat{\nabla}\tilde{v}$, starting with $J^{-T}\hat{\nabla}\tilde{u}$. Then

$$\begin{aligned}
J^{-T}\hat{\nabla}\tilde{u} &= \left(\frac{\hat{r}}{\eta}I - \frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\hat{x}\hat{x}^T\right)\hat{\nabla}\tilde{u} \\
&= \left(\frac{\hat{r}}{\eta}\hat{\nabla}\tilde{u}\right) - \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x}
\end{aligned} \tag{5.44}$$

and

$$\det(J)J^{-1} = \frac{\dot{\eta}\eta}{\hat{r}}\left(\left(\frac{\hat{r}}{\eta}\right)I - \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)\hat{x}\hat{x}^T\right)$$

$$= \dot{\eta}I - \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^3}\right)\hat{x}\hat{x}^T. \tag{5.45}$$

Using (5.44) and (5.45), we have

$$\det(J)J^{-1}J^{-1}\hat{\nabla}\tilde{u} = \frac{\dot{\eta}\hat{r}}{\eta}\hat{\nabla}\tilde{u} - \left(\frac{2\dot{\eta}\hat{r}-2\eta}{\hat{r}^2\eta}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x} + \left(\frac{(\dot{\eta}\hat{r}-\eta)^2}{\hat{r}^5\dot{\eta}\eta}\right)\hat{r}^2(\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x}$$

$$= \frac{\dot{\eta}\hat{r}}{\eta}\hat{\nabla}\tilde{u} - \left(\frac{2\dot{\eta}\hat{r}-2\eta}{\hat{r}^2\eta}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x} + \left(\frac{(\dot{\eta}\hat{r}-\eta)^2}{\hat{r}^3\dot{\eta}\eta}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x}$$

$$= \frac{\dot{\eta}\hat{r}}{\eta}\hat{\nabla}\tilde{u} + \left[-\frac{2\dot{\eta}\hat{r}-2\eta}{\hat{r}^2\eta} + \frac{(\dot{\eta}\hat{r}-\eta)^2}{\hat{r}^3\dot{\eta}\eta}\right](\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x}$$

$$= \frac{\dot{\eta}\hat{r}}{\eta}\hat{\nabla}\tilde{u} + \left[\frac{-2\dot{\eta}\hat{r}+2\eta}{\hat{r}^2\eta} + \frac{\dot{\eta}^2\hat{r}^2-2\dot{\eta}\eta\hat{r}+\eta^2}{\hat{r}^3\dot{\eta}\eta}\right](\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x}$$

$$= \frac{\dot{\eta}\hat{r}}{\eta}\hat{\nabla}\tilde{u} + \left[\frac{-2\dot{\eta}^2\hat{r}^2+2\eta\dot{\eta}\hat{r}}{\hat{r}^3\dot{\eta}\eta} + \frac{\dot{\eta}^2\hat{r}^2-2\dot{\eta}\eta\hat{r}+\eta^2}{\hat{r}^3\dot{\eta}\eta}\right](\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x}$$

$$= \frac{\dot{\eta}\hat{r}}{\eta}\hat{\nabla}\tilde{u} + \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\hat{r}^3\dot{\eta}\eta}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x}. \tag{5.46}$$

Next, we complete the work for computing the first integrand by computing the dot product with the gradient of a specially chosen test function (5.42). Using (5.43) and

(5.46), we have

$$\det(J)J^{-1}J^{-1}\hat{\nabla}\tilde{u}\cdot\hat{\nabla}\tilde{v} = \left(\frac{\dot{\eta}\hat{r}}{\eta}\hat{\nabla}\tilde{u} + \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\hat{r}^3\dot{\eta}\eta}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{x}\right)\left(\frac{\dot{\eta}v}{\hat{r}\hat{R}}\hat{x} + \frac{\eta}{\hat{R}}\hat{\nabla}v\right)$$

$$= \frac{\dot{\eta}^2}{\eta\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})v + \frac{\dot{\eta}\hat{r}}{\hat{R}}\hat{\nabla}\tilde{u}\cdot\hat{\nabla}v + \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\eta\hat{r}^4\hat{R}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})\hat{r}^2v$$
$$+ \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\dot{\eta}\hat{r}^3\hat{R}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})$$

$$= \frac{\dot{\eta}^2}{\eta\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})v + \frac{\dot{\eta}\hat{r}}{\hat{R}}\hat{\nabla}\tilde{u}\cdot\hat{\nabla}v + \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\eta\hat{r}^2\hat{R}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})v$$
$$+ \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\dot{\eta}\hat{r}^3\hat{R}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})$$

$$= \frac{\dot{\eta}^2\hat{r}^2}{\eta\hat{r}^2\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})v + \frac{\dot{\eta}\hat{r}}{\hat{R}}\hat{\nabla}\tilde{u}\cdot\hat{\nabla}v + \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\eta\hat{r}^2\hat{R}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})v$$
$$+ \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\dot{\eta}\hat{r}^3\hat{R}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})$$

$$= \frac{\dot{\eta}\hat{r}}{\hat{R}}\hat{\nabla}\tilde{u}\cdot\hat{\nabla}v + \left(\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\dot{\eta}\hat{r}^3\hat{R}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})$$
$$+ \frac{\eta^2}{\eta\hat{r}^2\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})v$$

$$= \frac{\dot{\eta}\hat{r}}{\hat{R}}\hat{\nabla}\tilde{u}\cdot\hat{\nabla}v + \frac{1}{\hat{R}}\left(\frac{\eta^2}{\dot{\eta}\hat{r}^3} - \frac{\dot{\eta}}{\hat{r}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})$$
$$+ \frac{\eta}{\hat{r}^2\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})v \tag{5.47}$$

This is the precise expression for the first integrand in the PML region Ω_{pml} as stated

in [27]. Next, we turn our attention to the integral involving the function $V = V(\hat{x}_1, \hat{x}_2)$. With the use of our specially chosen test functions $\tilde{v} \in \mathcal{V}$, we have that

$$\int_\Omega V\tilde{u}\tilde{v}\det(J)d\hat{x} = \int_{\Omega_{int}} V\tilde{u}v \cdot 1 d\hat{x} + \int_{\Omega_{pml}} 0 \cdot \tilde{u}\tilde{v}\det(J)d\hat{x}$$

$$= \int_{\Omega_{int}} V\tilde{u}v d\hat{x}$$

This term does not involve any factor of η or $\dot{\eta}$, and hence will be associated with other integrals involving the term Z^0. The right-hand side integral corresponding to the bilinear form b (and using the specially chosen test functions $\tilde{v}$) can be written as

$$\int_\Omega \tilde{u}\tilde{v}\det(J)d\hat{x} = \int_{\Omega_{int}} \tilde{u}\tilde{v}\det(J)d\hat{x} + \int_{\Omega_{pml}} \tilde{u}\tilde{v}\det(J)d\hat{x}$$

$$= \int_{\Omega_{int}} \tilde{u}v d\hat{x} + \int_{\Omega_{pml}} \tilde{u}v\frac{\eta}{\hat{R}}\frac{\dot{\eta}\eta}{\hat{r}}d\hat{x}$$

$$= \int_{\Omega_{int}} \tilde{u}v d\hat{x} + \int_{\Omega_{pml}} \tilde{u}v\frac{\dot{\eta}\eta^2}{\hat{r}\hat{R}}d\hat{x} \tag{5.48}$$

Our next step is to now expand powers of η and pull out factors of $\dot{\eta} = (1 + \alpha i)/Z$, as factors and powers of $\dot{\eta}$ will correspond to powers of the eigenvalues Z we wish to compute. We begin with the second integral in (5.48).

$$\int_{\Omega_{pml}} \tilde{u}v\frac{\dot{\eta}\eta^2}{\hat{r}\hat{R}}d\hat{x} = \int_{\Omega_{pml}} \tilde{u}v\frac{\dot{\eta}[\dot{\eta}(\hat{r}-\hat{R})+\hat{R}]^2}{\hat{r}\hat{R}}d\hat{x}$$

$$= \int_{\Omega_{pml}} \tilde{u}v\frac{\dot{\eta}[\dot{\eta}^2(\hat{r}-\hat{R})^2+2\dot{\eta}\hat{R}(\hat{r}-\hat{R})+\hat{R}^2]}{\hat{r}\hat{R}}d\hat{x}$$

$$= \int_{\Omega_{pml}} \tilde{u}v\frac{\dot{\eta}^3(\hat{r}-\hat{R})^2}{\hat{r}\hat{R}}d\hat{x} + \int_{\Omega_{pml}} \tilde{u}v\frac{2\dot{\eta}^2\hat{R}(\hat{r}-\hat{R})}{\hat{r}\hat{R}}d\hat{x} + \int_{\Omega_{pml}} \tilde{u}v\frac{\dot{\eta}\hat{R}^2}{\hat{r}\hat{R}}d\hat{x}$$

$$= \int_{\Omega_{pml}} \tilde{u}v\frac{\dot{\eta}^3(\hat{r}-\hat{R})^2}{\hat{r}\hat{R}}d\hat{x} + \int_{\Omega_{pml}} \tilde{u}v\frac{2\dot{\eta}^2(\hat{r}-\hat{R})}{\hat{r}}d\hat{x} + \int_{\Omega_{pml}} \tilde{u}v\frac{\dot{\eta}\hat{R}}{\hat{r}}d\hat{x} \tag{5.49}$$

Now recall that the first integral over Ω_{int} in (5.48) and the three integrals in (5.49) are multiplied by a power of Z^2. Then it follows that these integrals can be written as

$$Z^2\int_{\Omega_{int}} \tilde{u}vd\hat{x} \tag{5.50a}$$

$$Z^2\dot{\eta}^3\int_{\Omega_{pml}} \tilde{u}v\frac{(\hat{r}-\hat{R})^2}{\hat{r}\hat{R}}d\hat{x} = Z^{-1}(1+\alpha i)^3\int_{\Omega_{pml}} \tilde{u}v\frac{(\hat{r}-\hat{R})^2}{\hat{r}\hat{R}}d\hat{x} \tag{5.50b}$$

$$2Z^2\dot{\eta}^2\int_{\Omega_{pml}} \tilde{u}v\frac{(\hat{r}-\hat{R})}{\hat{r}}d\hat{x} = 2(1+\alpha i)^2\int_{\Omega_{pml}} \tilde{u}v\frac{(\hat{r}-\hat{R})}{\hat{r}}d\hat{x} \tag{5.50c}$$

$$Z^2\dot{\eta}\hat{R}\int_{\Omega_{pml}} \tilde{u}v\frac{1}{\hat{r}}d\hat{x} = Z\hat{R}(1+\alpha i)\int_{\Omega_{pml}} \tilde{u}v\frac{1}{\hat{r}}d\hat{x} \tag{5.50d}$$

Next, we revisit the integrand (5.47) in order to collect like powers of $\dot{\eta}$ and expand out powers of η. To begin, we have that

$$\int_{\Omega_{pml}} \frac{\dot{\eta}\hat{r}}{\hat{R}} \hat{\nabla}\tilde{u} \cdot \hat{\nabla} v d\hat{x} = \frac{\dot{\eta}}{\hat{R}} \int_{\Omega_{pml}} \hat{r}\hat{\nabla}\tilde{u} \cdot \hat{\nabla} v d\hat{x}$$

$$= Z^{-1} \frac{(1+\alpha i)}{\hat{R}} \int_{\Omega_{pml}} \hat{r}\hat{\nabla}\tilde{u} \cdot \hat{\nabla} v d\hat{x}. \tag{5.51}$$

To tackle the second integrand, we expand out the term adjacent to the product $(\hat{\nabla}\tilde{u} \cdot \hat{x})(\hat{\nabla} v \cdot \hat{x})$ and group like powers of $\dot{\eta}$.

$$\frac{1}{\hat{R}} \left(\frac{\eta^2}{\dot{\eta}\hat{r}^3} - \frac{\dot{\eta}}{\hat{r}} \right) = \frac{1}{\hat{R}} \left(\frac{[\dot{\eta}(\hat{r} - \hat{R}) + \hat{R}]^2}{\dot{\eta}\hat{r}^3} - \frac{\dot{\eta}}{\hat{r}} \right)$$

$$= \frac{1}{\hat{R}} \left(\frac{[\dot{\eta}^2(\hat{r} - \hat{R})^2 + 2\dot{\eta}\hat{R}(\hat{r} - \hat{R}) + \hat{R}^2]}{\dot{\eta}\hat{r}^3} - \frac{\dot{\eta}}{\hat{r}} \right)$$

$$= \frac{1}{\hat{R}} \left(\frac{\dot{\eta}^2(\hat{r} - \hat{R})^2}{\dot{\eta}\hat{r}^3} + \frac{2\dot{\eta}\hat{R}(\hat{r} - \hat{R})}{\dot{\eta}\hat{r}^3} + \frac{\hat{R}^2}{\dot{\eta}\hat{r}^3} - \frac{\dot{\eta}}{\hat{r}} \right)$$

$$= \frac{1}{\hat{R}} \left(\frac{\dot{\eta}(\hat{r} - \hat{R})^2}{\hat{r}^3} - \frac{\dot{\eta}}{\hat{r}} \right) + \frac{2(\hat{r} - \hat{R})}{\hat{r}^3} + \frac{\hat{R}}{\dot{\eta}\hat{r}^3}$$

$$= \frac{\dot{\eta}}{\hat{R}} \left(\frac{(\hat{r} - \hat{R})^2}{\hat{r}^3} - \frac{1}{\hat{r}} \right) + \frac{2(\hat{r} - \hat{R})}{\hat{r}^3} + \frac{\hat{R}}{\dot{\eta}\hat{r}^3} \tag{5.52}$$

The corresponding integrals written in terms of powers of Z, are

$$Z^{-1}(1+\alpha i)\int_{\Omega_{pml}} \frac{1}{\hat{R}}\left(\frac{(\hat{r}-\hat{R})^2}{\hat{r}^3}-\frac{1}{\hat{r}}\right)(\hat{\nabla}\tilde{u}\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})d\hat{x} \tag{5.53a}$$

$$Z^0\int_{\Omega_{pml}} \frac{2(\hat{r}-\hat{R})}{\hat{r}^3}(\hat{\nabla}\tilde{u}\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})d\hat{x} \tag{5.53b}$$

$$Z\frac{\hat{R}}{1+\alpha i}\int_{\Omega_{pml}} \frac{1}{\hat{r}^3}(\hat{\nabla}\tilde{u}\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})d\hat{x} \tag{5.53c}$$

The last term of (5.47) can be written as

$$\frac{\eta}{\hat{r}^2\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})v=\dot{\eta}\frac{\hat{r}-\hat{R}}{\hat{r}^2\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})v+\frac{\hat{R}}{\hat{r}^2\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})v,$$

and so the corresponding integrals are

$$Z^{-1}(1+\alpha i)\int_{\Omega_{pml}} \frac{\hat{r}-\hat{R}}{\hat{r}^2\hat{R}}(\hat{\nabla}\tilde{u}\cdot\hat{x})vd\hat{x} \tag{5.54a}$$

$$\int_{\Omega_{pml}} \frac{1}{\hat{r}^2}(\hat{\nabla}\tilde{u}\cdot\hat{x})vd\hat{x} \tag{5.54b}$$

Next, we take the results of (5.50), (5.52), (5.53), and (5.54) to define the following bilinear forms for $w, v \in \mathcal{V}$.

$$\begin{aligned} b_0(w,v) &= (1+\alpha i)\int_{\Omega_{pml}} \frac{\hat{r}}{\hat{R}}\hat{\nabla}w\cdot\hat{\nabla}v+\frac{1}{\hat{R}}\left(\frac{(\hat{r}-\hat{R})^2}{\hat{r}^3}-\frac{1}{\hat{r}}\right)(\hat{\nabla}w\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})d\hat{x} \\ &\quad+(1+\alpha i)\int_{\Omega_{pml}} \frac{\hat{r}-\hat{R}}{\hat{r}^2\hat{R}}(\hat{\nabla}w\cdot\hat{x})vd\hat{x}-(1+\alpha i)^3\int_{\Omega_{pml}} wv\frac{(\hat{r}-\hat{R})^2}{\hat{r}\hat{R}}d\hat{x} \end{aligned} \tag{5.55a}$$

$$\begin{aligned} b_1(w,v) &= \int_{\Omega_{int}} \hat{\nabla}w\cdot\hat{\nabla}v+Vwvd\hat{x} \\ &\quad+\int_{\Omega_{pml}} \frac{2(\hat{r}-\hat{R})}{\hat{r}^3}(\hat{\nabla}w\cdot\hat{x})(\hat{\nabla}v\cdot\hat{x})+\frac{1}{\hat{r}^2}(\hat{\nabla}w\cdot\hat{x})vd\hat{x} \end{aligned}$$

$$- 2(1+\alpha i)^2 \int_{\Omega_{pml}} wv \frac{(\hat{r} - \hat{R})}{\hat{r}} d\hat{x} \tag{5.55b}$$

$$b_2(w,v) = \frac{\hat{R}}{1+\alpha i} \int_{\Omega_{pml}} \frac{1}{\hat{r}^3} (\hat{\nabla} w \cdot \hat{x})(\hat{\nabla} v \cdot \hat{x}) d\hat{x} - \hat{R}(1+\alpha i) \int_{\Omega_{pml}} wv \frac{1}{\hat{r}} d\hat{x} \tag{5.55c}$$

$$b_3(w,v) = - \int_{\Omega_{int}} wv d\hat{x} \tag{5.55d}$$

The nonlinear eigenvalue problem can be stated as follows: Find a $\tilde{u} \in \mathcal{V}$ and $Z \in \mathbb{C}$ such that for all $v \in \mathcal{V}$,

$$\sum_{i=0}^{3} Z^{i-1} b_i(\tilde{u}, v) = 0. \tag{5.56}$$

Note that we have technically derived a rational eigenproblem with a pole at $Z = 0$ [32, 61]. Since we are not concerned with $Z = 0$ as an eigenvalue, we multiply (5.56) through by Z and obtain

$$\sum_{i=0}^{3} Z^{i} b_i(\tilde{u}, v) = 0. \tag{5.57}$$

To discretize this problem, let $\mathcal{T}_h$ be a geometrically conforming triangular mesh of Ω where $h > 0$ is the mesh size. For a polynomial degree $p > 0$, let W_{hp} denote the lagrange finite element space

$$W_{hp} = \{v \in V : v|_K \in \mathbb{P}_p \quad \forall K \in \mathcal{T}_h\} \tag{5.58}$$

where $\mathbb{P}_p$ is the space of polynomials in the variables $\hat{x}_1, \hat{x}_2$ of degree at most p. In this finite-dimensional setting, we seek to find a $\tilde{u}_{hp} \in W_{hp}$ satisfying

$$\sum_{i=0}^{3} Z^i b_i(\tilde{u}_{hp}, v) = 0 \quad \forall v \in W_{hp}. \tag{5.59}$$

Denote the basis of W_{hp} by $\{\phi_j\}_{j=1}^n$, and define the matrices A_i, $i = 0, 1, 2, 3$ by

$$[A_i]_{kl} = b_i(\phi_l, \phi_k).$$

letting $v = \phi_k$ for $k = 1, 2, \ldots, n$ and substituting the representation of u_{hp} in the basis of W_{hp}, we obtain

$$\begin{aligned} \sum_{i=0}^{3} Z^i b_i(\tilde{u}_{hp}, v) &= \sum_{i=0}^{3} Z^i b_i \left(\sum_{j=1}^{n} c_j \phi_j, \phi_k \right) \qquad (5.60) \\ &= \sum_{i=0}^{3} Z^i \sum_{j=1}^{n} b_i(\phi_j, \phi_k) c_j \\ &= \sum_{i=0}^{3} Z^i \sum_{j=1}^{n} A_{kj} c_j \qquad (5.61) \end{aligned}$$

for each $k = 1, 2, \ldots, n$. More concisely, we can state this problem as finding a $c \in \mathbb{C}^n$ and $Z \in \mathbb{C}$ satisfying

$$P(Z)c = 0 \tag{5.62}$$

where $P(Z) = A_0 + ZA_1 + Z^2A_2 + Z^3A_3$.

5.7.4 A Complex-Symmetric Weak Formulation

To go one step further, we can derive a cubic eigenproblem that is complex-symmetric by letting

$$\tilde{u} = \left(\frac{\eta(\hat{r})}{\hat{R}}\right)^{1/2} \breve{u}, \quad v = \left(\frac{\eta(\hat{r})}{\hat{R}}\right)^{1/2} \breve{v}.$$

in Ω_{pml}, with $\tilde{u} = \breve{u}$ and $v = \breve{v}$ in Ω_{int}. Consequently, we need to compute some quantities involving gradients, dot products, and products of functions in other integrals. To this end, we begin by computing the gradients of $\tilde{u}$ and v:

$$\begin{aligned}
\hat{\nabla}\tilde{u} &= \hat{\nabla}\left(\frac{\eta}{\hat{R}}\right)^{1/2} \breve{u} + \left(\frac{\eta}{\hat{R}}\right)^{1/2} \hat{\nabla}\breve{u} \\
&= \frac{1}{2}\left(\frac{\eta}{\hat{R}}\right)^{-1/2} \frac{\dot{\eta}}{\hat{r}\hat{R}}\hat{x}\breve{u} + \left(\frac{\eta}{\hat{R}}\right)^{1/2} \hat{\nabla}\breve{u} \\
&= \left(\frac{\eta}{\hat{R}}\right)^{-1/2} \frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\hat{x} + \left(\frac{\eta}{\hat{R}}\right)^{1/2} \hat{\nabla}\breve{u} \qquad (5.63a)
\end{aligned}$$

Similarly, we have that

$$\hat{\nabla}v = \left(\frac{\eta}{\hat{R}}\right)^{-1/2} \frac{\dot{\eta}\breve{v}}{2\hat{r}\hat{R}}\hat{x} + \left(\frac{\eta}{\hat{R}}\right)^{1/2} \hat{\nabla}\breve{v} \qquad (5.63b)$$

Next, we compute the quantities $J^{-T}\hat{\nabla}\tilde{u}$, as the computations for $J^{-T}\hat{\nabla}v$ are similar. Thus, we have

$$J^{-T}\hat{\nabla}\tilde{u} = \left(\frac{\hat{r}}{\eta}I - \frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\hat{x}\hat{x}^T\right)\left(\left(\frac{\eta}{\hat{R}}\right)^{-1/2}\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\hat{x} + \left(\frac{\eta}{\hat{R}}\right)^{1/2}\hat{\nabla}\breve{u}\right)$$

$$= \frac{\hat{r}}{\eta}\left(\frac{\eta}{\hat{R}}\right)^{-1/2}\left(\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\right)\hat{x} + \frac{\hat{r}}{\eta}\left(\frac{\eta}{\hat{R}}\right)^{1/2}\hat{\nabla}\breve{u}$$
$$- \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)\left(\frac{\eta}{\hat{R}}\right)^{-1/2}\left(\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\right)\hat{r}^2\hat{x} - \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)\left(\frac{\eta}{\hat{R}}\right)^{1/2}(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}$$

$$= \frac{\hat{r}}{\eta}\left(\frac{\eta}{\hat{R}}\right)^{-1/2}\left(\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\right)\hat{x} + \frac{\hat{r}}{\eta}\left(\frac{\eta}{\hat{R}}\right)^{1/2}\hat{\nabla}\breve{u}$$
$$- \left(\frac{\dot{\eta}\hat{r}-\eta}{\dot{\eta}\eta}\right)\left(\frac{\eta}{\hat{R}}\right)^{-1/2}\left(\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\right)\hat{x} - \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)\left(\frac{\eta}{\hat{R}}\right)^{1/2}(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}$$

$$= \frac{\hat{r}}{\eta}\left(\frac{\eta}{\hat{R}}\right)^{-1/2}\left(\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\right)\hat{x} - \left(\frac{\dot{\eta}\hat{r}-\eta}{\dot{\eta}\eta}\right)\left(\frac{\eta}{\hat{R}}\right)^{-1/2}\left(\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\right)\hat{x}$$
$$+ \frac{\hat{r}}{\eta}\left(\frac{\eta}{\hat{R}}\right)^{1/2}\hat{\nabla}\breve{u} - \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)\left(\frac{\eta}{\hat{R}}\right)^{1/2}(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}$$

$$= \left(\frac{\eta}{\hat{R}}\right)^{-1/2}\left(\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\right)\left[\frac{\dot{\eta}\hat{r}}{\dot{\eta}\eta} - \left(\frac{\dot{\eta}\hat{r}-\eta}{\dot{\eta}\eta}\right)\right]\hat{x}$$
$$+ \left(\frac{\eta}{\hat{R}}\right)^{1/2}\left[\frac{\hat{r}}{\eta}\hat{\nabla}\breve{u} - \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}\right]$$

$$= \left(\frac{\eta}{\hat{R}}\right)^{-1/2}\left(\frac{\dot{\eta}\breve{u}}{2\hat{r}\hat{R}}\right)\left[\frac{\eta}{\dot{\eta}\eta}\right]\hat{x}$$
$$+ \left(\frac{\eta}{\hat{R}}\right)^{1/2}\left[\frac{\hat{r}}{\eta}\hat{\nabla}\breve{u} - \left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}\right]$$

$$= \left(\frac{\eta}{\hat{R}}\right)^{-1/2} \left(\frac{\breve{u}}{2\hat{r}\hat{R}}\right) \hat{x} + \left(\frac{\eta}{\hat{R}}\right)^{1/2} \left[\frac{\hat{r}}{\eta}\hat{\nabla}\breve{u} - \left(\frac{\dot{\eta}\hat{r} - \eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}\right] \quad (5.64a)$$

Likewise, $J^{-T}\hat{\nabla}v$ is given by

$$J^{-T}\hat{\nabla}v = \left(\frac{\eta}{\hat{R}}\right)^{-1/2} \left(\frac{\breve{v}}{2\hat{r}\hat{R}}\right) \hat{x} + \left(\frac{\eta}{\hat{R}}\right)^{1/2} \left[\frac{\hat{r}}{\eta}\hat{\nabla}\breve{v} - \left(\frac{\dot{\eta}\hat{r} - \eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{v}\cdot\hat{x})\hat{x}\right]. \quad (5.64b)$$

We compute the product $J^{-T}\hat{\nabla}\tilde{u}\cdot J^{-T}\hat{\nabla}v$ term-by-term. Then

$$\left(\frac{\eta}{\hat{R}}\right)^{-1/2} \left(\frac{\breve{u}}{2\hat{r}\hat{R}}\right) \hat{x} \cdot \left(\frac{\eta}{\hat{R}}\right)^{-1/2} \left(\frac{\breve{v}}{2\hat{r}\hat{R}}\right) \hat{x} = \left(\frac{\eta}{\hat{R}}\right)^{-1} \left(\frac{\breve{u}\breve{v}}{4\hat{r}^2\hat{R}^2}\right) \hat{x}\cdot\hat{x}$$

$$= \left(\frac{\eta}{\hat{R}}\right)^{-1} \left(\frac{\breve{u}\breve{v}}{4\hat{r}^2\hat{R}^2}\right) \hat{r}^2$$

$$= \frac{\breve{u}\breve{v}}{4\eta\hat{R}} \quad (5.65a)$$

Next, we dot the first term of (5.64a) with the second term of (5.64b) to obtain

$$\left(\frac{\breve{u}}{2\hat{r}\hat{R}}\right) \hat{x} \cdot \left[\frac{\hat{r}}{\eta}\hat{\nabla}\breve{v} - \left(\frac{\dot{\eta}\hat{r} - \eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{v}\cdot\hat{x})\hat{x}\right] = \frac{(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}}{2\eta\hat{R}} - \left(\frac{\dot{\eta}\hat{r} - \eta}{2\hat{r}^3\hat{R}\dot{\eta}\eta}\right)(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}\hat{r}^2$$

$$= \frac{(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}}{2\eta\hat{R}} - \left(\frac{\dot{\eta}\hat{r} - \eta}{2\hat{r}\hat{R}\dot{\eta}\eta}\right)(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}$$

$$= \frac{(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}}{2\eta\hat{R}}\left[1 - \frac{\dot{\eta}\hat{r} - \eta}{\dot{\eta}\hat{r}}\right]$$

$$= \frac{(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}}{2\eta\hat{R}}\left[\frac{\eta}{\dot{\eta}\hat{r}}\right]$$

$$= \frac{(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}}{2\dot{\eta}\hat{r}\hat{R}} \tag{5.65b}$$

Likewise, we dot the second term of (5.64a) with the first term of (5.64b) to obtain

$$\left(\frac{\breve{v}}{2\hat{r}\hat{R}}\right)\hat{x}\cdot\left[\frac{\hat{r}}{\eta}\hat{\nabla}\breve{u}-\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}\right]=\frac{(\hat{\nabla}\breve{u}\cdot\hat{x})\breve{v}}{2\dot{\eta}\hat{r}\hat{R}} \tag{5.65c}$$

Finally, we dot the second term of (5.64a) with the second term of (5.64b) to obtain

$$\frac{\eta}{\hat{R}}\left[\frac{\hat{r}}{\eta}\hat{\nabla}\breve{u}-\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}\right]\cdot\left[\frac{\hat{r}}{\eta}\hat{\nabla}\breve{v}-\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{v}\cdot\hat{x})\hat{x}\right] \tag{5.65d}$$

Computing this dot product term-by-term yields

$$\frac{\eta}{\hat{R}}\left(\frac{\hat{r}}{\eta}\hat{\nabla}\breve{u}\right)\cdot\left(\frac{\hat{r}}{\eta}\hat{\nabla}\breve{v}\right)=\frac{\hat{r}^2}{\hat{R}\eta}(\hat{\nabla}\breve{u}\cdot\hat{\nabla}\breve{v}) \tag{5.66a}$$

$$-\frac{\eta}{\hat{R}}\left(\frac{\hat{r}}{\eta}\hat{\nabla}\breve{u}\right)\cdot\left(\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{v}\cdot\hat{x})\hat{x}\right)=-\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}\hat{R}\dot{\eta}\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x}) \tag{5.66b}$$

$$-\frac{\eta}{\hat{R}}\left(\frac{\hat{r}}{\eta}\hat{\nabla}\breve{v}\right)\cdot\left(\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}\right)=-\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}\hat{R}\dot{\eta}\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x}) \tag{5.66c}$$

$$\frac{\eta}{\hat{R}}\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}^2\dot{\eta}\eta}\right)^2(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{u}\cdot\hat{x})\hat{x}\cdot\hat{x}=\frac{\eta}{\hat{R}}\left(\frac{(\dot{\eta}\hat{r}-\eta)^2}{\hat{r}^4\dot{\eta}^2\eta^2}\right)\hat{r}^2(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x})$$

$$=\left(\frac{\dot{\eta}^2\hat{r}^2-2\dot{\eta}\eta\hat{r}+\eta^2}{\hat{r}^2\hat{R}\dot{\eta}^2\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x}) \tag{5.66d}$$

Combining the coefficients of $(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x})$ in (5.66b), (5.66c), and (5.66d) yield

$$-2\left(\frac{\dot{\eta}\hat{r}-\eta}{\hat{r}\hat{R}\dot{\eta}\eta}\right)+\left(\frac{\dot{\eta}^2\hat{r}^2-2\dot{\eta}\eta\hat{r}+\eta^2}{\hat{r}^2\hat{R}\dot{\eta}^2\eta}\right)=\frac{2\eta-2\dot{\eta}\hat{r}}{\hat{r}\hat{R}\dot{\eta}\eta}+\frac{\dot{\eta}^2\hat{r}^2-2\dot{\eta}\eta\hat{r}+\eta^2}{\hat{r}^2\hat{R}\dot{\eta}^2\eta}$$

$$=\frac{2\dot{\eta}\eta\hat{r}-2\dot{\eta}^2\hat{r}^2}{\hat{r}^2\hat{R}\dot{\eta}^2\eta}+\frac{\dot{\eta}^2\hat{r}^2-2\dot{\eta}\eta\hat{r}+\eta^2}{\hat{r}^2\hat{R}\dot{\eta}^2\eta}$$

$$=\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\hat{r}^2\hat{R}\dot{\eta}^2\eta}$$

Then (5.65d) can be written as

$$\frac{\hat{r}^2}{\hat{R}\eta}(\hat{\nabla}\breve{u}\cdot\hat{\nabla}\breve{v})+\frac{\eta^2-\dot{\eta}^2\hat{r}^2}{\hat{r}^2\hat{R}\dot{\eta}^2\eta}(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x})=\frac{\hat{r}^2}{\hat{R}\eta}(\hat{\nabla}\breve{u}\cdot\hat{\nabla}\breve{v})$$
$$+\frac{1}{\hat{R}}\left(\frac{\eta^2}{\hat{r}^2\dot{\eta}^2\eta}-\frac{\dot{\eta}^2\hat{r}^2}{\hat{r}^2\dot{\eta}^2\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x})$$

$$=\frac{\hat{r}^2}{\hat{R}\eta}(\hat{\nabla}\breve{u}\cdot\hat{\nabla}\breve{v})$$

$$+\frac{1}{\hat{R}}\left(\frac{\eta}{\hat{r}^2\dot{\eta}^2}-\frac{1}{\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x}). \tag{5.67}$$

Summing all terms of (5.65) and using the results of (5.67) yields

$$\frac{\hat{r}^2}{\hat{R}\eta}(\hat{\nabla}\breve{u}\cdot\hat{\nabla}\breve{v})+\frac{1}{\hat{R}}\left(\frac{\eta}{\hat{r}^2\dot{\eta}^2}-\frac{1}{\eta}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x})+\frac{1}{2\dot{\eta}\hat{r}\hat{R}}\left((\hat{\nabla}\breve{u}\cdot\hat{x})\breve{v}+(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}\right)$$
$$+\frac{\breve{u}\breve{v}}{4\eta\hat{R}} \tag{5.68}$$

Multiplying (5.69) through by $\det(J)=\dot{\eta}\eta/\hat{r}$ and simplifying the second term yields

$$\frac{\dot{\eta}\hat{r}}{\hat{R}}(\hat{\nabla}\breve{u}\cdot\hat{\nabla}\breve{v})+\frac{1}{\hat{r}^2\hat{R}}\left(\frac{\eta^2}{\dot{\eta}\hat{r}}-\dot{\eta}\hat{r}\right)(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x})+\frac{\eta}{2\hat{r}^2\hat{R}}\left((\hat{\nabla}\breve{u}\cdot\hat{x})\breve{v}+(\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}\right)$$
$$+\frac{\dot{\eta}}{4\hat{r}\hat{R}}\breve{u}\breve{v} \tag{5.69}$$

To this end, we just need to tackle the remaining integrals on the left and right-hand-sides. To this end, we have that

$$\int_\Omega V\tilde{u}v\det(J)d\hat{x}=\int_{\Omega_{int}}V\tilde{u}v\cdot 1d\hat{x}+\int_\Omega 0\cdot\tilde{u}v\det(J)d\hat{x}$$

$$=\int_{\Omega_{int}}V\breve{u}\breve{v}d\hat{x}. \tag{5.70}$$

On the right-hand-side, we have

$$\int_{\Omega} \tilde{u}v \det(J) d\hat{x} = \int_{\Omega_{int}} \breve{u}\breve{v} d\hat{x} + \int_{\Omega_{pml}} \left(\frac{\eta}{\hat{R}}\right)^{1/2} \breve{u} \left(\frac{\eta}{\hat{R}}\right)^{1/2} \breve{v} \frac{\dot{\eta}\eta}{\hat{r}} d\hat{x}$$

$$= \int_{\Omega_{int}} \breve{u}\breve{v} d\hat{x} + \int_{\Omega_{pml}} \left(\frac{\eta}{\hat{R}}\right) \breve{u}\breve{v} \frac{\dot{\eta}\eta}{\hat{r}} d\hat{x}$$

$$= \int_{\Omega_{int}} \breve{u}\breve{v} d\hat{x} + \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{\dot{\eta}\eta^2}{\hat{r}\hat{R}} d\hat{x}$$

$$= \int_{\Omega_{int}} \breve{u}\breve{v} d\hat{x} + \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{\dot{\eta}[\dot{\eta}(\hat{r} - \hat{R}) + \hat{R}]^2}{\hat{r}\hat{R}} d\hat{x} \tag{5.71}$$

The last integral in (5.71) can be split into three terms, which we show below.

$$\int_{\Omega_{pml}} \breve{u}\breve{v} \frac{\dot{\eta}[\dot{\eta}(\hat{r} - \hat{R}) + \hat{R}]^2}{\hat{r}\hat{R}} d\hat{x} = \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{\dot{\eta}[\dot{\eta}^2(\hat{r} - \hat{R})^2 + 2\dot{\eta}\hat{R}(\hat{r} - \hat{R}) + \hat{R}^2]}{\hat{r}\hat{R}} d\hat{x}$$

$$= \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{\dot{\eta}^3(\hat{r} - \hat{R})^2 + 2\dot{\eta}^2\hat{R}(\hat{r} - \hat{R}) + \dot{\eta}\hat{R}^2}{\hat{r}\hat{R}} d\hat{x} \tag{5.72}$$

Next, we split (5.72) into three terms to obtain

$$\int_{\Omega_{pml}} \breve{u}\breve{v} \frac{\dot{\eta}^3(\hat{r} - \hat{R})^2}{\hat{r}\hat{R}} d\hat{x} = Z^{-3} \frac{(1 + \alpha i)^3}{\hat{R}} \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{(\hat{r} - \hat{R})^2}{\hat{r}} d\hat{x} \tag{5.73a}$$

$$\int_{\Omega_{pml}} \breve{u}\breve{v} \frac{2\dot{\eta}^2\hat{R}(\hat{r} - \hat{R})}{\hat{r}\hat{R}} d\hat{x} = Z^{-2}(1 + \alpha i)^2 \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{2(\hat{r} - \hat{R})}{\hat{r}} d\hat{x} \tag{5.73b}$$

$$\int_{\Omega_{pml}} \breve{u}\breve{v}\frac{\dot{\eta}\hat{R}^2}{\hat{r}\hat{R}}d\hat{x} = Z^{-1}\hat{R}(1+\alpha i)\int_{\Omega_{pml}} \breve{u}\breve{v}\frac{1}{\hat{r}}d\hat{x} \tag{5.73c}$$

Since all integrals on the right-hand-side have to be multiplied by a power of Z^2, applying this to the first integral of (5.71) and to the integrals in (5.73) yields

$$Z^2\int_{\Omega_{int}} \breve{u}\breve{v}d\hat{x} \tag{5.74a}$$

$$Z^{-1}\frac{(1+\alpha i)^3}{\hat{R}}\int_{\Omega_{pml}} \breve{u}\breve{v}\frac{(\hat{r}-\hat{R})^2}{\hat{r}}d\hat{x} \tag{5.74b}$$

$$(1+\alpha i)^2\int_{\Omega_{pml}} \breve{u}\breve{v}\frac{2(\hat{r}-\hat{R})}{\hat{r}}d\hat{x} \tag{5.74c}$$

$$Z\hat{R}(1+\alpha i)\int_{\Omega_{pml}} \breve{u}\breve{v}\frac{1}{\hat{r}}d\hat{x} \tag{5.74d}$$

Combining the results of (5.69), (5.70), (5.74), we define the following bilinear forms $b_i : \mathcal{V}\times\mathcal{V}\to\mathbb{C}$ for $i = 0,1,2,3$ and $w, v \in \mathcal{V}$.

$$\begin{aligned} b_0(w,v) = (1+\alpha i)\Bigg[&\int_{\Omega_{pml}} \frac{\hat{r}}{\hat{R}}(\hat{\nabla}\breve{u}\cdot\hat{\nabla}\breve{v}) + \frac{\hat{R}-2\hat{r}}{\hat{r}^3}(\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x})d\hat{x} \\ &+ \int_{\Omega_{pml}} \frac{(\hat{r}-\hat{R})}{2\hat{r}^2\hat{R}}\Big((\hat{\nabla}\breve{u}\cdot\hat{x})\breve{v} + (\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u}\Big) + \frac{\breve{u}\breve{v}}{4\hat{r}\hat{R}}d\hat{x}\Bigg] \end{aligned}$$

$$- \frac{(1+\alpha i)^3}{\hat{R}} \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{(\hat{r}-\hat{R})^2}{\hat{r}} d\hat{x} \tag{5.75a}$$

$$\begin{aligned} b_1(w,v) = &\int_{\Omega_{pml}} \frac{2(\hat{r}-\hat{R})}{\hat{r}^3} (\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x}) d\hat{x} \\ &+ \int_{\Omega_{pml}} \frac{1}{2\hat{r}^2} \left((\hat{\nabla}\breve{u}\cdot\hat{x})\breve{v} + (\hat{\nabla}\breve{v}\cdot\hat{x})\breve{u} \right) d\hat{x} \\ &+ \int_{\Omega_{int}} V\breve{u}\breve{v} d\hat{x} - (1+\alpha i)^2 \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{2(\hat{r}-\hat{R})}{\hat{r}} d\hat{x} \end{aligned}$$

$$b_2(w,v) = \frac{\hat{R}}{1+\alpha i} \int_{\Omega_{pml}} \frac{1}{\hat{r}^3} (\hat{\nabla}\breve{u}\cdot\hat{x})(\hat{\nabla}\breve{v}\cdot\hat{x}) d\hat{x} - \hat{R}(1+\alpha i) \int_{\Omega_{pml}} \breve{u}\breve{v} \frac{1}{\hat{r}} d\hat{x} \tag{5.75b}$$

$$b_3(w,v) = -\int_{\Omega_{int}} \breve{u}\breve{v} d\hat{x} \tag{5.75c}$$

The complex-symmetric nonlinear eigenvalue problem can be stated as follows: Find a $\breve{u} \in \mathcal{V}$ and $Z \in \mathbb{C}$ such that for all $\breve{v} \in \mathcal{V}$,

$$\sum_{i=0}^{3} Z^{i-1} b_i(\breve{u}, \breve{v}) = 0. \tag{5.76}$$

As before, we have technically derived a rational eigenproblem with a pole at $Z = 0$ [32, 61]. Since we are not concerned with $Z = 0$ as an eigenvalue, we multiply (5.56) through by Z and obtain

$$\sum_{i=0}^{3} Z^{i} b_i(\breve{u}, \breve{v}) = 0. \tag{5.77}$$

Our next step, as before, is to let $\mathcal{T}_h$ be a geometrically conforming triangular mesh of Ω where $h > 0$ is the mesh size. For a polynomial degree $p > 0$, let W_{hp} denote the lagrange finite element space as defined in (5.58). In this finite-dimensional setting, we seek to find a $\breve{u}_{hp} \in W_{hp}$ satisfying

$$\sum_{i=0}^{3} Z^i b_i(\breve{u}_{hp}, \breve{v}) = 0 \quad \forall \breve{v} \in W_{hp}. \tag{5.78}$$

Denote the basis of W_{hp} by $\{\phi_j\}_{j=1}^n$, and define the matrices A_i, $i = 0, 1, 2, 3$ by

$$[A_i]_{kl} = b_i(\phi_l, \phi_k).$$

letting $\breve{v} = \phi_k$ for $k = 1, 2, \ldots, n$ and substituting the representation of $\breve{u}_{hp}$ in the basis of W_{hp}, we obtain

$$\begin{aligned} \sum_{i=0}^{3} Z^i b_i(\breve{u}_{hp}, \breve{v}) &= \sum_{i=0}^{3} Z^i b_i \left(\sum_{j=1}^{n} c_j \phi_j, \phi_k \right) \qquad (5.79) \\ &= \sum_{i=0}^{3} Z^i \sum_{j=1}^{n} b_i(\phi_j, \phi_k) c_j \\ &= \sum_{i=0}^{3} Z^i \sum_{j=1}^{n} A_{kj} c_j \end{aligned} \tag{5.80}$$

for each $k = 1, 2, \ldots, n$. More concisely, we can state this problem as finding a $c \in \mathbb{C}^n$ and $Z \in \mathbb{C}$ satisfying

$$P(Z)c = 0 \tag{5.81}$$

where $P(Z) = A_0 + ZA_1 + Z^2A_2 + Z^3A_3$.

Chapter 6

Applications to Fiber Optics

6.1 Introduction

In this chapter, we provide numerical results to verify the correctness of finding guided modes using the DPG discretization, as well as verification of the correctness of Algorithm 4 in Chapter 5 with an application to finding leaky modes of an ytterbium-doped step-index fiber. In addition, we explore the efficacy of our algorithm when applied to the task of computing confinement losses, including an experiment that shows our computed confinement losses remain stable upon varying parameters affecting the PML. This is followed by showing that there is a nontrivial sensitivity to computed confinement losses arising from perturbations in the geometry of a six-capillary microstructure fiber. We begin with a verification of our polynomial eigensolver in computing leaky modes and propagation constants for a step index fiber.

6.2 Step-Index Fiber Guided and Leaky Modes

6.2.1 Guided Mode Verification using the DPG Discretization of the Resolvent

In this last section, we look at applying the FEAST algorithm using the DPG discretization to the task of computing guided modes of a large mode area (LMA) fiber.

Such fibers usually support multiple guided modes, and the location of the corresponding propagation consants is dictated in the optics literature [43, 52]. In combination with known solutions of the form (1.15), this serves as an ideal problem to test the use of the DPG discretization in the FEAST algorithm.

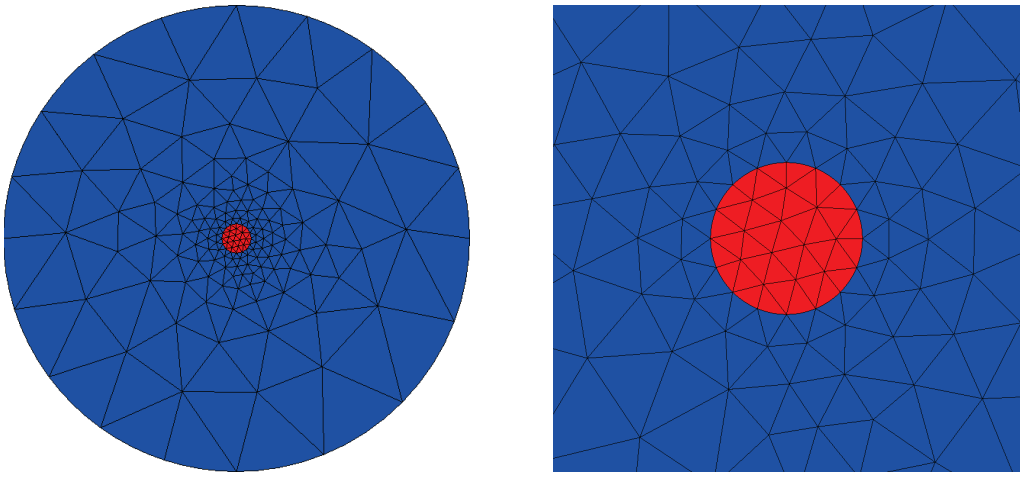
Our computational domain for this problem will correspond to the circular cross-section similar to 1.1. For such fibers, we have a core radius of $r_{core} > 0$ and outer cladding radius $r_{clad} > 0$ corresponding to R_0 and R_{fin}, respectively (see figure 1.1). In this case, our computational domain in $\mathbb{R}^2$ is a cross-section of an optical fiber, whose longitudinal direction (i.e. z-axis) is orthogonal to the circular cross-section. For the optical fiber whose modes we wish to find, we assume that $n(x)$ for $x = (x_1, x_2)^T \in \mathbb{R}^2$ is piecewise constant, taking a value of $n_1(x) \equiv n_{core} > 0$ in the core region $\{x \in \mathbb{R}^2 : x_1^2 + x_2^2 < r_{core}^2\}$ and n_{clad} in the cladding region $\{x \in \mathbb{R}^2 : r_{core}^2 < x_1^2 + x_2^2 < r_{clad}^2\}$, with $n_{clad} < n_{core}$.

The guided modes $\varphi_l(x, y)$ of the fiber are nontrivial functions satisfying, together with propagation constants $\beta_l \in \mathbb{R}$

$$\Delta\varphi_l + k^2 n^2 \varphi_l = \beta_l^2 \varphi_l, \tag{6.1}$$

where k is the wavenumber of the light propagating through the optical fiber. Since the guided modes of the fiber decay exponentially outside of the core region, and because the cladding radius is sufficiently large, we apply a zero-dirichlet boundary condition at the outer radius r_{clad}, i.e. $\varphi_l = 0$ for $||x|| = r_{clad}$. Furthermore, the optics literature directly states that the propagation constants β_l satisfy [43, 52]

Figure 6.1: Step-Index Fiber Mesh



(a) The mesh with curved elements adjacent to the core and cladding boundaries.

(b) Zoomed-in view of the mesh in Figure 6.1a near the core.

Figure 6.1: The mesh used for computing modes of the ytterbium-doped step-index fiber.

$$k^2 n_{clad}^2 < \beta_l^2 < k^2 n_{core}^2,$$

so up to a scaling, we have a search interval to provide to the FEAST algorithm. To this end, the parameters for our commercially-available ytterbium-doped fiber [24] are given by $n_{core} = 1.45097$, $n_{clad} = 1.44973$, $r_{core} = 12.5 \times 10^{-6}$ m, and $r_{clad} = 16 r_{core} = 200 \times 10^{-6}$ m [24]. Due to the small dimensions of the fiber and the large magnitude of the propagation constants we wish to compute, we nondimensionalize the PDE to the disk $\hat{\Omega} = \{\hat{x} \in \mathbb{R}^2 : ||\hat{x}|| < 16\}$ and compute $\hat{\varphi} : \hat{\Omega} \to \mathbb{C}$ satisfying $(\Delta + r_{core}^2 k^2 n^2)\hat{\varphi}_l = r_{core}^2 \beta^2 \hat{\varphi}_l$ on $\hat{\Omega}$ and $\hat{\varphi}_l = 0$ on $\partial\hat{\Omega}$.

In figure 6.1 [24], we provide two different views of our computational domain. Notice that in both subfigures 6.1a and 6.1b, we use isoparametrically curved elements on the boundary of the computational domain and at the interface of the core and cladding regions. This is to minimize error that comes from representing a curved boundary using otherwise polygonal elements [24]. The modes computed using the

FEAST algorithm are shown in figure 6.2 [24]. In each subfigure, we capture the nontrivial behavior of each mode in core region, which we outline with a dashed black circle. The mode in subfigure 6.2f is what in the physics community as the LP01 or fundamental mode [52].

To test our implementation of the FEAST algorithm, we perform a convergence study with a mesh whose mesh size is $h_c = 1/16$ in the core region. The mesh is finer in this region by design, as the salient features of guided modes are contained within this region. We perform three mesh refinements, after which we curve the elements whose boundary is either $\partial\hat{\Omega}$ or the interface of the core and cladding regions. With $N = 16$ quadrature points, we compute six eigenvalues $\hat{\lambda}_l^h$ and corresponding eigenfunctions φ_l^h for $l = 1, 2, \ldots, 6$ shown in figure 6.2. The exact eigenvalues, $\hat{\lambda}_l = r_{core}^2 \beta^2$, are given to seven places after the decimal point by

$$\hat{\lambda}_1 = 2932065.0334243,$$
$$\hat{\lambda}_2 = \hat{\lambda}_3 = 2932475.1036310,$$
$$\hat{\lambda}_4 = \hat{\lambda}_5 = 2934248.1978369,$$
$$\hat{\lambda}_6 = 2935689.8561775.$$

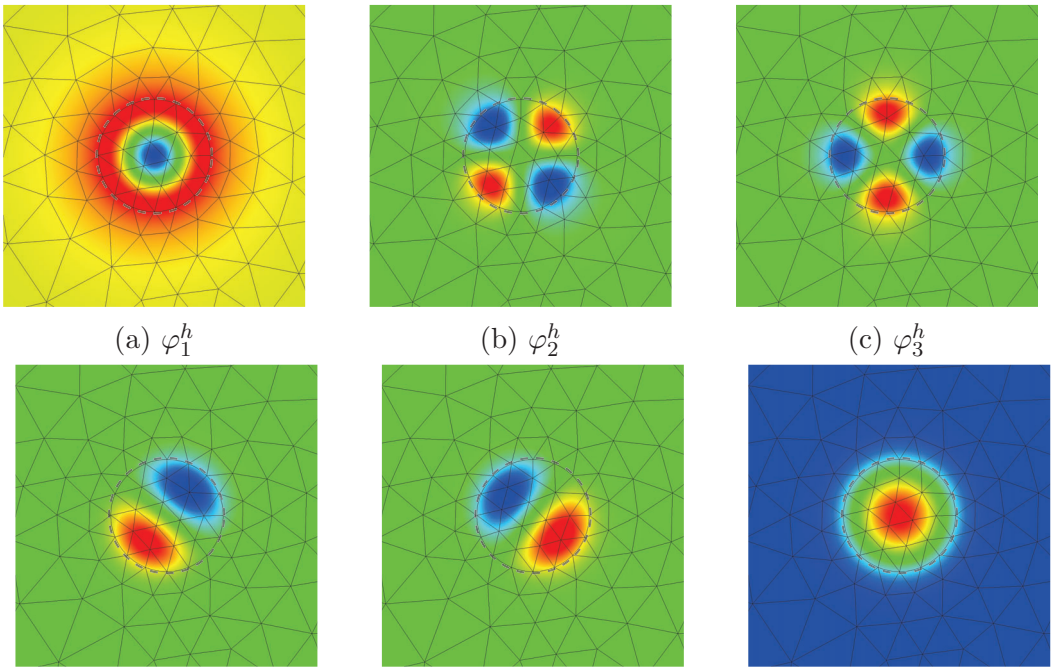
We then fix the polynomial degree at $p = 3$, and report the error $e_l = |\hat{\lambda} - \hat{\lambda}_l^h|/|\hat{\lambda}_l^h|$ for $l = 1, 2, \ldots, 6$ in figure 6.1. While the convergence rates approach order $2p = 6$, they do not match up as closely as we would expect from the examples we have seen in previous sections. Since we approached error within a few order of magnitudes of machine precision, no further refinements were performed.

Table 6.1: Step-Index Fiber Convergence Rates

core h	e_1	NOC	e_2	NOC	e_3	NOC	e_4	NOC	e_5	NOC	e_6	NOC
h_c	1.26e-07	–	2.01e-07	–	1.81e-07	–	4.99e-08	–	4.37e-08	–	1.72e-08	–
$h_c/2$	9.42e-09	3.7	1.63e-08	3.6	1.32e-08	3.8	6.46e-09	3.0	4.84e-09	3.2	3.38e-09	2.4
$h_c/4$	1.17e-10	6.3	2.13e-10	6.3	1.80e-10	6.2	7.03e-11	6.5	4.84e-11	6.6	3.64e-11	6.5
$h_c/8$	9.16e-14	10.3	1.33e-12	7.3	3.06e-13	9.2	3.75e-13	7.6	6.87e-13	6.1	6.69e-14	9.1

Table 6.1: Convergence rates for the eigenvalues of the ytterbium-doped step-index fiber.

Figure 6.2: Guided Mode Intensities



(a) φ_1^h (b) φ_2^h (c) φ_3^h

(d) φ_4^h (e) φ_5^h (f) φ_6^h

Figure 6.2: A close view of the approximate eigenfunctions φ_j^h computed by FEAST for the ytterbium-doped fiber. The boundary of the fiber core region is marked by dashed black circles.

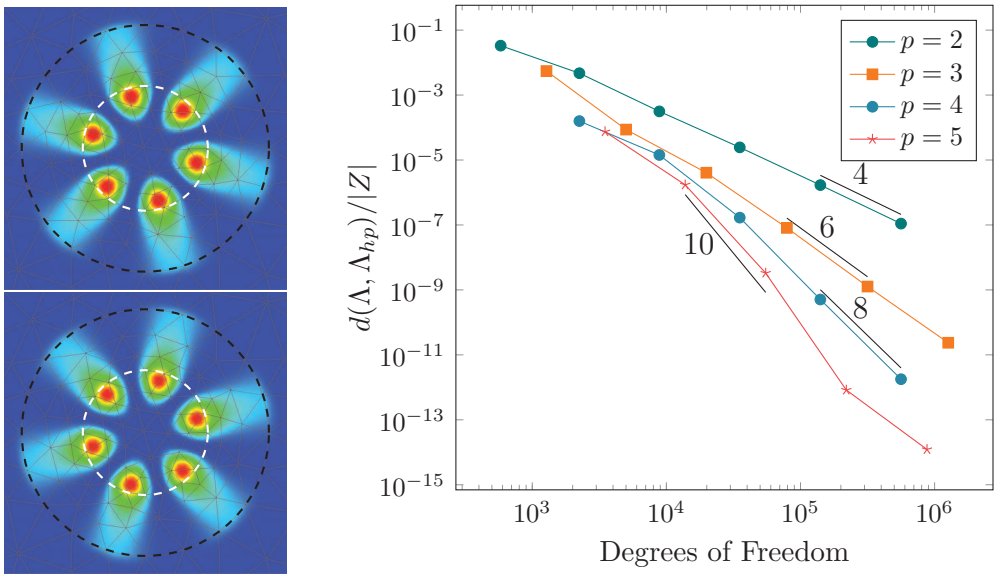
6.2.2 Leaky Mode Verification using the Polynomial Eigensolver

Revisiting the step-index fiber, we now transition to using Algorithm 4 to compute leaky modes. We compute a known set of leaky modes for an ytterbuim-doped step-index fiber with an index of refraction that is piecewise constant and defined as in (1.2) with a constant refractive index n_1 that is slightly higher than the refractive index n_0 in the cladding region. To verify our results, we compare a computed solution using Algorithm 4 to an analytic solution we derive in a similar fashion to the analytic solution (1.15) for guided modes in Chapter 1.

Now that we have an exact solution to check against the results of applying Algorithm 4, we test our method on a commercially available step-index fiber (see [24] for further details). The parameters needed for our simulation include the core radius of the fiber $R_0 = 12.5 \times 10^{-6}$ m, core index $n_1 = 1.45097$, and a cladding index $n_0 = 1.44973$. Since the operating wavelength of such a fiber is typically 1064 nanometers (nm), we take $\lambda = 1.064 \times 10^{-6}$ m, and $k = 2\pi/\lambda$ m^{-1} [24].

To solve this problem computationally, we implemented algorithm 4 using the open source finite element library NGSolve [59], with nondimensional geometric parameters $\hat{R}_0 = 1, \hat{R} = 2$, and $\hat{R}_{fin} = 4$ [27]. Parameters for our implementation of FEAST include specifying circular contour with center $y = 1.9 - 0.2i$, radius $\gamma = 10^{-1}$, an initial span of $m = 5$ random vectors, and a PML decay strength of $\alpha = 8$. The algorithm is then run until convergence is achieved. Our target eigenvalue from the analytic solution for $\ell = 3$ is given by $Z_3 \approx 1.957793 - 0.185432i$.
For our verification, we let the polynomial degree for the Lagrange finite element space 5.58 vary as $p = 2, 3, 4, 5$. In Figure 6.3a, we show the desired modes computed with algorithm 4 using $p = 10$ and no initial mesh refinement. In the acompanying figure

Figure 6.3: Leaky Mode Intensities and Convergence Results



(a) Mode intensities

(b) Convergence of corresponding eigenvalues

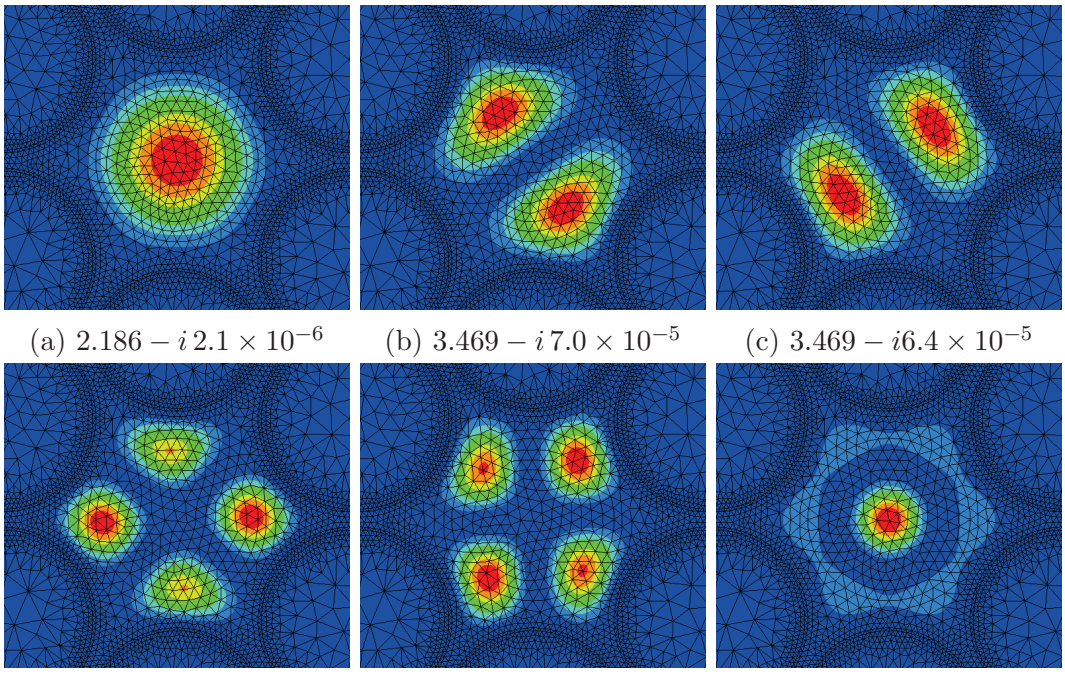
Figure 6.3: *Left* (6.3a): Intensities of computed step-index leaky modes corresponding to two eigenvalues in $\Lambda_{hp} = \{Z_{hp}^{(1)}, Z_{hp}^{(2)}\}$ are shown. The white and dark dashed curves indicate the core-cladding interface and the start of the PML, respectively. *Right* (6.3b): Log-scale plot of the distance between exact and approximate eigenvalue cluster Λ_{hp} for polynomial degrees $p = 2, \ldots, 5$ and uniform mesh refinements.

6.3b, we see that our eigenvalues are converging at the approximate desired rates $O(h^{2p})$ [27] for $p = 2, 3$, and eventually for polynomial degrees $p = 4, 5$. Note that the error does bottom out for $p = 5$: This corresponds to the computed eigenvalues becoming closer in accuracy to the target eigenvalue, which is computed to a precision of $O(10^{-14})$.

6.3 Computed Modes for the Six-Capillary Microstructure Fiber

Analogous to the guided and leaky modes of step-index fibers, we take a brief look at the computed modes for a six-capillary microstucture fiber, including the analogous fundamental mode and some higher-order modes. Pictured in Figure 6.4 (see also [27]) are the approximate modes computed using our polynomial eigensolver for a microstructure fiber with operating wavelength 1000 nanometers, and labeled with their corresponding non-dimensional eigenvalues Z computed using FEAST. These modes were computed with no initial mesh refinement and polynomial degree $p = 20$. In subsequent sections, we will look more closely at confinement loss computations using the computed propagation constant at different wavelengths for the fundamental mode pictured in Figure 6.4a.

Figure 6.4: Computed Mode Intensities of a Microstructure Fiber



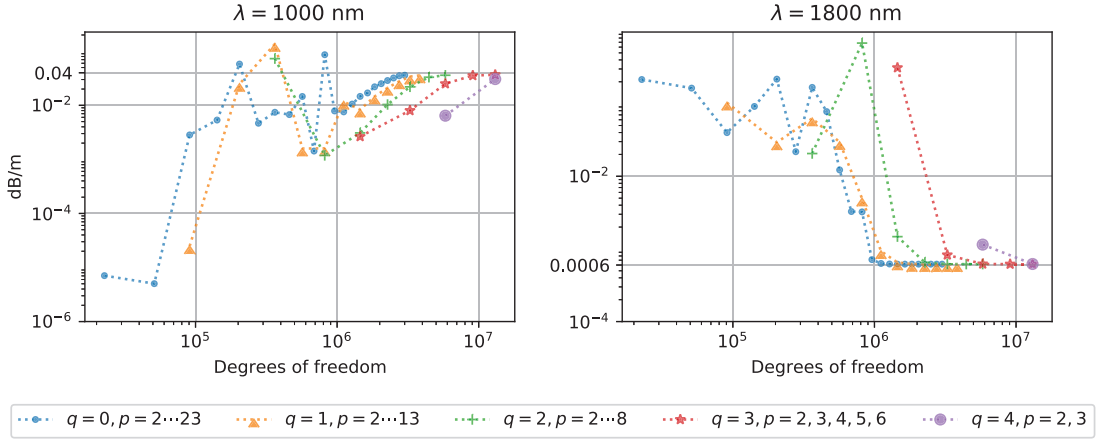
(a) $2.186 - i\,2.1 \times 10^{-6}$ (b) $3.469 - i\,7.0 \times 10^{-5}$ (c) $3.469 - i6.4 \times 10^{-5}$

(d) $4.637 - i\,2.4 \times 10^{-3}$ (e) $4.637 - i\,2.2 \times 10^{-3}$ (f) $4.961 - i\,8.6 \times 10^{-4}$

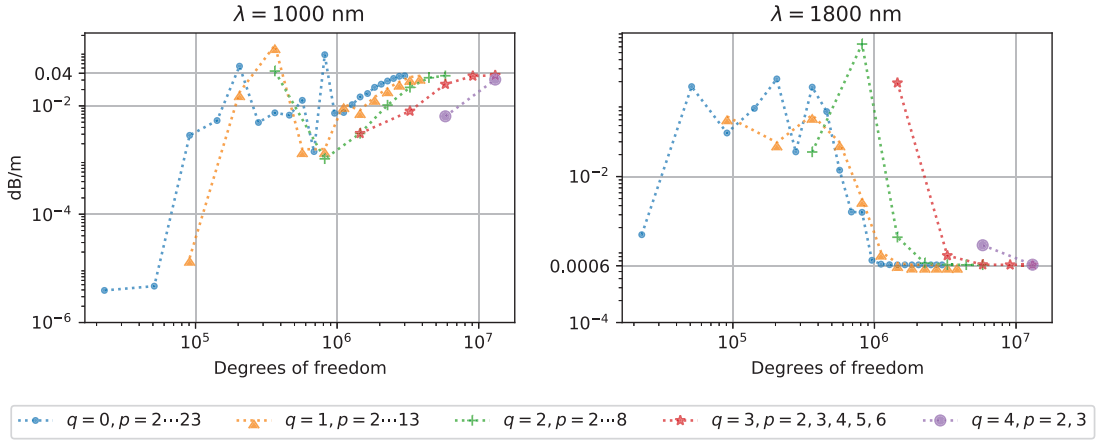
Figure 6.4: Intensities of computed modes are shown zoomed into a rectangle covering the hollow core (the region $r < R_{core}$ of Figure 1.2, where $r = ||x||$ for $x \in \mathbb{R}^2$), labeled with their approximate nondimensional Z values for $\lambda = 10^{-6}$ m.

6.4 Confinement Losses for Fixed Geometric Parameters

Figure 6.5: Confinement Loss Convergence Studies



(a) Polynomial eigensolver results.



(b) Linear eigensolver results.

Figure 6.5: Convergence study results for confinement losses at two different wavelengths using the polynomial eigensolver (Figure 6.5a) and standard linear FEAST (Figure 6.5b). Computational results show a large preasymptotic regime before computed losses apear to converge.

This section covers the computation of the cladding losses for the LP_{01}-like (fundamental) mode of a six-ring Antiresonent Nodeless Fiber (ANF), analogous to that in Poletti's work [50]. Our goal is to find a regime of mesh refinements and polynomial degrees in which the computed confinement losses of this mode stabilize. Recall

from Chapter 1 that confinement losses are computed using Equation (1.35), i.e. $\text{CL} = \frac{20}{\log(10)}\Im(\beta)$ dB/m, where $\Im(\beta)$ is the imaginary part of the propagation constant β.

For this set of experiments, we perform up to four uniform mesh refinements with increasing polynomial degree. Simulations were limited to available computational resources, as a single application of our eigensolver to compute confinement losses for the fundamental mode consumed well over 150 gigabytes of memory as we pushed our solver past the preasymptotic regime. From a recently submitted work [27], we show two simulations in Figure 6.5a at two different vacuum wavelengths of 1000 nm and 1800 nm, respectively. The experiments at the two respective wavelengths used circular contours with centers near 2.18 and 2.24, respectively. In both experiments, the outgoing medium is taken to be air with an approximate refractive index of $n_{air} = 1.00028$. We conducted a similar experiment with standard linear FEAST using the NGSolve auto PML and shown in Figure 6.5b. In each set of experiments, the polynomial degree of the underlying finite element discretization is given by $p \geq 2$, and the number of uniform mesh refinements performed on the initial coarse mesh is given by the integer $q \geq 0$.

The results of our experiments are given in Figure 6.5. These experiments stress the importance of pushing the limits of our discretization to confidently state the losses we are computing. In the asymptotic regime of both experiments for the linear and polynomial eigensolvers, confinement losses appear to stabilize near 0.04 dB/m and 0.0006 dB/m, respectively. Furthermore, the results of Figure 6.5 show the quickest path to the asymptotic region is most easily acheived with high p and no initial mesh refinements, or moderate p and a single uniform mesh refinement. While our

results in the asymptotic regime do not match the precise results in the optics literature [50], our hope is that our convergence study provides a baseline to accurately compute and compare confinement losses in the future. We do, however, see that our algorithm achieves comparable results with the standard linear FEAST algorithm, thus giving us confidence in the consistency of our results.

6.5 Variation of Geometric Parameters: Outer PML Thickness

In addition to verifying the convergence towards target confinement loss values, we are interested in ensuring that our computed confinement losses remain stable as we vary the decay strength $\alpha > 0$ in the PML region and the thickness of the PML region itself. To this end, we fix the polynomial degree for this study at $p = 10$ and perform $q = 1$ uniform mesh refinement. In addition, we fix the wavelength for the fiber at 1800 nm, and provide FEAST with a center of $y = 2.247 + 0i$ and a radius of $\gamma = 0.005$.

Table 6.2: PML Parameter Variation Results

PML width (μm)	Degrees of freedom	CL (dB/m)		
		$\alpha = 1$	$\alpha = 5$	$\alpha = 10$
50	2270641	0.000630	0.000629	0.000629
100	2603121	0.000628	0.000628	0.000629
150	2482041	0.000628	0.000628	0.000628

Table 6.2: Computed confinement losses for varying PML widths and PML parameters α (fixing $q = 1$ and $p = 10$).

As we see in Table 6.2, we vary the thickness of the PML region between 50 and 150 micrometers. The strength of the decay in the PML region, measured by $\alpha > 0$, is also varied. We see that as our parameters vary, we still maintain the same approximate mantissa. The exception to this is the case of a 50 micron PML width using a decay strength of $\alpha = 1$, where the solution may not have decayed significantly.

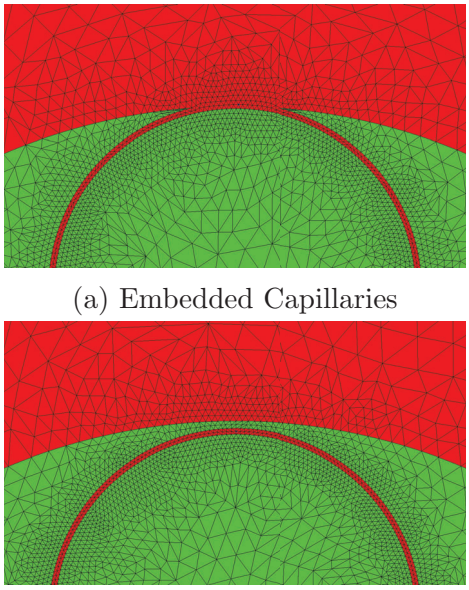
6.6 Displacement of Capillary Tubes and Confinement Losses

The other study we perform is looking at the displacement of capillary tubes and how this influences the computed losses. For the experiment, we let the position of the capillaries vary between being embedded into the outer glass jacket of the microstructure fiber, followed by allowing the capillaries to sit in the interior of the hollow core. Our FEAST algorithm for polynomial eigenvalue problems was used to compute the fundamental mode and corresponding propagation constant every 0.1 micrometers, from which we computed the corresponding confinement loss. The medium in the PML is taken to be air, the refractive index for which is approximately $n_{air} = 1.00028$.

When embedding the capillaries into the glass jacket, we let the embedding distance into the glass jacket vary between 0.01 and 0.42 micrometers. Similarly, we pulled the capillaries into the hollow core a maximum distance of 0.5 micrometers. Experiments could not be performed with the capillaries perfectly tangent to the air-cladding interface, as the NGSolve package cannot mesh boundaries that are potentially tangent at a single point. Hence, the blue and red curves are connected via a black dashed line in Figure 6.8. For all computed losses, we performed one uniform mesh refinement and varied the polynomial degrees for our finite element spaces between $p = 10, 11, 12, 13$. No significant qualitative differences were observed when plotting the losses across various values of p, so we show one just plot corresponding to the highest chosen polynomial degree $p = 13$ in Figure 6.8. We also show the corresponding real part of the effective index $n_{eff} = \beta/k$, denoted $\Re(\beta/k)$, in Figure 6.9.

Notice in Figure 6.8 that there are two peaks in the computed losses: One in the embedded mesh case, and one in the freestanding mesh case. Furthermore, there is a

Figure 6.6: Embedded and Freestanding Meshes



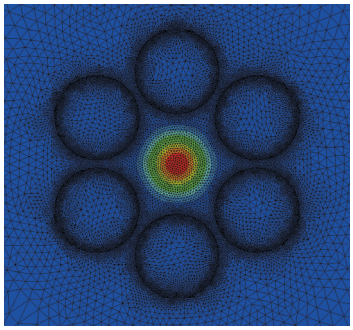
(a) Embedded Capillaries

(b) Freestanding Capillaries

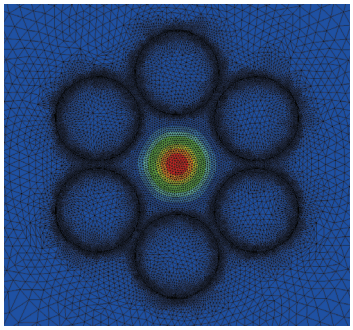
Figure 6.6: A zoomed-in view of the topmost capillary for the mesh with embedded capillaries (Figure 6.6a) and freestanding capillaries 6.6b).

significant change in the computed losses as we displace the capillaries: Losses vary from being $O(10^{-4})$ to $O(1)$, indicating a nontrivial sensitivity to the displacement of the capillaries. In Figure 6.9, however, we see that the real part of the effective index appears to remain constant in the case of the embedded mesh, but decreases as the capillaries move closer to the center of the hollow core region in the freestanding case. The corresponding imaginary part of the effective index for the embedded and freestand meshes, of course, exhibits the same behavior (up to a linear scaling) as the confinent losses in Figure 6.8, so we omit its plot.

Figure 6.7: The Fundamental Mode on Two Meshes



(a) Embedded Capillaries



(b) Freestanding Capillaries

Figure 6.7: The computed fundamental mode in the cases of embedded and freestanding capillaries. Each such mode is computed on a mesh that has been refined once uniformly, with modes approximated using polynomial degree $p = 13$. Each mode corresponds to the maximum confinement losses computed for each mesh type indicated in Figure 6.6.

Figure 6.8: Confinement Losses for Various Capillary Displacements

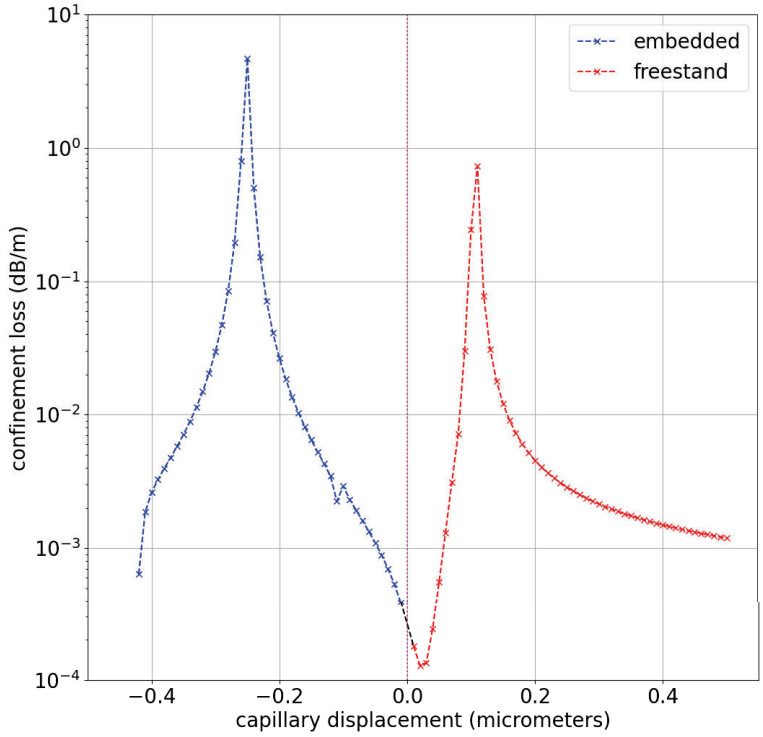


Figure 6.8: Computed losses as the capillary displacement varies. Negative displacements correspond to embedding the capillaries further into the outer cladding layer. Positive displacements correspond to moving the capillaries closer to the center of the hollow core.

Figure 6.9: Fundamental Mode Real Effective Index

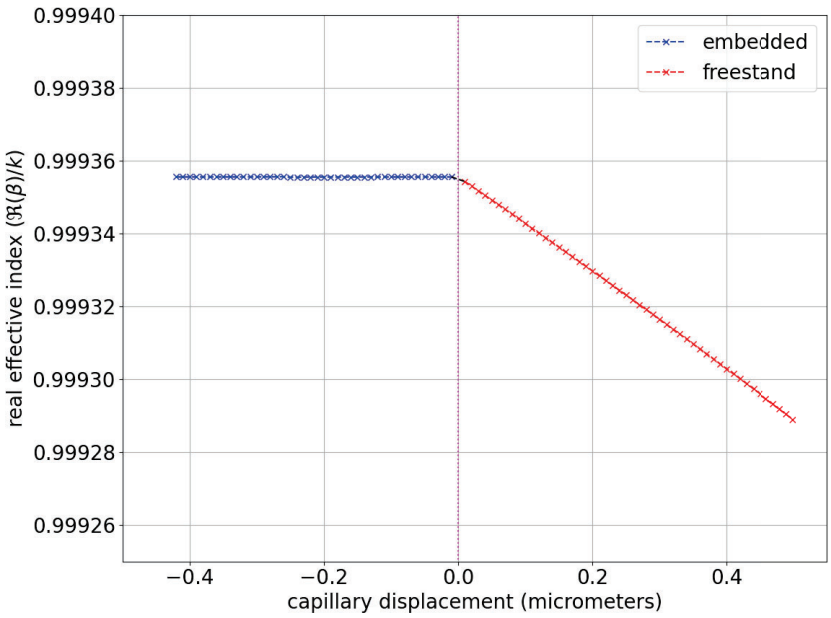


Figure 6.9: The real part of the effective index $n_{eff} = \beta/k$ over various capillary displacements. Negative displacements correspond to embedding the capillaries further into the outer cladding layer. Positive displacements correspond to moving the capillaries closer to the center of the hollow core.

Chapter 7

Conclusions and Future Work

In this work, we explored the use of the FEAST algorithm as a computational tool for solving eigenvalue problems in fiber optics. We developed two extensions of the FEAST algorithm for solving eigenvalue problems through the use of the finite element method. The first of these extensions was the application of the resolvent operator through the use of the DPG discretization, through which we connected errors in the FEAST approximations of eigenvalues and eigenspaces to the discretization parameters used in the DPG method. Our second extension was the adaptation of the FEAST algorithm to solving polynomial eigenvalue problems arising from a specific formulation of a perfectly matched layer to compute confinement losses in microstructure fibers. Our adaptation of the FEAST algorithm used the underlying structure of a linearization to efficiently compute eigenvalues and eigenspaces while handling large problem sizes.

Our results from Chapter 6 show a large preasymptotic regime when the classical finite element method is used for our underlying discretization, but reveals a path to the asymptotic regime by restricting the number of mesh refinements to a low value while keeping the polynomial degree high. Furthermore, we see that there is a nontrivial sensitivity in the computed losses as the geometry of the fiber changes,

specifically due to placement of the glass capillaries. This certainly warrants further study, as well as a connection to confinement losses that one can see reported in the optics literature for six-capillary and eight-capillary microstructure fibers [38,50]. Furthermore, our results stress the importance of converged confinement losses, which we did not see specified explicitly in our review of the optics literature. Such an endeavor is particularly important to our collaborators at the Air Force Research Laboratory (AFRL). From a conversation with one of our collaborators at the AFRL (Jacob Grosek, oral communication, 2021), his and others' experience with various mode solvers (i.e. eigensolvers) in conjunction with our own literature search have revealed inconsistencies in reported confinement losses for similar models of optical fibers. Most importantly, there has not been a clear indication of convergence studies performed across various mode solvers to unify reported confinement losses.

Future research from this work comprises many possible directions. One avenue to explore would be an implementation of our polynomial eigensolver using the DPG method instead of the classical finite element method, especially since the classical finite element method demonstrated a large preasymptotic regime in our convergence studies. A comparison to the linear solvers we have used already with the classical FEM case would be appropriate here to see if there is any benefit gained from using this discretization. Similarly, a more fine-grained approach to the experiments shown in the Appendix would be of interest in order to show which algorithm gives a clearer picture of the spectrum we wish to see. One addition to this experiment would be to explore this throught the lense of the complex-symmetric formulation given in Chapter 5.

As seen in Chapter 5, we formulated a polynomial eigenproblem using a frequency-

dependent PML, where the PML was implemented beginning with a complex coordinate transformation. This led to the problem of finding eigenvalues and eigenspaces for a polynomial eigenvalue problem. A natural extension for our polynomial solver would be to look at solvers for rational eigenproblems, as our formulation originally yielded a matrix with a negative power of the nondimensional eigenvalue Z. Works such as [61] explore solutions to such problems using rank-revealing factorizations, and such works could be used as a departure point for extending FEAST further.

Yet another direction to pursue is the development of a distributed memory FEAST algorithm, such as in [36]. This would allow for convergence studies and other experiments where FEAST pushes the free memory limits on currently available computing resources. This is especially important in light of our own experimental results, for which the accurate computation of confinement losses required upwards of millions of unknowns.

References

[1] M. Abramowitz and I.E. Stegun. *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. Number 55 in Applied Mathematics Series. U.S. Department of Commerce National Bureau of Standards, Washington, D.C., 1972.

[2] P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computation Physics*, 114(2):185–200, 1994.

[3] W. Beyn. An integral method for solving nonlinear eigenvalue problems. *Linear Algebra and its Applications*, 436(10):3839–3863, 2012.

[4] C.B. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*. Number 15 in Texts in Applied Mathematics. Springer-Verlag, New York, 3rd edition, 2008.

[5] T. Bühler and D.A. Salamon. *Functional Analysis*, volume 191 of *Graduate Studies in Mathematics*. American Mathematical Society, 2018.

[6] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. A posteriori error control for DPG methods. *SIAM J. Numer. Anal.*, 52(3):1335–1353, 2014.

[7] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. Breaking spaces and forms for the DPG method and applications including Maxwell equations. *Computers and Mathematics with Applications*, 72(3):494–522, 2016.

[8] J.S. Chiang. Analysis of leaky modes in photonic crystal fibers using the surface integral equation method. *Crystals*, 8(4):177, 2018.

[9] F. Collino and P. Monk. The perfectly matched layer in curvilinear coordinates. *SIAM J. Sci. Comput.*, 19(6):2061–2090, 1998.

[10] J.B. Conway. *A Course in Functional Analysis*. Springer-Verlag, New York, second edition, 1990.

[11] W.Y. Crutchfield, H. Cheng, and L. Greengard. Sensitivity analysis of photonic crystal fiber. *Opt. Express*, 12(18):4220–4226, 2004.

[12] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. *Comput. Methods Appl. Mech. Engrg.*, 199:1558–1572, 2010.

[13] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions. *Numer. Methods for Partial Differential Equations*, 27(1):70–105, 2011.

[14] L. Demkowicz, J. Gopalakrishnan, and A.H. Niemi. A class of discontinuous Petrov–Galerkin methods. Part III: Adaptivity. *Appl. Numer. Math.*, 62:396–427, 2012.

[15] V. Finazzi, T.M. Monro, and D.J. Richardson. Small-core silica holey fibers: nonlinearity and confinement loss trade-offs. *J. Opt. Soc. Am. B*, 20(7):1427–1436, 2003.

[16] G.N. Gatica. *A Simple Introduction to the Mixed Finite Element Method*. SpringerBriefs in Mathematics. Springer International Publishing, 1 edition, 2014.

[17] B. Gavin. *Inexact and nonlinear extension of the feast eigenvalue algorithm*. Dissertation, University of Massachusetts Amherst, Amherst, MA, 2018.

[18] B. Gavin, A. Miedlar, and E. Polizzi. FEAST eigensolver for nonlinear eigenvalue problems. *Journal of Computational Science*, 27:107–117, 2018.

[19] B. Gavin and E. Polizzi. Krylov eigenvalue strategy using the FEAST algorithm with inexact system solves. *Numer. Linear Algebra Appl.*, 25(5), 2018.

[20] I. Goldberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*. Classics in Applied Mathematics. SIAM, Philadelphia, 1982.

[21] J. Gopalakrishnan. Five lectures on DPG methods. *ArXiv e-prints*, 2014. [math.NA].

[22] J. Gopalakrishnan, L. Grubišić, and J. Ovall. Filtered Subspace Iteration for Self-adjoint Operators. *Portland Institute for Computational Science Publications*, 3, 2017.

[23] J. Gopalakrishnan, L. Grubišić, and J. Ovall. Spectral discretization errors in filtered subspace iteration. *Mathematics of Computation*, 89:203–228, 2020.

[24] J. Gopalakrishnan, L. Grubišić, J. Ovall, and B.Q. Parker. Analysis of FEAST spectral approximations using the DPG discretization. *Computational Methods in Applied Mathematics*, 19(2):251–266, 2019.

[25] J. Gopalakrishnan, S. Moskow, and F. Santosa. Asymptotic and numerical techniques for resonances of thin photonic structures. *SIAM J. Appl. Math.*, 69(1):37–63, 2008.

[26] J. Gopalakrishnan and B.Q. Parker. Pythonic FEAST. `https://bitbucket.org/jayggg/pyeigfeast`, 2021.

[27] J. Gopalakrishnan, B.Q. Parker, and P. VandenBerge. Computing leaky modes of optical fibers using a FEAST algorithm for polynomial eigenproblems. *Wave Motion*, 108:102826, 2022.

[28] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Mathematics of Computation*, 83(286):537–552, 2014.

[29] T. Goswami. Confinement loss in a hollow core fiber. Internship Report, 2018.

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

[31] S. Güttel, E. Polizzi, P.T.P. Tang, and G. Viaud. Zolotarev quadrature rules and load balancing for the FEAST eigensolver. *SIAM J. Sci. Comput.*, 37(4):A2100–A2122, 2015.

[32] S. Güttel and F. Tisseur. The nonlinear eigenvalue problem. *Acta Numerica*, 26:1–94, 2017.

[33] W.W. Hager. Updating the inverse of a matrix. *SIAM Rev.*, 31(2):221–239, 1989.

[34] A. Horning and A. Townsend. FEAST for differential eigenvalue problems. *SIAM J. Numer. Anal.*, 58(2):1239–1262, 2020.

[35] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, 2nd edition, 1976.

[36] J. Kestyn, V. Kalantzis, E. Polizzi, and Y. Saad. PFEAST: A high performance sparse eigenvalue solver using distributed-memory linear solvers. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 178–189, Salt Lake City, UT, 2016. IEEE.

[37] J. Kestyn, E. Polizzi, and P.T.P. Tang. FEAST eigensolver for non-hermitian problems. *SIAM J. Sci. Comput.*, 38:S772–S799, 2016.

[38] A.N. Kolyadin, A.F. Kosolapov, A.D. Pryamikov, A.S. Biriukov, V.G. Plotnichenko, and E.M. Dianov. Light transmission in negative curvature hollow core fiber in extremely high material loss region. *Opt. Express*, 21(8):9514–9519, 2013.

[39] B.T. Kuhlmey, H.C. Nguyen, M.J. Steel, and B.J. Eggleton. Confinement loss in adiabatic photonic crystal fiber tapers. *J. Opt. Soc. Am. B*, 23(9):1965–1974, 2006.

[40] P. Lancaster and P. Rózsa. The spectrum and stability of a vibrating rail supported by sleepers. *Computers Math. Applic.*, 31(4/5):201–213, 1996.

[41] S. Larsson and T. Vidar. *Partial Differential Equations with Numerical Methods*. Number 45 in Texts in Applied Mathematics. Springer-Verlag, Berlin, 2003.

[42] D. Marcuse. *Light Transmission Optics*. Bell Laboratories Series. Van Nostrand Reinhold, New York, 1972.

[43] D. Marcuse. *Theory of Dielectric Optical Waveguides*. Academic Press, Inc., Boston, second edition, 1991.

[44] T. McComb. *Power scaling of large mode area thulium fiber lasers in various spectral and temporal regimes*. Electronic Theses and Dissertations, University of Central Florida, 2004.

[45] R.C. McOwen. *Partial Differential Equations: Methods and Applications*. Prentice Hall, 2nd edition, 2003.

[46] L. Nannen and M. Wess. Computing scattering resonances using perfectly matched layers with frequency dependent scaling functions. *BIT Numer. Math.*, 58(2):373–395, 2018.

[47] A. Neumaier. Residual inverse iteration for the nonlinear eigenvalue problem. *SIAM J. Numer. Anal.*, 11(5):914–923, 1985.

[48] R. Parini. cxroots: A Python module to find all the roots of a complex analytic function within a given contour. `https://github.com/rparini/cxroots`, 2018.

[49] H.M. Pask, R.J. Carman, D.C. Hanna, A.C. Tropper, C.J. Mackechnie, P.R. Barber, and J.M. Dawes. Ytterbium-doped silica fiber lasers: versatile sources for the 1-1.2 mu/m region. *IEEE J. Sel. Top. Quantum Electron.*, 1(1):2–13, 1995.

[50] F. Poletti. Nested antiresonant nodeless hollow core fiber. *Opt. Express*, 22:23807–23838, 2014.

[51] E. Polizzi. Density-matrix-based algorithm for solving eigenvalue problems. *Phys. Rev. B.*, 79(11), 2009.

[52] G.A. Reider. *Photonics: An Introduction*. Springer International Publishing, Switzerland, 2016.

[53] A. Ruhe. Algorithms for the nonlinear eigenvalue problem. *SIAM Journal on Numerical Analysis*, 10(4):674–689, 1973.

[54] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Society for Industrial and Applied Mathematics, 2011.

[55] E.B. Saff and A.D. Snider. *Fundamentals of Complex Analysis with Applications to Engineering and Science*. Pearson Education, Inc., New Jersey, 3rd edition, 2003.

[56] K. Saitoh and M. Koshiba. Confinement losses in air-guiding photonic bandgap fibers. *IEEE Photonics Technology Letters*, 15(2):236–238, 2003.

[57] R.T. Schermer and J.H. Cole. Improved bend loss formula verified for optical fiber by simulation and experiment. *IEEE Journal of Quantum Electronics*, 43(10):899–909, 2007.

[58] K. Schmüdgen. *Unbounded Self-adjoint Operators on Hilbert Space*, volume 265 of *Graduate Texts in Mathematics*. Springer Netherlands, 1 edition, 2012.

[59] J. Schöberl. NGSolve. `https://ngsolve.org/`, 2019.

[60] F.E. Seraji and F. Asghari. Determination of refractive index and confinement losses in photonic crystal fibers using FDFD method: A comparative analysis. *International Journal of Optics and Photonics*, 3(1), 2009.

[61] Y. Su and Z. Bai. Solving rational eigenvalue problems via linearization. *SIAM J. Matrix Anal. Appl.*, 32(1):201–216, 2012.

[62] P.T.P. Tang and E. Polizzi. FEAST as a subspace iteration eigensolver accelerated by approximate spectral projection. *SIAM J. Matrix Anal. Appl.*, 35(2):354–390, 2014.

[63] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Review*, 43(2):235–286, 2001.

[64] L.N. Treffethen and D. Bau, III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 1997.

[65] D.S. Watkins. *Fundamentals of Matrix Computations*. John Wiley & Sons, Inc., Hoboken, NJ, 3rd edition, 2010.

[66] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V.M. Calo. A class of discontinuous Petrov–Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D. *J. Comput. Phys.*, 230:2406–2432, 2011.

Appendix: A Comparison of Approximate Spectra

In this section, we include a comparison of the spectra we compute when using our polynomial eigensolver versus the standard FEAST algorithm applied to the linear eigenproblem of computing leaky modes for the microstructure fiber in previous sections. In each case, we wish to illuminate the differences in computed non-dimeniosnal eigenvalues Z between each approach to gain a qualitivate understanding of their differences. In Figures 1 and 2, $\Re(Z^2)$ and $\Im(Z^2)$ corresponding to the real and imaginary parts of the computed Z^2-values, respectively, when using standard linear FEAST. Likewise, $\Re(Z)$ and $\Im(Z)$ corresponding to the real and imaginary parts of the computed Z-values, respectively.

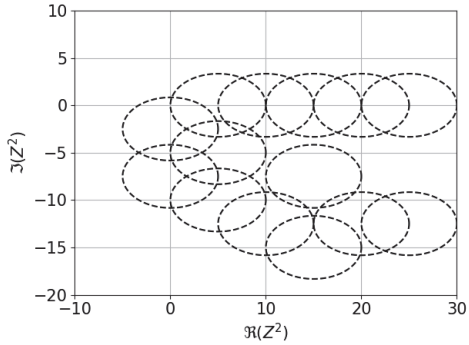
In Figures 1 and 2, we computed an approximate picture of the spectrum using the Lagrange finite element space with polynomial degree $p = 5$ and one uniform mesh refinement. In both cases, we fixed the PML strength at $\alpha = 5$ and started with spans of fifty random vectors for each circular contour used to compute the approximate eigenvalues. We chose a coarse relative stopping tolerance of $\epsilon = 10^{-7}$ for the computed eigenvalues Z, and a maximum of thirty FEAST iterations. For the application of standard linear FEAST, whose results are shown in Figure 1, we used the NGSolve auto PML mentioned in Chapter 2.

Concerning fiber properties, we take the operating operating wavelength of the microstructure fiber to be $\lambda = 1.8 \times 10^{-6}$ m, and the medium in the PML to be that of air. This means that the index of refraction in the PML region is approximately $n_0 \approx 1.00028$, and a corresponding index of refraction of Silica glass given by $n_1 \approx 1.43882$. The geometric properties of the fiber, including the core radius, inner and outer radii of the capillaries, capillary thickness, and thickness of Silica layer outside of the core region are detailed in [27, §4].

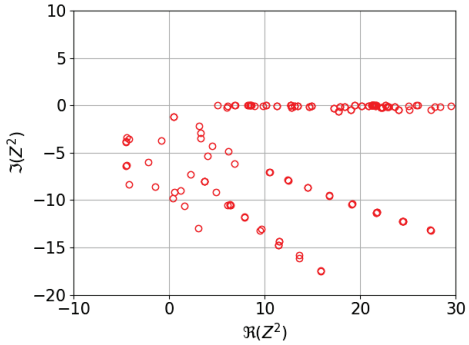
For both standard linear FEAST and its extension to polynomial eigenvalue problems, we used elliptical contours shown in Figures 1a and 2a, with the corresponding computed Z-values in Figures 1c and 2. For the quadrature approximation of the spectral projector, we used the N-point shifted elliptical trapezoid quadrature rule [22]:

$$z_j = y + \frac{\eta}{2}\left(\rho e^{i(\theta_j+\phi)} + \rho^{-1}e^{-i(\theta_j+\phi)}\right) \tag{1a}$$

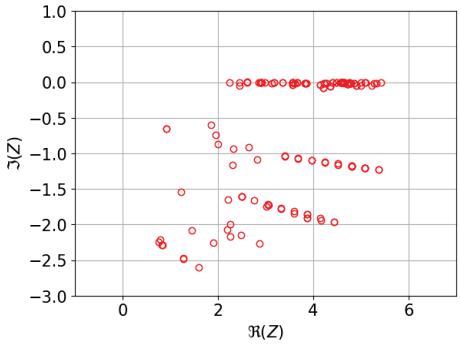
Figure 1: Linear Eigensolver Spectrum Results



(a) Contours used by FEAST.



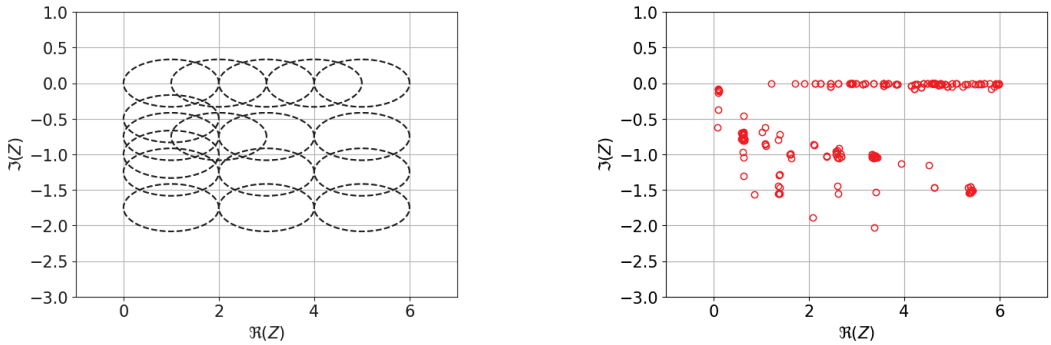
(b) Computed eigenvalues in the Z^2-plane.



(c) Computed eigenvalues in the Z-plane.

Figure 1: Approximate spectra for the microstructure fiber problem using standard linear FEAST. Figure 1a shows the contours in the Z^2-plane used by FEAST to compute the eigenvalue approximations. Figure 1b shows the results computed by FEAST in the Z^2-plane. To compare with the results of the polynomial eigensolver, Figure 1c shows the results of standard linear FEAST in the Z-plane.

Figure 2: Polynomial Eigensolver Spectrum Results



(a) Contours used by FEAST.

(b) Computed eigenvalues in the Z-plane.

Figure 2: Approximate spectra for the microstructure fiber problem using the polynomial eigensolver. The figure on the left includes the contours used to compute the approximate Ritz values. The figure on the right shows the same window (sans contours) over which the Ritz values were computed.

$$w_j = \frac{\eta}{2N}\left(\rho e^{i(\theta_j+\phi)} - \rho^{-1}e^{-i(\theta_j+\phi)}\right). \tag{1b}$$

For the quadrature rule (1), z_j are the quadrature nodes, w_j are the quadrature weights, $\theta_j = 2\pi j/N$ for $j = 0, 1, \ldots, N-1$, and $\phi = \pm\pi/N$ is a shift that moves quadrature points from the real line to prevent a quadrature node from potentially coinciding with any eigenvalues Z we wish to compute. In addition, $\rho > 1$ is fixed, and we set $\eta = 2\gamma/(\rho + \rho^{-1})$. Note that by our choice of η, the semimajor axis of the ellipse is given by $\eta(\rho + \rho^{-1})/2 = \gamma$ and the semiminor axis is given by $\eta(\rho - \rho^{-1})/2 = \gamma(\rho - \rho^{-1})/(\rho + \rho^{-1})$. Thus, we are in a position to specify the length of the semiminor axis as a fraction $0 < \hat{q} < 1$ of the length of the semimajor axis. We can do so by solving $\gamma(\rho - \rho^{-1})/(\rho + \rho^{-1}) = \gamma\hat{q}$ for ρ, given a fixed choice of $\gamma > 0$. Solving for ρ in terms of $\hat{q}$, we have that $\rho = \sqrt{(1+\hat{q})/(1-\hat{q})}$.

When computing the spectrum using the standard FEAST algorithm for linear eigenproblems, we choose for our contours a semimajor axis of $\gamma = 5$ and a semiminor axis of length $10/3$ corresponding to a choice of $\hat{q} = 2/3$. The corresponding contours are show in Figure 1, along with the corresponding non-dimensional eigenvalues Z^2 computed using standard linear FEAST. The centers used for the corresponding contours are given in Table 1a. We then obtain a picture of the approximate spectrum in the complex Z-plane by computing the square roots of the results, thus yielding the desired figure in Figure 1c. For the extension of FEAST to polynomial eigenvalue problems, we used elliptical contours specified with semimajor axis $\gamma = 1$ and semiminor axis of length $1/6$. The centers used in this case are given in Table 1b

In Figures 1c and 2b, the locations for which we typically compute the fundamen-

tal mode and higher order modes (see, for example, Figure 6.7a, or Figure 6.4 in the $\lambda = 10^{-6}$ m case) are located close to the real-axis in the neighborhood of $\{Z \in \mathbb{C} : \Re(Z) \in [2, 6]\}$. In both cases, we see that there are numerous other Z-value close by, indicating that one may need to use tight circular or elliptical contours to compute a particular mode of interest. In addition, computes Z-values further down in the fourth quadrant of Figure 1c show two rays that appear to emanate from the origin when using standard linear FEAST, versus a less structured set of Z-values in Figure 2b. At this stage, these results give a prelimary look into the difference in computed eigenvalues when using standard linear FEAST versus the extension of FEAST to polynomial eigenproblems in Chapter 5. The future goal with such experiments is to see what version of FEAST performs better for such problems, specifically by determining which algorithm gives a clearer picture of the portion of the spectrum we wish to compute.

Table 1: Contour Centers for Linear and Polynomial Eigensolvers

NGSolve Auto PML Centers
$0.00 - 2.50i$
$0.00 - 7.50i$
$5.00 - 0.00i$
$5.00 - 5.00i$
$5.00 - 10.00i$
$10.00 - 0.00i$
$10.00 - 12.50i$
$15.00 - 0.00i$
$15.00 - 7.50i$
$15.00 - 15.00i$
$20.00 - 0.00i$
$20.00 - 12.50i$
$25.00 - 0.00i$
$25.00 - 12.50i$

(a) Centers used with standard linear FEAST.

Polynomial FEAST Centers
$1.00 - 0.00i$
$1.00 - 0.50i$
$1.00 - 0.75i$
$1.00 - 1.00i$
$1.00 - 1.25i$
$1.00 - 1.75i$
$2.00 - 0.00i$
$2.00 - 0.75i$
$3.00 - 0.00i$
$3.00 - 0.75i$
$3.00 - 1.25i$
$3.00 - 1.75i$
$4.00 - 0.00i$
$5.00 - 0.00i$
$5.00 - 0.75i$
$5.00 - 1.25i$
$5.00 - 1.75i$

(b) Centers used with polynomial FEAST.

Table 1: Tables for the centers of the contours used for computing the approximate spectrum in each approach.