A Multivariate Spline Approach to the Maxwell Equations

by

Clayton Mersmann

(Under the direction of Dr. Ming-Jun Lai)

Abstract

We investigate the application of multivariate splines (2D and 3D) to the Maxwell equations. Basic properties of spline functions and various traditional finite element formulations of the Maxwell equations for numerical analysis are reviewed. We find that a Helmholtz-type formulation is well suited for traditional node-based spline analysis. Consequently, we study multivariate spline solutions to the Helmholtz equation with high wave number, a context which poses numerical challenges which are well met by a new implementation of multivariate splines of arbitrary degree.

We extend this study to solve Maxwell boundary value problems in real-world contexts in both potential and Helmholtz-type formulations. We modify the traditional spline smoothness conditions to deal with domain inhomogeneities in a novel way. Our spline implementation with arbitrary degree and modified smoothness conditions has the potential to address a variety of difficulties left unsolved by traditional nodal-based finite element methods.

Index words: Numerical, splines, partial differential equations, Helmholtz, Maxwell

A Multivariate Spline Approach to the Maxwell Equations

by

Clayton Mersmann

B.A., University of Georgia, 2013

M.A.M.S., University of Georgia, 2016

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2019

A Multivariate Spline Approach to the Maxwell Equations

by

Clayton Mersmann

Approved:

Major Professor:	Ming-Jun Lai
Committee:	Juan Gutierrez Alexander Petukhov Jingzhi Tie Robert Varley

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

Acknowledgments

I am grateful to Dr. Ming-Jun Lai for his guidance and throughout this process; to my family for their unconditional support and encouragement throughout my life; to my wife Laura for her patience and love; to my friends in the UGA math department who have helped me many times and in many ways; and to all my teachers who have taught me well over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation of Study

The importance of Maxwell's equations is hard to overstate. Groundbreaking physicist Richard Feynman has this to say:

"From a long view of the history of mankind–seen from, say, ten thousand years from now–there can be little doubt that the most significant event of the 19th century will be judged as Maxwell's discovery of the laws of electrodynamics."

The equations have proved invaluable since their discovery, and have helped engineers make great improvements, for example, in circuit design and efficiency, the invention and performance of electric generators, and in understanding and use of electromagnetic waves for communication. Their contribution is not finished, either; even today, engineers rely on numerical models of full field solutions of Maxwell's equations to aid in the design of a new generation of electromagnetic devices.[39]

One exciting example is the goal of designing devices that will enable wireless energy transfer. The applications of such a device would be nearly endless. Electric vehicles could be charged while they sit in a parking spot, without the hassle of plugging them in; or if such a device could be implanted into a road, cars could charge while they wait at a busy intersection. Medical patients with electronic implants could recharge them wirelessly while they sit comfortably, avoiding the need to design such devices around the many of the current constraints of modern batteries.

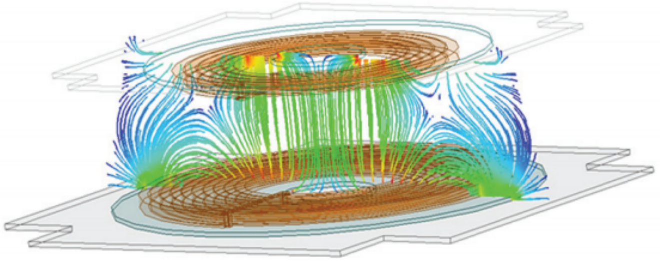


Figure 1.1: Flux lines resulting from inductively coupled wire coils. Calculated by ANSYS Maxwell[1].

Some short-distance wireless energy transfer already exists today[38], enabled by various strategies arising from the Maxwell equations. technologies. One type is inductive coupling[26]. Here, power is transmitted though the coupling of two wire coils (transmitter and receiver) via an induced magnetic field. Roughly, an oscillating current is fed into the “transmitter" coil; this changing current distribution induces a magnetic field which in turn, creates an electrical force on the “receiver" coil. If the receiver coil is part of an electrical circuit, the electric force on the receiver coil can cause current to flow, allowing a device to be powered or a battery to be charged.

The performance of such a device depends almost entirely on the mutual inductance between the two coils, which can vary substantially depending on circuit design, magnetic core materials used, and the geometry of the coils themselves. As such, accurate computer models are absolutely necessary for faster, cost-effective advancement in this field. Indeed, proprietary software packages like ANSYS Maxwell that offer full field finite element solutions of Maxwell's equations (integrated with circuitry models) are already in use in the industry[1].

Wireless energy transfer is just example that demonstrates that computational electromagnetics (CEM) remains an active area of research. The work of this dissertation does not yet address such exciting applications, but is rather a new and fundamental approach to the the challenges of CEM. This work might have better been done 20 years ago, before or alongside the establishment of edge elements in the community. But,we hope it may still make a contribution today as a simple and effective approach to numerical solutions of the Maxwell equations.

1.2 Opportunities for a Multivariate Spline Approach

Solving Maxwell's equations with bivariate and trivariate splines offers potential advantages over the existing finite element framework:

(i) **Inherent and favorable numerical properties.** Most of these properties are laid out in detail in [2]. The spline subspaces we implement numerically have stable local bases, which afford them full approximation power[23]. Exact formulas for inner products and triple products of spline functions are known; there is no need for quadrature in our numerical scheme. The degree of the basis functions used by our code is easily adjustable for problems where more or fewer degrees of freedom are required. In short, multivariate splines may be successfully implemented in application where *hp*-FEM is. The Maxwell equations, then, are due for a look through the spline lens.

(ii) **Continuous or smooth approximations of field quantities arising from the potential formulation of Maxwell's equations.** As detailed below, the introduction of scalar and vector potentials for the field quantities in Maxwell's equations can lead to a simpler, decoupled system of PDEs. In the electrostatic and magnetostatic cases, they reduce to the well-understood Poisson equation with Dirichlet or mixed boundary conditions. Because our spline functions allow us to specify global smoothness of arbitrary order, we expect greater retention of accuracy after differentiation of the potential functions to obtain the electric and magnetic fields. Most common finite element schemes use linear or quadratic elements that are only C^0, so their approximations of the field quantities in question may not even be continuous.

Certain specially constructed C^1 finite elements[15][4] may offer the same advantage as a particular spline space like $S_5^1(\triangle)$ or $S_9^1(\text{◭})$, but the spline implementation is flexible. We can require C^1 smoothness (and higher) simply by changing one parameter in our code. There is obvious utility in being able to apply the same numerical formulation to solve boundary value problems in different contexts, and some contexts may particularly benefit from approximation by C^r elements for $r \geq 2$. After all, the field quantities $\mathbb{E}$ and $\mathbb{H}$ are infinitely differentiable in homogeneous regions.

(iii) **Simple and explicit enforcement of continuity conditions for problems with inhomogeneous domains.** Electromagnetic fields satisfy certain interface conditions along the junctions of materials with different electrical and magnetic properties. These conditions must be accounted for, then, in a numerical analysis of the Maxwell equations in an inhomogeneous region. Traditional nodal finite elements implement different strategies to take these constraints into account. For specificity, let us suppose that there is an an inhomogeneity in the electric permittivity of materials filling the computational domain of a boundary value problem. Then, the laws of physics require that components of the electric field normal to that material interface suffer a discontinuity related to the ratio of permittivities. If an unknown of the boundary value problem is normal to the interface, nodal elements can allow the necessary discontinuity to occur by introducing multiple nodes on either side of the interface[30].

Spline functions have the flexibility to take this approach, too (e.g., using in the formulation in [2], we could simply not impose a continuity condition along the interface) but have the flexibility to impose the required discontinuity explicitly by modifying the standard C^0 condition accordingly. Similarly, if the problem is in potential formulation, then in the normal derivative of the scalar potential across the material interface experiences a jump. Here again, using multivariate splines, we can explicitly enforce the appropriate discontinuity in the normal derivative of the unknown using modified C^1 smoothness conditions. (The same holds true in problems where the BVP unknown is transverse to the material interface, like in waveguide analysis.) Traditional nodal finite elements do not have this control; these conditions are satisfied only naturally in via the standard variational formulations[20]. This is problematic, and can lead to spurious, non-physical finite element solutions[20][17]. To our knowledge, the explicit enforcement of these interface conditions via modified smoothness constraints is original to this work.

(iv) **A better response to the problem of spurious solutions?** The problem of non-physical solutions in finite element solutions to Maxwell problems has been a source of difficulty and disagreement[32] for more than 40 years. In his first paper on edge elements, Nédélec wrote[16] of his hope that these new elements would have great utility in “approximating Maxwell’s equations while exactly verifying one of the physical laws." Twenty years later, there was still fundamental disagreement about the root cause of the spurious solutions[19][32][20], and whether the vector edge elements effectively addressed this cause or not. In [31], Mur demonstrates that edge elements do allow spurious solutions, and, moreover, explained other problems with edge elements, noting that they are less efficient than nodal elements and inflexible in their deployment. Other, more recent works concur with these assessments[4][21], and yet edge elements have become entrenched in the computational finite element community, appearing ubiquitously in standard texts[22].

Do multivariate splines offer a straightforward way to eliminate the appearance of non-physical solutions in nodal-based numerical analysis of Maxwell problems? Of course, there is agreement that a “correct" numerical formulation must be used[32][19], but it seems that the modified spline smoothness conditions original to this work may be able to rectify more traditional formulations without the need for a new element framework. Indeed, this proposed solution would satisfy the conditions required by Mur and Lager in [32]: i) the discretized field should be expanded by functions that can ensure the continuity of the field inside interface-free subdomains, and ii)“the expansions functions should *explicitly* satisfy the interface conditions" and boundary conditions. Similarly, modified smoothness conditions would address what Jin claims is the root cause of spurious modes in an inhomogeneous waveguide problem in[20]. So, while we certainly have more numerical work to do to verify that the spline method eliminates spurious solutions in various contexts and formulations, there is reason to feel optimistic about the chances for success.

Of course, there have been many other approaches ([4][21], etc.) to address the problem of spurious solutions, but none seem to have caught on as widely as the edge elements. We do not try to give an exhaustive accounting of these approaches, but instead concern ourselves with developing the theoretical underpinnings and numerical tools for multivariate spline functions.

In the view of the author, the main contributions of this work are twofold.

(a) **Improving and expanding the Matlab implementation of multivariate spline code and extending the scope of application** In 2007, Ming-Jun Lai, G. Awanou, and P. Wenston copyrighted a Matlab package for splines of arbitrary degree and smoothness over arbitrary triangulations for applications to data fitting and numerical solutions of PDEs[2]. Since that time, many of Dr. Lai's students have used this package in their research, making modifications and improvements as needed[7][6][14][28][12]. In particular, G. Slavov wrote code to generate a C^0 bivariate spline basis over an arbitrary triangulation for use with his time-stepping application in [37]. His ideas helped me to refine my own vectorized implementation of code that generates a C^0 basis for bi- or trivariate splines, and applicable to boundary value problems with Dirichlet, Neumann, Robin, or mixed boundary conditions.

Additionally, I implemented a new vectorized conceptualization of spline code for data fitting and solutions to PDEs. This includes vectorized generation of mass, stiffness, and even smoothness matrices in the 2D and 3D setting. The vectorized implementation scales well with refinement, up to the limits of the computer's RAM, and is generalized in that a spline of arbitrary degree and smoothness may be produced simply by changing the appropriate parameters.

The result is a far more efficient Matlab implementation, whose runtimes compare favorably with some vectorized finite element implementations found in the literature (e.g. [34]). The improved efficiency is often drastic; for example, using the implementation found in [2], computing a spline solution $s \in S_9^1(\triangle)$ to a Poisson equation over a given tetrahedral partition took a laptop (4gb RAM, 2.53GHz) more than 10 minutes to compute; the same computer and the new vectorized implementation produces the “same" solution in under 3 seconds. The new implementation extends the scope of multivariate spline functions to new, more numerically challenging settings, like solving the (indefinite) Helmholtz equation with high wavenumber4, and enables splines to be applied competitively in other settings where other finite elements are already established.

(b) **Bringing multivariate splines to the Maxwell Equations and the Maxwell equations to multivariate splines** It is the opinion of the author that the application of spline functions to the Maxwell equations is a step forward both for splines and for the study of problems arising from the equations. The potential for multivariate spline functions to address some of the challenges that arise when solving the Maxwell equations numerically has been mentioned above, and will be discussed in more detail in 3. On the other hand, the potential utility that the spline modified conditions offer a mathematical scientist gives a convincing reason for the world to care about multivariate splines over any other finite element.

Chapter 2

Mulitvariate Splines and Their Properties

2.1 Bivariate Splines

2.1.1 Barycentric Coordinates in $\mathbb{R}^2$ and the Bernstein Basis

Consider a triangle $T = [v_1, v_2, v_3]$, $v_i \in \mathbb{R}^2$. We define the barycentric coordinates (b_1, b_2, b_3) of a point $(x_o, y_o) \in \mathbb{R}^2$. These coordinates are the solution to the following system of equations

$$\begin{aligned} &b_1 + b_2 + b_3 = 1 \\ &b_1 v_{1,x} + b_2 v_{2,x} + b_3 v_{3,x} = x_o \\ &b_1 v_{1,y} + b_2 v_{2,y} + b_3 v_{3,y} = y_o, \end{aligned}$$

and are nonnegative if (x_o, y_o) lies in the interior of T. The barycentric coordinates are then used to define the Bernstein basis polynomials of degree d. These polynomials arise from the terms in the expansion

$$1 = (b_1 + b_2 + b_3)^d \tag{2.1.1}$$

and take the form

$$B^d_{i,j,k}(x,y) := \frac{d!}{i!j!k!} b^i_1(x,y) b^j_2(x,y) b^k_3(x,y), \qquad i+j+k=d. \tag{2.1.2}$$

In light of 2.1.1, it is clear that the B^d_{ijk} form a partition of unity. Associated with each basis function is a special point ξ_{ijk} in triangle T where B^d_{ijk} finds its maximum. These points are called *domain points*

$$\mathcal{D}_{d,T} := \{\xi_{ijk} := \frac{iv_1 + jv_2 + kv_3}{d}\}_{i+j+k=d}. \tag{2.1.3}$$

Each B^T_{ijk} is a polynomial of degree d, and collectively, they form a basis for the space $\mathcal{P}_d$ of polynomials of degree d over T. Therefore we can represent all $P \in \mathcal{P}_d$ in B-form:

$$P = \sum_{i+j+k=d} p_{ijk} B_{ijk}, \tag{2.1.4}$$

where the B-coefficients p_{ijk} are uniquely determined by P. The basis formed by B^d_{ijk} is stable in that $||P_T||_\infty$ is “comparable" [23] to the infinity norm of the coefficient vector $\{p_{ijk}\}$ of P_T:

Theorem 2.1.1. *Let P_T be a polynomial written in B-form 2.1.4 with coefficient vector p. Then*

$$\frac{||p||_\infty}{K} \leq ||p||_T \leq ||p||_\infty, \tag{2.1.5}$$

where K is a constant depending only on d.

The constant K is easily computable given d. This stability leads to desirable numerical properties, and important approximation results like 2.1.22.

2.1.2 Bivariate Splines on Triangulations

Given a polygonal region $\Omega \subset \mathbb{R}^2$, a collection $\Delta := \{T_1, ..., T_n\}$ of triangles is an ordinary triangulation of Ω if $\Omega = \cup_{i=1}^{n} T_i$ and if any two triangles T_i, T_j intersect at most at a common vertex or a common edge.

We are now ready to define the spline space

$$S_d^0 := \{s \in C^0(\Omega) : s|_{T_i} \in \mathcal{P}_d\}, \tag{2.1.6}$$

where T_i is a triangle in a triangulation $\triangle$ of Ω, and give a parametrization for $s \in S_d^0$ using the concept of domain points.

The set of domain points over $\triangle$ is

$$\mathcal{D}_{d,\triangle} := \bigcup_{T \in \triangle} \mathcal{D}_{d,T}, \tag{2.1.7}$$

where points on vertices and edges shared by adjacent triangles are included only once. Each spline $s \in S_d^0$ is uniquely associated with its set of coefficients $\{c_\xi\}_{\xi \in \mathcal{D}_{d,\triangle}}$

$$s|_T = \sum_{\xi \in \mathcal{D}_{d,T}} c_\xi B_\xi^{T,d}, \tag{2.1.8}$$

where the superscript T indicates that B_ξ^d is generated from triangle T.

By specifying an order to the set of triangles and domain points, we can think of this coefficent set as a vector. The rule for ordering domain points is different in this dissertation than in [23], which uses lexicographical order. Our rule to get the “next" domain point after ξ_{ijk} is to decrement i while incrementing j; if this is not possible, increment k while resetting i to d-k and j to 0. For example, the domain points and coefficients for $d = 3$ are ordered thusly:

$$c_{300}, c_{210}, c_{120}, c_{030}, c_{201}, c_{111}, c_{021}, c_{102}, c_{012}, c_{003}. \tag{2.1.9}$$

We use the continuous spline space 2.1.6 to define

$$S_d^r(\triangle) := C^r(\Omega) \cap S_d^0(\triangle), \tag{2.1.10}$$

the spline space of degree d and smoothness $r \geq 0$ over triangulation $\triangle$. Spline functions in $S_d^r(\triangle)$ are expressible in B-form as in 2.1.8, but their coefficients c_ξ are subject to additional relations. We include more detailed information about spline smoothness here because of its relevance to the dissertation in dealing with inhomogeneous domains, particularly at the junction of materials with different electromagnetic properties.

The de Casteljau algorithm is helpful for computing the derivatives of polynomials in B-form, and for understanding how to enforce continuity in the derivative of a piecewise polynomial across a shared triangle edge. It begins with the recurrence relation

$$B_{ijk}^d = b_1 B_{i-1,j,k}^{d-l} + b_2 B_{i,j-1,k}^{d-l} + b_3 B_{i,j,k-1}^{d-l}, \quad \text{for all } i+j+k=d, \tag{2.1.11}$$

where all term with negative subscripts are taken to be 0. We can then define

$$c_{ijk}^{(0)} := p_{ijk}, \tag{2.1.12}$$

and, for $\ell = 1, ..., d$, we have

$$c_{ijk}^{(\ell)} = b_1 c_{i+1,j,k}^{(\ell-1)} + b_2 c_{i,j+1,k}^{(\ell-1)} + b_3 c_{i,j,k+1}^{(\ell-1)}.$$

Letting $u = (x, y)$, we can write

$$P(u) = \sum_{i+j+k=d-\ell} c_{ijk}^{(\ell)} B_{ijk}^{d-\ell}(u).$$

Suppressed in the notation here and in [23], but crucial for application, is the fact that the $c_{ijk}^{(\ell)}$ expressed in 2.1.12 are functions of the vector (b_1, b_2, b_3), whose components themselves are functions of position (x, y).

Let $u, v \in \mathbb{R}^2$ be represented in barycentric coordinates by $(\alpha_1, \alpha_2, \alpha_3)$ and $(\beta_1, \beta_2, \beta_3)$ respectively. Then the vector $a = u - v$ is given in barycentric coordinates by $a_i = \alpha_i - \beta_i$, and the derivative in that direction is given by

$$D_a B_{ijk}^d = d\big(a_1 B_{i-1,j,k}^{d-l} + a_2 B_{i,j-1,k}^{d-l} + a_3 B_{i,j,k-1}^{d-l}\big) \tag{2.1.13}$$

for any $i + j + k = d$. A straightforward proof is in [23]. It follows immediately that

$$D_a P = d \sum_{i+j+k=d-1} \big(c_{ijk}^{(1)}(a) B_{ijk}^{d-1}\big). \tag{2.1.14}$$

Theorem 2.1.2 gives linear conditions for two Bernstein polynomials to join smoothly across the edge between two adjacent triangles. It is taken almost verbatim from [23] which also contains an elegant proof and using ideas from the de Castlejau algorithm. I formulate the following corollary, which is utilized throughout the numerical results in the following.

Theorem 2.1.2. *Let $T_1 := [v_1, v_2, v_3]$ and $T_2 := [v_2, v_1, v_4]$ be triangles sharing the edge $e = [v_1, v_2]$. Let*

$$P := \sum_{i+j+k=d} c_{ijk} B_{ijk}^d \tag{2.1.15}$$

and

$$Q := \sum_{i+j+k=d} r_{ijk} R^d_{ijk} \tag{2.1.16}$$

be the degree d polynomials defined over each triangle and B_{ijk} and R_{ijk} be the Bernstein basis polynomials defined over T_1 and T_2 respectively. Suppose a is any direction not parallel to e and $n = 0, ..., r \leq d$. Then

$$D_a^{(n)} P(v) = D_a^{(n)} Q(v), \quad \forall v \in e \tag{2.1.17}$$

if and only if

$$r_{ijn} = \sum_{\nu+\mu+\kappa=n} c_{j+\nu,i+\mu,\kappa} B^n_{\nu\mu\kappa}, \quad j+k=d-n \tag{2.1.18}$$

Corollary 2.1.3. Let $\alpha := (\alpha_1, \alpha_2, \alpha_3)$ be the point v_4 expressed in in the barycentric coordinates of T_1. Then the function S formed by the joining of P and Q across e will be C^0 if and only if

$$r_{ij0} = c_{ji0} \tag{2.1.19}$$

and C^1 if and only if 2.1.19 holds and

$$r_{ij1} = \alpha_1 c_{j+1,i,0} + \alpha_2 c_{j,i+1,0} + \alpha_3 c_{j,i,1}. \tag{2.1.20}$$

The C^1 condition has a beautiful geometric interpretation. Take the points formed by considering c_{ijk} as a graph over ξ_{ijk}. Then 2.1.20 is equivalent to requiring these (3-dimensional) points to be coplanar.

As is apparent in 2.1.18, smoothness conditions across a given edge for any r are linear constraints. Thus, for a matrix A whose rows are determined by the linear constraints arising from 2.1.19 and 2.1.20, a spline s with coefficient vector c over triangulation $\triangle$ which satisfies a set of smoothness conditions $\mathcal{T}$ belongs to the set

$$S_d^{\mathcal{T}} = \{s \in S_d^0(\triangle) : Ac = 0\}. \tag{2.1.21}$$

This matrix representation of smoothness conditions is used repeatedly in our numerical experiments.

For applications to PDEs, we also need information about the approximation properties of spline functions. The following is Theorem 5.19 in [23].

Theorem 2.1.4. *Suppose that $\triangle$ is a regular triangulation of a polygonal domain Ω. For every $u \in W_2^{m+1}(\Omega)$, there exists a quasi-interpolatory spline function $Q_d(u) \in S_d^0(\triangle)$ such that*

$$\|D_x^\alpha D_y^\beta(u - Q_d(u))\|q, \Omega \leq K|\triangle|^{m+1-\alpha-\beta}|u|_{m+1,q,\Omega} \tag{2.1.22}$$

for $0 \leq \alpha + \beta \leq m \leq d$, where K is a positive constant dependent only on d, the domain Ω, and the triangulation[1].

The theorem shows that the space $S_d^0(\triangle)$ has *full approximation power in the q-norm* as there exists a constant C depending only on the triangulation $\triangle$ such that

$$\mathrm{inf}_{s \in S_d^0}\|f - s\|_q \leq C\mathrm{inf}_{p \in \mathcal{PP}_d}\|f - p\|_q, \tag{2.1.23}$$

where $\mathcal{PP}_d$ is the space of piecewise polynomials of degree d on $\triangle$.

The proof of Theorem 4.3.1 relies on the concept of a stable minimal determining set, or *MDS*. For C^0 splines, the domain points $\mathcal{D}_{d,\triangle}$ determine a stable *MDS*. It is not the case that all splines with $d > r$ have optimal approximation power (and therefore

[1] The constant depends on the triangulation in two ways: 1)the minimum angle of the triangulation and 2)the integer constant ℓ which describes how much a change in a coefficient in one triangle propagates throughout the triangulation. Details can be found in Chapter 5 of [23].

do not have stable *MDS*), but in Chapter 10 of [23], Lai and Schumaker construct a stable MDS for a superspline subspace of $S_d^r(\triangle)$ for $d \geq 3r+2$. This shows that S_d^1 has full approximation power when $d \geq 5$, a fact important to the majority of the numerical experiments that follow.

As solutions to the Helmholtz equation with impedance boundary condition are often complex, let us finally define a complex spline space by

$$\mathbb{S}_d^r(\triangle) = \{s = s_r + \mathbf{i}s_i, s_i, s_r \in S_d^r(\triangle)\}. \tag{2.1.24}$$

This definition is equivalent to letting the B-coefficients p_{ijk} as in 2.1.4 be complex.

2.2 Trivariate Splines

2.2.1 Barycentric Coordinates in $\mathbb{R}^3$ and the Bernstein Basis

For a tetrahedron $T \subset \mathbb{R}^3$, $T = [v_1, v_2, v_3, v_4]$, we define the barycentric coordinates (b_1, b_2, b_3, b_4) of a point $(x_o, y_o, z_o) \in \mathbb{R}^3$. These coordinates are the solution to the following system of equations

$$\begin{aligned}
&b_1 + b_2 + b_3 + b_4 = 1\\
&b_1 v_{1,x} + b_2 v_{2,x} + b_3 v_{3,x} + b_4 v_{4,x} = x_o\\
&b_1 v_{1,y} + b_2 v_{2,y} + b_3 v_{3,y} + b_4 v_{4,y} = y_o\\
&b_1 v_{1,z} + b_2 v_{2,z} + b_3 v_{3,z} + b_4 v_{4,z} = z_o,
\end{aligned}$$

and are nonnegative if (x_o, y_o, z_o) is in T. The barycentric coordinates are then used to define the Bernstein polynomials of degree d at $v = (x, y, z)$:

$$B^T_{i,j,k,l}(v) := \frac{d!}{i!j!k!l!} b_1^i(v) b_2^j(v) b_3^k(v) b_4^l(v), \qquad i+j+k+l = d. \tag{2.2.1}$$

which are again a partition of unity as in 2.1.2. They also form a stable basis for the space $\mathcal{P}_d$ of trivariate polynomials of degree d. Therefore we can represent all $P \in \mathcal{P}_d$ in B-form:

$$P_T = \sum_{i+j+k+l=d} p_{ijkl} B^T_{ijkl}, \tag{2.2.2}$$

where the B-coefficients p_{ijk} are uniquely determined by P. The stability of the B-form is expressible by a theorem analogous to Theorem 2.1.1.

2.2.2 Trivariate Splines on Tetrahedral Partitions

Given a polyhedral region Ω, a collection $\triangle := T_1, ..., T_n$ of tetrahedra is a tetrahedral partition of Ω if $\Omega = \cup_{i=1}^n T_i$, and if any two tetrahedra T_i, T_j intersect at a common vertex, edge, or face. (We acknowledge the overloading of some notation $\triangle, T_i$, but the meaning should be clear in context.)

As above, we define the spline space $S_d^0 := \{s \in C^0(\Omega) : s|_{T_i} \in \mathcal{P}_d\}$, where T_i is a tetrahedron in a triangulation $\triangle$ of Ω, and then

$$S_d^r := C^r(\Omega) \cap S_d^0(\triangle), \tag{2.2.3}$$

the spline space of degree d and smoothness $r \geq 0$ over tetrahedral partition $\triangle$. The domain points

$$\mathcal{D}_{d,T} := \{\xi_{ijkl} := (iv_1 + jv_2 + kv_3 + lv_4)/d\}_{i+j+k+l=d}. \tag{2.2.4}$$

play an analogous role here, and we can represent any trivariate spline in B-form as in 2.1.8.

As in Section 2.1, we do not use lexicographical order in this dissertation, but, given the m^{th} domain point ξ_{ijkl}, the multi-index for the $m + 1^{st}$ domain point is given by incrementing j: $(i-1, j+1, k, l)$; or if $i - 1 < 0$, then increment k: $(d - k -$

$l-1,0,k+1,l)$; or if $(k+1)+l > d$, then increment l: $(d-l-1,0,0,l+1)$. For example, the domain points and coefficients for d=2 are ordered

$$(c_{2000}, c_{1100}, c_{0200}, c_{1010}, c_{0110}, c_{0020}), (c_{1001}, c_{0101}, c_{0011}), (c_{0002}).$$

The grouping shows that, for a fixed l (say $l = a$), the ordering for C_{ijka} is consistent with the bivariate indexing for (ijk).

The de Casteljau algorithm again plays a role in [23] in establishing the smoothness relations necessary to ensure that a trivariate spline is C^r. With 4 barycentric coordinates, the recurrence relation takes the form

$$B^d_{ijk} = b_1 B^{d-l}_{i-1,j,k,l} + b_2 B^{d-l}_{i,j-1,k,l} + b_3 B^{d-l}_{i,j,k-1,l} + b_4 B^{d-1}_{i,j,k,l+1}, \quad \text{for all } i+j+k=d. \tag{2.2.5}$$

We define $c^{(0)}_{ijkl} := p_{ijkl}$ Then for $\ell = 1, ..., d$, we have

$$c^{(\ell)}_{ijkl} = b_1 c^{(\ell-1)}_{i+1,j,k,l} + b_2 c^{(\ell-1)}_{i,j+1,k,l} + b_3 c^{(\ell-1)}_{i,j,k+1,l} + b_4 c^{(\ell-1)}_{i,j,k,l+1},$$

so, letting $v = (x, y, z)$, we can write

$$P(v) = \sum_{i+j+k+l=d-\ell} c^{(\ell)}_{ijkl} B^{d-\ell}_{ijkl}(v).$$

As in the bivariate case, we can depress the directional derivative of a Bernstein basis function by expressing the direction vector in barycentric coordinates

$$D_a B^d_{ijkl} = d\big(a_1 B^{d-l}_{i-1,j,k,l} + a_2 B^{d-l}_{i,j-1,k,l} + a_3 B^{d-l}_{i,j,k-1,l} + a_4 B^{d-l}_{i,j,k,l-1}\big), \tag{2.2.6}$$

and thus can compactly represent $D_a P$ using de Casteljau.

In the following chapters we are interested in the smoothness (or not) of trivariate spline functions across a common face of two adjoining tetrahedra. I report the general result from [23] and then formulate a corollary which is germane to later numerical results. There is no proof of the theorem 2.2.1 in [23], although it follows from the bivariate case; here I prove the corollary directly using only the properties of the basis functions.

Theorem 2.2.1. *Let $T_1 := [v_1, v_2, v_3, v_4]$ and $T_2 := [v_1, v_2, v_3, v_5]$ be tetrahedra sharing the face $f = [v_1, v_2, v_3]$. Let*

$$P := \sum_{i+j+k+l=d} c_{ijkl} B^d_{ijkl} \tag{2.2.7}$$

and

$$Q := \sum_{i+j+k+l=d} r_{ijkl} R^d_{ijkl} \tag{2.2.8}$$

be the degree d polynomials defined over each tetrahedron and B_{ijkl} and R_{ijkl} be the trivariate Bernstein basis polynomials defined over T_1 and T_2 respectively. Suppose a is any direction not in the plane of f and $n = 0, ..., r \leq d$. Then

$$D^{(n)}_a P(v) = D^{(n)}_a Q(v), \quad \forall v \in f \tag{2.2.9}$$

if and only if

$$r_{ijkn} = \sum_{\nu+\mu+\kappa+\delta=n} c_{j+\nu,i+\mu,k+\kappa,\delta} B^n_{\nu\mu\kappa\delta}, \quad j+k+l = d-n. \tag{2.2.10}$$

Corollary 2.2.2. Let $\alpha := (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ be the point v_5 expressed in in the barycentric coordinates of T_1. Then the function S formed by the joining of P and Q across

f will be C^1 if and only if

$$r_{ijk0} = c_{ijk0} \tag{2.2.11}$$

and

$$r_{ijk1} = \alpha_1 c_{j+1,i,k,0} + \alpha_2 c_{j,i+1,k,0} + \alpha_3 c_{j,i,k+1,0} + \alpha_4 c_{j,i,k,1}. \tag{2.2.12}$$

Proof. Let v be a point on edge f. Then for any Bernstein basis function with $l > 0$ we have

$$B^d_{ijkl}(v) = R^d_{ijkl}(v) = 0, \tag{2.2.13}$$

since the fourth barycentric coordinate for each tetrahedron is identically zero on f. Then continuity $P(v) = Q(v)$ requires

$$P = \sum_{i+j+k=d} c_{ijk0} B^d_{ijk0} = \sum_{i+j+k=d} r_{ijk0} R^d_{ijk0} = Q. \tag{2.2.14}$$

But, v is necessarily expressible as some weighted average of v_1, v_2, and v_3; the barycentric coordinates for v with respect to either tetrahedron *are* those weights. Thus, on f,

$$B_{ijk0} = R_{ijk0}, \tag{2.2.15}$$

so requiring 2.2.11 is sufficient (and necessary) for continuity across the face.

Thus we see that $S|_f$ is a bivariate polynomial of degree d, and so the directional derivative of S in any direction in the span of $\{v_2 - v_1, v_3 - v_1\}$ exists. To enforce C^1 smoothness across this interface, we need only require

$$D_a P(v) = D_a Q(v) \tag{2.2.16}$$

for a direction a with some (nonzero) component normal to f. Then, appropriate linear combinations of the directional derivatives yield the partials S_x, S_y, and S_z which are continuous on a neighborhood of any point v in f.

Drawing from [23], we choose a in the direction $v_5 - v_1$. In T_1's coordinates, this is $(\alpha_1 - 1, \alpha_2, \alpha_3, \alpha_4)$; in T_2's, it's simply $(-1, 0, 0, 1)$. We apply formula 2.2.6 and set the directional derivative of Q and P equal

$$\begin{aligned}\sum_{i+j+k+l=d} r_{ijkl}\big(-1(R^{d-1}_{i-1,j,k,l}) + 0(R^{d-1}_{i,j-1,k,l}) + 0(R^{d-1}_{i,j,k-1,l}) + 1(R^{d-1}_{i,j,k,l-1})\big) = \\ \sum_{i+j+k+l=d} c_{ijkl}\big((\alpha_1 - 1)(B^{d-1}_{i-1,j,k,l}) + \alpha_2(B^{d-1}_{i,j-1,k,l}) + \alpha_3(B^{d-1}_{i,j,k-1,l}) + \alpha_4(B^{d-1}_{i,j,k,l-1})\big).\end{aligned} \tag{2.2.17}$$

We again use the fact that, on f, the only nonzero basis functions are those with for $k = 0$. Thus the sums may be grouped as

$$\begin{aligned}-\sum_{i+j+k=d} r_{ijk0}R^d_{i-1,j,k,0} + \sum_{i+j+k=d-1} r_{ijk1}R^d_{ijk0} = \\ \sum_{i+j+k=d} c_{ijk0}[(\alpha_1 - 1)B^d_{i-1,j,k,0} + \alpha_2 B^d_{i,j-1,k,0} + \alpha_3 B^d_{i,j,k-1,0}] + \alpha_4 \sum_{i+j+k=d-1} c_{ijk1}B^d_{ijk0}\end{aligned} \tag{2.2.18}$$

Making use of 2.2.11 and 2.2.15, we simplify

$$\begin{aligned}\sum_{i+j+k=d-1} r_{ijk1}B^d_{ijk0} =& \alpha_1 \sum_{i+j+k=d} c_{ijk0}B^d_{i-1,j,k,0} + \alpha_2 \sum_{i+j+k=d} c_{ijk0}B^d_{i,j-1,k0} + \alpha_3 \sum_{i+j+k=d} c_{ijk0}B^d_{i,j,k-1,0} + \\ & \alpha_4 \sum_{i+j+k=d-1} c_{ijk1}B^d_{ijk0}.\end{aligned} \tag{2.2.19}$$

Reindexing yields

$$\sum_{i+j+k=d-1} r_{ijk1} B^d_{ijk0} = \sum_{i+j+k=d-1} \big(\alpha_1 c_{i+1,j,k,0} + \alpha_2 c_{i,j+1,k,0} + \alpha_3 c_{i,j,k+1,0} + \alpha_4 c_{i,j,k,1}\big) B^d_{ijk0}, \tag{2.2.20}$$

from which 2.2.12 follows. □

As in section2.1, there is geometric interpretation of 2.2.12, too–it is the requirement that the (4-dimensional) points formed by the domain points and the corresponding coefficient value lie in the same hyperplane.

Lastly, we report approximation results for trivariate splines. Like the bivariate case, proving that a spline space has full approximation power relies on the ability to define a stable *MDS*. The set of domain points is such a determining set for $S^0_d(\triangle)$, which therefore has the best approximation order, but [23] does not contain a general result for $d \geq f(r)$. Still, for the trivariate spline subspace

$$\mathcal{S}_1(\triangle) := \{s \in S^1_9(\triangle) : s \in C^2(e) \text{ and } s \in C^4(v), \forall e, v \in \triangle\}, \tag{2.2.21}$$

a construction for a stable local minimal determining set is given. Thus, $\mathcal{S}_1$ has optimal approximation power, as does S^1_d for any $d >= 9$.

Theorem 2.2.3. *For all u in $W^{m+1}_q(\Omega)$ with $1 \leq q \leq \infty$, there exists a spline s in $\mathcal{S}_1(\triangle)$ such that*

$$\|D^\alpha(u-s)\|_{q,\Omega} \leq K|\triangle|^{m+1-|\alpha|}|u|_{m+1,q,\Omega}, \tag{2.2.22}$$

for all $0 \leq |\alpha| \leq m \leq 9$. The constant K depends only on d and the tetrahedral partition.

Finally, we remark that the definition of the two dimensional complex spline space 4.3.3 also holds in the trivariate setting. More details about the properties of spline functions can be found in [23] and [35].

Chapter 3

The Maxwell Equations

3.1 A Brief History

The groundwork for the theory of electrodynamics was begun in 1819 when Danish scientist Hans Christain Ørsted performed an experiment in which he held a compass near a wire. When he ran current through the wire, the compass needle moved, revealing a previously undiscovered relationship between electrical and magnetic phenomena. After hearing about Ørsted's findings in 1820, it took the Frenchman André Ampère just one week to hypothesize a mathematical theory to describe them. He predicted that the usual orientation of a compass's needle could be explained by electrical currents within the earth, and hypothesized and later verified attractive and repulsive forces between current carrying wires. He published his work in 1821, and the equation therein would eventually become the fourth of the Maxwell equations.

The primary contributions of Ørsted and Ampère occurred a decade before the birth of James Maxwell in 1831. From this time until Maxwell's work began in the 1850s, most of the progress in the field was made by chemist Michael Faraday. He performed an astounding number of careful experiments, and although he never translated his findings into mathematical models, he was incredibly productive. Faraday discovered the principle of electromagnetic induction and used the idea to build the first generator, the first transformer, and the first electric motor.

There were two key experiments that led to the most important parts of Faraday's work. One experiment involved wrapping coils of wire around an iron ring. The wires were electrically insulated from one another, and yet, when a current was passed through one of the wires, a current in the other was briefly detected. This investigation was later extended; a current could also be induced in the wires by passing a magnet through the center of the iron ring. In a second consequential experiment, Faraday discovered that he could generate current in a closed circuit simply by varying the distance between the circuit and a magnet. This evidence of a relationship between a changing a magnetic field and electrical phenomena eventually led to the third of Maxwell's equations–Faraday's Law.

Because he was not mathematically sophisticated (though Maxwell himself believed that Faraday was still a "mathematician of a very high order[9]") , Faraday's discoveries were largely ignored by the physics community at the time. Clerk Maxwell successfully converted Faraday's work into mathematical theory. He was born in 1931, educated at Edinburhg Univeristy (1847-50) and Cambridge University (1850-1854), and became a Fellow of the Royal Society of Edinburugh in 1856. His first contribution after earning his graduate degree was a detailed explanation of an idea from Faraday's work. It was published in 1855, entitled *On Fararday's Lines of Force.* He and Faraday hypothesized that electrical and magnetic phenomena did not arise from "action at a distance" (this was the accepted notion at the time, developed at least in part by Weber, Neumann, Riemann, and Lorentz, and referred to as the *German Theory*), but instead were propogations of electromagnetic disturbances traveling at the speed of light. In 1862, he published *On Physical Lines of Force*, and commented on the similarity between the speed of electromagnetic "undulations" and the speed of light as measured in Fizeau's contemporary optical experiments.

This text contained what was perhaps Maxwell's most important contribution to electrodynamics—the so-called "displacement current" addition to Ampère's Law **??**. In fact, Maxwell's motivation for the addition of the $\frac{\partial D}{\partial t}$ term was based on a model of ether rather than on sound physical principles, but it remains today an essential component of the equations. Along the continuity equation which specifies the conservation of charge in a system (Eq. **??**), the displacement current guarantees that both sides of Eq. **??** are divergence free, even for a time-varying electric field. It is also allows for the derivation of the electromagnetic wave euqations from Maxwell's laws.

Maxwell himself hypothesized that light was itself an electromagnetic disturbance in his 1865 publication *A Dynamical Theory of the Electromagnetic Field*, in which the famous Maxwell equations appeared for the first time. Collectively, they consist of 20 equations and 20 variables. Table 3.1 summarizes the quantities involved [36].

Taking advantage of the vector notation, we can represent Maxwell's original 20 equations more compactly. Below are six vector equations (each made up of three component equations)

$$\mathbf{J}_T = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{3.1.1}$$

$$\mu\mathbf{H} = \nabla \times \mathbf{A} \tag{3.1.2}$$

$$\nabla \times \mathbf{H} = 4\pi\mathbf{J}_T \tag{3.1.3}$$

$$\mathbf{E} = \mu\mathbf{v} \times \mathbf{H} - \frac{\partial \mathbf{A}}{\partial t} - \nabla\psi \tag{3.1.4}$$

$$\mathbf{E} = k\mathbf{D} \tag{3.1.5}$$

$$\mathbf{E} = \rho'\mathbf{J} \tag{3.1.6}$$

$$\tag{3.1.7}$$

Table 3.1: Table summarizing the quantities involved in Maxwell's original expression of his equations

Maxwell Variable Name	Maxwell Symbol	Modern Variable Name	Modern Symbol
Electromagnetic Momentum	F, G, H	Magnetic Vector Potential	$\mathbf{A}$
Magnetic Force	α, β, γ	Magnetic Field Intensity	$\mathbf{H}$
Electromotive Force	P, Q, R	Electric Field Intensity	$\mathbf{E}$
Current Due to True Conduction	p, q, r	Conduction Current Density	$\mathbf{J}$
Electric Displacement	f, g, h	Electric Flux Density	$\mathbf{D}$
Total Current Including Variation of Displacement	$p^l = p + \frac{df}{dt}$ $q^l = q + \frac{dg}{dt}$ $r^l = r + \frac{dh}{dt}$	Conduction plus Displacement Current Density	$\mathbf{J}_T$
Quantity of Free Electricity	e	Volume Density of Electric Charge	ρ
Electric Potential	ψ	Electric Scalar Potential	ψ

where $\mathbf{v}$ is the velocity of a conductor moving in an isotropic medium, μ is what Maxwell called the coefficient of magnetic induction (we now refer to this quantity as the permeability of the medium, and set the flux density vector $\mathbf{B} = \mu\mathbf{H}$), k is the coefficient of electric elasticity (related to what is now the permittivity of the medium), and ρ' is the resistivity of the medium. The remaining two equations are the scalar equations

$$\nabla \cdot \mathbf{D} = \rho \tag{3.1.8}$$

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t}. \tag{3.1.9}$$

Decades later, in the 1880s, Heinrich Hertz and Oliver Heaviside both independently reformulated these into a set of four equations involving the field vectors $\mathbf{E}$, $\mathbf{B}$,

D, **H**. One reason for the delay in the advancement of the theory was Maxwell's use of quaternions in his original work, a concept which was unfamiliar to most physicists of the time[25]. Below are the modern forms of Maxwell's equations in a vacuum:

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\rho \qquad \text{Gauss' Law} \tag{3.1.10}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{Faraday's Law of Induction} \tag{3.1.11}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{Gauss' Law for Magnetism} \tag{3.1.12}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \qquad \text{Ampère's Law.} \tag{3.1.13}$$

In the time since their conception, Maxwell's field equations have had a profound impact, not just in the development of electromagnetic theory, but also in the design of electromagnetic devices. In the late 1800s, device makers based their designs only on circuit theory, and although they were aware of some additional interference from electromagnetic fields, they disregarded these contributions, as 1) their devices were low frequency and the field effects were minimal, and 2) the device manufacturing was not precise enough to correct or counteract these effects. As device manufacturing improved and more powerful electric machines were built, efforts were made to calculate the effect of electromagnetic fields in important regions of the device. Maxwell's equations applied to the specific geometries of the machines in question were far too difficult to solve analytically. Engineers had to resort to hand-plotting techniques to solve a simplified version of the problem; they converted Maxwell's equations into uncoupled Poisson equations, and estimated fluxes in specific regions of the machines in question.

In the early 1900s, as demand for improved electronic devices increased, engineers began to use idealized models for parts of their machines, and solved analytically for the fields in those regions. It also became common practice to use other methods like fluid flow systems or resistive networks to model the electromagnetic effects. Methods like these have since become known as classical design method, and were largely ad hoc ways of dealing with the field interference.

New devices invented in the mid 1900s like linear induction motors, axial flux machines, internal permanent magnet machines (IPM) involved strong magnetic fields which rendered the classical design methods obsolete. Advances in the theory happened too, however. In 1959 Hammond presented algebraic methods for solving for field distributions in simple electric machines; in 1960, Carpenter published a paper on calculating forces on magnetized iron components of machines using the Maxwell stress tensor; and most importantly, the development of computer-based methods grew to allow for full field solutions of Maxwell's equations. By the late 1970s, computation power allowed for solutions of simple 2D magnetostatic approximations in complex geometries. As computers improved, so did solutions; fewer and fewer approximations were needed until about 2004, when it was possible to solve a fully coupled, dynamical Maxwell system. Today, the ability to model the full fields is used to perform virtual experiments aimed to identify flaws in design. [25]

3.2 Modern Formulation of Maxwell's Equations

The equations (3.2.1-3.2.4) are the Maxwell equations in matter. They written in the traditional way, emphasizing the curl and divergence of the field quantities on the left-hand side. It is important to remember, however, that the source terms in the equations above are the free charge density ρ_f and the free current density $\mathbf{J}_f$. Certain relations hold between the field quantities $\mathbf{E}$ and $\mathbf{B}$ and the auxillary fields $\mathbf{D}$ and $\mathbf{H}$, respectively, depending on what type of medium the fields are passing through. In linear media, for example, we have the constitutive relations

$$\mathbf{D} = \epsilon \mathbf{E} \tag{3.2.1}$$

$$\mathbf{H} = \frac{1}{\mu} \mathbf{B} \tag{3.2.2}$$

where $\epsilon = \epsilon_0(1 + \chi_e) = \epsilon_0 \epsilon_r$ is the *permittivity* of the material and $\mu = \mu_0(1 + \chi_m) = \mu_0 \mu_r$ is its *permeability*. Roughly, the permittivity of a material describes its susceptibility to (electrical) polarization, and the permeability describes a material's magnetic susceptibility. In a vacuum, $\epsilon = \epsilon_0 \approx = 8.854 \times 10^{-12}$ farads per meter and $\mu_0 = 4\pi \times 10^{-7}$ henries per meter. The difference in the magnitudes of these constants points towards he fact that electric forces are typically much larger than magnetic forces[11]. In general, the materials involved in electrodynamic computations may be inhomogenous an anisotropic. In the following, we restrict our attention to isotropic materials.

If we are working in a dielectric or polarizable medium,it is convenient to distinguish between bound charge ρ_b and free charge ρ_f and between bound, polarization, or free current densities $\mathbf{J}_b$, $\mathbf{J}_p$, $\mathbf{J}_f$. Then the Maxwell equations can be written in the following form:

$$\nabla \cdot \mathbf{D} = \rho_f \qquad \text{Gauss' Law} \tag{3.2.3}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{Faraday's Law of Induction} \tag{3.2.4}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{Gauss' Law for Magnetism} \tag{3.2.5}$$

$$\nabla \times \mathbf{H} = \mathbf{J}_f + \frac{\partial \mathbf{D}}{\partial t} \qquad \text{Ampère's Law} \tag{3.2.6}$$

Perhaps the most basic form of Maxwell's equations are the equations for electro- and magnetostatics. The electrostatic equations describe the curl and divergence of a stationary electric field–that is, the field arising from a collection of stationary charges. In a vacuum, static electric fields ($\frac{\partial \boldsymbol{D}}{\partial t} = 0$), lack of free current ($\boldsymbol{J}_f = 0$), and Theorem **??** implies $\boldsymbol{B} = 0$. Then the Maxwell equations take the form

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\rho \tag{3.2.7}$$

$$\nabla \times \mathbf{E} = 0 \tag{3.2.8}$$

where ϵ_0 is the electric permittivity of free space and ρ is the source charge distribution. With the condition that $\mathbf{E} \to 0$ as the distance from the source charge distribution, $\mathbf{r} \to \infty$, the above equations determine the electric field, given ρ[11].

Similarly, the magnetostatic equations arise from physical situations involving a constant flow of current, $\mathbf{J}$, or $\frac{\partial \mathbf{J}}{\partial t} = 0$. Then the system of equations decouples, and with the condition that $\mathbf{B} \to 0$ as the distance from the currents grows to infinity, the equations

$$\nabla \cdot \mathbf{B} = 0 \tag{3.2.9}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}. \tag{3.2.10}$$

determine the magnetic field.

The physical content of these static equations is clear; electric fields diverge away from stationary point charges, while magnetic fields curl around the flow of a steady current.

Another important formulation of the Maxwell system comes from the the time-harmonic regime. These can be derived via Fourier transform (assuming the fields admit the integration) or by assuming the fields behave periodically (with the same frequency ω) in time, or because we simply wish to study the field behavior at a particular frequency [**?**]. In any case, we assume the field quantities in question take the form

$$\boldsymbol{E}(\mathbf{x},t) = \mathrm{Re}\left(e^{-i\omega t}\hat{\boldsymbol{E}}(\mathbf{x})\right) \tag{3.2.11}$$

(similarly for $\boldsymbol{H}$, $\boldsymbol{D}$, and $\boldsymbol{B}$), and that the source terms ρ and $\boldsymbol{J}$ can likewise be written

$$\rho(\mathbf{x},t) = \mathrm{Re}\left(e^{i\omega t}\hat{\rho}(\mathbf{x})\right) \tag{3.2.12}$$

$$\boldsymbol{J}(\mathbf{x},t) = \mathrm{Re}\left(e^{i\omega t}\hat{\boldsymbol{\mathcal{J}}}(\mathbf{x})\right). \tag{3.2.13}$$

Then the time-harmonic Maxwell equations are given by

$$\nabla \cdot \hat{\boldsymbol{D}} = \hat{\rho} \tag{3.2.14}$$

$$\nabla \times \hat{\boldsymbol{E}} - i\omega\hat{\boldsymbol{B}} = 0 \tag{3.2.15}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{3.2.16}$$

$$\nabla \times \hat{\boldsymbol{H}} - i\omega\hat{\boldsymbol{D}} = \hat{\boldsymbol{\mathcal{J}}}. \tag{3.2.17}$$

Notably, the electromagnetic wave equation can be easily derived from the Maxwell Equations in a vacuum. Taking the curl of Faraday's Law, and applying the vector identity

$$\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2\mathbf{A}, \tag{3.2.18}$$

we have

$$\nabla \times (\nabla \times \boldsymbol{E}) = \nabla\left(\nabla \cdot \boldsymbol{E}\right) - \nabla^2\boldsymbol{E} = \nabla \times (-\frac{\partial \boldsymbol{B}}{\partial t}). \tag{3.2.19}$$

Gauss' Law means $\nabla \cdot \boldsymbol{E} = 0$, and, interchanging the order of differentiation to substitute Ampère's Law on the right-hand side, we have

$$-\Delta \boldsymbol{E} = -\frac{\partial}{\partial t}\left(\mu_0\epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}\right) \tag{3.2.20}$$

$$\implies \quad -\Delta \boldsymbol{E} + \mu_0\epsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = 0. \tag{3.2.21}$$

The same analysis for $\boldsymbol{B}$ yields the same equation. Evidently, electromagnetic phenomenon move as waves through space at a speed

$$c = \frac{1}{\sqrt{\epsilon_0\mu_0}} \approx 3 \times 10^8 \text{ m/s}. \tag{3.2.22}$$

Of course, if the waves propagate at a single frequency ω, the wave equation may be reduced to the Helmholtz equation with wavenumber $k = \sqrt{(}\mu_0\epsilon_0)\omega = \frac{\omega}{c}$.

Consideration of the Maxwell equations in potential form also leads to a Helmholtz-type equation. If $\mathbf{B} = \nabla \times \mathbf{A}$ for some vector field $\mathbf{A}$, then $\nabla \cdot \mathbf{B} = \nabla \cdot (\nabla \times \mathbf{A}) = 0$; it is also true (using the Helmholtz decomposition) that if $\mathbf{B} \in \mathcal{C}^2$ and $\mathbf{B} \to 0$ faster than $1/r$, then we indeed have $\mathbf{B} = \nabla \times \mathbf{A}$. Substituting this into (3) we have

$$\nabla \times \mathbf{E} = -\frac{\partial(\nabla \times \mathbf{A})}{\partial t} \implies \nabla \times (\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}) = 0$$

With the same limiting assumptions on $\mathbf{E}+\frac{\partial \mathbf{A}}{\partial t}$ as above, this implies that for some scalar function $-\phi$, we have $-\nabla\phi = (\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t})$. This means that $\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}$, so from (2) we have

$$\nabla \cdot (-\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}) = \frac{1}{\epsilon_0}\rho$$

$$\Delta\phi + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = -\frac{1}{\epsilon_0}\rho. \tag{3.2.23}$$

Similarly, substituting into Eq. 3.1.4 and making use of the identity **??** again, we get

$$\begin{aligned}
\nabla \times (\nabla \times \mathbf{A}) &= \mu_0 J + \mu_0\epsilon_0 \frac{\partial}{\partial t}(-\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}) \\
\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} &= \mu_0 J - \mu_0\epsilon_0 \left(\nabla \frac{\partial \phi}{\partial t} + \frac{\partial^2 \mathbf{A}}{\partial t^2}\right) \\
-\mu_0 \mathbf{J} &= \left(\nabla^2 \mathbf{A} - \mu_0\epsilon_0 \frac{\partial^2 \mathbf{A}}{\partial t^2}\right) - \nabla\left(\nabla \cdot \mathbf{A} + \mu_0\epsilon_0 \frac{\partial \phi}{\partial t}\right) \qquad (3.2.24)
\end{aligned}$$

At this point, we have succeeded in rewriting Equations 3.1.1-3.1.4 in terms of the potential functions $\mathbf{A}$ and ϕ. We can now take advantage of the opportunity to impose extra conditions on the scalar potential ϕ and the vector potential $\mathbf{A}$. This is referred to as *gauge freedom*, and it refers to the fact that there might be multiple potential functions which correspond to the same electric and magnetic fields. Let $\mathbf{A}'$ and ϕ' be such functions, with $\mathbf{A}' - \mathbf{A} = \mathbf{a}$ and $\phi' - \phi = p$. Then we must have $\nabla \times \mathbf{a} = 0$, so $\mathbf{a} = \nabla a$; similarly,

$$-\nabla p = -\nabla(\phi' - \phi) = (\mathbf{E} + \frac{\partial \mathbf{A}'}{\partial t}) - (\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}) = \frac{\partial \mathbf{a}}{\partial t}. \qquad (3.2.25)$$

This implies $\nabla(p + \frac{\partial a}{\partial t}) = 0$, so we conclude $p + \frac{\partial a}{\partial t}$ is a function of time only. We can call this function $k(t)$, and absorb it into the arbitrary potential difference p. Then we have $p(t) = -\frac{\partial a}{\partial t}$, so we discover that we're free to add a gradient of a scalar function a to $\mathbf{A}$ as long as we subtract $\frac{\partial a}{\partial t}$ from ϕ. These changes are called *gauge transformations*, and the widely used Lorentz gauge

$$\nabla \cdot \mathbf{A} = -\mu_0\epsilon_0 \frac{\partial \phi}{\partial t} \qquad (3.2.26)$$

allows us to recast equations (**??**) and (**??**). The rightmost term in (**??**) vanishes, and we have

$$\Delta\phi - \mu_0\epsilon_0\frac{\partial^2\phi}{\partial t^2} = -\frac{1}{\epsilon_0}\rho \tag{3.2.27}$$

$$\Delta\mathbf{A} - \mu_0\epsilon_0\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu_0\mathbf{J} \tag{3.2.28}$$

Using this gauge allows us to solve for the vector and scalar potential in the same way; both are acted on by the same differential operator. Now both equations are in the form $\Box^2 u = -f$, and, assuming the quantities involved admit a Fourier transform in time, the problem to be solved takes Helmholtz form.

$$\Delta\hat{u} + k^2\hat{u} = -\hat{f}.$$

In the following, we are particularly interested in boundary value problems of the time-harmonic Maxwell equations (most of our numerical examples arise from this setting). For exterior domain problems, see [?]

3.3 Boundary Conditions

Solving the Maxwell equations in a domain of interest amounts to solving a boundary value problem. Here we explore the boundary conditions that arise in many practical electromagnetic problems involving interfaces between conducting and nonconducting materials.[11]

3.3.1 Interface Conditions Arising from the Divergence Equations

We can derive conditions on the field quantities $\boldsymbol{E}$ and $\boldsymbol{H}$ by considering the integral form of equations 3.2.1 and 3.2.3:

$$\begin{aligned} &\oint_S \boldsymbol{D} \cdot d\boldsymbol{a} = Q_{f_{enc}} && \text{Gauss' Law in Integral Form} \\ &\oint_S \boldsymbol{B} \cdot d\boldsymbol{a} = 0 && \text{Gauss' Law for Magnetism in Integral Form} \end{aligned}$$

where the integrals in question may be done over any closed surface S, enclosing total charge $Q_{f_{enc}}$. At an interface between two surfaces, we imagine S to be the surface of a thin box whose thickness just barely allows it to extend into both materials. The top and bottom of the box have non-negligible surface area, but in the limit that the thickness of the box goes to 0, the sides contribute nothing to the integral. At the same time, however, the surface charge cσ_f contained within the box, at the interface itself, does not change. For a box which is small enough so that the fields $\boldsymbol{D}_i$ and normal $\boldsymbol{n}$ of the interface are approximately constant, we have

$$\oint_S \boldsymbol{D} \cdot d\boldsymbol{a} = a(\boldsymbol{D}_1 \cdot \boldsymbol{n} - \boldsymbol{D}_2 \cdot \boldsymbol{n}) = \sigma_f a$$

which leads to the boundary condition

$$D_1^\perp - D_2^\perp = \sigma_f.$$

This condition tells us that at an interface between two materials the normal component of the electric displacement is discontinuous if there is any surface charge present. For linear media the boundary condition takes the form $\epsilon_1 E_1^\perp - \epsilon_2 E_2^\perp = \sigma_f$.

If, as is often the case, there is no charge present at the interface, we have

$$\epsilon_1 \boldsymbol{E}_1 \cdot \boldsymbol{n} = \epsilon_2 \boldsymbol{E}_2 \cdot \boldsymbol{n}, \tag{3.3.1}$$

where $\boldsymbol{n}$ points from material 2 into material 1.

For the same reason, we see that the perpendicular component of $\boldsymbol{B}$ is continuous across an interface:

$$B_1^{\perp} - B_2^{\perp} = 0.$$

Our integrals are line integrals now, instead of surface integrals, so

3.3.2 Interface Conditions Arising from the Curl Equations

We first consider the integral form of the curl equations:

$$\begin{aligned} \oint_P \boldsymbol{E} \cdot d\boldsymbol{l} &= -\frac{d}{dt} \int_S \boldsymbol{B} \cdot d\boldsymbol{a} && \text{Faraday's Law of Induction} \\ \oint_P \boldsymbol{H} \cdot d\boldsymbol{l} &= I_{f_{enc}} + \frac{d}{dt} \int_S \boldsymbol{D} \cdot d\boldsymbol{a} && \text{Ampere's Law,} \end{aligned}$$

Since the integrals in question are now line integrals, we consider a narrow rectangular loop (Amperian loop) P, much broader than tall, which extends into the materials forming the interface. As the height of this loop approaches zero, the integral on the left is dominated by the segments which run parallel, rather than through, the interface. This suggests

$$\boldsymbol{E}_1 \cdot \boldsymbol{l} - \boldsymbol{E}_2 \cdot \boldsymbol{l} = -\frac{d}{dt} \int_S \boldsymbol{B} \cdot d\boldsymbol{a}.$$

But, as the height of the loop goes to zero, so too does its cross sectional area S and the flux of the the magnetic field through S. Therefore we have the boundary condition

$$\boldsymbol{E}_1^{||} - \boldsymbol{E}_2^{||} = 0.$$

Similarly, for $\int_S \boldsymbol{J}_f \cdot d\boldsymbol{a} = I_{f_{enc}}$ where $\boldsymbol{J}_f$ is the free surface current and $I_{f_{enc}}$ is the current flowing through the loop, we have

$$\boldsymbol{H}_1 \cdot \boldsymbol{l} - \boldsymbol{H}_1 \cdot \boldsymbol{l} = I_{f_{enc}}.$$

For the vector $\boldsymbol{n} \times \boldsymbol{l}$ perpendicular to the loop and free surface current density $\boldsymbol{K}_f$, we have

$$I_{f_{enc}} = \boldsymbol{K}_f \cdot (\boldsymbol{n} \times \boldsymbol{l}) = (\boldsymbol{K}_f \times \boldsymbol{n}) \cdot \boldsymbol{l}$$

The interface condition on $\boldsymbol{H}$ follows:

$$\boldsymbol{H}_1^{||} - \boldsymbol{H}_2^{||} = \boldsymbol{K}_f \times \boldsymbol{n};$$

for linear media, this amounts to

$$\frac{1}{\mu_1}\boldsymbol{B}_1^{||} - \frac{1}{\mu_2}\boldsymbol{B}_2^{||} = \boldsymbol{K}_f \times \boldsymbol{n};$$

if there is no surface current, then the condition is

$$\frac{1}{\mu_1}\boldsymbol{B}_1 \times \boldsymbol{n} = \frac{1}{\mu_2}\boldsymbol{B}_2 \times \boldsymbol{n}.$$

3.3.3 Conductors

Conducting materials, utilized in a variety of electronic applications, are often of interest in electrodynamic boundary value problems. In a conductor, electrons are free to travel throughout the material. In practice, conductors are idealized as perfect conductors; we suppose there are an unlimited supply of electrons in the material that can flow around in reaction to electric forces. This physical property leads to the following conditions on electrostatic electric fields and charges in and near a conductor.

1)$\boldsymbol{E} = 0$ *inside* a conductor.

If (momentarily), there is a nonzero electric field within a conductor, the free electrons within the conductor migrate in response to the force that they experience. The result is an accumulation of net charge at the surface of the conductor, arranged in a way that creates an electric field within the conductor that is exactly counter to the external field.

2)$\boldsymbol{E}$ is perpendicular to the surface just outside the surface of a conductor.

Suppose this weren't true; then the electric field would have a tangential component at some point on the surface. This field would cause electrons to flow along the surface of the conductor until they no longer experienced a net force, canceling out the tangential component. In actuality, the charge on the surface (if there is any) spreads out “evenly" along the surface; it can be shown that this configuration is the minimal energy configuration of a net charge on a conductor.

3)A conductor is an equipotential surface.

This follows from 2); we have $V(b) - V(a) = -\int_L \boldsymbol{E} \cdot d\boldsymbol{l}$, but since L is a path along the surface of the conductor, $\boldsymbol{E} \cdot d\boldsymbol{l} = 0$ everywhere. Thus, for two points $\boldsymbol{a}, \boldsymbol{b}$ on the surface, we must have $V(a) = V(b)$.

Chapter 4

The Helmholtz Equation

4.1 Introduction

The following partial differential equation, referred to as Helmholtz equation or reduced wave equation is well known:

$$\begin{cases} -\Delta u - k^2 u &= f, \quad \text{in } \Omega \\ \mathbf{n} \cdot \nabla u + \mathbf{i} k u &= g \quad \text{in } \partial\Omega, \end{cases} \tag{4.1.1}$$

where $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$, is a bounded domain with Lipschitz boundary, $\mathbf{i} = \sqrt{-1}$ denotes the imaginary unit, $\mathbf{n}$ is the unit normal to $\partial\Omega$, and k is the wave number. This Helmholtz problem arises from many application areas: acoustic scattering, electromagnetic fields, etc.. In particular, the solution to (4.1.1) provides an approach for numerical solution of Maxwell's equations in a special case. Over many years, the finite element method, discontinuous Galerkin methods, weak-Galerkin methods, and their variants have been used to tackle the numerical solution of the Helmholtz equation (4.1.1) when wave number k is large. See literature in [8, 29**? ? ? ? ? ? ?**], and etc.. Theoretical study on the existence, uniqueness, stability of the Helmholtz problem (4.1.1) has been carried out extensively. See existence and uniqueness of the weak solution of (4.1.1) in [?]. See [?] and [?] for the stability of the weak solution under the assumption of the domain which is strictly star-shaped (see its definition below). However, similar results for other domains have not been established so far. In addition, the numerical computation of solutions to (4.1.1) remains challenging when the wave number k is large.

Let us first present a quick review of the study of finite element method, discontinuous Galerkin method, weak Galerkin method and their variations in [8, 29? ? ? ? ? ?] for numerical solution to (4.1.1). In all references mentioned above, the underlying domain Ω has to be a strictly star-shaped domain , which means that there exist a point $\mathbf{x}_0 \in \Omega$ and a positive constant γ_Ω depending only on Ω such that

$$(\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{n} \geq \gamma_\Omega > 0, \quad \forall \mathbf{x} \in \partial\Omega. \tag{4.1.2}$$

If $\gamma_\Omega = 0$, Ω is said to be star-shaped domain. Nevertheless, all the computational methods work well for non-convex domains as well as domains which are not strictly star-shaped. Mathematically, it is interesting to have a theory for more general domains.

The convergence analysis of many existing numerical methods has been carried out in the literature. To explain the analysis, we shall use the following norm over a complex-valued Sobolev space $\mathbb{H}^1(\Omega)$ over Ω in the paper:

$$|||u|||_{1,k,\Omega} := (\|\nabla u\|^2_{L^2(\Omega)} + k^2\|u\|^2_{L^2(\Omega)})^{1/2}. \tag{4.1.3}$$

This is equivalent to the standard H^1-norm on $\mathbb{H}^1(\Omega)$ with constants dependent on k. In [?], the following result was established.

Theorem 4.1.1 (Proposition 8.2.7 in [?]). *Let Ω be a bounded star-shaped domain with smooth boundary (or a bounded convex domain). Let $S_h \subset \mathbb{H}^1(\Omega)$ be the finite element space. Then there exists a positive constant C_3 dependent on Ω such that,* *under the assumption that $1 + k^2 h \leq C$,*

$$|||u - u_{FE}|||_{1,k,\Omega} \leq C_3 \inf_{s \in S_h} |||u - s|||_{1,k,\Omega}. \tag{4.1.4}$$

This result was improved several times in [27**?**] and recently in [**?**]. That is, letting $S_h \subset \mathbb{H}^1(\Omega)$ be the higher order finite element space of degree p over triangulation $\triangle$ with size $h = |\triangle|$, a subspace of complex-valued Sobolev space $\mathbb{H}^1(\Omega)$, Du and Wu proved the following result in [**?**]:

Theorem 4.1.2 (Theorem 5.1 in [**?**])**.** *Let u and u_h be the weak solutions satisfying (4.1.1) and (4.3.4), respectively. Then there exists a constant C_0 independent of k and h such that if*

$$k(kh)^p \leq C_0 \tag{4.1.5}$$

then the following estimate holds:

$$|||u - u_h|||_{1,k,\Omega} \leq (1 + k(kh)^p) \inf_{s \in S_h} |||u - s|||_{1,k,\Omega}. \tag{4.1.6}$$

For the internal penalty discontinuous Galerkin (IPDG) method using the spline $S_p^{-1}(\triangle)$ of discontinuous piecewise polynomials of degree p over triangulation $\triangle$ with an internal penalty, Du and Zhu in [**?**] obtained the following

Theorem 4.1.3 (Theorem 1 in [**?**])**.** *Let u be the weak solutions satisfying (4.1.1) and u_h be the IPDG solution based on $S_p^{-1}(\triangle)$. Then there exists a constant C_0 independent of k and h such that if*

$$k(kh)^{2p} \leq C_0 \tag{4.1.7}$$

then the following estimate holds:

$$\|\!|u - u_h\|\!|_E \leq (1 + k(kh)^p) \inf_{s \in S_h} \|\!|u - s\|\!|_E, \tag{4.1.8}$$

where $S_h = S_p^{-1}(\triangle) \cap H^1(\Omega)$ *and* $\|\!|u - u_h\|\!|_E$ *is the norm in terms of jumps of function values as well as derivative values.*

In addition, the results above improve the study in [?] which has a very complicated proof. Not only does the convergence analysis of numerical methods require $k^{1+q}h \leq C < \infty$ for some $q > 0$, but also numerical experiments show the so-called pollution phenomenon which ruins the accuracy of numerical solution when wave number k is large, even $k = 200$ or $k = 300$.

In this paper, we shall provide a new way to establish the existence, uniqueness and stability of the weak solution to the Helmholtz equation under a new assumption. To replace the assumption that the domain is strictly star-shaped, we assume that for wave number k such that k^2 is not a Dirichlet eigenvalue of the Laplace operator over Ω. Under this new assumption, we are able to establish the coercivity of the sesquilinear form $B(u, v)$ and use the Lax-Milgram theorem to establish the existence and uniqueness of the weak solution to the Helmholtz equation in (4.1.1). Due to the new assumption, the study leads to the new stability estimate of the weak solution which does not require the classic assumption of strictly star-shaped domains. The new stability estimate enables us to give a new convergence analysis. Although we are not able to find out how the coercivity constant is dependent on k, we are able to show that the coercivity constant will not go to zero when $k \to \infty$ and hence the desired approximation order will be achieved when $kh \leq C < \infty$. These is much better than the assumptions $(1 + k^2h) \leq C < \infty$ and $k(kh)^p \leq C < \infty$ mentioned in (4.1.4), and (4.1.5) and (4.1.7) above.

In addition to the theoretical analysis of the new stability estimate and convergence analysis, in this paper we shall explain how to use bivariate spline functions to numerically solve (4.1.1) and demonstrate that our spline method can be more efficient and effective to find the numerical solution of Helmholtz equations with large wave numbers, e.g. $k = 500$— 1500 in Example 5.1.4 in a later section. We will present a large amount of numerical evidence to demonstrate the convergence of our bivariate spline method. No pollution phenomenon is observed in our computational experiments. Numerical results show that bivariate spline method is much better than the weak-Galerkin(WG) method in [29] and hybridized DG and WG methods in [**?**], [**?**] in the sense that we are able to achieve high accuracy and for larger wave numbers. More numerical evidence on spline solution to Helmholtz equation over inhomogeneous media and Maxwell equations with time harmonic source term will be reported elsewhere. See, e.g. [**?**] and [**?**].

The paper is organized as follows. We first explain the existence, uniqueness, and stability of the weak solution in $H^1(\Omega)$ and in a spline space and a convergence analysis of spline weak solution in §2 and §3 and in §4. Next we present our numerical results in §5, where we present some simulation results by using Bessel functions as a known solution and check how accurate our spline solutions are in convex and nonconvex settings. Mainly, our spline solutions with various degrees $p = 5, 6, \cdots, 17$ will be used for wave numbers 5— 1500. Finally, we make a few remarks on many unsolved research problems in §6 as the study generates some interesting mathematical problems on the behavior of Dirichlet eigenvalues and coercivity constants L.

4.2 The Well-Posedness of the Helmholtz BVP

Let $\mathbb{L}^2(\Omega)$ be the space of all complex-valued square integrable functions over Ω and $\mathbb{H}^1(\Omega) \subset \mathbb{L}^2(\Omega)$ be the complex-valued square integrable functions over Ω such that their gradients are in $\mathbb{L}^2(\Omega)$. We introduce the following sesquilinear form:

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla \overline{v} dxdy, \quad u, v \in \mathbb{H}^1(\Omega),$$

where $\overline{u}$ stands for the complex conjugate of the complex-valued function u. Also we need two different inner products. let

$$\langle u, v\rangle_\Omega = \int_\Omega u\overline{v}dxdy \quad \forall u, v \in \mathbb{L}^2(\Omega) \text{ and } \langle u, v\rangle_\Gamma = \int_\Gamma u\overline{v}d\Gamma, \quad \forall u, v \in \mathbb{L}^2(\Gamma),$$

be the standard inner products in $\mathbb{L}^2(\Omega)$ and in $\mathbb{L}^2(\Gamma)$, respectively, where $\Gamma = \partial\Omega$. The variational formulation to the Helmholtz problem (4.1.1) is to find $u \in \mathbb{H}^1(\Omega)$ such that

$$a(u,v) - k^2\langle u, v\rangle_\Omega + \mathbf{i}k\langle u, v\rangle_\Gamma = \langle f, v\rangle_\Omega + \langle g, v\rangle_\Gamma, \quad \forall v \in \mathbb{H}^1(\Omega) \tag{4.2.1}$$

which is the weak formulation of (4.1.1). If a function $u \in \mathbb{H}^1(\Omega)$ satisfies the above equation, u is called the weak solution.

In this section, we mainly discuss the existence, uniqueness, and stability of the weak solution to the Helmholtz problem (4.1.1). We first present a proof of the existence and uniqueness based on the Fredholm alternative theorem and then we use Lax-Milgram theorem to give another proof. Consider the following two second order partial differential equations:

$$\begin{cases} \Delta u + \lambda u & = 0, \text{ in } \Omega \\ \mathbf{n} \cdot \nabla u + \mathbf{i}ku & = 0, \text{ in } \partial\Omega \end{cases} \tag{4.2.2}$$

and

$$\begin{cases} \Delta u + \lambda u & = f, \text{ in } \Omega \\ \mathbf{n} \cdot \nabla u + \mathbf{i} k u & = 0, \text{ in } \partial\Omega, \end{cases} \tag{4.2.3}$$

where $\lambda > 0$ is a constant, $f \in L^2(\Omega)$. We have the following well-known result:

Theorem 4.2.1 (Fredholm Alternative Theorem)**.** *Fix $\lambda > 0$. Precisely one of the following two statements holds: Either (4.2.2) has a nonzero weak solution $u \in \mathbb{H}^1(\Omega)$ or there exists a unique weak solution $u_f \in \mathbb{H}^1(\Omega)$ satisfying (4.2.3).*

We refer to [**?**] for a proof for the case with Dirichlet boundary condition. A similar argument works for Theorem 4.2.1 and the detail is left to the interested reader. The following existence and uniqueness is well-known (cf. e.g. [**?**]). For convenience, we present another proof.

Theorem 4.2.2. *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^2$. Then there exists a unique weak solution $u \in \mathbb{H}^1(\Omega)$ to (4.1.1) in the sense that it satisfies (4.2.1).*

Proof. By Fredholm Alternative Theorem 4.2.1, let us show that k^2 is not an eigenvalue of (4.2.2). Otherwise, if there exists a nonzero eigenfunction $u_{k^2} \in \mathbb{H}^1(\Omega)$ satisfying (4.2.2) with $\lambda = k^2$, then the weak formulation of (4.2.2), i.e.

$$a(u_{k^2}, v) - k^2 \langle u_{k^2}, v \rangle_\Omega + \mathbf{i} k \langle u_{k^2}, v \rangle_\Gamma = 0, \forall v \in \mathbb{H}^1(\Omega),$$

shows that $u_{k^2} = 0$ on Γ by using $v = u_{k^2}$. It follows that $\mathbf{n} \cdot \nabla u_{k^2} = 0$ on Γ by using the boundary condition of (4.2.2). That is, u_{k^2} is also an eigenfunction of Laplacian operator over Ω associated with Neumann boundary condition. By using the following Lemma 4.2.1, $u_{k^2} \equiv 0$ as $u_{k^2} \in H_0^1(\Omega)$. This is a contradiction and hence, k^2 is not an eigenvalue of (4.2.2). Fredholm Alternative theorem implies that (4.2.3) has a unique solution. □

In the proof above, we have used the result of Lemma 4.2.1. Let us introduce some notation. We first recall that the standard eigenvalue problem associated with Laplacian operator Δ:

$$\begin{cases} \Delta u + \lambda u & = 0, \text{ in } \Omega \\ \quad u & = 0, \text{ in } \partial\Omega. \end{cases} \tag{4.2.4}$$

If (4.2.4) has a nonzero solution, λ is called an eigenvalue (or Dirichlet eigenvalue) of the Laplace operator Δ over the underlying domain Ω. It is known that all eigenvalues are positive and there is an infinitely many eigenvalues which increase to infinity. Let us write $\lambda_i, i = 1, \cdots, \infty$ for the eigenvalues and ϕ_i for a normalized eigenfunction associated with λ_i. Similarly, let $v_\nu \in H^1(\Omega)$ be an eigenfunction associated with Neumann eigenvalue ν, i.e. v_ν satisfies the following

$$\begin{cases} -\Delta u - \nu u & = 0, \text{ in } \Omega \\ \quad \mathbf{n} \cdot \nabla u & = 0, \text{ on } \partial\Omega. \end{cases} \tag{4.2.5}$$

For convenience, let us write $\ker(-\Delta - \nu I)$ be the eigen-space associated with Neumann eigenvalue ν, i.e. the collection of all eigen-function $v_\nu \in H^1(\Omega)$ satisfying (4.2.5). It is known that the sequence of the Neumann eigenvalues is unbounded, nonnegative, and countably infinite. We are now ready to prove the following

Lemma 4.2.1 (Filonov, 2004[**?**]). *For each Neumann eigenvalue $\nu > 0$,*

$$H_0^1(\Omega) \cap ker(-\Delta - \nu I) = \{0\},$$

where I is the identity operator.

Proof. The proof is short and we include it here for convenience. Let $v_\nu \in H^1(\Omega)$ be an eigenfunction associated with Neumann eigenvalue ν, i.e. $v_\nu \in \ker(-\Delta - \nu I)$. If $v_\nu \in H_0^1(\Omega)$, we extend v_ν by zero outside Ω and call it w. Then $w \in H_0^1(\mathbb{R}^2)$ and we have

$$\int_{\mathbb{R}^2} \nabla w \nabla u = \int_{\Omega} \nabla v_\nu \nabla u = -\nu \int_{\Omega} v_\nu u = -\nu \int_{\mathbb{R}^2} wu$$

for all $u \in H_0^1(\mathbb{R}^2)$. That is, w is an eigenfunction of the Laplacian operator over $\mathbb{R}^2$ and hence, $w \equiv 0$. □

Next let us use the well-known Lax-Milgram theorem to establish the existence, uniqueness, and stability of the weak solution u. To this end, we need some preparatory results. For convenience, let us define a sesquilinear form:

$$B(u,v) = a(u,v) - k^2\langle u,v\rangle_\Omega + \mathbf{i}k\langle u,v\rangle_\Gamma. \tag{4.2.6}$$

Also, we define

$$|||u|||_{1,k,\Omega} := \left(\|\nabla u\|^2_{L^2(\Omega)} + k^2\|u\|^2_{L^2(\Omega)}\right)^{1/2}.$$

It is easy to see $|||\cdot|||_{1,k,\Omega}$ is a norm on $\mathbb{H}^1(\Omega)$. Associated with this norm, we let $\langle u,v\rangle_A = a(u,v) + k^2\langle u,v\rangle$ be the inner product on $\mathbb{H}^1(\Omega)$ in the rest of the paper. The following continuity condition of the sesquilinear form $B(u,v)$ is known (cf. [**?**]).

Lemma 4.2.2. *Suppose that $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$. Then*

$$|B(u,v)| \le C_B |||u|||_{1,k,\Omega}\, |||v|||_{1,k,\Omega}, \tag{4.2.7}$$

where C_B is a positive constant dependent on Ω only.

Recall ϕ_i is a H^1-normalized eigenfunction associated with Dirichlet eigenvalue $\lambda_i, i = 1, \cdots, \infty$. To explain the coercivity of $B(u,v)$, however, we must point out a basic fact that $B(u,v)$ is not coercive when k^2 is a Dirichlet eigenvalue. Indeed, let $u = \phi_i$ be an eigenfunction associated with Dirichlet eigenvalue λ_i. If $k^2 = \lambda_i$, we will have $B(u,v)=0$ for all $v \in \mathbb{H}_0^1(\Omega)$ while $u \neq 0$. In particular, $B(\phi_i,\phi_i)=0$ if $k^2=\lambda_i$. Thus, in the rest of the paper, we shall often make an assumption that k^2 is not a Dirichlet eigenvalue. Write $Y_i = \text{span}\{\phi_1,\cdots,\phi_i\} \subset H_0^1(\Omega)$. Using Rayleigh-Ritz approximation, it is known (cf. [?]) that

$$\lambda_{i+1} = \min\{\frac{\|\nabla w\|^2}{\|w\|^2} : w \in Y_i^\perp\}, \tag{4.2.8}$$

where $Y_i^\perp$ is the orthogonal complement of Y_i in $H_0^1(\Omega)$ under the inner product $\int_\Omega \nabla w \cdot \nabla v$. Note that over $H_0^1(\Omega)$, the inner product $\langle w,v\rangle_A$ is equivalent to $\int_\Omega \nabla w \cdot \nabla v$ if k^2 is not an eigenvalue. Indeed for any $v \in Y_i$, say $v=\phi_j$ for some $1 \le j \le i$ and $w \in Y_i^\perp$,

$$\int_\Omega \nabla w \cdot \nabla \bar{v} = -\int_\Omega w\Delta\bar{v} = -\lambda_j \int_\Omega w\bar{v}.$$

Thus, $\langle v,w\rangle_A = (1-k^2/\lambda_j)\int_\Omega \nabla w \cdot \nabla \bar{v}$. Furthermore, let $X_i^\perp$ be the orthogonal complement of Y_i in $\mathbb{H}^1(\Omega)$. We shall point out another fact that $\mathbb{H}_0^1(\Omega)$ is not dense in $\mathbb{H}^1(\Omega)$. Otherwise, the testing space in (4.2.1) could be replaced by $\mathbb{H}_0^1(\Omega)$. Then when $k^2=\lambda_i$, we could have more than one solution to (4.2.1) as we know $a(\phi_i,v)-k^2(\phi_i,v)+\mathbf{i}\langle\phi_i,v\rangle_\Gamma = 0$ which can be added into (4.2.1) for another solution. This would violate Theorem 4.2.2. Therefore, $X_i^\perp \neq Y_i^\perp$. Also for convenience, we shall use $\lambda_0=0$ in the rest of the paper although λ_0 is not a Dirichlet eigenvalue.

Theorem 4.2.3. *Let Ω be a domain with Lipschitz boundary. Let λ_{i+1} be the first eigenvalue of the Laplacian operator over Ω such that $k^2 < \lambda_{i+1}$. Then there exists a lower bound $L > 0$ such that*

$$|B(u,u)| \ge L\|\!|u|\!\|_{1,k,\Omega}^2, \quad \forall u \in X_i^\perp. \tag{4.2.9}$$

Furthermore, L does not go to 0 as $k \to \infty$.

Proof. If (4.2.9) is not true, then there exists a sequence $u_n \in X_i^{\perp}$ such that $|||u_n|||_{1,k,\Omega}^2 = 1$ and $|B(u_n, u_n)| \leq 1/n$ for $n \geq 1$. The boundedness of u_n in $X_i^{\perp} \subset \mathbb{H}^1(\Omega)$ implies that there exists a $u^* \in \mathbb{H}^1(\Omega)$ such that a subsequence, say the whole sequence $\{u_n, n \geq 1\}$ converges to u^* in $L^2(\Omega)$ norm and converges to u^* weakly in $H^1(\Omega)$ semi-norm by Rellich-Kondrachov Theorem (cf. [**?**]). Indeed, the boundedness of $u_n \in H^1(\Omega)$ implies that there exists a subsequence which is weakly convergent to $u^* \in H^1(\Omega)$ and then the subsequence contains a subsequence which is strongly convergent to u^* in L^2 norm by Rellich-Kondrachov Theorem. It follows that

$$a(u_n, u^*) - k^2\langle u_n, u^*\rangle \longrightarrow a(u^*, u^*) - k^2\langle u^*, u^*\rangle,$$

$\|\nabla u_n\| \to \|\nabla u^*\|$, and $\langle u_n, u^*\rangle_\Gamma \to \langle u^*, u^*\rangle_\Gamma$ by using the Sobolev trace theorem. That is,

$$|B(u^*, u^*)| = 0$$

In other words, the real and imaginary parts of $B(u^*, u^*)$ implies that $\|\nabla u^*\|_{L^2(\Omega)}^2 = k^2\|u^*\|_{L^2(\Omega)}^2$ and $\int_\Gamma |u^*|^2 d\Gamma = 0$. Thus, $u^* \in \mathbb{H}_0^1(\Omega)$. Furthermore, since u_n is orthogonal to Y_i, so is u^*. It follows that $u^* \in Y_i^{\perp}$. If $u^* \neq 0$, the inequality in (4.2.8) implies $\lambda_{i+1} \leq \dfrac{\|\nabla u^*\|^2}{\|u^*\|^2} = k^2 < \lambda_{i+1}$ which is a contradiction. Thus, we have $u^* \equiv 0$.

On the other hand, $|||u^*|||_{1,k,\Omega} = 1$ because of $|||u_n|||_{1,k,\Omega} = 1$. We get a contradiction again. Therefore, there exists a positive number $L > 0$ satisfying (4.2.9).

Next we claim that $L \not\to 0$ as $k \to \infty$. For convenience, let $L_k > 0$ be the largest constant on the right-hand side of (4.2.9) for each k. Since $L_k > 0$, there exists a $u_k \in \mathbb{H}^1(\Omega)$ with $|||u_k|||_{1,k,\Omega} = 1$ such that

$$|B(u_k, u_k)| \leq 2L_k.$$

Since $|||u_k|||_{1,1,\Omega} \leq |||u_k|||_{1,k,\Omega} = 1$, we know there exists $u^* \in \mathbb{H}^1(\Omega)$ and a subsequence which is weakly convergent to u^* in $\mathbb{H}^1(\Omega)$ and strongly convergent to u^* in L_2 norm by using Rellich-Kondrachov Theorem. As $\|u_k\|_{L^2(\Omega)} \leq 1/k^2$, we see that $\|u^*\| = 0$ and hence, $u^* = 0$ almost everywhere. Thus, $u|_\Gamma = 0$ and $\nabla u^* = 0$. Note that $\|\nabla u_k\| \to \|\nabla u^*\| = 0$ as $k \to \infty$. Now if $L_k \to 0$, we would have $|B(u_k, u_k)| \to 0$ or $|\|\nabla u_k\| - k^2\|u_k\|| \to 0$. It follows that $k^2\|u_k\| \to 0$ which together with $\|\nabla u_k\| \to 0$ proved above contradicts to the fact that $|||u_k|||_{1,k,\Omega} = 1$. □

We are now ready to establish the following existence and uniqueness result by using the Lax-Milgram theorem.

Theorem 4.2.4. *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^2$. Then there exists a unique weak solution $u \in H^1(\Omega)$ to (4.1.1) satisfying (4.2.1).*

Proof. We decompose $\mathbb{H}^1(\Omega) = X_i^\perp \oplus Y_i$, where $X_i^\perp$ is the orthogonal complement of Y_i in $\mathbb{H}^1(\Omega)$ for each $i \geq 0$ with $Y_0 = \mathbb{H}_0^1(\Omega)$ and $X_0 = \mathbb{H}^1(\Omega)$. Suppose that for an integer i, $\lambda_i < k^2 \leq \lambda_{i+1}$, where $\lambda_0 = 0$ although it is not an eigenvalue. We first project the solution onto Y_i which can be done as follows. We compute the projection of f onto Y_i, i.e.

$$f_i = \sum_{j=0}^{i} \langle f, \phi_j \rangle \phi_j. \tag{4.2.10}$$

Then we can choose $u_i \in \mathbb{H}_0^1(\Omega)$ by

$$u_i = -\sum_{j=1}^{i} \frac{1}{-\lambda_j + k^2} \langle f, \phi_j \rangle \phi_j. \tag{4.2.11}$$

Then it is easy to see that u_i satisfies $\Delta u_i + k^2 u_i = -f_i$.

Next we consider $v \in X_i^\perp$ to be the solution

$$\begin{cases} -\Delta v - k^2 v & = f - f_i, \quad \text{in } \Omega \subset \mathbb{R}^2 \\ \mathbf{n} \cdot \nabla v + \mathbf{i} k v & = g - \mathbf{n} \cdot \nabla u_i \quad \text{on } \partial\Omega. \end{cases} \tag{4.2.12}$$

Consider its weak formulation and it is easy to see that the right-hand side of the weak formulation is a continuous linear functional. The continuity of $B(u, v)$ and the coercivity (4.2.9) proved above enable us to use the Lax-Milgram theorem and conclude the existence and uniqueness of the weak solution v of (4.2.12). Now we can easily check $u = v + u_i$ is the solution of (4.1.1) satisfying (4.2.1). Indeed, for any $w \in \mathbb{H}^1(\triangle)$,

$$\begin{aligned}
&B(u, w) = B(u_i, w) + B(v, w) = \langle \nabla u_i, \nabla w\rangle - k^2\langle u_i, w\rangle + B(v, w)\\
&= -\langle \Delta u_i + k^2 u_i, w\rangle + \langle \mathbf{n}\cdot\nabla u_i, w\rangle_\Gamma + \langle f - f_i, w\rangle + \langle g - \mathbf{n}\cdot\nabla u_i, w\rangle_\Gamma\\
&= \sum_{j=0}^{i} \frac{\langle f, \phi_j\rangle}{-\lambda_j + k^2}\langle(-\lambda_j + k^2)\phi_j, w\rangle + \langle f - f_i, w\rangle + \langle g, w\rangle_\Gamma = \langle f, w\rangle + \langle g, w\rangle_\Gamma.
\end{aligned}$$

That is, $u \in \mathbb{H}^1_p(\Omega)$ is the weak solution. The argument of the proof of Theorem 4.2.2 can be used to establish the uniqueness of this solution by using Lemma 4.2.1. □

Furthermore, the weak solution is stable in the following senses.

Theorem 4.2.5. *Suppose that Ω has a $C^{1,1}$ smooth boundary or Ω is convex. Suppose that k^2 is not a Dirichlet eigenvalue of the Laplacian operator over Ω. Let us say $\lambda_i < k^2 \le \lambda_{i+1}$ for some $i \ge 0$. Let $u \in \mathbb{H}^1(\Omega)$ be the unique weak solution to (4.1.1) as explained above. Then there exists a constant $C > 0$ independent of f, g such that*

$$|||u|||_{1,k,\Omega} \le C(\|f\| + \|g\|_\Gamma) \tag{4.2.13}$$

for $k \ge 1$, where C is dependent on $\dfrac{1}{1 - \lambda_i/k^2}$ and the constant L which is the low bound in (4.2.9). Furthermore, suppose Ω is convex and $g \in \mathbb{H}^{3/2}(\Gamma)$. Then

$$|u|_{2,2,\Omega} \le C(1 + k)\left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)}\right) + \|\nabla_T g\|_{L^2(\Gamma)} \tag{4.2.14}$$

for any $k \ge 0$, where ∇_T stands for the tangential derivative on Γ.

Proof. By using the proof of Theorem 4.2.4, we use the orthonormality of ϕ_i to have

$$\|\nabla u_i\|^2_{L^2(\Omega)} = \sum_{j=1}^{i} \left(\frac{\lambda_j}{k^2 - \lambda_j} \right)^2 |\langle f, \phi_j \rangle|^2 \text{ and } k^2 \|u_i\|^2_{L^2(\Omega)} = \sum_{j=1}^{i} \left(\frac{k}{k^2 - \lambda_j} \right)^2 |\langle f, \phi_j \rangle|^2.$$

Hence, we have

$$|||u_i|||_{1,k,\Omega} \leq C_1 \|f\|, \tag{4.2.15}$$

where $C_1 > 0$ is a constant dependent on

$$\max\{\frac{k + \lambda_j}{k^2 - \lambda_j}, j = 1, \cdots, i\} \leq \frac{k + k^2}{k^2(1 - \lambda_i/k^2)} \leq \frac{2}{1 - \lambda_i/k^2}$$

as $k \geq 1$ and ϕ_j are orthogonal to each other, and we have used the Bessel inequality $\sum_{j=1}^{i} |\langle f, \phi_i \rangle|^2 = \|f_i\|^2 \leq \|f\|^2$. For convenience, let $C_1 = \dfrac{2}{1 - \lambda_i/k^2}$ which will be referred a few times later.

Since v is a weak solution satisfying (4.2.12) in its weak formulation, we have

$$B(v, v) = \langle f - f_i, v \rangle + \langle g - \mathbf{n} \cdot \nabla u_i, v \rangle_\Gamma.$$

The right-hand side of the above equality can be bounded as follows: letting $\hat{g} = g - \mathbf{n} \cdot \nabla u_i$,

$$\begin{aligned}
&|\langle f - f_i, v \rangle| + |\langle \hat{g}, v \rangle| \leq \|f - f_i\| \|v\| + \|\hat{g}\|_\Gamma \|v\|_\Gamma \\
\leq & \frac{1}{2\epsilon k^2} \|f - f_i\|^2 + \frac{\epsilon}{2} k^2 \|v\|^2 + \frac{1}{2\epsilon k} \|\hat{g}\|^2_\Gamma + \frac{\epsilon}{2} k \|v\|^2_\Gamma \\
\leq & \frac{1}{2\epsilon k^2} \|f - f_i\|^2 + \frac{\epsilon}{2} |||v|||^2_{1,k,\Omega} + \frac{1}{2\epsilon k} \|\hat{g}\|^2_\Gamma + \frac{\epsilon}{2} C_\Omega k \|v\|_{L^2(\Omega)} \cdot \|\nabla v\|_{L^2(\Omega)} \\
\leq & \frac{1}{2\epsilon k^2} \|f - f_i\|^2 + \frac{1}{2\epsilon k} \|\hat{g}\|^2_\Gamma + \epsilon_1 |||v|||^2_{1,k,\Omega}
\end{aligned}$$

for $\epsilon > 0$ with $\epsilon_1 = \epsilon/2 + C_\Omega \epsilon/2$, where we have used the Sobolev trace theorem (cf. Lemma 1.5.1.9 in [?]). Now we use the lower bound in (4.2.9) to have the inequality in (4.2.16) by choosing $\epsilon_1 = m/2$ and $\|f_i\| \le \|f\|$ by the Bessel inequality.

$$|\!|\!| v |\!|\!|_{1,k,\Omega} \le \frac{C}{k}\|f\| + \frac{C}{\sqrt{k}}\|\hat{g}\|_\Gamma \tag{4.2.16}$$

for $k \ge 1$.

Next $\|\hat{g}\|_\Gamma^2 \le 2\|g\|_\Gamma^2 + 2\|\nabla u_i\|_\Gamma^2$ and although $u_i = 0$ over Γ, we have to estimate ∇u_i over Γ. Let us first use Sobolev trace inequality to have

$$\|\nabla u_i\|_\Gamma^2 \le C_\Omega \|\nabla u_i\|_{L^2(\Omega)} |\nabla u_i|_{1,2,\Omega} = C_\Omega \|\nabla u_i\|_{L^2(\Omega)} |u_i|_{2,2,\Omega} \tag{4.2.17}$$

for a positive constant C_Ω dependent on Ω, where $|\cdot|_{\ell,2,\Omega}$ is the ℓth semi-norm for $H^\ell(\Omega)$ for $\ell = 1, 2$. As estimated above, $\|\nabla u_i\|_{L^2(\Omega)} \le |\!|\!| u_i |\!|\!|_{1,k,\Omega} \le C_1\|f\|$. So let us concentrate on an estimate for $|u_i|_{2,2,\Omega}$. When Ω has $C^{1,1}$ smooth boundary or Ω is convex, we know that each eigenfunction ϕ_j is in $H^2(\Omega)$ and $|\phi_j|_{2,2,\Omega} \le C_\Omega\|\Delta\phi_j\| = C_\Omega \lambda_j \|\phi_j\|$ for a positive constant C_Ω dependent only on Ω. For simplicity, we write $u_i = \sum_{j=1}^i c_j\phi_j$ to have

$$\begin{aligned} |u_i|_{2,2,\Omega} &\le \sum_{j=1}^i |c_j||\phi_j|_{2,2,\Omega} \le C_\Omega \sum_{j=1}^i |c_j| \|\Delta\phi_j\|_{L^2(\Omega)} \\ &\le C_\Omega \sum_{j=1}^i |c_j|\lambda_j \|\phi_j\|_{L^2(\Omega)} = C_\Omega \sum_{j=1}^i |c_j|\lambda_j. \end{aligned} \tag{4.2.18}$$

As above, $c_j = \dfrac{1}{k^2 - \lambda_j}\langle f, \phi_j\rangle$ and thus,

$$\sum_{j=1}^i |c_j|\lambda_j \le \frac{1}{1 - \lambda_i/k^2} \sum_{j=1}^i \frac{\lambda_j}{k^2}|\langle f, \phi_j\rangle| \le \frac{1}{1-\lambda_i/k^2}\|f_i\| \left(\sum_{j=1}^i \frac{\lambda_j^2}{k^4}\right)^{1/2}.$$

Let $C_2 = \sqrt{\sum_{j=1}^i \lambda_j^2/k^4}$ which can be estimated by using the so-called Weyl law on the number of Dirichlet eigenvalues over polygonal domain. Indeed, let $N(a)$ be the number of eigenvalues counting the multiplicities less or equal to $a > 0$. The Weyl law says that

$$N(a) = \frac{A_\Omega}{4\pi} a + O(\sqrt{a}) \tag{4.2.19}$$

(cf. e.g. [**?**]), where A_Ω stands for the area of Ω. Then $C_2^2 = \frac{1}{k^4}\sum_{j=1}^{i} \lambda_j^2 \leq \frac{1}{k^4}\lambda_i^2 N(k^2) = B\lambda_i^2/k^2 \leq Bk^2$ for another positive constant B. That is, $C_2 \leq \sqrt{B}k$. Hence, we have

$$|u_i|_{2,2,\Omega} \leq C_1\sqrt{B}k\|f_i\| \leq C_1\sqrt{B}k\|f\| \tag{4.2.20}$$

and together with (4.2.15), the terms on the right-hand side of (4.2.17) can be simplified to be

$$\|\nabla u_i\|_\Gamma^2 \leq C_\Omega^2 C_1\|f\| C_1\sqrt{B}k\|f_i\| \leq C_\Omega^2 C_1^2\sqrt{B}\|f\|^2 k \tag{4.2.21}$$

and hence from (4.2.16),

$$|||v|||_{1,k,\Omega} \leq \frac{C}{k}\|f\| + \frac{C}{\sqrt{k}}\|g\|_\Gamma + C_1 C_\Omega B^{1/4}\|f\|. \tag{4.2.22}$$

Therefore, we summarize the discussion above to have

$$\begin{aligned} |||u|||_{1,k,\Omega} \leq |||v|||_{1,k,\Omega} + |||u_i|||_{1,k,\Omega} &\leq \frac{C}{k}\|f\| + \frac{C}{\sqrt{k}}\|g\|_\Gamma + C_\Omega C_1\|f\|B^{1/4} \\ &= C_3(\|f\| + \frac{1}{\sqrt{k}}\|g\|_\Gamma) \end{aligned}$$

for a positive constant C_3 dependent on $2/(1-\lambda_i/k^2)$ and the lower bound L.

Finally, to establish (4.2.14) we follow the standard approach and apply the formula in Chapter 3, [**?**] to v. That is, for any $u \in H^2(\Omega)$, we use $\mathbf{v} = \nabla u$ in Theorem

3.1.1.1. in [?] to have

$$\sum_{i,j=1}^{2}\int_{\Omega}(\partial_{ij}u)^2 = \int_{\Omega}(\Delta u)^2 d\mathbf{x} + 2\int_{\partial\Omega}\nabla_T u\cdot\nabla_T(\nabla u\cdot\mathbf{n})d\sigma + \int_{\partial\Omega}\left[\mathcal{B}(\nabla_T u,\nabla_T u) + \mathrm{tr}(\mathcal{B})(\nabla u\cdot\mathbf{n})^2\right]d\sigma, \qquad (4.2.23)$$

where T and $\mathbf{n}$ stand for the tangential and normal direction of Γ, $\mathcal{B}$ is the bilinear form, i.e. the Hessian matrix and tr is the trace operator. Due to the convexity, the last two terms involving the Hessian of the boundary Γ are negative. For our solution v, the first term on the right-hand side above can be estimated as follows: by using the Helmholtz equation,

$$\begin{aligned}\int_{\Omega}|\Delta v|^2 d\mathbf{x} = \int_{\Omega}|f-u_i-k^2v|^2 d\mathbf{x} &\le 2\|f-u_i\|^2 + 2k^4\|v\|^2\\ &\le C(\|f\|^2+\|u_i\|^2) + 2k^2|\!|\!|v|\!|\!|^2_{1,k,\Omega}\\ &\le C(\|f\|^2+\|f\|^2/k^2) + 2k^2(\|f\|^2/k^2+\|g\|^2_{\Gamma}/k+\sqrt{B}\|f\|^2)\\ &\le Ck^2(\|f\|^2+\|g\|^2_{\Gamma})\end{aligned}$$

for a positive constant C, where we have used (4.2.15) and (4.2.16). Next, by using the Robin boundary condition, the second term on the right-hand side of (4.2.23) is estimated as follows:

$$|\int_{\partial\Omega}\nabla_T v\cdot\nabla_T(\nabla v\cdot\mathbf{n})d\sigma| \le \|\nabla_T v\|^2_{\Gamma} + |\int_{\Gamma}\nabla_T v\nabla_T g d\sigma| \le \frac{3}{2}\|\nabla v\|^2_{\Gamma} + \frac{1}{2}\|\nabla_T g\|^2_{\Gamma}.$$

Furthermore, by using Sobolev trace inequality, the first term above on the right-hand side can be estimated by

$$\|\nabla v\|^2_{\Gamma} \le C_{\Omega}\|\nabla v\|^2 + \frac{1}{2}|v|^2_{2,2,\Omega} \le C_{\Omega}|\!|\!|v|\!|\!|^2_{1,k,\Omega} + \frac{1}{2}|v|^2_{2,2,\Omega}.$$

Therefore, it follows from (4.2.23) that

$$\frac{1}{2}|v|^2_{2,2,\Omega} \le Ck^2(\|f\|^2 + \|g\|^2_\Gamma) + \frac{3C_\Omega}{2}\|\!|v|\!\|^2_{1,k,\Omega} + \frac{1}{2}\|\nabla g\|^2_\Gamma$$

Together with (4.2.20) and (4.2.22), we have obtained (4.2.14). □

Note that there are two different stability conditions in Theorem 4.2.5, mainly two different stability constants: one is dependent on $1/(1 - \lambda_i/k^2)$ as well as L and the other is dependent on $(1 + k)$. It is interesting to know if the lower bound L in (4.2.9) is dependent on k or not. To this end, we decompose a weak solution u into three parts: $u = u_i + v_i + w$ with $u_i \in Y_i, v_i = Y_i^\perp$ and $w \in (\mathbb{H}^1_0(\Omega))^\perp$. Let us begin with the following

Lemma 4.2.3. *There exists a positive constant L_1 such that*

$$|B(u,u)| \ge L_1\|\!|u|\!\|^2_{1,k,\Omega}, \quad \forall u \in (H^1_0(\Omega))^\perp. \tag{4.2.24}$$

Proof. Suppose that we do not have $L_1 > 0$ for (4.2.24). For each $n > 1$, we have $u_n \in (H^1_0(\Omega))^\perp$ with $\|\!|u_n|\!\|_{1,k,\Omega} = 1$ such that $|B(u_n,u_n)| \le 1/n$. First of all, the boundedness of u_k in $\mathbb{H}^1(\Omega)$ implies that there is a function $u^* \in \mathbb{H}^1(\Omega)$ and a convergent subsequence, say the whole sequence which converges weakly to u^* in $\mathbb{H}^1(\Omega)$ and $\|\!|u_n|\!\|_{1,k,\Omega} \to \|\!|u^*|\!\|_{1,k,\Omega}$. By Rellich-Kontrachov's theorem, without loss of generality, let us say $u_k \to u^*$ in $\mathbb{L}^2(\Omega)$ strongly. It follows that $|B(u^*,u^*)| = 0$. Thus, $\langle u^*, u^*\rangle_\Gamma = 0$, i.e. $u^* \in H^1_0(\Omega)$. However, $u_n \in (H^1_0(\Omega))^\perp$ implies that $u^* \in (H^1_0(\Omega))^\perp$. That is, $u^* \in H^1_0(\Omega) \cap (H^1_0(\Omega))^\perp = \{0\}$ which contradicts to the fact $\|\!|u^*|\!\|_{1,k,\Omega} = 1$. Therefore, we have $L_1 > 0$ for (4.2.24). □

Lemma 4.2.4. *Suppose that k^2 is not a Dirichlet eigenvalue of $-\Delta$ over Ω. Let us say $\lambda_i < k^2 < \lambda_{i+1}$ for some $i \ge 0$. Then there exists a positive constant $L_2 > 0$ such* *that*

$$|B(u,u)| \geq L_2 |\!|\!| u |\!|\!|^2_{1,k,\Omega}, \quad \forall u \in Y_i. \tag{4.2.25}$$

Proof. To prove (4.2.25), we assume otherwise. There exists a nonzero $u^* \in Y_i$ such that $B(u^*, u^*) = 0$. It follows that $\|\nabla u^*\|^2 = k^2 \|u^*\|^2$. Let us write $u^* = \sum_{j=1}^i c_j \phi_j \in Y_i$. Then we have $\|u^*\|^2 = \sum_{j=1}^i |c_j|^2$ by using the orthonormality of ϕ_j's and similarly, $\|\nabla u^*\|^2 = \sum_{j=1}^i |c_j|^2 \lambda_j$. Since $\lambda_j < k^2$ for $j = 1, \cdots, i$, we have $\|\nabla u^*\|^2 < k^2 \|u^*\|^2$ which is a contradiction to the eigenvalue property: $k^2\|u^*\|^2 = \|\nabla u^*\|^2$.

In fact, L_2 can be found as follows. For any $u = \displaystyle\sum_{j=1}^{i} c_j \phi_j \in Y_i$, we have

$$\begin{aligned} |B(u,u)| &= |\|\nabla u\|^2 - k^2\|u\|^2| = \sum_{j=1}^{i} |c_j|^2 (k^2 - \lambda_j) \geq \frac{k^2 - \lambda_i}{k^2 + \lambda_i} \sum_{j=1}^{i} (k^2 + \lambda_j)|c_j|^2 \\ &= L_2(\|\nabla u\|^2 + k^2\|u\|^2) \end{aligned}$$

with $L_2 = \dfrac{k^2 - \lambda_i}{k^2 + \lambda_i} = \dfrac{1 - \lambda_i/k^2}{1 + \lambda_i/k^2}$. □

Finally, we have

Lemma 4.2.5. *Suppose that k^2 is not a Dirichlet eigenvalue of $-\Delta$ over Ω. Let us say $\lambda_i < k^2 < \lambda_{i+1}$ for some $i \geq 0$. Then there exists a positive constant $L_3 > 0$ such that*

$$B(u,u) \geq L_3 |\!|\!| u |\!|\!|^2_{1,k,\Omega}, \quad \forall u \in Y_i^\perp. \tag{4.2.26}$$

Proof. For $u \in Y_i^\perp$, we have

$$\begin{aligned} |\!|\!| u |\!|\!|^2_{1,k,\Omega} &= B(u,u) + 2k^2 \|u\|^2_{L^2(\Omega)} \leq B(u,u) + 2k^2 \|u\|_{L^2(\Omega)} \|\nabla u\|_{L^2(\Omega} \frac{\|w\|_{L^2(\Omega)}}{\|\nabla u\|_{L^2(\Omega)}} \\ &\leq B(u,u) + k |\!|\!| u |\!|\!|^2_{1,k,\Omega} \frac{1}{\sqrt{\lambda_{i+1}}} \end{aligned}$$

by using the Cauchy-Schwarz inequality and the Rayleigh-Ritz approximation of the eigenvalues. It follows that

$$(1-\frac{k}{\sqrt{\lambda_{i+1}}})|||u|||^2_{1,k,\Omega} \leq B(u,u) \tag{4.2.27}$$

with $L_3 = 1 - \frac{k}{\sqrt{\lambda_{i+1}}} > 0$. □

4.3 On Spline Weak Solution to Helmholtz Equation

In this section, we mainly explain bivariate spline spaces which will be useful in the study later. We refer to [23] and [3] for detail. Given a polygonal region Ω, a collection $\triangle := \{T_1, ..., T_n\}$ of triangles is an ordinary triangulation of Ω if $\Omega = \cup_{i=1}^n T_i$ and if any two triangles T_i, T_j intersect at most at a common vertex or a common edge. We also assume that triangulation $\triangle$ is quasi-uniform, that is, there exists a positive constant $\gamma > 0$ such that

$$\sup_{T\in\triangle} \frac{|T|}{\rho_T} \leq \gamma < \infty \tag{4.3.1}$$

where $|T|$ stands for the minimal diameter of the circle containing triangle T and ρ_T the largest radius of the circle contained inside T. E.g. a triangulation $\triangle$ which is the nth uniform refinement of a fixed triangulation $\triangle_0$ of Ω is quasi-uniform. Also, $|\triangle|$ is the largest of diameters of triangles $T \in \triangle$. For $r \geq 0$ and $d > r$, let

$$S^r_p(\triangle) = \{s \in C^r(\triangle) : s|_T \in \mathbb{P}_p, \forall T \in \triangle\} \tag{4.3.2}$$

be the spline space of degree p and smoothness $r \geq 0$ over triangulation $\triangle$.

As solutions to the Helmholtz equation will be a complex solution, let us use a complex spline space in this paper defined by

$$\mathbb{S}^r_p(\triangle) = \{s = s_r + \mathbf{i}s_i, s_i, s_r \in S^r_p(\triangle)\}. \tag{4.3.3}$$

A spline solution $u_\triangle \in \mathbb{S}_p^r(\triangle)$ with $r \geq 1$ is a weak solution of (4.1.1) if $u_\triangle \in \mathbb{S}_p^r(\triangle)$ satisfies

$$a(u_\triangle, v) - k^2\langle u_\triangle, v\rangle + \mathbf{i}k\langle u_\triangle, v\rangle_{\partial\Omega} = \langle f, v\rangle_\Omega + \langle g, v\rangle_\Gamma, \quad \forall v \in \mathbb{S}_p^r(\triangle) \tag{4.3.4}$$

which consists with a standard finite element formulation for $r \geq 0$.

The spline space $\mathbb{S}_p^r(\triangle)$ has the similar approximation properties as the standard real-valued spline space $S_p^r(\triangle)$. The following theorem can be established by the same constructional techniques (cf. [**?**] or [23] for spline space $S_p^r(\triangle)$ for real valued functions):

Theorem 4.3.1. *Suppose that $\triangle$ is a γ-quasi-uniform triangulation of polygonal domain Ω. Let $p \geq 3r+2$ be the degree of spline space $\mathbb{S}_p^r(\triangle)$. For every $u \in \mathbb{H}^{m+1}(\Omega)$, there exists a quasi-interpolatory spline function $Q_p(u) \in \mathbb{S}_p^r(\triangle)$ such that*

$$\sum_{T\in\triangle} \|D_x^\alpha D_y^\beta(u - Q_p(u))\|_{2,T}^2 \leq K_5|\triangle|^{2(m+1-s)}|u|_{2,m+1,\Omega}^2 \tag{4.3.5}$$

for $\alpha + \beta = s, 0 \leq s \leq m+1$, where $0 \leq m \leq p$, K_5 is a positive constant dependent only on γ, Ω, and p.

We can show the existence and uniqueness of spline weak solution.

Theorem 4.3.2. *Let Ω be a polygonal domain and $\triangle$ be a triangulation of Ω. Let $\mathbb{S}_d^r(\triangle)$ with $d \geq 3r+2$ be a complex-valued spline space of degree d and smoothness 1 over $\triangle$. Then the spline weak solution to (4.1.1), i.e. satisfying (4.3.4) and Robin boundary condition exists and is unique.*

Proof. Let us consider a spline solution $u \in \mathbb{S}_p^1(\triangle) \subset \mathbb{H}^1(\Omega)$ which satisfies the weak formulation (4.3.4) with $r = 1$ for all $v \in \mathbb{S}_p^1(\triangle)$. Indeed, since $v \in \mathbb{S}_p^1(\triangle)$, i.e., v is continuously differentiable over Ω. In particular, the inward normal derivative $-\mathbf{n}\cdot\nabla v$ is well defined along the boundary of Ω which will be converted to the desired outward normal derivative in the obvious way. Then it leads to a system of linear equations due to the finite dimensionality of $\mathbb{S}_p^1(\triangle)$. To see the linear system of equations has a unique solution, we need to show that the solution u has to be zero if the right-hand side is zero, i.e., $f = 0 = g$. That is, we need to show that the solution $u \in \mathbb{S}_p^1(\triangle)$ satisfying the following

$$\int_\Omega \nabla u \cdot \nabla \overline{v} dxdy - k^2 \int_\Omega u \overline{v} dxdy + \mathbf{i}k \int_\Gamma u\overline{v} d\Gamma = 0, \quad \forall v \in \mathbb{S}_p^1(\triangle) \tag{4.3.6}$$

has to be zero. Let $v = u$ in the above equation to have

$$\int_\Omega |\nabla u|^2 dxdy - k^2 \int_\Omega |u|^2 dxdy + \mathbf{i}k \int_\Gamma |u|^2 d\Gamma = 0.$$

We conclude that $\int_\Gamma |u|^2 d\Gamma = 0$ and hence, $u \equiv 0$ on $\Gamma = \partial\Omega$. Hence, it follows from (4.3.6) that

$$\int_\Omega \nabla u \cdot \nabla \overline{v} dxdy - k^2 \int_\Omega u \overline{v} dxdy = 0, \quad \forall v \in \mathbb{S}_p^1(\triangle). \tag{4.3.7}$$

That is, if $u \neq 0$, u is an eigenfunction in $\mathbb{S}_p^1(\triangle)$ corresponding to eigenvalue k^2.

Furthermore, $\mathbf{n} \cdot \nabla u \equiv 0$ along Γ by the Robin boundary condition due to $g \equiv 0$ and $u \equiv 0$ on Γ. Without loss of generality, we may assume that Ω contains 0. Let $\alpha \in (0, 1)$ and $\Omega \subset \Omega_\alpha$ as in Lemma 4.3.1. In addition, let $\triangle_\alpha$ be a triangulation of Ω_α by adding triangles to the existing $\triangle$. Then the zero boundary conditions of u enable us to extend u outside of Ω by zero and hence, $u \in \mathbb{S}_p^1(\triangle_\alpha)$ because both $u \equiv 0$ and $\mathbf{n} \cdot \nabla u \equiv 0$ along Γ . Hence, u is also an eigenfunction in $\mathbb{S}_p^1(\triangle_\alpha)$ with eigenvalue k^2. By Lemma 4.3.1, $k^2 = \alpha^2 \lambda_i$ for some $\lambda_i \in \Lambda_1$, the collection of all eigenvalues of Laplacian operator over spline space $\mathbb{S}_p^1(\triangle)$. Since Λ_1 has finitely many eigenvalues; however, $\alpha \in (0, 1)$ can be infinitely many and we know that different α implies different λ_i. This is a contradiction. Therefore, we conclude that $u \equiv 0$ and hence, there exists a unique solution to the spline weak equation (4.3.4). □

In the proof above, we have used the following

Lemma 4.3.1. *Let $\Omega \subset \mathbb{R}^2$ be a domain with Lipschitz boundary. Without loss of generality, we may assume $0 \in \Omega$. For each $\alpha \in (0,1)$, we let $\Omega_\alpha = \{(x,y) : (\alpha x, \alpha y) \in \Omega\}$. Let*

$$\Lambda_1 = \{0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq \cdots\} \tag{4.3.8}$$

be the collection of all eigenvalues of Laplace operator $-\Delta$ over Ω. Similarly, let

$$\Lambda_\alpha = \{0 < \lambda_1(\alpha) \leq \lambda_2(\alpha) \leq \cdots \leq \lambda_n(\alpha) \leq \cdots\} \tag{4.3.9}$$

be the collection of all eigenvalues of $-\Delta$ on Ω_α. Then each eigenvalue $\lambda_i(\alpha) \in \Lambda_\alpha$ is equal to

$$\lambda_i(\alpha) = \alpha^2 \lambda_i, i = 1, 2, \cdots, n, \cdots . \tag{4.3.10}$$

Proof. For any function $u \in H_0^1(\Omega)$, let $u_\alpha(x,y) = u(\alpha x, \alpha y)$ which is a function in $H_0^1(\Omega_\alpha)$. If u is an eigenfunction of $-\Delta$ over Ω with eigenvalue $\lambda \in \Lambda_1$, we have $-\Delta u = \lambda u$. Thus,

$$-\Delta u_\alpha(x,y) = -\alpha^2 \Delta u(\alpha x, \alpha y) = \alpha^2 \lambda u(\alpha x . \alpha y) = \alpha^2 u_\alpha(x,y).$$

Thus, $\alpha^2 \lambda \in \Lambda_\alpha$ with eigenfunction u_α. Similarly, we can show each eigenvalue $\lambda_\alpha \in \Lambda_\alpha$, $\lambda_\alpha/\alpha^2 \in \Lambda_1$. This completes the proof. □

Next we need to show that the spline weak solution $u_\triangle$ are bounded independent of $\triangle$. Following the proof of Theorem 4.2.3 in the previous section, we have

Theorem 4.3.3. *Let Ω be a convex domain with Lipschitz boundary satisfying $\lambda_i < k^2 < \lambda_{i+1}$. Let $u_\triangle \in \mathbb{S}_p^1(\triangle)$ be the spline weak solution satisfying (4.3.4) and suppose* *that $u_\triangle \in X_i^\perp \cap \mathbb{S}_p^1(\triangle)$. Then there exists a constant M independent of $\triangle$ such that*

$$|||u_\triangle|||_{1,k,\Omega} \leq M(\|f\| + \|g\|_\Gamma). \tag{4.3.11}$$

Proof. We simply use the proof of Theorem 4.2.3 to have

$$L|||u_\triangle|||_{1,k,\Omega}^2 \leq |B(u_\triangle, u_\triangle)|.$$

Since $B(u_\triangle, u_\triangle) = \langle f, u_\triangle\rangle + \langle g, u_\triangle\rangle_\Gamma$, we use Cauchy-Schwarz inequality to obtain

$$\begin{aligned}|B(u_\triangle, u_\triangle)| &\leq \frac{1}{2k\epsilon}\|f\|^2 + \frac{\epsilon}{2}|||u_\triangle|||_{1,k,\Omega}^2 + \frac{1}{2k\epsilon}\|g\|_\Gamma^2 + \frac{\epsilon}{2}k\|u_\triangle\|^2 \\ &\leq \frac{1}{2k\epsilon}\|f\|^2 + \frac{1}{2k\epsilon}\|g\|_\Gamma^2 + (\frac{\epsilon}{2} + \frac{C_\Omega\epsilon}{2})|||u_\triangle|||_{1,k,\Omega}^2.\end{aligned}$$

By choosing $\epsilon > 0$ small enough, e.g., $(\frac{\epsilon}{2} + \frac{C_\Omega\epsilon}{2}) \leq L/2$, we have (4.3.11) with $M = 2/L$. □

Similarly, we can show that $u_\triangle$ is bounded when $u_\triangle \in \mathbb{S}_p^1(\triangle) \cap Y_i$ by using Lemma 4.2.4. We leave it to the interested reader. Following the same arguments of Theorem 4.2.5, we can construct a spline solution $u_\triangle$ based on spline approximations of eigenfunctions ϕ_i's and then due to the C^1 smoothness of the spline solution $u_\triangle$, we can prove the similar result to that of Theorem 4.2.5. That is, we have

Theorem 4.3.4. *Suppose that Ω has a $C^{1,1}$ smooth boundary or Ω is convex. Suppose that k^2 is not a Dirichlet eigenvalue of the Laplacian operator over Ω. Let us say $\lambda_i < k^2 < \lambda_{i+1}$ for some $i \geq 0$. Let $u_\triangle \in \mathbb{S}_p^1(\triangle) \cap \mathbb{H}^1(\Omega)$ be a spline weak solution to (4.1.1) according to the construction in the proof of Theorem 4.2.5. Then there exists a constant $C > 0$ independent of f, g such that*

$$|||u_\triangle|||_{1,k,\Omega} \leq C(\|f\| + \|g\|_\Gamma) \tag{4.3.12}$$

for $k \geq 1$, where C is dependent on $\dfrac{1}{1-\lambda_i/k^2}$ and the constant L which is the low bound in (4.2.9). Furthermore, suppose Ω is convex and $g \in \mathbb{H}^{3/2}(\Gamma)$. Then

$$|u_\triangle|_{2,2,\Omega} \leq C(1+k)\left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)}\right) + \|\nabla_T g\|_{L^2(\Gamma)}. \tag{4.3.13}$$

where ∇_T stands for the tangential derivative on Γ.

Proof. For convenience, let us give an outline of proof. Let $\phi_{j,\triangle} \in S^1_p(\triangle)$ be the spline approximation of ϕ_j, $j = 1, \cdots, i$ and $\lambda_{j,\triangle}$ be the numerical approximation of λ_j. It is known that $\lambda_{j,\triangle}$ approximates λ_j very well for $j \leq i$ when $|\triangle| \to 0$. We project the right-hand side f to $Y_i \cap S^1_p(\triangle)$ to have

$$f_{i,\triangle} = \sum_{j=0}^{i} \langle f, \phi_{j,\triangle}\rangle \phi_{j,\triangle}.$$

Let $u_{i,\triangle} = -\sum_{j=0}^{i} \dfrac{\langle f, \phi_{j,\triangle}\rangle}{-\lambda_{j,\triangle}+k^2}\phi_{j,\triangle}$. It is easy to see $-(\Delta u_i + k^2 u_i) = f_i$. Let $v_{i,\triangle}$ be the weak solution in $\mathbb{S}^1_p(\triangle)$ satisfying (4.2.12) with f_i and u_i replaced by $f_{i,\triangle}$ and $u_{i,\triangle}$, respectively.

Let us write $u_\triangle = u_{i,\triangle} + v_{i,\triangle}$. Then for any $w \in \mathbb{S}^1_p(\triangle)$,

$$\begin{aligned}
&B(u_\triangle, w) = B(u_{i,\triangle}, w) + B(v_{i,\triangle}, w)\\
&= \langle \nabla u_{i,\triangle}, \nabla w\rangle - k^2\langle u_{i,\triangle}, w\rangle + B(v_{i,\triangle}, w)\\
&= -\langle \Delta u_{i,\triangle} + k^2 u_{i,\triangle}, w\rangle + \langle \mathbf{n}\cdot\nabla u_{i,\triangle}, w\rangle_\Gamma + \langle f - f_{i,\triangle}, w\rangle + \langle g - \mathbf{n}\cdot\nabla u_{i,\triangle}, w\rangle_\Gamma\\
&= \sum_{j=0}^{i} \frac{\langle f, \phi_{j,\triangle}\rangle}{-\lambda_{j,\triangle}+k^2}\langle(-\lambda_{j,\triangle}+k^2)\phi_{j,\triangle}, w\rangle + \langle f - f_i, w\rangle + \langle g, w\rangle_\Gamma = \langle f, w\rangle + \langle g, w\rangle_\Gamma.
\end{aligned}$$

That is, $u_\triangle \in \mathbb{S}^1_p(\triangle)$ is the weak solution.

Now we can use the proof of Theorem 4.2.5 to conclude (4.3.12). In the same fashion of the proof of Theorem 4.2.5, we can establish (4.3.13). The detail is left to the interested reader. □

4.4 Convergence of Spline Weak Solutions

In this section, we first use the coercivity in Theorem 4.2.3 to establish

Lemma 4.4.1. *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary. Let u be the unique weak solution in $\mathbb{H}^1(\Omega)$ satisfying (4.2.1) and $u_\triangle \in \mathbb{S}^r_p(\triangle), p \geq 3r+2$ be the spline weak solution to (4.1.1) satisfying (4.3.4). Suppose that $u \in (H^1_0(\Omega))^\perp$ and $u_\triangle \in (H^1_0(\Omega))^\perp \cap \mathbb{S}^1_p(\triangle)$. Then if $u \in \mathbb{H}^s(\Omega)$ with $1 \leq s \leq p$, there exists $C > 0$ independent of k such that*

$$|||u - u_\triangle|||_{1,k,\Omega} \leq C(1 + k|\triangle|)|\triangle|^{s-1}|u|_{s,2,\Omega}, \tag{4.4.1}$$

where $|u|_{s,2,\Omega}$ is the semi-norm in $\mathbb{H}^s(\Omega)$.

Proof. We use Lemma 4.2.3 to have

$$L_1 |||u - u_\triangle|||^2_{1,k,\Omega} \leq |B(u - u_\triangle, u - u_\triangle)|.$$

It follows from (4.2.1) and (4.3.4) the orthogonality condition:

$$a(u - u_\triangle, w) - k^2\langle u - u_\triangle, w\rangle + \mathbf{i}\langle u - u_\triangle, w\rangle_{\partial\Omega} = 0, \quad \forall w \in \mathbb{S}^r_p(\triangle). \tag{4.4.2}$$

That is, $B(u - u_\triangle, w) = 0$ for all $w \in \mathbb{S}^r_p(\triangle)$. By choosing $w = Q_p(u)$, the quasi-interpolatory spline of u as in the previous section, we have

$$|B(u - u_\triangle, u - u_\triangle) = |B(u - u_\triangle, u - Q_p(u)\rangle)| \leq C_B |||u - u_\triangle|||_{1,k,\Omega} |||u - Q_p(u)|||_{1,k,\Omega}.$$

In other words,

$$|||u - u_\triangle|||_{1,k,\Omega} \leq \frac{C_B}{L_1} |||u - Q_p(u)|||_{1,k,\Omega}. \tag{4.4.3}$$

Finally, we use the approximation property of spline space $\mathbb{S}_p^r(\triangle)$, i.e. (4.3.5). For $u \in \mathbb{H}^s(\Omega)$ with $1 \leq s \leq p$, we use the quasi-interpolatory operator $Q_p(u)$ of u to have

$$|||u - Q_p(u)|||_{1,k,\Omega} \leq C(1 + k|\triangle|)|\triangle|^{s-1}|u|_{s,2,\Omega}$$

for a constant C dependent on Ω, p and the smallest angle of $\triangle$ only. Therefore, the combination of (4.4.3) and the estimate above yields (4.4.1). □

Similarly, if $u \in H_0^1(\Omega)$, we can find spline approximation $u_\triangle$ satisfying (4.3.4) for $v \in \mathbb{S}_p^1(\triangle) \cap H_0^1(\Omega)$. Using Lemma 4.2.5, we have

Lemma 4.4.2. *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary. Suppose that Ω is a domain such that k^2 is not a Dirichlet eigenvalue of the Laplacian over Ω, say $\lambda_i < k^2 < \lambda_{i+1}$ for some $i \geq 0$. Let u be the unique weak solution in $H^1(\Omega)$ satisfying (4.2.1) and $u_\triangle \in \mathbb{S}_p^r(\triangle), p \geq 3r + 2$ be the spline weak solution to (4.1.1) satisfying (4.3.4). Suppose that $u \in Y_i^\perp$ and $u_\triangle \in Y_i^\perp \cap \mathbb{S}_p^1(\triangle)$. Then if $u \in \mathbb{H}^s(\Omega)$ with $1 \leq s < p$, there exists $C > 0$ dependent on $1 - k/\sqrt{\lambda_{i+1}}$ such that*

$$|||u - u_\triangle|||_{1,k,\Omega} \leq C(1 + k|\triangle|)|\triangle|^{s-1}|u|_{s,2,\Omega}. \tag{4.4.4}$$

In general, we do not know if the solution u is in $H_0^1(\Omega)$ or in $(H_0^1(\Omega))^\perp$. However, we can check if k^2 is an eigenvalue or not.

Theorem 4.4.1. *Let $\Omega \subset \mathbb{R}^2$ be a bounded convex domain or a bounded domain with $C^{1,1}$ boundary. Suppose that Ω is a domain such that k^2 is not a Dirichlet eigenvalue of the Laplacian over Ω. Let u be the unique weak solution in $\mathbb{H}^1(\Omega)$ satisfying (4.2.1) and $u_\triangle \in \mathbb{S}_p^r(\triangle), p \geq 3r + 2$ be the spline weak solution to (4.1.1) satisfying (4.3.4). Then if $u \in \mathbb{H}^s(\Omega)$ with $1 \leq s \leq p$, there exists $C > 0$ independent of $|\triangle|$, f and g*

such that

$$|||u - u_\triangle|||_{1,k,\Omega} \le C(1 + k|\triangle|)|\triangle|^{s-1}(|u|_{s,2,\Omega} + |u_i|_{s,2,\Omega}), \tag{4.4.5}$$

where u_i is the projection of u in Y_i.

Proof. We simply decompose u to be $v + u_i$, where $u_i \in Y_i$ and $v \in X_i^\perp$. As the domain Ω is convex or has a $C^{1,1}$ boundary, the regularity theory of Poisson's equation implies that each eigenfunction ϕ_j is very smooth and so is u_i. Thus, v has the same regularity as that of u. For v, we use the coercive condition, i.e. Theorem 4.2.3 to have

$$L|||v - v_\triangle|||^2_{1,k,\Omega} \le |B(v - v_\triangle, v - v_\triangle)|,$$

where $v_\triangle$ is the spline weak solution to v. Similar to the proof of Theorem 4.4.1, we have

$$|||v - v_\triangle|||_{1,k,\Omega} \le C(1 + k|\triangle|)|\triangle|^{s-1}|v|_{s,2,\Omega} \le C(1 + k|\triangle|)|\triangle|^{s-1}(|u|_{s,2,\Omega} + |u_i|_{s,2,\Omega}) \tag{4.4.6}$$

for another positive constant C dependent on L in (4.2.9).

Next we discuss the spline approximation $u_{i,\triangle}$ of u_i. The classic theory (cf. [?] and [?]) says that letting $\phi_{j,\triangle} \in S^1_p(\triangle)$ be the spline approximation of eigenfunction ϕ_j using Rayleigh-Ritz approximation method, $\phi_{j,\triangle} \to \phi_j$ very well for each $j = 1, \cdots, i$ in the sense that for $0 \le \ell \le s$,

$$|\phi_j - \phi_{j,\triangle}|_{\ell,2,\Omega} \le C|\triangle|^{s-\ell}|\phi_j|_{s,2,\Omega} \tag{4.4.7}$$

for a positive constant C independent of $\triangle$, since the spline space $S^1_p(\triangle)$ has the desired approximation power required in the proof of (4.4.7) (cf. [?]). It follows that

$u_{i,\triangle} \to u_i$ and

$$|||u_{i,\triangle} - u_i|||_{1,k,\Omega} \leq C|\triangle|^{s-1}(1 + k|\triangle|)\|f\|. \tag{4.4.8}$$

Indeed, we recall the Weyl law on the number $N(k^2)$ of Dirichlet eigenvalues less or equal to k^2 from [?] and use the formula for u_i in (4.2.11) to have

$$\begin{aligned}\|u_i - u_{i,\triangle}\| &\leq \sum_{j=1}^{i} \frac{\|f\|}{k^2 - \lambda_j}\|\phi_j - \phi_{j,\triangle}\| \\ &\leq \frac{1}{k^2}C_1 N(k^2)C|\triangle|^s \max_{j=1,\cdots,i} |\phi_j|_{s,2,\Omega} \leq B_1|\triangle|^s \max_{j=1,\cdots,i} |\phi_j|_{s,2,\Omega}\end{aligned}$$

for a positive constant B_1 dependent on $1 - \lambda_i/k^2$. Similarly, we have

$$\begin{aligned}\|\nabla(u_i - u_{i,\triangle})\| &\leq \sum_{j=1}^{i} \frac{\|f\|}{k^2 - \lambda_j}|\phi_j - \phi_{j,\triangle}|_{1,2,\Omega} \\ &\leq \frac{1}{k^2}C_1 N(k^2)C|\triangle|^{s-1} \max_{j=1,\cdots,i} |\phi_j|_{s,2,\Omega} \leq B_1|\triangle|^{s-1} \max_{j=1,\cdots,i} |\phi_j|_{s,2,\Omega}\end{aligned}$$

which leads to (4.4.8). Combining (4.4.6) and (4.4.8) completes the proof of Theorem 4.4.1. □

Let us point out that more detail on computation of eigenvalues and eigenfunctions of $-\Delta$ by using bivariate splines can be found in [?]. Mainly we can show that $\phi_{i,\triangle}$ is a spline weak solution to the eigenfunction equation.

In addition to the lower bound we have established in Theorem 4.2.3, we can also find an estimate for the inf-sup condition. That is, we estimate the following inf-sup condition of $B(u, v)$. The following result was well-known. See, e.g. [?] for the domain which is strictly star-shaped.

Theorem 4.4.2. *Let $\Omega \subset \mathbb{R}^2$ be a bounded strictly star-shaped domain. Then there exists $C > 0$ (independent of k) such that*

$$\inf_{v\in\mathbb{H}^1(\Omega)} \sup_{u\in\mathbb{H}^1(\Omega)} \frac{Re(B(u,v))}{|||u|||_{1,k,\Omega}\,|||v|||_{1,k,\Omega}} \geq \frac{C}{1+k}. \tag{4.4.9}$$

For convenience, we explicitly write down all the detail of a proof based on a standard approach for establishing the inf-sup condition in (4.4.9). That is, let us first prove the following

Lemma 4.4.3. *For each $v \in \mathbb{H}^1(\Omega)$, there exists a $w_v \in \mathbb{H}^1(\Omega)$ such that*

$$Re(B(w_v, v)) \geq \alpha |||v|||^2_{1,k,\Omega} \text{ and } |||w_v|||_{1,k,\Omega} \leq \beta |||v|||_{1,k,\Omega} \tag{4.4.10}$$

for positive constants α and β independent of v, w_v.

Once we have the result in (4.4.10), we can establish the inf-sup condition (4.4.9). Indeed,

Proof of Theorem 4.4.2. It follows from (4.4.10) we have

$$\mathrm{Re}(B(w_v, v)) \geq \alpha |||v|||_{1,k,\Omega}\,|||w_v|||_{1,k,\Omega}/\beta$$

or

$$\sup_{u\in\mathbb{H}^1(\Omega)} \frac{\mathrm{Re}(B(u,v))}{|||u|||_{1,k,\Omega}|||v|||_{1,k,\Omega}} \geq \frac{\mathrm{Re}(B(w_v,v))}{|||w_v|||_{1,k,\Omega}\,|||v|||_{1,k,\Omega}} \geq \frac{\alpha}{\beta}.$$

Taking the inf both sides of the inequality above, we conclude the proof of (4.4.9). □

We now spend some time to prove Lemma 4.4.3.

Proof of Lemma 4.4.3. By Theorem 4.2.2, for each $v \in \mathbb{H}^1(\Omega)$, let $z_v \in \mathbb{H}^1(\Omega)$ be the solution to the Helmholtz equation (4.1.1) with $f = 2k^2 v$ and $g = 0$ satisfying

$$B(z_v, u) = 2k^2\langle v, u\rangle, \quad \forall u \in \mathbb{H}^1(\Omega).$$

We let $w_v = v + z_v \in \mathbb{H}^1(\Omega)$. To see the first inequality in (4.4.10), we have

$$\mathrm{Re}(B(w_v, v)) = \mathrm{Re}(B(v,v)) + \mathrm{Re}(B(z_v, v)) = a(v,v) - k^2\langle v, v\rangle + 2k^2\langle v, v\rangle = |||v|||_{1,k,\Omega}^2.$$

That is, the first inequality in (4.4.10) holds with $\alpha = 1$.

Next by using the stability in [?], i.e. $|||z_v|||_{1,k,\Omega} \le C2k^2\|v\|$ for a positive constant C independent of k when $k \ge 1$, we have

$$|||w_v|||_{1,k,\Omega} \le |||v|||_{1,k,\Omega} + |||z_v|||_{1,k,\Omega} \le |||v|||_{1,k,\Omega} + Ck^2\|v\| \le C(1+k)|||v|||_{1,k,\Omega}$$

which is the second inequality in (4.4.10) with $\beta = C(1+k)$. □

It is interesting to know the estimate for the inf-sup condition when domain Ω is not a strictly star-shaped domain. Using the Dirichlet eigenvalues, we can establish the following

Theorem 4.4.3. *Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain. Suppose that k^2 is not a Dirichlet eigenvalue of $-\Delta$ over Ω. Then there exists $C > 0$ such that*

$$\inf_{v\in\mathbb{H}^1(\Omega)} \sup_{u\in\mathbb{H}^1(\Omega)} \frac{Re(B(u,v))}{|||u|||_{1,k,\Omega}\,|||v|||_{1,k,\Omega}} \ge L_4. \tag{4.4.11}$$

Furthermore, L_4 does not go to zero when $k \to \infty$.

Proof. Suppose (4.4.11) does not hold. Then there exists $v_n \in \mathbb{H}^1(\Omega)$ such that $|||v_n|||_{1,k,\Omega} = 1$ and

$$\sup_{u\in\mathbb{H}^1(\Omega)} \frac{\mathrm{Re}(B(u,v_n))}{|||u|||_{1,k,\Omega}} \le \frac{1}{n}$$

for $n = 1, \cdots, \infty$. The boundedness of v_n in $\mathbb{H}^1(\Omega)$ implies that there exists a weakly convergent subsequence. By the boundedness of the weakly convergent subsequence, we can find another subsequence which is convergent strongly in L^2 norm by Rellich-Kondrachov Theorem. Without loss of generality we may assume that $v_n \to v^*$ in

$\mathbb{L}^2(\Omega)$ norm and in the semi-norm on $\mathbb{H}^1(\Omega)$ with $\|\!|v^*|\!\|_{1,k,\Omega} = 1$. It follows that for each $u \in \mathbb{H}^1(\Omega)$ with $\|\!|u|\!\|_{1,k,\Omega} = 1$, $\text{Re}(B(u, v_n)) \to 0$. Hence, $\text{Re}(B(u, v^*)) = 0$. By using $u = -\mathbf{i}v^*$, we see that $\text{Re}(B(u, v^*)) = \langle v^*, v^* \rangle_\Gamma = 0$. So $v^* = 0$ on Γ. That is, $v^* \in \mathbb{H}^1_0(\Omega)$. It follows that $\text{Re}(B(u, v^*)) = 0$ for all $u \in \mathbb{H}^1_0(\Omega)$. So v^* is an eigenfunction with eigenvalue k^2 which contradicts to the assumption. Hence, we have $L_4 > 0$ in (4.4.11).

Next let us show that $L_4 \not\to 0$ as $k \to \infty$. As L_4 is dependent on k, let us write L_k for convenience. Since the lower bound $L_k > 0$, we can find v_k with $\|\!|v_k|\!\|_{1,k,\Omega} = 1$ such that

$$\sup_{u \in \mathbb{H}^1(\Omega)} \frac{\text{Re}(B(u, v_k))}{\|\!|u|\!\|_{1,k,\Omega}} \le 2L_k. \tag{4.4.12}$$

That is, $\text{Re}(B(v_k, v_k)) \le 2L_k$. Since $\|\!|v_k|\!\|_{1,k,\Omega} = 1$, we use Rellich-Kondrachov Theorem again to conclude that there exists a $u^* \in \mathbb{H}^1(\Omega)$ such that $v_k \to u^*$ in L^2 norm and $\|\nabla v_k\| \to \|\nabla u^*\|$ without loss of generality. As $k^2\|v_k\|^2 \le 1$, i.e. $\|v_k\| \le 1/k$, we have $\|u^*\| \le 2/k$ for $k > 0$ large enough. It follows that $u^* \equiv 0$. That is, $\nabla u^* \equiv 0$ and hence, $\|\nabla v_k\| \to 0$.

If $L_k \to 0$, we use (4.4.12) have $|\|\nabla v_k\|^2 - k^2\|v_k\|^2| = |\text{Re}((B(v_k, v_k))| \to 0$. Since $\|\nabla v_k\| \to 0$ mentioned above, it follows that $k^2\|v_k\|^2 \to 0$. However, since $\|\!|v_k|\!\|_{1,k,\Omega} = 1$, we should have $k^2\|v_k\|^2 \to 1$. That is, we got a contradiction. Therefore, L_k does not go to zero when $k \to \infty$. □

Again, it is difficult to determine how the constant L_4 in (4.4.11) is dependent on k. According to the study in the above, L_4 may be dependent only on $1 - k/\sqrt{\lambda_{i+1}}$ or $1 - \lambda_i/k^2$ instead of $1/(k+1)$ in (4.4.9). Next we need one more critical estimate.

Lemma 4.4.4. *Let Ω be a bounded Lipschitz domain. Suppose that k^2 is not a Dirichlet eigenvalue of $-\Delta$ over Ω. Let u be the unique weak solution in $\mathbb{H}^1(\Omega)$ satisfying (4.2.1) and $u_\triangle \in \mathbb{S}^r_p(\triangle), p \ge 3r+2, r \ge 1$ be the spline weak solution to*

(4.1.1) satisfying (4.3.4). Then there exists a positive constant $K > 0$ such that

$$|||u - u_\triangle|||_{1,k,\Omega} \leq K|||u|||_{1,k,\Omega}, \tag{4.4.13}$$

where K is independent of u and will not go to ∞ as $k \to \infty$.

Proof. Recall that Y_i is the finite dimensional subspace of $H_0^1(\Omega)$ spanned by eigenfunctions associated with eigenvalues $\lambda_j < k^2, j = 1, \cdots, i$. $X_i^\perp$ is the orthogonal complement of Y_i in $\mathbb{H}^1(\Omega)$. We first decompose $u = u_1 + u_2$ with $u_1 \in Y_i, u_2 \in X_i^\perp$. Similarly, we write $u_\triangle = u_{1,\triangle} + u_{2,\triangle}$. Then Theorem 4.2.3 implies that there exists a positive constant L (see 4.2.24) such that

$$L|||u_2 - u_{2,\triangle}|||^2_{1,k,\Omega} \leq |B(u_2 - u_{2,\triangle}, u_2 - u_{2,\triangle})|.$$

By the orthogonality condition (4.4.2), $B(u_2 - u_{2,\triangle}, w) = 0$ for all $w \in \mathbb{S}_p^1(\triangle)$. We have

$$|B(u_2 - u_{2,\triangle}, u_2 - u_{2,\triangle})| = ||B(u_2 - u_{2,\triangle}, u_2)| \leq C_B|||u_2 - u_{2,\triangle}|||_{1,k,\Omega}\, |||u_2|||_{1,k,\Omega}.$$

Together with the estimate above of the estimate above, $L|||u_2 - u_{2,\triangle}|||_{1,k,\Omega} \leq C_B|||u_2|||_{1,k,\Omega}$.

Similarly, using Lemma 4.2.4, we can have $L_2|||u_1 - u_{1,\triangle}|||_{1,k,\Omega} \leq C_B|||u_1|||_{1,k,\Omega}$. Let us put these two estimates together to have

$$\begin{aligned} |||u - u_\triangle|||_{1,k,\Omega} \leq & |||u_1 - u_{1,\triangle}|||_{1,k,\Omega} + |||u_2 - u_{2,\triangle}|||_{1,k,\Omega} \\ \leq & C_B/L_2|||u_1|||_{1,k,\Omega} + C_B/L|||u_2|||_{1,k,\Omega}. \end{aligned} \tag{4.4.14}$$

Finally we recall that the decomposition of Y_i and X_i are based on the inner produce $\langle u, v\rangle_A = \int_\Omega \nabla u \cdot \nabla \bar{v} + k^2 \int_\Omega u\bar{v}$. It follows that

$$|||u_1|||^2_{1,k,\Omega} + |||u_2|||^2_{1,k,\Omega} = |||u_1 + u_2|||^2_{1,k,\Omega} = |||u|||^2_{1,k,\Omega}.$$

Combining the above estimate with (4.4.14), we conclude the desired result with $K = C_B\sqrt{1/L_2^2 + 1/L^2}$. □

Finally, let us establish the main result in this paper.

Theorem 4.4.4. *Let Ω be a bounded Lipschitz domain. Suppose that k^2 is not a Dirichlet eigenvalue of $-\Delta$ over Ω. Let u be the unique weak solution in $\mathbb{H}^1(\Omega)$ satisfying (4.2.1) and $u_\triangle \in \mathbb{S}^r_p(\triangle), p \geq 3r+2, r \geq 1$ be the spline weak solution to (4.1.1) satisfying (4.3.4). Then if $u \in \mathbb{H}^s(\Omega)$ with $1 \leq s \leq p$, there exists $C > 0$ such that*

$$|||u - u_\triangle|||_{1,k,\Omega} \leq C(1 + k|\triangle|)|\triangle|^{s-1}|u|_{s,2,\Omega}, \tag{4.4.15}$$

where C is dependent on $1/L_4$ which does not go to ∞ when $k \to \infty$.

If $\Omega \subset \mathbb{R}^2$ is a bounded strictly star-shaped domain and has Lipschitz boundary, then the approximation constant C in (4.4.15) can be more precisely written as $C = c(1+k)$ for a positive constant c independent k.

Proof. We simply use the inf-sup condition, i.e. Theorem 4.4.3. For each $v \in \mathbb{H}^1(\Omega)$,

$$\sup_{u \in \mathbb{H}^1(\Omega)} \frac{\mathrm{Re}(B(u,v))}{|||u|||_{1,k,\Omega}\,|||v|||_{1,k,\Omega}} \geq L_4.$$

By the continuity of the sesquilinear $B(\cdot,\cdot)$, the left-hand side is bounded above by the constant C_B. For each v, there exists a $w \in \mathbb{H}^1(\Omega)$ dependent on v which is larger than one third of the upper limit of the left-hand side above, i.e.

$$\frac{\mathrm{Re}(B(w,v))}{|||w|||_{1,k,\Omega}\,|||v|||_{1,k,\Omega}} \geq \frac{1}{3}L_4. \tag{4.4.16}$$

In particular, by choosing $v = u - u_\triangle$ in (4.4.16), we have

$$\mathrm{Re}(B(w, u - u_\triangle)) \geq \frac{L_4}{3} |||w|||_{1,k,\Omega} \, |||u - u_\triangle|||_{1,k,\Omega}.$$

for $w \in \mathbb{H}^1(\Omega)$ dependent on $u - u_\triangle$. Note that from (4.2.1) and (4.3.4), we have the orthogonality condition:

$$a(u - u_\triangle, v) - k^2 \langle u - u_\triangle, v\rangle + \mathbf{i}\langle u - u_\triangle, v\rangle_{\partial\Omega} = 0, \quad \forall v \in \mathbb{S}^1_p(\triangle). \tag{4.4.17}$$

That is, $B(u - u_\triangle, v) = 0$ for all $v \in \mathbb{S}^1_p(\triangle)$. By using $v = w_\triangle$, the spline weak solution in $\mathbb{S}^1_p(\triangle)$ to the Helmholtz equation (4.1.1) whose weak solution is w, we have

$$\mathrm{Re}(B(w - w_\triangle, u - u_\triangle)) \geq \frac{L_4}{3} |||w|||_{1,k,\Omega} \, |||u - u_\triangle|||_{1,k,\Omega}$$

By using $v = u_\triangle - Q(u)$, where $Q(u)$ is the quasi-interpolatory spline of u, we have $B(w - w_\triangle, u_\triangle - Q(u)) = 0$ and add it to the inequality above which yields

$$Re(B(w - w_\triangle, u - Q(u))) \geq \frac{L_4}{3} |||w|||_{1,k,\Omega} \, |||u - u_\triangle|||_{1,k,\Omega}.$$

It follows that

$$|||u - u_\triangle|||_{1,k,\Omega} \, |||w|||_{1,k,\Omega} \leq \frac{3}{L_4} |||u - Q_p(u)|||_{1,k,\Omega} |||w - w_\triangle|||_{1,k,\Omega}. \tag{4.4.18}$$

Since $|||w|||_{1,k,\Omega} \neq 0$ and $|||w - w_\triangle|||_{1,k,\Omega} \leq K |||w|||_{1,k,\Omega}$ for a positive constant K by Lemma 4.4.4, the inequality in (4.4.18) can be simplified to be

$$|||u - u_\triangle|||_{1,k,\Omega} \leq \frac{3K}{L_4} |||u - Q_p(u)|||_{1,k,\Omega}.$$

Finally, we use the approximation property of spline space $\mathbb{S}_p^r(\triangle)$, i.e. (4.3.5). For $u \in \mathbb{H}^s(\Omega)$ with $1 \leq s \leq p$, we use the quasi-interpolatory operator $Q_p(u)$ of u to have

$$|||u - Q_p(u)|||_{1,k,\Omega} \leq C(1 + k|\triangle|)|\triangle|^{s-1}|u|_{s,2,\Omega}$$

for another constant C dependent on Ω, p and the smallest angle of $\triangle$ only. With the term above, we can rewrite (4.4.18) as follows:

$$|||u - u_\triangle|||_{1,k,\Omega} \leq \frac{C}{L_4}(1 + k|\triangle|)|\triangle|^{s-1}|u|_{s,2,\Omega}. \tag{4.4.19}$$

for another positive constant C.

If we use Theorem 4.4.2 in the place of Theorem 4.4.3 above, we can get the estimate in (4.4.15) with a constant dependent on $c(1+k)$. These complete the proof of Theorem 4.4.4. □

From the results above, we can see that the estimate in (4.4.15) is better when C dependent on $1/L_4$ than the one with $C = c(1 + k)$ which is a traditional estimate which accounts for the pollution error in numerical experiments. Our proof of Theorem 4.4.4 removes the dependence of the constant C on k. However, we still do not know how L_4 is dependent on k although L_4 does not go to 0 when $k \to \infty$. We leave the problem to the interested reader.

4.5 Future Research Problems

Remark 4.5.1. *We have assumed Ω is convex or is a bounded domain with $C^{1,1}$ boundary. This requirement can be weakened by using the new condition called domain with positive reach as explained in [**?**]. Under the positive reach condition, the solution of Poisson equation will be in $H^2(\Omega)$. Similarly, the solution to Helmholtz equation will be in $H^2(\Omega)$. We leave the details for future study.*

Remark 4.5.2. *We are working to extend these results to the 3D setting by using trivariate splines of arbitrary degree and smoothness. Numerical results will be reported soon. See [*? *].*

Remark 4.5.3. *When extending the study in the 3D setting, a major difficulty is the approximation order of trivariate spline spaces. That is, the similar results to the bivariate spline setting in [*? *] are not available. We shall leave them to our study on trivariate splines for Helmholtz equation and Maxwell's equations.*

Remark 4.5.4. *As pointed out in several places in previous sections, the dependence of constants L and L_4 on wave number k is not clear when the domain Ω is not a strictly star-shaped domain. It is interesting to find out. The authors will continue their study on this issue.*

Remark 4.5.5. *Several estimates discussed in previous sections are dependent on whether the number k^2 is a Dirichlet eigenvalue or not. As the theory of the existence and uniqueness to Helmholtz equation (4.1.1) has no such requirement, it is interesting to remove such a condition. For example, it is also interesting to extend the stability result in Theorem 4.4.3 when k^2 is a Dirichlet eigenvalue.*

Chapter 5

Numerical Solutions of the Helmholtz Equation

In this section, we shall present our computational method and then report some numerical results. Our computational algorithm is given as follows. For spline space $\mathbb{S}_p^1(\triangle)$, let $\mathbf{c}$ be the coefficient vector associated with each spline function $s \in \mathbb{S}_p^1(\triangle)$. In the implementation explained in [3], $\mathbf{c}$ is a stack of the polynomial coefficients over each triangle in $\triangle$. Let H be the smoothness matrix such that $s \in \mathbb{S}_p^1(\triangle)$ if and only if $H\mathbf{c} = 0$. Next let $\mathbf{f}$ and $\mathbf{g}$ be the vectors of coefficients for the spline approximations for the source functions f and g, respectively. Let M and K be the mass and stiffness matrices as in [3]. Then the spline solution to the Helmholtz equation in weak form can be given in terms of these matrices as follows:

$$\overline{\mathbf{c}}^\top K\mathbf{c}_\triangle - k^2\overline{\mathbf{c}}^\top M\mathbf{c}_\triangle + \mathbf{i}\overline{\mathbf{c}}^\top M_\Gamma\mathbf{c}_\triangle = \overline{\mathbf{c}}^\top M\mathbf{f} + \overline{\mathbf{c}}^\top M_\Gamma\mathbf{g}, \quad \forall \mathbf{c} \in \mathbb{R}^N \tag{5.0.1}$$

for $\mathbf{c}$ and $\mathbf{c}_\triangle$ which satisfies $H\mathbf{c}_\triangle = 0$ and $H\mathbf{c} = 0$, where M_Γ is the mass matrix over the boundary such that $\int_\Gamma u\bar{v} = \mathbf{c}_\triangle^\top M_\Gamma\bar{\mathbf{c}}$. Note that $\overline{\mathbf{c}}$ is the standard conjugate of $\mathbf{c}$ and $N = (p+1)(p+2)N_\triangle/2$ and $N_\triangle$ is the number of triangles in $\triangle$. To solve this constrained system of linear equations, we use the so-called the constrained iterative minimization method described in [3]. That is, we solve the following constrained minimization:

$$\min_{\mathbf{c}} \frac{1}{2}(\overline{\mathbf{c}}^\top K\mathbf{c} - k^2\overline{\mathbf{c}}^\top M\mathbf{c} + \mathbf{i}\overline{\mathbf{c}}^\top M_\Gamma\mathbf{c}) - \overline{\mathbf{c}}^\top M\mathbf{f} - \overline{\mathbf{c}}^\top M_\Gamma\mathbf{g}, \tag{5.0.2}$$

subject to $H\mathbf{c} = 0$. The constrained iterative minimization method in [3] provides an efficient way to find the solution of the minimization above. In fact, we only do 3 iterations. This is a difference between our spline method and the standard high order finite element methods. It makes spline solutions very accurate for the Helmholtz problem even with high wave numbers.

Next we report numerical results based on our bivariate spline functions. Solving the Helmholtz equations with bivariate splines offers advantages over the existing finite element framework including high order FEM, interior penalty and hybridized discontinuous Galerkin methods, and weak Galerkin methods. Our implementation is relatively straightforward and allows us to find accurate spline solutions of arbitrary degree and smoothness for problems involving large wave numbers. More precisely,

(1) we are able to solve the Helmholtz problem with large wave number $1 \leq k \leq 500$ by using our splines of degree $p \geq 5$ and $h = 1/64$ on a laptop computer; using a large memory (1000GB) node from the Sapelo 2 cluster at University of Georgia, we are able to find accurate solutions to the Helmholtz equation with wave numbers from 500–1500 by using spline functions of degree 12 and $h = 1/100$. For example, the calculation Example 5.1.4 for $k = 1500$ used just under 630GB of memory and our program ran for 30 hours.

(2) our numerical evidence strongly shows that $hk \leq p/2$ will enable us to find very accurate approximation in H^1 norm for various wave numbers $k = 500 - 1000$ and degrees $p = 5, \cdots, 17$, and no pollution errors were observed. See Example 5.1.3.

(3) we are able to solve the Helmholtz problem with accurate numerical solutions over domains which are not strictly star-shaped or not convex. See Example 5.1.4.

(4) we are able to use the same implementation for exterior domain problem of Helmholtz equation.

(5) although our theory established in the previous sections requires C^1 smooth spline functions, our numerical experiments show that our algorithm works using C^0 spline functions with a slightly more accurate solutions. In order to be consistent with the theory, we report our numerical results based on C^1 smooth splines.

We shall present some simulation results based on a known exact solution and see how accurate our bivariate spline method can be. In particular, spline solutions to the Helmholtz problem for large wave numbers k will be demonstrated. It is known (cf. [27]) that higher-order methods are less sensitive to pollution. We shall use spline functions of large degrees. Our MATLAB code was written based a general degree of splines. That is, the degree of splines is an input variable. We can choose a large degree as long as a computer can handle. In the following examples, we solve the following 2D Helmholtz problem:

$$\begin{cases} -\Delta u - k^2 u & = f, \text{ in } \Omega, \\ \alpha(\nabla u \cdot \mathbf{n}) + \beta u & = g, \text{ on } \Gamma = \partial\Omega \end{cases} \tag{5.0.3}$$

over various domains Ω and for various $k \geq 1$.

5.1 2D Numerical Results

Example 5.1.1. We take Ω to be unit regular hexagon with center $(0,0)$ as seen in [8] and [29]. Here we take $f = \frac{\sin(kr)}{r}$, $\alpha = 1$, $\beta = \mathbf{i}k$, and g is chosen so that the exact solution is given by:

$$u = \frac{\cos(kr)}{k} - \frac{\cos(k) + \mathbf{i}\sin(k)}{k(J_0(k) + \mathbf{i}J_1(k))} J_0(kr)$$

in polar coordinates, where $J_\nu(z)$ are Bessel function of the first kind and $r = \sqrt{x^2 + y^2}$.

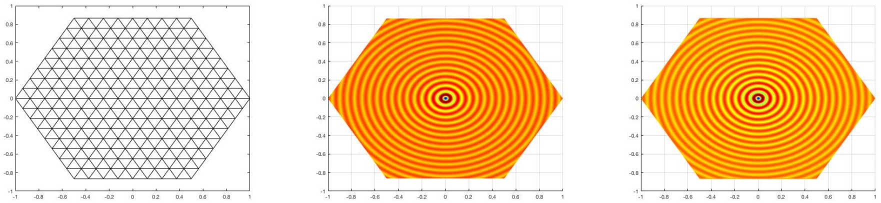


Figure 5.1: Example 5.1.1: Real and imaginary part of the spline solution $u_s \in S^1_9$ with wave number 100. Real part shown middle and imaginary part shown right.

In Fig. 5.1 we show plots of the spline solution $u_s \in S^1_9$ (real and imaginary parts) to Eq. (5.0.3) with wave number $k = 100$. We also use spline functions in S^1_d degree $d = 5, ..., 17$ to approximate the solution over the domain shown in Fig. 5.1, left. The relative errors in the L^∞ norm as well as the root mean square error based on 67201 equally-spaced points within Ω are shown in Table 5.1 ($k = 200$). It is clear from Table 5.1, when the degree of splines increases, the errors get better.

Table 5.1: Example 5.1.1: Relative and maximum L^2 and H^1 seminorm errors for C^1 spline solutions of various degrees to the Helmholz BVP with wave number k=200

d	h	rel. L2 error	rel. H1 error	ℓ_∞ error	$\|u\|_{1,\infty}$
5	0.063	1.3054e+00	1.4632e+00	2.9566e-02	1.0032e+01
7	0.063	1.0274e+00	1.0207e+00	5.4012e-02	1.6182e+01
9	0.063	5.0853e-02	5.7129e-02	1.8468e-03	5.6149e-01
11	0.063	1.1197e-03	1.5021e-03	4.3664e-05	1.2068e-02
13	0.063	4.8710e-05	1.0916e-04	1.7632e-06	9.6094e-04
15	0.063	2.2405e-06	4.8859e-06	7.6330e-08	3.4396e-05
17	0.063	1.0917e-07	2.5422e-07	3.8012e-09	2.5664e-06

Example 5.1.2. We next solve (5.0.3) again over the unit regular hexagon with center at $(0, 0)$ (as shown in the left graph of Fig. 5.1) for large wave number $k = 500$. We

use uniformly refined triangulations to find spline solutions of (5.0.3) which accurately approximate the exact solution as shown in Table 5.2. The errors decrease as the sizes of triangulation decrease.

Table 5.2: Example 5.1.2: Accuracy of spline solutions in S^1_{12} to the Helmholtz equation with wave number $k = 500$

wave no.	h	d	rel. L2 error	rel. H1 error	ℓ_∞ error	$\lvert u\rvert_{1,\infty}$ error
500	0.125	12	1.4541e+00	1.2967e+00	1.0588e-02	9.0560e+00
500	0.062	12	1.1921e+00	1.1743e+00	1.4517e-02	7.6721e+00
500	0.031	12	6.3515e-03	8.6685e-03	7.6444e-05	7.8923e-02
500	0.016	12	9.8523e-08	8.7072e-07	1.5062e-09	7.3869e-06

Table 5.3: Example 5.1.3: Accuracy of spline approximation for fixed $kh = 4$

wave no. k	h	kh	rel. L2 error	rel. H1 error	ℓ_∞ error	$\lvert u\rvert_{1,\infty}$ error
60	0.067	4	3.3041e-06	1.3658e-05	3.5912e-07	6.4930e-05
120	0.033	4	3.6576e-06	1.1103e-05	1.5460e-07	6.5157e-05
180	0.022	4	1.6801e-06	1.4977e-05	7.8387e-08	5.8730e-05
240	0.017	4	1.6748e-06	1.5134e-05	6.8373e-08	6.9122e-05
300	0.013	4	1.6735e-06	1.5231e-05	6.1516e-08	7.8338e-05
360	0.011	4	1.6738e-06	1.5294e-05	5.6416e-08	8.6716e-05
420	0.010	4	1.6725e-06	1.5342e-05	5.2400e-08	9.4442e-05

Example 5.1.3. In this example, we report the relative L^2 and H^1 error results with the size of the triangulation h chosen so that that the wave number and size of mesh satisfy $kh = p/2$ in Table 5.3. As we use the spline space S^1_8, so that $kh = 4$ for various k. Our numerical results in Table 5.3 suggest that there is no pollution error phenomenon or the pollution error is well controlled in our spline method. The numerical results demonstrate that the approximation constant C in Theorem 4.4.4 is indeed independent of k.

Example 5.1.4. In this example, we show the accuracy of spline solutions for high wave numbers. Again, we solve (5.0.3) over the unit hexagon. We report the relative errors for our spline solutions with $p = 10$, $r = 1$ with high wave numbers ($k =$

$500-1000$) in the top part of Table 5.4 and $p=12$ and $r=1$ ($k=1100-1500$) in the bottom part of Table 5.4.

Table 5.4: Example 5.1.4: Accuracy of spline solution in S^1_{10} for various large wave numbers

wavenumber k	h	rel. L2 error	rel. H1 error	ℓ_∞ error	$\|u\|_{1,\infty}$ error
500	0.016	5.4581e-06	6.7180e-05	1.1751e-07	5.9545e-04
600	0.016	1.3501e-04	4.0429e-04	1.7397e-06	2.6984e-03
700	0.016	2.0630e-03	2.5532e-03	3.3968e-05	1.6688e-02
800	0.008	8.0733e-07	7.6723e-06	1.3186e-08	6.6425e-05
900	0.008	1.9449e-06	2.3920e-05	3.5339e-08	1.8619e-04
1000	0.008	7.3629e-06	6.7027e-05	9.8781e-08	8.5276e-04

Table 5.5: Example 5.1.4 Accuracy of spline solution in S^1_{12} for various large wave numbers

wavenumber k	h	rel. L2 error	rel. H1 error	ℓ_∞ error	$\|u\|_{1,\infty}$ error
1100	0.0100	2.9026e-05	4.9725e-05	1.8495e-07	4.6325e-04
1200	0.0100	8.5032e-05	1.3023e-04	4.5849e-07	9.1309e-04
1300	0.0100	3.8509e-04	4.3707e-04	5.2119e-06	2.0724e-03
1400	0.0100	1.9326e-03	1.9489e-03	2.3369e-05	1.7926e-02
1500	0.0100	8.3163e-03	8.2431e-03	5.3503e-05	1.2119e-01

Example 5.1.5. To see the degrees of freedom when solving (5.0.3), let us present two tables for our spline method with the weak Galerkin method in [29]. For wave number $k=1$, we compare the accuracy of spline solutions from the space S^1_5 to piecewise constant weak Galerkin solutions (relative error results from [29]) along with degree of freedom counts. For the piecewise constant WG method, we calculate the degrees of freedom by $dof_{cwg} = \#(E) + \#(T)$. For splines in $S^1_5(\triangle)$, we report an only upper bound on the degrees of freedom for convenience; $dof_{S^1_5} < 2\#(V) + \#(E)(d-1) + \#(E)(d-3)$. We write $\#(V), \#(E)$, and $\#(T)$ to denote the number of vertices, edges, and triangles in a given triangulation. The numerical results are shown in Table 5.6.

In Table 5.7, a comparison of relative errors of the solutions of spline S^1_5 and piecewise linear weak Galerkin solutions from [29] is shown, along with degree of

Table 5.6: Comparison of the accuracy of spline method with piecewise constant weak Galerkin method

	piecewise constant WG			Spline S_5^1		
$\lvert\triangle\rvert$	rel. L2 error	rel. H1 error	dof	rel. L2 error	rel. H1 error	dof
1.000	-	-	-	9.285e-07	1.948e-05	98
0.500	4.170e-03	2.490e-02	66	1.672e-08	6.819e-07	332
0.250	1.050e-03	1.110e-02	252	6.635e-10	2.224e-08	1214
0.125	2.630e-04	5.380e-03	984	-	-	-
0.062	6.580e-05	2.670e-03	3888	-	-	-
0.031	1.650e-05	1.330e-03	15456	-	-	-
0.016	4.110e-06	6.650e-04	61632	-	-	-

freedom counts. For piecewise linear WG, we calculate $dof_{lwg} = 2\#(E)+3\#(T)$. The spline method provides a more accurate solution using far fewer degrees of freedom. Here the wave number is $k = 5$.

Table 5.7: Comparison of the accuracy of spline solution with piecewise constant linear weak Galerkin method

	linear WG			Spline S_5^1		
$\lvert\triangle\rvert$	rel. L2 error	rel. H1 error	dof	rel. L2 error	rel. H1 error	dof
1.000	-	-	-	4.287e-03	1.489e-02	98
0.500	-	-	-	1.183e-04	7.197e-04	332
0.250	2.580e-04	9.480e-03	600	2.019e-06	2.546e-05	1214
0.125	3.460e-05	2.310e-03	2352	3.525e-08	8.609e-07	4634
0.062	4.470e-06	5.740e-04	9312	2.411e-09	2.866e-08	18098
0.031	5.640e-07	1.430e-04	37056	-	-	-
0.016	7.060e-08	3.580e-05	147840	-	-	-
0.008	8.790e-09	8.960e-06	590592	-	-	-

Example 5.1.6. Let us consider the Helmholtz boundary value problem over a non-convex domain, shown left in Fig. 5.2. For this example, $\alpha = 1$, $\beta = \mathbf{i}k$ and source functions f and g are chosen so that the analytic solution to (5.0.3) is given by

$$u = J_\xi(kr)\cos(\xi\theta).$$

As above, r and θ are the usual polar coordinates, k is the wavenumber, and J_ξ is a Bessel function of the first kind. This is another standard testing function studied in [29] and [10]. We study three situations where $\xi = 1$, $3/2$, and $2/3$. Plots of the spline solutions from S^1_5 for $k = 4$ and $k = 20$ are shown in Fig. 5.2– 5.4. We summarize numerical results for each of these three cases in Tables 5.8, 5.10, and 5.9.

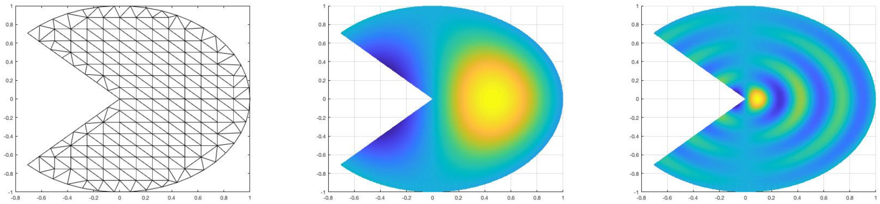


Figure 5.2: Example 5.1.6: Spline solution $s \in S^1_5$ to non-convex Helmholtz problem with exact solution $u = J_\xi(kr)\cos(\xi\theta)$, with $\xi = 1$. The underlying triangulation is shown left, solution with wave number $k = 4$ center, and solution with wave number $k = 20$ right.

Table 5.8: Example 5.1.6: Numerical results of spline approximation $\in S^1_5$ over non-convex domain with $\xi = 1$

	wavenumber=4		wavenumber=20	
$\lvert\Delta\rvert$	rel. L2 error	rel. H1 error	rel. L2 error	rel. H1 error
1.0000	1.1242e-03	4.6766e-03	1.3420e+00	1.6892e+00
0.5000	2.0562e-04	8.2798e-04	8.9020e-01	8.7483e-01
0.2500	3.4424e-06	3.0885e-05	1.0677e-01	1.1434e-01
0.1250	8.1231e-08	1.1162e-06	1.6385e-03	4.1769e-03
0.0625	-	-	2.0492e-05	1.3421e-04
0.0312	-	-	6.5958e-06	8.2274e-06

Example 5.1.7. Certainly, we are interested in exploring numerical solution to a nonconvex domain with larger wave numbers $k = 100, 200, 300$. As referenced in [29], the computation for $\xi = 2/3$ is more challenging than the case where $\xi = 1$ and $\xi = 3/2$. However, the spline solution in S^1_{10} is nonetheless highly accurate. In Fig. 5.5, graphs of the spline solutions in $S^1_{10}(\triangle)$ to the difficult BVP with $\xi = 2/3$ are shown for higher wave numbers $k = 200$ and $k = 300$. Relative errors are given in the plots; all relative L^2 and H^1 errors are on the order of 10^{-2}.

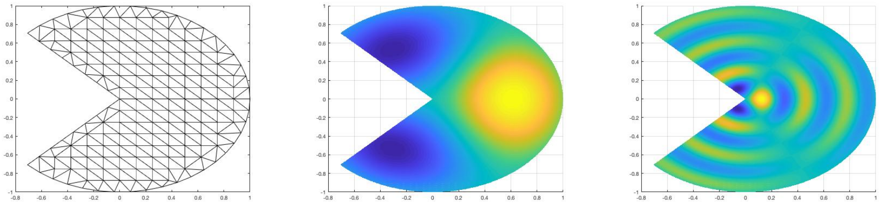


Figure 5.3: Example 5.1.6: Spline solution $s \in S_5^1$ to non-convex Helmholtz problem with exact solution $u = J_\xi(kr)\cos(\xi\theta)$, where $\xi = 3/2$. The underlying triangulation is shown left, solution with wave number $k = 4$ center, and solution with wave number $k = 20$ right.

Table 5.9: Example 5.1.6: Numerical results of spline approximation $\in S_5^1$ over non-convex domain with $\xi = 3/2$

	wavenumber=4		wavenumber=20	
$\lvert\Delta\rvert$	rel. L2 error	rel. H1 error	rel. L2 error	rel. H1 error
1.0000	1.0993e-01	1.4599e-01	1.9511e+00	2.3335e+00
0.5000	2.7023e-03	2.1935e-02	1.0055e+00	1.0333e+00
0.2500	1.0796e-03	5.7067e-03	4.4131e-02	6.5030e-02
0.1250	2.2659e-04	2.0220e-03	6.4977e-03	1.1583e-02
0.0625	4.9490e-05	7.1555e-04	1.3156e-03	3.4129e-03
0.0312	1.0926e-05	2.2876e-04	2.8728e-04	1.0494e-03

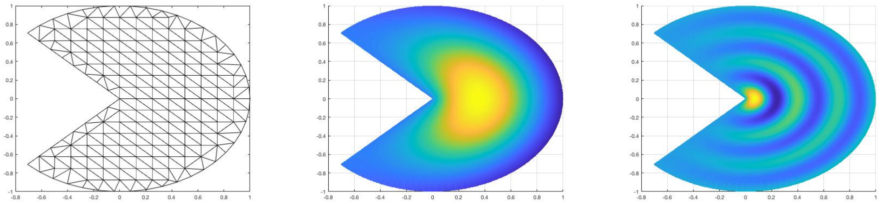


Figure 5.4: Example 5.1.6: Spline solution to non-convex Helmholtz problem with exact solution $u = J_\xi(kr)\cos(\xi\theta)$, with $\xi = 2/3$. The underlying triangulation is shown left, solution with wave number $k = 4$ center, and solution with wave number $k = 20$ right.

Table 5.10: Example 5.1.6: Numerical results of spline approximation $\in S_5^1$ over nonconvex domain with $\xi = 2/3$

	wavenumber=4		wavenumber=20	
$\lvert\Delta\rvert$	rel. L2 error	rel. H1 error	rel. L2 error	rel. H1 error
0.5000	9.1279e-03	5.3100e-02	1.4984e+00	1.5024e+00
0.2500	3.3169e-03	3.0808e-02	9.5122e-01	9.4475e-01
0.1250	1.2753e-03	1.8893e-02	8.1904e-03	2.6594e-02
0.0625	4.9854e-04	1.0827e-02	3.1416e-03	1.4909e-02
0.0312	1.9433e-04	5.6974e-03	1.2276e-03	7.7787e-03

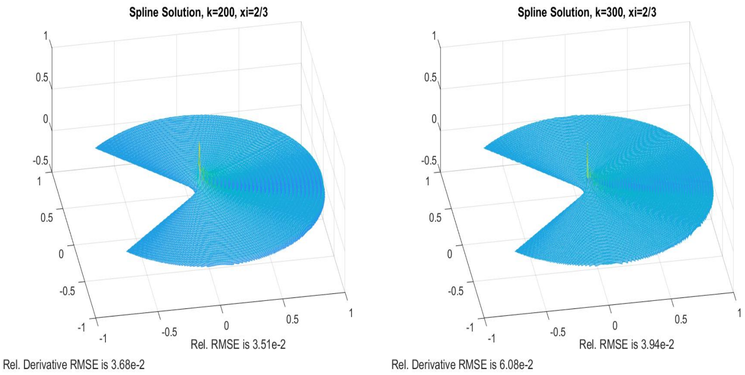


Figure 5.5: Example 5.1.7: Spline solution $\in S^1_{10}$ to the non-convex number Helmholtz problem with large wave number. The exact solution is $u = J_\xi(kr)\cos(\xi\theta)$ and $\xi = 2/3$, with wave numbers $k = 200$ left and $k = 300$ right.

5.2 3D Numerical Results

In this section, we take Ω to be unit regular cube with center $(0.5, 0.5, 0.5)$ and consider the Helmholtz equation in three dimensions. The splines solutions are generated in the same way as described in 5.1, but producing accurate numerical solutions for large wave numbers is more challenging in the 3D setting. An experiment demonstrating this difficulty is shown in Example **??**.

Example 5.2.1. This is a generalization of Example 5.1.1 from 5.1. As above, we choose the boundary condition with $\beta = \mathbf{i}k$, and g is chosen so that the exact solution is given by:

$$u = \sin(kz)\big(\frac{\cos(kr)}{k} - \frac{\cos(k) + \mathbf{i}\sin(k)}{k(J_0(k) + \mathbf{i}J_1(k))}J_0(kr)\big)$$

in cylindrical coordinates, where $J_\nu(z)$ are Bessel function of the first kind and $r = \sqrt{x^2 + y^2}$.

Table 5.11: Relative and maximum errors for C^1 spline solutions of various degrees to the 3-dimensional Helmholz BVP with wave number k=25. The errors shown are in the L^2 norm and H^1 seminorm.

d	h	rel. L2 error	rel. H1 error	ℓ_∞ error	$\lvert u\rvert_{1,\infty}$
5	0.125	9.3057e-03	2.6901e-02	2.0714e-03	1.2002e-01
6	0.125	2.4927e-03	8.0630e-03	7.6470e-04	7.0878e-02
7	0.125	4.6380e-04	1.9359e-03	1.0437e-04	1.7896e-02
8	0.125	7.9027e-05	3.9022e-04	2.1745e-05	4.9318e-03
9	0.125	1.7481e-05	5.6743e-05	4.2484e-06	7.0103e-04
10	0.125	2.2924e-06	1.0117e-05	5.9486e-07	8.9463e-05

Chapter 6

Numerical Solutions to the Maxwell Equations

6.1 Shielded Microstrip

Here we present a calculation of the potential and electric field resulting from a shielded mircrostrip operating at low frequency.

A microstrip a kind of “waveguide" or transmission line. When discussing electrical circuits, a transmission line is simply some structure designed to deliver an electrical signal from one part of the circuit to another. The microstrip is the most common type of planar transmission line, and is “quasi-TEM", (a TEM wave is a transverse electromagnetic wave), which means that we can use TEM analysis when the microstrip circuit operates at low frequencies. Striplines and coaxial lines are two other common types of TEM transmission lines. For microstrip circuit elements, the cutoff frequency for TEM versus non-TEM analysis occurs at low microwave frequencies of around 5 GHz. For higher frequency currents, the longitudinal components of the electric field cannot be ignored.

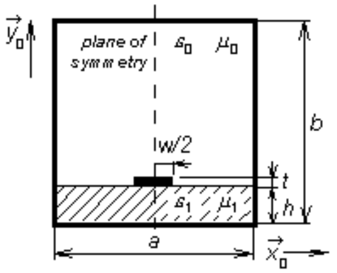


Figure 6.1: A schematic of a shielded microstrip wavequide. The rectangular boundary of the planar section is a conducting surface and so the electric potential V=0 there. The shaded lower section is a dielectric with electric permittivity ϵ_1 and magnetic permeability μ_1; the upper region is simply air, so its permittivity and permeability are simply defined to be the vacuum constants. The black region is the current-carrying strip, maintained at a certain electric potential.

We seek to reproduce an example originally presented by Jin in [20] of electrostatic analysis of a shielded microstrip operating at low frequency. Jiang also addresses the problem in [18] using a first-order div-curl formulation. In both cases, the analysis is done by solving a boundary value problem over a domain like the one shown in Figure 6.1.

First, we assume that the inner conductor is held at a constant potential V_{imp} so that electrostatic analysis is warranted; second, we make use of the symmetry of the domain shown in Fig. 6.1 to cut the size of the domain in half. That is, rather than solve over the whole rectangle, we instead bisect the domain vertically and impose a Neumann boundary condition on this plane of symmetry Γ_s (the dashed line in the figure above).

The first-order formulation of an electrostatic boundary value problem is

$$\begin{aligned} \nabla \times \boldsymbol{E} &= 0 \quad \text{in } \Omega \\ \nabla \cdot (\epsilon \boldsymbol{E}) &= \rho \quad \text{in } \Omega \qquad (6.1.1) \\ \boldsymbol{E} &= -\nabla V \text{ in } \Omega \qquad (6.1.2) \end{aligned}$$

where V is the electric potential (which is known on the boundary), ρ is the charge distribution and ϵ is the electric permittivity over all of Ω. Here, we have $\rho = 0$, and ϵ is defined piecewise. The required external boundary conditions are

$$V = V_{imp} \quad \text{on } \Gamma_c c \tag{6.1.3}$$

$$\boldsymbol{n} \times \boldsymbol{E} = 0 \quad \text{on } \Gamma_c \tag{6.1.4}$$

$$\boldsymbol{n} \cdot \boldsymbol{E} = 0 \quad \text{on } \Gamma_s. \tag{6.1.5}$$

where Γ_c is the shielding conducting surface and Γ_{cc} is portion of the boundary that actually touches the current-carrying portion of the microstrip (the boundary of the small black rectangle in Fig. 6.1). Let $\Omega = \Omega_1 \cup \Omega_0$ where Ω_1 is the lower shaded region which contains the dielectric material and Ω_0 is the upper air-filled region; then we must impose the following additional boundary conditions at the junction $\Gamma_{int} := \bar{\Omega}_1 \cap \bar{\Omega}_0$, as justified in 3.2.1:

$$V^+ = V^- \quad \text{on } \Gamma_{int} \tag{6.1.6}$$

$$\boldsymbol{n} \times \boldsymbol{E}^+ = \text{n} \times \boldsymbol{E}^- \quad \text{on } \Gamma_{int} \tag{6.1.7}$$

$$\boldsymbol{n} \cdot (\epsilon_0 \boldsymbol{E}^+) = \quad \boldsymbol{n} \cdot (\epsilon_1 \boldsymbol{E}^-). \tag{6.1.8}$$

Jiang's formulation in [18] uses the least squares method, which amounts to the quadratic functional from Equation **??**, $I(\mathbf{E}) = ||\nabla \times \mathbf{E}||^2 + ||\nabla \cdot \mathbf{E} - \rho/\epsilon||^2$ with the additional internal boundary conditions enforced “naturally" by adding the term

$$(\nabla \boldsymbol{E}^+ - \nabla \boldsymbol{E}^-)^2 + (\boldsymbol{E}_x^+ - \boldsymbol{E}_x^-)^2 + (\epsilon_0 \boldsymbol{E}_y^+ - \epsilon_1 \boldsymbol{E}_y^-)^2$$

to the functional for each node that borders Γ_{int}.

We make use of the potential formulation detailed in Section **??**, and solve for V and differentiate to get $\boldsymbol{E}$. We substitute Equation 6.1.2 into the rest of the first-order formulation and equation 6.1.1 reduces to

$$-\Delta(\epsilon V) = 0 \quad \text{in } \Omega \quad \text{ or } \quad \begin{cases} -\Delta(\epsilon_0 V) = 0 & \text{in } \Omega_0 \\ -\Delta(\epsilon_1 V) = 0 & \text{in } \Omega_1. \end{cases} \tag{6.1.9}$$

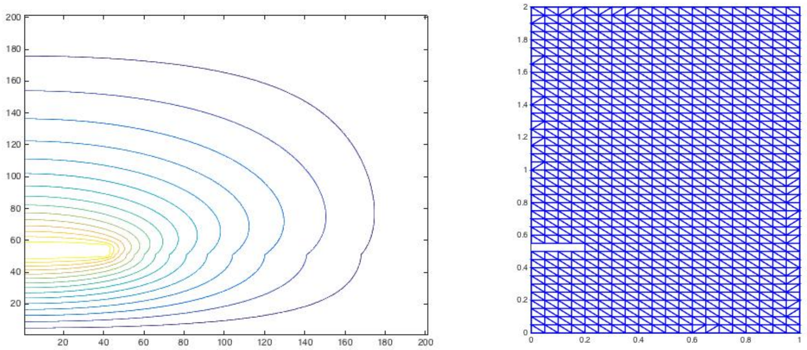


Figure 6.2: Shielded Microstrip: A contour plot of the electric potential and its underlying triangulation. The potential, V for a shielded microstrip, held at constant voltage (V=1), left. The triangulation used to compute the spline solution, right.

We must also convert the boundary conditions to be compatible with this formulation. Condition 6.1.3 is automatically satisfied; for condition 6.1.4 we consider $-\boldsymbol{n} \times (\nabla V) = 0$. In the potential formulation, this is simply the requirement that the derivative of V in the tangent direction at the boundary is zero; it will be exactly satisfied if V is constant on Γ_c. In the case of the shielded microstrip, Γ_c is a grounded conductor, so $V \equiv 0$ on Γ_c. The Dirichlet condition 6.1.5 becomes a Neumann boundary condition after the substitution $-\nabla V = \boldsymbol{E}$:

$$\boldsymbol{n} \cdot \boldsymbol{E} = 0 \implies \boldsymbol{n} \cdot \nabla V = 0 \implies \frac{\partial V}{\partial \boldsymbol{n}} = 0 \quad \text{on } \Gamma_s.$$

Then internal boundary conditions at Γ_{int} but also be converted. Condition 6.1.6 will be satisfied as an essential boundary condition since our numerical solution belongs to a subspace of S_0^d. By choosing a triangulation Δ so that Γ_{int} only coincides with edges of triangles in Δ, we can also easily enforce condition 6.1.7. To see this, let edge E_{int} be an edge of the triangulation that lies on Γ_{int}, and let T^+ and T^- be the triangles above and below the edge respectively. Since V is globally continuous, $V_{T^+}|_{E_{int}} = V_{T^-}|_{E_{int}}$; on the edge E_{int}, V_{T^+} and V_{T^-} reduce to the same univariate polynomial. Therefore, their derivatives in the direction tangent to E_{int} will match. Therefore we have

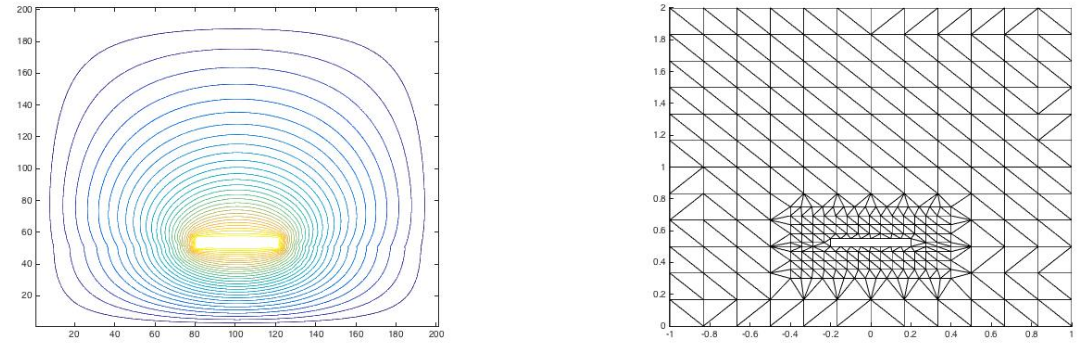
$$\boldsymbol{t} \cdot \nabla V_{T^+} = \boldsymbol{t} \cdot \nabla V_{T^-}\Big|_{E_{int}} \iff \boldsymbol{n} \times \nabla V_{T^+} = \boldsymbol{n} \times V_{T^-}\Big|_{E_{int}} \iff \boldsymbol{n} \times \boldsymbol{E}_{T^+} = \mathrm{n} \times \boldsymbol{E}_{T^-}\Big|_{E_{int}}.$$

Since this will hold for all such edges E_{int}, as long as we cover Γ_{int} with edges of Δ, condition 6.1.7 will be satisfied.

Condition 6.1.8 is somewhat more difficult. In the potential formulation, it becomes a Neumann-type boundary condition $\epsilon_0 \frac{\partial V^+}{\partial \boldsymbol{n}} = \epsilon_1 \frac{\partial V^-}{\partial \boldsymbol{n}}$. The simple linear condition described in section **??** guarantees that the derivatives of, say V_{T^+} and V_{T^-}, in the direction of an (unshared) edge of T^+ match. As described above, since the spline function is continuous, the derivatives of these polynomial pieces in the direction tangent to the shared edge also match. The linear independence of these directions then gives $\mathcal{C}^1$ smoothness across the edge in question.

We impose condition 6.1.8 across the appropriate edges of the triangulation, by altering the smoothness conditions on the domain points near the edge in question. Instead of enforcing matching derivatives in an edge direction, we directly require continuity of the normal derivatives. This is accomplished by calculating an edge's normal direction using barycentric direction vectors of the neighboring triangles. The details can be found in Section **??**, but once formulated, we can simply multiply one side of the linear constraint equation by ϵ_1/ϵ_0 to guarantee 6.1.8 is satisfied.

Figure 6.2 shows level curves of the calculated electric potential over the triangulated domain on the right. Note the change in the shape of these curves at the interface Γ_{int} between the dielectric material and air. Figure **??** shows the compu-



Shielded Microstrip: A contour plot of the electric potential and its underlying triangulation over the full cross-section]Shielded Microstrip: A contour plot of the electric potential and its underlying triangulation over the full cross-section. The potential, V for a shielded microstrip, held at constant voltage (V=1), left. The triangulation used to compute the spline solution, right.

tation done over the entire domain, where we use a nonuniform triangulation for improved efficiency.

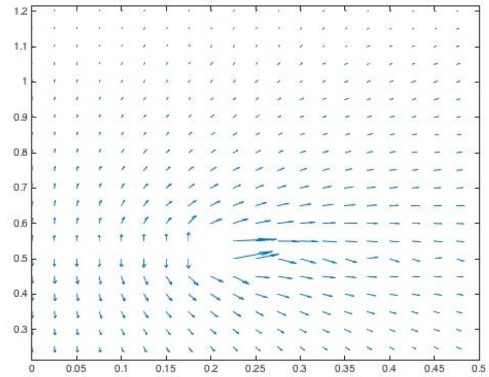


Figure 6.3: Shielded Microstrip: Computed Electric Field. We take the negative gradient of the numerical solution of the potential equation (area near the microstrip shown for clarity).

We can then differentiate to obtain the electric field at all points in the domain. The resulting vector field is shown in Figure 6.3. To reproduce Jiang's calculation, we compute the electric field at each vertex in the triangulation, and give the electric

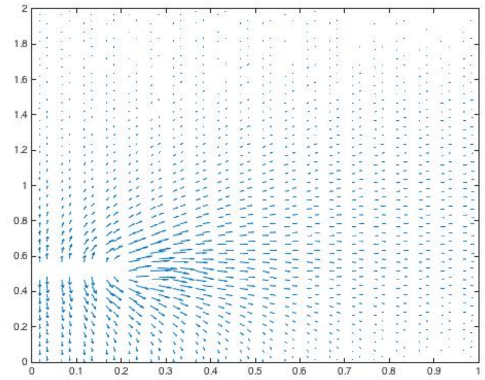


Figure 6.4: Shielded Microstrip: Averaged Electric Field. We average the computed electric field over each triangle to match Jiang's calculation in [18]

.

field vector at the center of each triangle as the average of the field at the triangle's vertices. This is Figure 6.4.

6.2 Coaxial Join

Here we explore a three dimensional problem in which the symmetry of the solution domain can be exploited to reduce the analysis to 2 dimensions. Consider a join of two coaxial waveguides of different inner radii. Each coaxial cable has an inner, current carrying conductor, and an outer, grounded conductor. In between these cylindrical conductors lies a layer of dielectric material. We wish to calculate the electrostatic potential and the electric field in this region.

Let the cable be running along the $z-$axis, with the join occurring at the origin. For the leftmost coaxial waveguide, we take the radius of the outer conductor to be 1.2, and inner radius 0.2; the guide it is joined to has the same outer radius, but an inner radius of 0.7. We assume the inner conducting surfaces are held at a constant

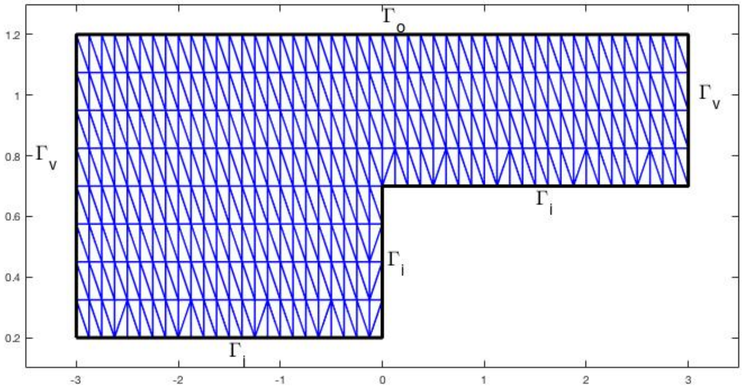


Figure 6.5: Coaxial Join: Triangulation of region of interest. The vertical axis here measures r; the horizontal is the z-axis. Far away from the join, on grounds of symmetry, we impose $\frac{\partial u}{\partial \boldsymbol{n}} = 0$ on the vertical boundaries and Dirichlet conditions (informed by 3.2.1) on the conductors in question.

potential of 1 V, while the outer conductor is grounded at 0 V. These become Dirichlet boundary conditions for the potential formulation of the boundary value problem.

To take advantage of the axial symmetry, we reformulate the electrostatic problem using cylindrical coordinates; $r := (x^2 + y^2)^{1/2}$, $\theta := \arctan(y/x)$, and $x := z$. The region in question does not vary with θ, so we consider a slice of the coaxial waveguide in the z-direction, perpendicular to θ. We also slice across the cylinder, parallel to θ, at a sufficient distance away from the join; along these edges, because of the symmetry in the z-direction away from the join, we expect $\frac{\partial u}{\partial n} = 0$. This gives us a portion of a plane Ω with boundary Γ on which we can perform the electrostatic analysis in two dimensions. Let Γ_o be the edge of Ω corresponding to the outer conductor, Γ_i the edges of the inner conductor, and Γ_v the vertical edges where the Neumann conditions hold. A triangulation Δ of the region is shown in Figure 6.5.

Clearly, this change in coordinates affects the equation to be solved. Instead of solving 6.1.9, we must consider the Poisson Equation in cylindrical coordinates

$$-\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) - \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} - \frac{\partial^2 u}{\partial z^2} = f(r, \theta, z).$$

In this case, we have by symmetry $\frac{\partial u}{\partial \theta} \equiv 0$, and because there is no charge in the domain of interest, $f \equiv 0$. Thus the equation we wish to solve is

$$-\frac{\partial}{\partial r}\left(\epsilon_r r\frac{\partial u}{\partial r}\right) - \frac{\partial}{\partial z}\left(\epsilon_r r\frac{\partial u}{\partial z}\right) = 0.$$

Since our analysis can now take place in the plane, we substitute y for r and x for z, and using the standard *del* operator the problem to be solved is

$$\nabla \cdot (\epsilon_r y \nabla u) = 0$$

with boundary conditions

$$\begin{aligned} u &= 0 && \text{on } \Gamma_o \\ u &= 1 && \text{on } \Gamma_i \\ \frac{\partial u}{\partial \boldsymbol{n}} &= 0 && \text{on } \Gamma_v \end{aligned}$$

We multiply through by a test function ϕ, and integrate by parts:

$$\int_\Omega \nabla \cdot (\epsilon_r y \nabla u)\phi d\Omega = \epsilon_r \int_\Gamma y(\nabla u \cdot \boldsymbol{n})\phi d\boldsymbol{\Gamma} - \epsilon_r \int_\Omega y \nabla u \cdot \nabla \phi d\Omega = 0$$

For a careful formulation, the boundary value problem should be reformulated as in Section **??** so that the integral over Γ becomes an integral only over Γ_v, where we impose the Neumann boundary conditions. We quickly notice that the construction of the stiffness matrix will be different than it was for the standard Poisson problem

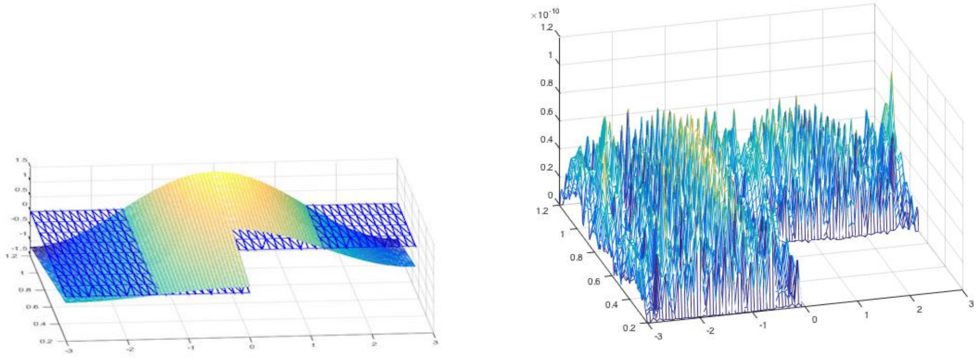


Figure 6.6: Plots of numerical solution to BVP with exact solution $u = y\sin(\frac{\pi}{3}x)$ and error. The spline approximation is shown left; the spatial distribution of errors is shown right. The plot demonstrates that the solution is approximated well at all boundaries and in the domain interior.

in Cartesian coordinates. Now, for the entries corresponding to a particular triangle $T \in \Delta$, we have

$$K_T = \left[\int_T y \nabla B^T_{ijk} \nabla B^T_{l,m,n}\right]_{\substack{i+j+k=d \\ l+m+n=d}};$$

i.e. the integral is weighted by the coordinate y. To address this in practice, we represent the function y as a degree d Bernstein-Bezier polynomial so that the product and the integral can be performed using the convenient formulas arising from the de Casteljau algorithm.

To test the accuracy of the code for this new formulation, we consider a test problem with exact solution $g(x, y) = y\sin(\frac{\pi}{3}x)$. This produces a nonhomogeneous case with source function $f(x, y) = ((\frac{\pi}{3}y)^2 - 1)\cos(\frac{\pi}{3}x)$; we impose $u = g(x, y)$ on Γ_o and Γ_i, and $\frac{\partial u}{\partial \boldsymbol{n}} = \frac{\partial g}{\partial \boldsymbol{n}} = 0$ on Γ_v, and solved using the approximation space S^1_5. The error between the approximate spline solution and the exact solution was calculated on a grid of 10000 points spread over Ω. The maximum error of the solution was 1.5e-10, and the maximum error in the first order derivatives was less than 9e-9. A plot of

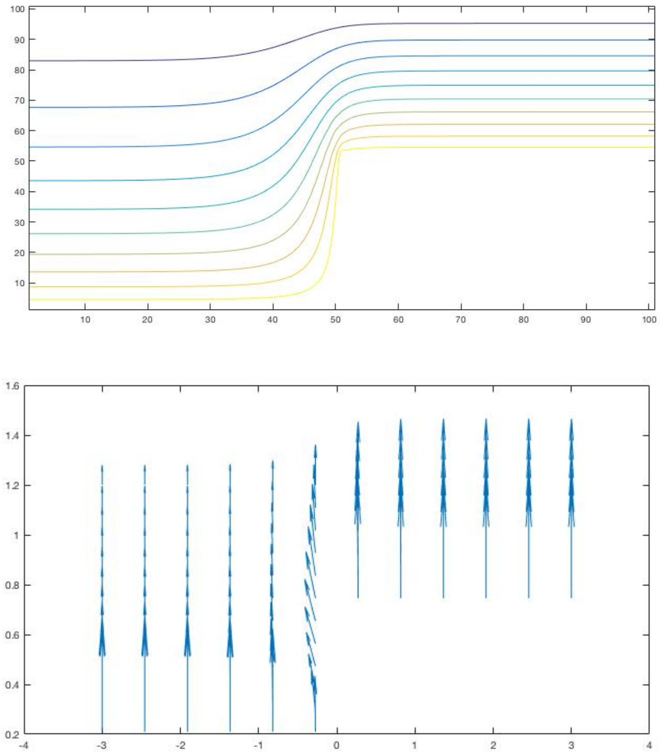


Figure 6.7: Coaxial Join: Contour plot of equipotential lines, top, and computed electric field, bottom

the numerical solution and a plot of the error is shown in Figure 6.6 respectively. This matches what we expect from the theory, and so we can trust our calculations for the coaxial join problem even though there is no exact solution for us to test against.

Thus we are ready to calculate the potential in the dielectric material surrounding the join in the coaxial waveguides. A contour plot of the potential surface and the computed electric field is shown in Figure **??**. The contour plot visually matches an example from Jin in [20], and the electric field satisfies the appropriate boundary conditions for a field near a conducting surface. Namely, the electric field is orthogonal to the surface of the conductors at or near the surface in question.

6.3 A Bivariate Spline Analysis of the TEM mode of a Parallel Plate Waveguide

Our goal is to characterize the disturbance to the TEM mode of a plane wave caused by a material discontinuity in a parallel plate waveguide. We will replicate a numerical experiment found in [20] to verify the validity of our numerical analysis, and then extend the existing analysis in the literature by varying the frequency of the waves in the waveguide and the shape of the dielectric discontinuities. We begin with a thorough explanation of the physics involved.

We consider two, perfectly conducting (sometimes referred to as PEC) metal plates parallel to each other and to the $yz-$plane as in Fig. 6.8. The dimensions of the plates are far greater than their separation d, assuring the effect of fringing fields is negligible [13].

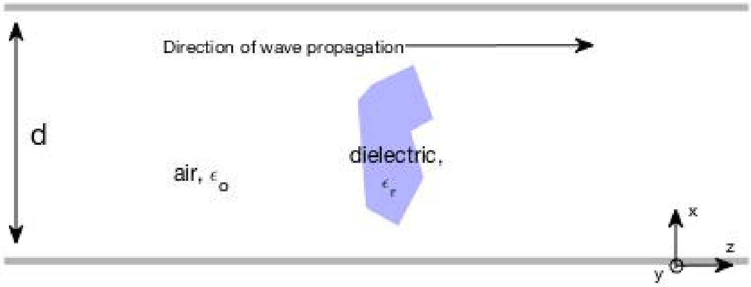


Figure 6.8: Schematic of a parallel plate waveguide with a material discontinuity. The obstruction and surrounding material may have different electric permittivities, resulting in interesting electromagnetic behavior.

The assumed geometry will lead to plane waves which propagate through the guide in the $z-$direction. The propagating waves are composed of 3 basic wave types: transverse electric (TE) waves, which have no electric field in the direction of propagation; and transverse magnetic (TM) waves, which have no magnetic field in the direction of propagation; transverse electromagnetic (TEM) waves, which have no electric or magnetic field component in the direction of propagation [33]. We assume that some source, outside our region of interest, is driving electromagnetic waves to propagate from left to right in Fig. 6.8. We further assume that the excitation of the conducting plates is uniform in the y-direction. Then the complete set of the TE, TM, and TEM modes allows for the representation the waves resulting from a source of arbitrary frequency [13]. In [20], Jin asserts that as long as the waveguide operates at a low enough frequency, the propagating wave will take the form

$$\boldsymbol{H}(x, y, z) = H_o e^{-ik_z z} \hat{\boldsymbol{y}}, \tag{6.3.1}$$

after accounting for differences in orientation.

But why must the wave take only that form if and (only if) the electromagnetic wave is low frequency? At what frequency threshold does this model breakdown? We expand Jin's justification for his analysis below. The following analysis is not new to the literature, but it is unique, and as it informs the numerical experiments which follow, we include it for understanding and completeness. Our analysis assumes the same orientation as depicted in Fig. 6.8; we refer to the interior of the waveguide as Ω.

We begin with the assumption that we have monochromatic plane waves propagating down the waveguide so that

$$\tilde{\boldsymbol{H}}(x, y, z, t) = \boldsymbol{H}(x, y, z) e^{-i\omega t} = \boldsymbol{\mathcal{H}}(x, y) e^{i(k_z z - \omega t)} \tag{6.3.2}$$

$$\tilde{\boldsymbol{E}}(x, y, z, t) = \boldsymbol{E}(x, y, z) e^{-i\omega t} = \boldsymbol{\mathcal{E}}(x, y) e^{i(k_z z - \omega t)}. \tag{6.3.3}$$

[13]. Then, away from the dielectric discontinuity, the (time-harmonic) Maxwell equations give

$$\begin{aligned} \nabla \cdot \boldsymbol{E} &= 0 & \qquad \nabla \times \boldsymbol{E} &= i\omega\mu\boldsymbol{H} \\ \nabla \cdot \boldsymbol{B} &= 0 & \qquad \nabla \times \boldsymbol{H} &= -i\omega\epsilon\boldsymbol{E}. \end{aligned}$$

Interpreting these equations component-wise, we discover that each of the components of the field quantities may be written in terms of their z-components only; therefore, our goal is simply to solve for those components. The full analysis is below, where we let, for example, $\boldsymbol{H} =< Hx, Hy, Hz >$.

$$\nabla \times \boldsymbol{E} = i\omega\mu\boldsymbol{H} \implies \begin{cases} \partial_y E_z - \partial_z E_y = i\omega\mu H_x & (6.3.4a) \\ \partial_z E_x - \partial_x E_z = i\omega\mu H_y & (6.3.4b) \\ \partial_x E_y - \partial_y E_x = i\omega\mu H_z & (6.3.4c) \end{cases}$$

$$\nabla \times \boldsymbol{H} = -i\omega\epsilon\boldsymbol{E} \implies \begin{cases} \partial_y H_z - \partial_z H_y = -i\omega\epsilon E_x & (6.3.5a) \\ \partial_z H_x - \partial_x H_z = -i\omega\epsilon E_y & (6.3.5b) \\ \partial_x H_y - \partial_y H_x = -i\omega\epsilon E_z. & (6.3.5c) \end{cases}$$

The assumed z-dependence from 6.3.2 and 6.3.3 then yields

$$\nabla \times \boldsymbol{E} = i\omega\mu\boldsymbol{H} \implies \begin{cases} \partial_y E_z - ik_z E_y = i\omega\mu H_x & (6.3.6a) \\ ik_z E_x - \partial_x E_z = i\omega\mu H_y & (6.3.6b) \\ \partial_x E_y - \partial_y E_x = i\omega\mu H_z & (6.3.6c) \end{cases}$$

$$\nabla \times \boldsymbol{H} = -i\omega\epsilon\boldsymbol{E} \implies \begin{cases} \partial_y H_z - ik_z H_y = -i\omega\epsilon E_x & (6.3.7a) \\ ik_z H_x - \partial_x H_z = -i\omega\epsilon E_y & (6.3.7b) \\ \partial_x H_y - \partial_y H_x = -i\omega\epsilon E_z. & (6.3.7c) \end{cases}$$

We combine 6.3.6b and 6.3.7a together to conclude

$$E_x = \frac{i}{\omega^2\mu\epsilon - k_z^2}(k_z\partial_x E_z + \omega\mu\partial_y H_z) \tag{6.3.8}$$

$$H_y = \frac{i}{\omega^2\mu\epsilon - k_z^2}(k_z\partial_y H_z + \omega\epsilon\partial_x E_z) \tag{6.3.9}$$

Similarly, 6.3.6a and 6.3.7b yield

$$E_y = \frac{i}{\omega^2\mu\epsilon - k_z^2}(k_z\partial_y E_z - \omega\mu\partial_x H_z) \tag{6.3.10}$$

$$H_x = \frac{i}{\omega^2\mu\epsilon - k_z^2}(k_z\partial_x H_z - \omega\epsilon\partial_y E_z). \tag{6.3.11}$$

Finally, we derive the scalar-valued PDEs to be solved by combining 6.3.6c with 6.3.8 and 6.3.10, and 6.3.7c with 6.3.11 and 6.3.9 respectively. This gives

$$-\Delta H_z - (\omega^2\epsilon\mu - k_z^2)H_z = 0 \tag{6.3.12}$$

$$-\Delta E_z - (\omega^2\epsilon\mu - k_z^2)E_z = 0 \tag{6.3.13}$$

As does Jin in [20], our analysis below will emphasize the magnetic field component H_z. For electromagnetic waves oscillating at a microwave frequency regime or lower, we can assume that the tangential electric field at the perfectly conducting parallel plates is 0. [24] That is, $E_y = E_z = 0$ at $x = 0$ and $x = d$. Equation 6.3.13 is supplemented by these Dirichlet boundary conditions. Applying this to 6.3.10, we discover the Neumann boundary condition imposed in [20], or

$$\begin{aligned} \partial_x H_z &= 0 && \text{at } x = 0, d \iff \\ \nabla H_z \cdot \boldsymbol{n} &= 0 && \text{at } x = 0, d, \end{aligned}$$

where $\boldsymbol{n}$ is the unit normal pointing out of the plate boundary.

Let us consider TE ($E_z = 0$) waves. By the aforementioned geometric symmetry or the waveguide, we know that (at least away from the dielectric discontinuity), we have $\partial_y H_z = 0$. We define

$$k_c := \sqrt{\omega^2 \mu \epsilon - k_z^2} \tag{6.3.14}$$

and return to solve the boundary value problem

$$\begin{cases} -\dfrac{\partial^2}{\partial x^2} H_z = & k_c^2 H_z \qquad \in \Omega \\ \partial_x H_z = & 0 \qquad x = 0, d, \end{cases} \tag{6.3.15}$$

which has general solution $H_z(x, y) = Ae^{ik_c x} + Be^{-ik_c x}$. Imposing the boundary conditions leads to the relation

$$k_c = \frac{n\pi}{d}, \quad n = 1, 2, 3... \tag{6.3.16}$$

and infinitely many solutions

$$H_z^n(x, y, z) = H_o \cos\big(\frac{n\pi}{d} x\big) e^{ik_z z}.$$

Similarly, we can consider the TM modes ($H_z = 0$) and solve

$$\begin{cases} -\dfrac{\partial^2}{\partial x^2} E_z = k_c^2 E_z \qquad \in \Omega \\ E_z = 0, \qquad x = 0, d, \end{cases} \tag{6.3.17}$$

for E_z. We again have that

$$k_c = \frac{n\pi}{d}, \quad n = 1, 2, 3... \tag{6.3.18}$$

for the TM modes, and get infinitely many solutions

$$E_z^n(x,y,z) = E_o \sin(\frac{n\pi}{d}x)e^{ik_z z}.$$

The other components of the TE and TM modes may be derived using relations 6.3.8–6.3.11.

Given a source or driving frequency ω, we are interested in the propagation behavior that results. The constant k_z governs this behavior, and, for both the TE and the TM modes, can be now be determined by using the relation $k_c = \frac{n\pi}{d}$. We have

$$k_z = \pm\sqrt{\omega^2\mu\epsilon - \big(\frac{n\pi}{d}\big)^2}. \tag{6.3.19}$$

The fact that k_z can be positive or negative is reflective of the fact that waves can travel down the waveguide in both directions. For a fixed n, if ω is such that $\omega^2\mu\epsilon > \frac{n\pi}{d}$, the corresponding mode will propagate without attenuation.

However, the mathematics raises the possibility that the wave number k_z might be an imaginary constant. If if ω is such that $\omega^2\mu\epsilon < \frac{n\pi}{d}$, then, for an appropriate α, we have $k_z = \pm i\alpha$. It may be surprising that the imaginary wave number still leads to a physically meaningful solution, but, at least for the positive root, this is indeed the case. For example, the z-component of the magnetic field takes the form

$$H_z^n(x,y,z) = H_o \cos\big(\frac{n\pi}{d}x\big)e^{-\alpha z}.$$

This wave decays exponentially as distance from its source increases. If the waveguide is long enough (one wavelength is sufficient according to [20]), these types of waves are can be omitted from the propagation analysis. The quantity k_c is referred to as the *cutoff frequency* of a particular waveguide. If ω is such that $\omega^2\mu\epsilon > k_c$, the corresponding wave modes propagate; if not, they decay exponentially. Note that

6.3.16 and 6.3.18 show that the cutoff frequencies are the same for the corresponding TE and TM modes for a parallel-plate waveguide.

The TEM mode has $H_z = E_z = 0$, which, referring to 6.3.8–6.3.11, implies that either all tangential field components are also 0 (no waves propagating), or that

$$k_z = \pm\omega\sqrt{\mu\epsilon}.$$

Consequently, we see from 6.3.14 that the cutoff frequency for any nontrivial TEM mode is 0. We investigate the existence of such a mode by first assuming that the waves are driven at a frequency low enough so that the conductor may be modeled as an equipotential surface. This is a standard and reasonable assumption [33], since our goal is to study the dominant mode of the waveguide–that mode with the lowest cutoff frequency. Let the potential of the top and bottom plate be 0 and V_o, respectively.

For the TEM mode, we have from 6.3.3 that $\mathcal{E}_z = 0$. With this, we see that $\nabla \times \mathcal{E} = 0$, and so we can write $\mathcal{E}$ as the (negative) gradient of a scalar potential function ϕ:

$$\mathcal{E} = -\nabla\phi.$$

The fact that no charges are present (so Gauss' Law gives $\nabla \cdot \boldsymbol{E} = 0$) indicates that this potential function satisfies Laplace's equation

$$\begin{cases} \Delta\phi = 0 & \in \Omega & \quad (6.3.20a) \\ \phi = 0 & x = 0, \forall y, \forall z & \quad (6.3.20b) \\ \phi = V_o & x = d, \forall y, \forall z. & \quad (6.3.20c) \end{cases}$$

The solution of this boundary value problem is $\phi = V_o x$; then we have

$$\mathcal{E}(x,y) =< -V_o, 0, 0 > \qquad \Longrightarrow \tag{6.3.21}$$

$$\boldsymbol{E}(x,y,z) = -V_o e^{ik_z z}\hat{\boldsymbol{x}}. \tag{6.3.22}$$

Finally, we can calculate H from 6.3.6b to conclude

$$\boldsymbol{H}(x,y,z) = \hat{\boldsymbol{y}} H_y = \frac{-V_o z_z}{\mu\omega} e^{ik_z z}\hat{\boldsymbol{y}} = H_o e^{-ik_z z}\hat{\boldsymbol{y}}, \tag{6.3.23}$$

with $H_o := \dfrac{-V_o z_z}{\mu\omega}$, in agreement with 6.3.1.

We now return to the problem posed by Jin in [20] of a parallel-plate waveguide with a dielectric discontinuity. Jin assumes that the waveguide functions at low frequency so that only the dominant mode of the wave propagates–the TEM mode; the previous analysis shows that this is valid as long as the wavenumber $k_z < \pi/d$.

At a distance (far enough) to the left of the discontinuity, Jin approximates the (y-component of the) wave as the sum of the incident wave and the part of the wave reflected by the dielectric:

$$u = u^{inc} + u^{ref} = H_o e^{-ikz} + RH_o e^{ikz}, \tag{6.3.24}$$

where H_o is a known constant related to the amplitude of the wave, k is the wave number, and R is the reflection coefficient. Similarly, to the right of the discontinuity, the part of the wave that continues to propagate is that which is not reflected, but is transmitted past the junction with the dielectric rod:

$$u = u^{trans} = TH_o e^{-ikz}, \tag{6.3.25}$$

where T is the transmission coefficient. Again, the previous analysis shows that this is reasonable; if the driving frequency is below the $n = 1$ cutoff frequency, only the TEM mode of the given form will propagate without attenuation, and thus all other modes are negligible at the left and right boundary.

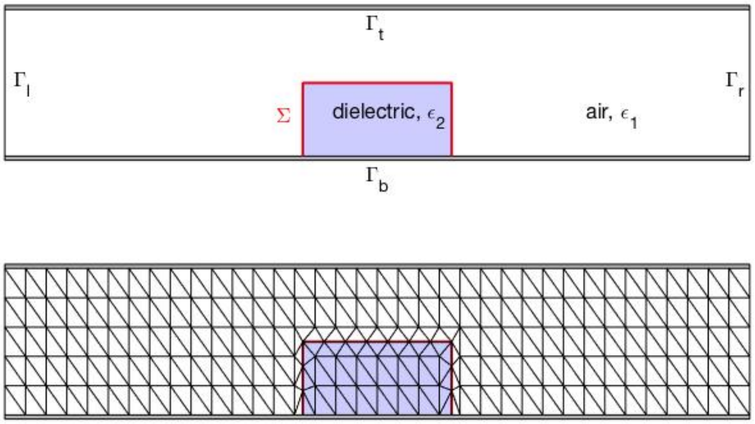


Figure 6.9: A schematic of the waveguide considered in 6.3.32 from Jin[20], top; a triangulation of the domain, with triangle boundaries along the material interface Σ. The width of the waveguide is taken to be 25cm; its height is 3.5cm. The width of the dielectric rod is 5cm; its height varies in the experiments below.

To determine how the propagating wave will interact with the dielectric discontinuity, we must solve the reduced-wave equation that results,

$$\nabla \cdot (\frac{1}{\epsilon_r}\nabla u) + k^2\mu_r u = 0 \qquad \in \Omega, \tag{6.3.26}$$

subject to boundary and continuity conditions arising from the physics of the setup.

At the waveguide walls (upper and lower boundary Γ_1), we have $\frac{\partial u}{\partial n} = 0$; on the far right boundary Γ_r, only the transmitted wave travels, so $\frac{\partial u}{\partial x} = -ikTH_o e^{-ikz} = -iku$. On the left boundary Γ_ℓ, we similarly calculate $\frac{\partial u}{\partial x} = -ikH_o e^{-ikz} + ikRH_o e^{ikz} = iku - 2ikH_o e^{-ikz}$. At the interface between the air and the dielectric rod Θ, the electromagnetic wave must be continuous, as is the component of its derivative that is parallel to the interface; but the perpendicular components on either side of the junction suffer a discontinuity related to the difference in the dielectric constant of the two materials:

$$\frac{1}{e_r^+}\frac{\partial u^+}{\partial n} = \frac{1}{e_r^-}\frac{\partial u^-}{\partial n}, \tag{6.3.27}$$

where the ± indicates the two sides of the material interface. This condition follows from the more standard continuity condition 6.3.28 applied to the time-harmonic Maxwell equations. For $\boldsymbol{n}$ pointing in the "positive" direction (from positive to negative), the condition is

$$\boldsymbol{n} \times (\boldsymbol{E}^+ - \boldsymbol{E}^-) = 0. \tag{6.3.28}$$

In the harmonic case $\nabla \times \boldsymbol{H} = -i\omega\epsilon\boldsymbol{E}$, so we have

$$\frac{1}{\epsilon^+}\big(\boldsymbol{n} \times \nabla \times \boldsymbol{H}^+\big) = \frac{1}{\epsilon^-}\big(\boldsymbol{n} \times \nabla \times \boldsymbol{H}^-\big).$$

Here, when there is no y-variation in the fields in question, and for $\boldsymbol{n} = [n_1; n_2; n_3]$; we compute

$$\nabla \times \boldsymbol{H} = \begin{bmatrix} -\partial_z H_y \\ \partial_z H_x - \partial_x H_z \\ \partial_x H_y \end{bmatrix}, \tag{6.3.29}$$

so

$$\boldsymbol{n} \times \nabla \times \boldsymbol{H} = \begin{bmatrix} n_2\partial_x H_y - n_3(\partial_z H_x - \partial_x H_z) \\ -n_1\partial_x H_y - n_3\partial_z H_y \\ n_1(\partial_z H_x - \partial_x H_z) + n_2\partial_z H_y \end{bmatrix}. \tag{6.3.30}$$

For the all of the dielectric obstructions discussed here, we have $n_2 \equiv 0$; this means that the only condition on H_y comes from the second component of 6.3.30. With the condition on n_2, this can be compactly written as $-\boldsymbol{n} \cdot \nabla H_y$. With 6.3.29, this leads to

$$\frac{1}{\epsilon^+}\frac{\partial H_y^+}{\partial n} = \frac{1}{\epsilon^-}\frac{\partial H_y^-}{\partial n}, \tag{6.3.31}$$

which is the condition 6.3.27 from [20].

In traditional finite element schemes, condition 6.3.27 is satisfied variationally [20]. Using spline functions allows the flexibility to enforce this condition explicitly as a modified smoothness condition. The implementation is straightforward and requires only that we triangulate the domain so that the interface boundary does not cross the interior of any triangles; that is, we require that this interior boundary be covered by edges of triangles in our triangulation. Note that the triangulation in Fig. 6.9 satisfies this property. Summarizing, and with reference to the figure, we have

$$\begin{cases} \nabla \cdot (\frac{1}{\epsilon_r}\nabla u) + k^2\mu_r u = 0 & \text{in } \Omega \quad (6.3.32a) \\ \frac{\partial u}{\partial n} = 0 & \text{on } \Gamma_1 \quad (6.3.32b) \\ \frac{\partial u}{\partial n} + iku = 2ikH_o e^{-ikz} & \text{on } \Gamma_\ell \quad (6.3.32c) \\ \frac{\partial u}{\partial n} + iku = 0 & \text{on } \Gamma_r \quad (6.3.32d) \\ \frac{1}{e_r^+}\frac{\partial u^+}{\partial n} = \frac{1}{e_r^-}\frac{\partial u^-}{\partial n} & \text{on } \Sigma, \quad (6.3.32e) \end{cases}$$

where ϵ_r is a discontinuous function giving the relative permittivity of the material throughout Ω.

Jin's first experiment is to determine the behavior of the electromagnetic field near the dielectric obstruction. He assumes that the waveguide is driven so that the electromagnetic wave propagates with wavelength $\lambda = 10cm$ (so the wavenumber

$k = 2\pi/10$). He takes the dielectric rod to have a rectangular cross-section of height 1.75cm, and considers dielectrics with 3 distinct relative permittivities: $\epsilon_2 = 4$, $4+1i$, $4+10i$. We seek to replicate his results.

We begin by demonstrating that our numerical scheme is accurate by performing the experiment with $H_o = 1$, $\mu_r = 1$, and $\epsilon_r = 1$, in which case 6.3.32 has the exact analytic solution

$$u(x, y) = e^{-ikz}.$$

We used the same wavenumber $k = \frac{2\pi}{10}$ as described above, and solve in the complex spline space $\mathbb{S}_5^1(\Omega)$ over a triangulation with 2011 triangles. The maximum error as evaluated over a grid of over one million points is 1.1517×10^{-5}; the root mean square error is 6.2924×10^{-6}. Contour plots of the real and imaginary part of the spline solution are shown in Fig. 6.10.

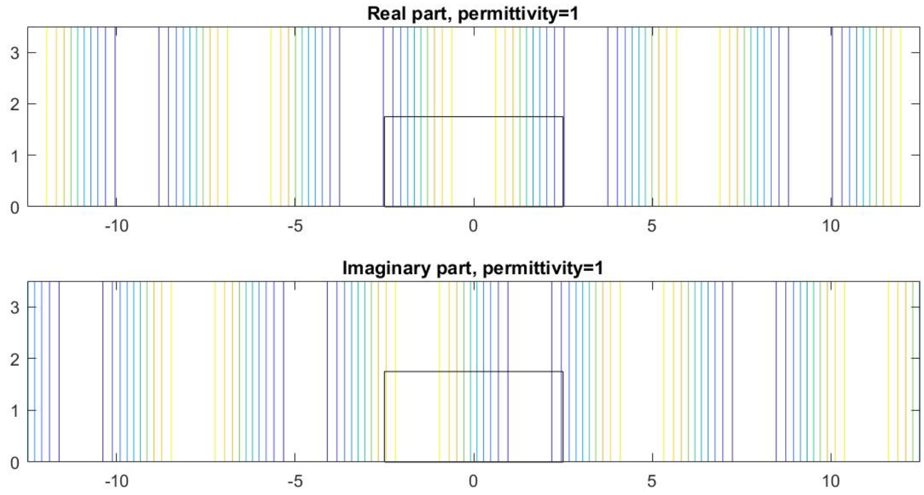


Figure 6.10: Contour plots of the real and imaginary part of the spline solution to boundary value problem 6.3.32 with $\epsilon_r = 1$, which has analytic solution $u = e^{-ikz}$. The spline solution in $\mathbb{S}_5^1$ has maximum pointwise error 1.1517×10^{-5}.

We now return to the case where the permittivity ϵ_2 is different from the permittivity ϵ_1 of the surrounding air. We include countour plots of the real and imaginary parts of Jin's finite element solutions in Fig. 6.11 for comparison with our spline solutions in Fig. 6.12, Fig. 6.13, and Fig. 6.14.

We note that the real part of the spline solutions up well with the imaginary parts of Jin's, and vice versa. We suspect that any differences in the contours shown are mostly due to the contouring algorithm rather than any substantial numerical differences. We posit that Jin has mislabeled his plots in [20], and refer again to the the accuracy of the spline solution seen in Fig. 6.10 to support our claim that the spline figures are correctly labeled.

We also present numerical data to demonstrate that the condition 6.3.27 is exactly and correctly enforced by the spline method, and compare the level of accuracy to that of a continuous finite element where the condition is enforced only variationally. Letting u_s be the computed numerical solution to the boundary value problem 6.3.32, the shows the difference between the ratio of normal derivatives along each edge of Σ and the ratio of electric permittivities. That is, referring to 6.3.27, we calculate

$$\left|\left(\frac{\partial u_s^+}{\partial n}\right)/\left(\frac{\partial u_s^-}{\partial n}\right) - \frac{\epsilon_r^+}{\epsilon_r^-}\right|. \tag{6.3.33}$$

Of course, if 6.3.27 is exactly satisfied, 6.3.33 will be exactly zero. The numerical results shown in Table 6.1 demonstrate that the spline method with modified smoothness condition satisfies the continuity condition almost exactly, and with much more accuracy than the variational approach. This explicit enforcement of the continuity condition is new, to our knowledge, and should produce a more accurate solution globally.

Next, Jin investigates the reflectance and transmittance of the electromagnetic wave with as the height of the dielectric rod in the waveguide *varies* from 0 to 3.5cm,

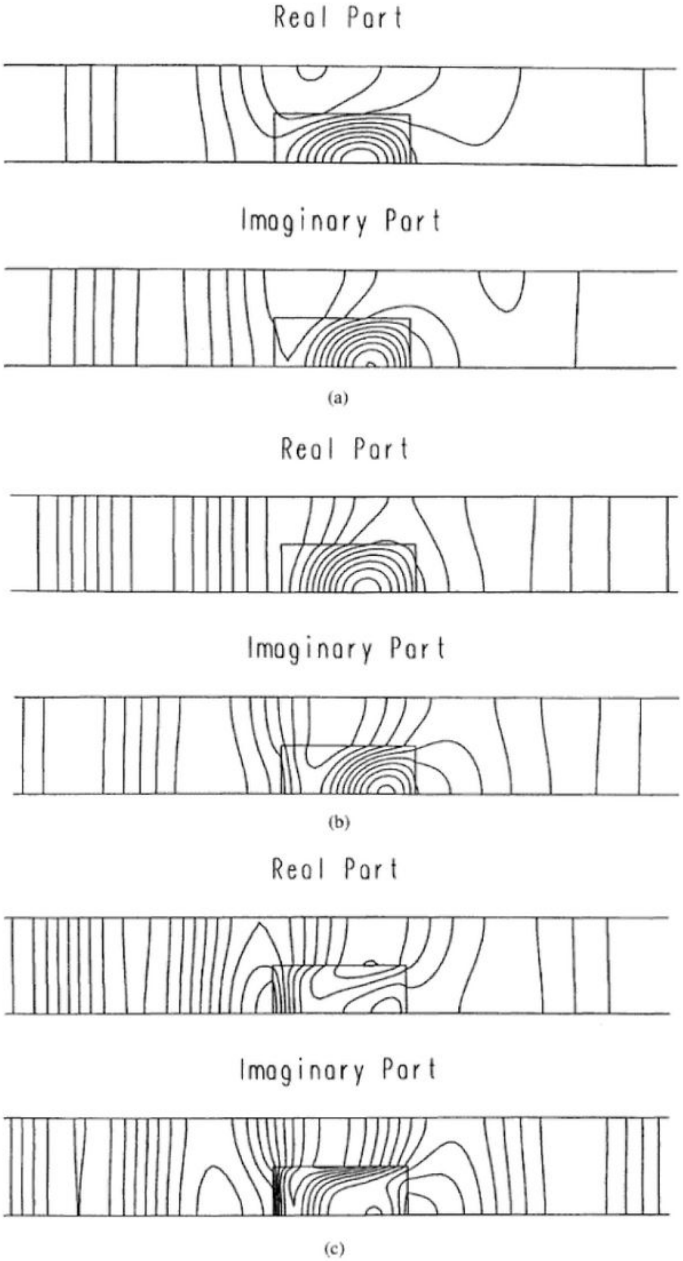


Figure 6.11: The finite element solutions to 6.3.32 from [20]. The coutour plots of the solutions where $\epsilon_2 = 4$, $4 - 1i$, and $4 - 10i$ appear in subfigures a), b), and c) respectively.

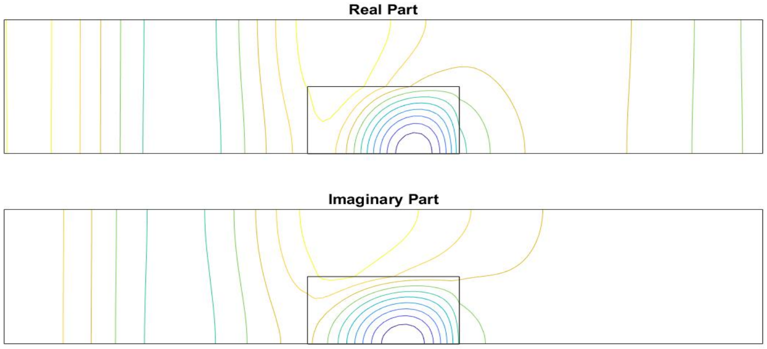


Figure 6.12: Contour plots of the real and imaginary parts of the spline solution in $\mathbb{S}_5^1$ to 6.3.32 for $\epsilon_2 = 4$. Compare to subfigure a) of Fig. 6.11.

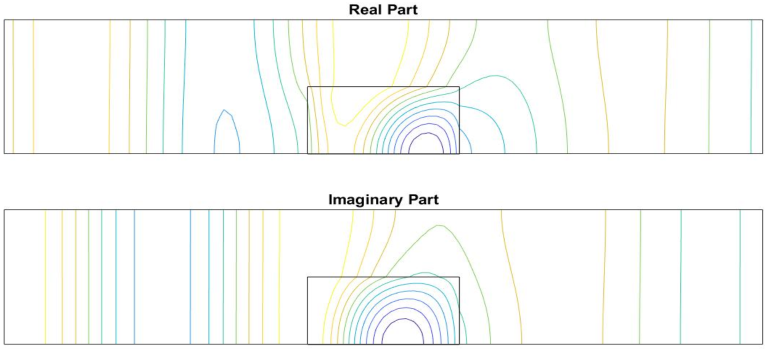


Figure 6.13: Contour plots of the real and imaginary parts of the spline solution $s \in \mathbb{S}_5^1$ to 6.3.32 for $\epsilon_2 = 4 - 1i$. Compare to subfigure b) of Fig. 6.11.

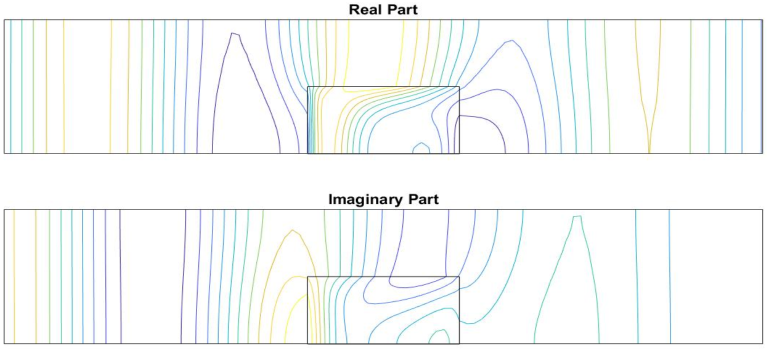


Figure 6.14: Contour plots of the real and imaginary parts of the spline solution $s \in \mathbb{S}_5^1$ to 6.3.32 for $\epsilon_2 = 4 - 10i$. Compare to subfigure c) of Fig. 6.11.

Table 6.1: Comparison of the accuracy of the interface condition enforced explicitly via modified spline smoothness conditions and variationally. The table on the left contains values corresponding to a spline solution to 6.3.33 with explicit enforcement. The table on the right shows the same results, but for a spline solution to 6.3.32 where 6.3.27 is enforced only variationally, as in [20]. The values are computed at the midpoints of the three edges of the dielectric.

	$\epsilon_r = 4$	$\epsilon_r = 4 - 1i$	$\epsilon_r = 4 - 10i$
Top	1.460e-13	1.354e-13	6.416e-14
Right	3.151e-13	2.004e-13	1.281e-13
Left	1.220e-13	2.330e-13	1.384e-13

	$\epsilon_r = 4$	$\epsilon_r = 4 - 1i$	$\epsilon_r = 4 - 10i$
Top	4.054e-04	2.261e-04	3.136e-04
Right	6.284e-05	5.600e-05	8.614e-05
Left	1.277e-04	1.889e-04	2.355e-04

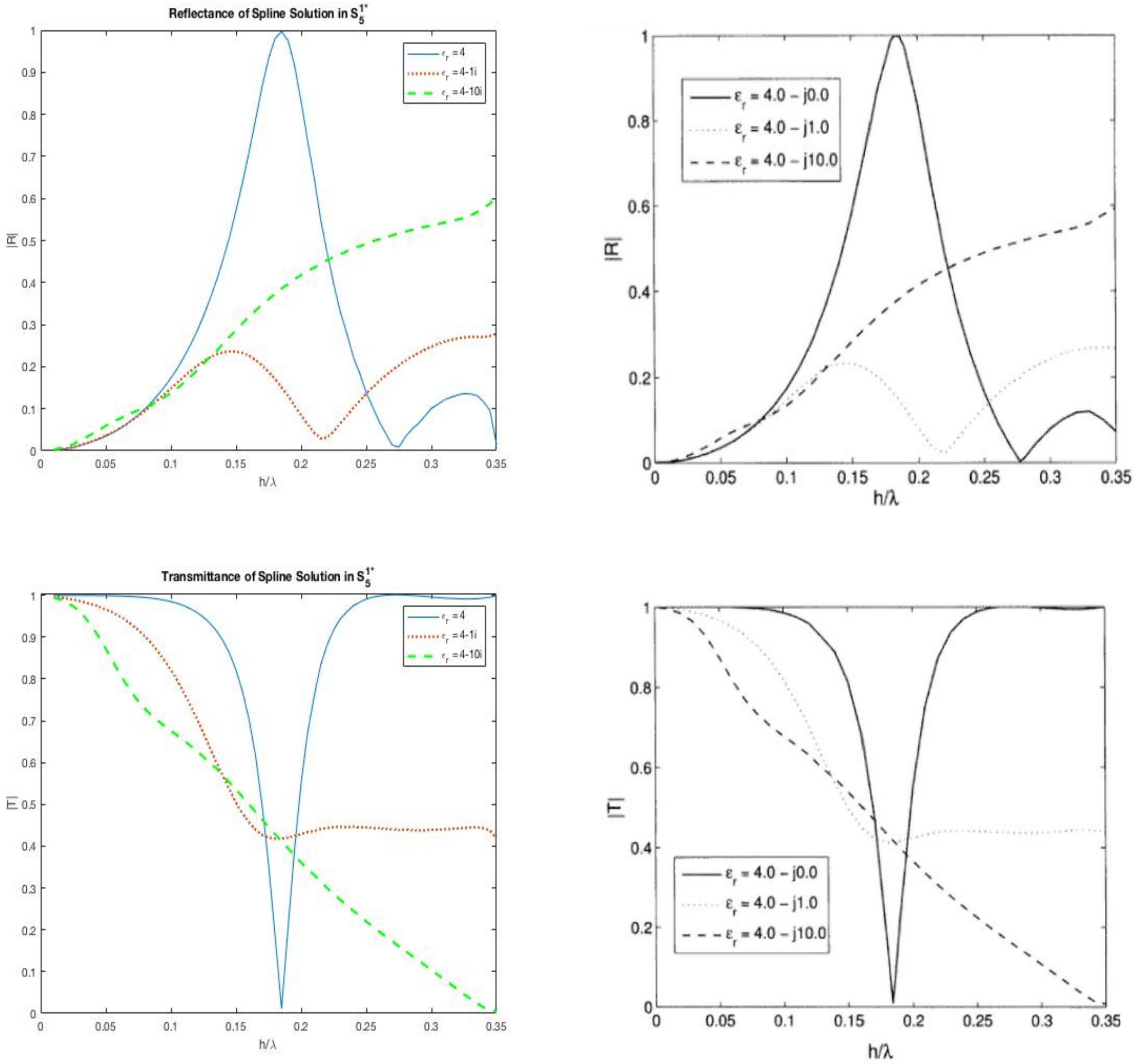
which is the height of the waveguide. Once H_y is determined by solving the boundary value problem 6.3.32, the coefficients can be calculated from 6.3.24 and 6.3.25, evaluated at the left- and right-hand sides of the waveguide, respectively. The experiment is repeated for dielectrics of the three relative permittivities mentioned previously. The dielectric material is called *lossless* if $\Im(\epsilon) = 0$, and, in that situation, the reflection coefficient R and transmission coefficient T satisfy

$$|R|^2 + |T|^2 = 1. \tag{6.3.34}$$

This relation gives us a method by which we can verify our calculations in the lossless case; computing the difference as in 6.3.33:

$$\left||R|^2 + |T|^2 - 1\right| \tag{6.3.35}$$

The plots of the magnitudes of the reflection and transmisison coefficiens of the spline solutions can be found in Fig. 6.20. For comparison, we have included the corresponding plots from [20].


Reflectance of Spline Solution in $S_5^{1^*}$

Transmittance of Spline Solution in $S_5^{1^*}$

Figure 6.15: Comparison of the plots of $|R|$ and $|T|$ from the spline solution $s \in \mathbb{S}_5^{1^*}$ and the plots from Jin. The spline plots from $\mathbb{S}_5^{1^*}$ are on the left, where the 1^* indicates that the spline solution is $\mathcal{C}^1$ everywhere except along the interface Σ, where 6.3.27 holds. The plots from the literature [20] are on the right. The images show that the spline solution reproduces the established result quite well.

The only discernible differences between the spline plots and Jin's come as the height of the dielectric bar approaches the height of the waveguide itself, particularly in the reflection coefficient in the case where $\epsilon_2 = 4$. Even as Jin's $|T|$ approaches 1 as the ratio h/λ approaches 0.35, it seems that the value of $|R|$ computed from the finite element solution hovers around $|R| = 0.1$, so it is unlikely that 6.3.34 would exactly or even nearly hold. When $h/\lambda = .35$, the spline solution yields R_s and T_s such that $\left|1 - (|R_s|^2 + |T_s|^2)\right| = 8.0087 \times 10^{-9}$.

Next, we extend the existing analysis to investigate the reflection/transmission phenomenon for electromagnetic waves of varying frequency. For the moment, we consider a parallel-plate waveguide with dielectric discontinuity of the same dimensions as the one seen in Fig. 6.9. If the boundary value problem described in 6.3.32 is to continue to guide our analysis, we must refer to 6.3.14 to find an upper limit for the wavenumbers we consider. Since we only wish to investigate the waveguide's dominant TEM mode, we must have

$$k < \frac{\pi}{d} = \frac{\pi}{3.5} \approx .8976. \tag{6.3.36}$$

We remark that in this particular case, the numerics themselves led us to the condition in 6.3.36. Experimenting with $k > .89$ in the case where $\epsilon_2 = 4$ led to solutions with $|R|$ and $|T|$ that came nowhere close to satisfying 6.3.34. We hypothesize that the dielectric material excites higher modes when the wavenumber is this large, and those modes propagate down the waveguide, making the boundary conditions 6.3.32c and 6.3.32d invalid. This is a good question to investigate with future research.

The reflection and transmission coefficients generated from spline solutions in $\mathbb{S}_5^{1^*}$ are displayed in plots below. We allow the wavenumber to vary from $k = .2$ to $k = .89$, corresponding to wavelengths varying from as large as 35 to as small as 7cm. In Fig. 6.16 we display the $|R|$ and $|T|$ plots for the lossless case; in Fig. 6.17 we assume the

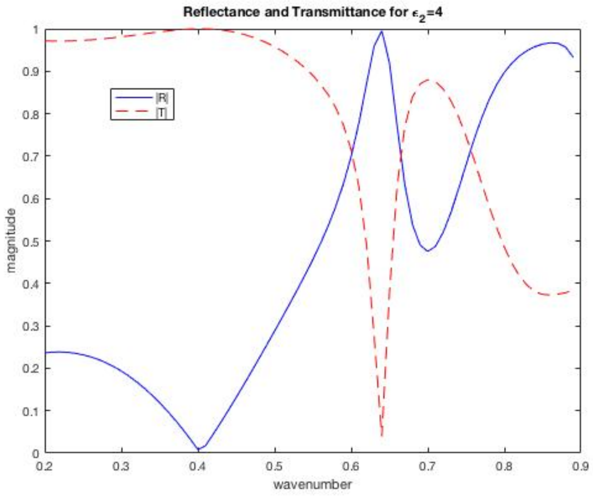


Figure 6.16: The plots of the $|R|$ and $|T|$ computed from the spline solutions in $\mathbb{S}_5^{1^*}$ as the wavenumber k varies from 0.2 to 0.9 with $\epsilon_r = 4$.

dielectric is a lossy material with the the same complex permittivities as the previous experiment.

In Table 6.2 we show how close the reflectance and transmittance of the spline approximation to 6.3.32 come to satisfying relation in 6.3.34. We have great agreeance as long as the wavenumber is small enough so that the wave's frequency is below the cutoff frequency for this waveguide. In the absence of any analytic solution or established results to compare against, this table valuable evidence that the data presented in plots Fig. 6.16 and Fig. 6.17 is accurate. After wavenumber crosses the cutoff threshold, the relation is not nearly satisfied, signifying the breakdown of this numerical approach. This is also a positive outcome; we can detect strange numerical behavior in a situation where our spline solution *should not* describe the physics of the waveguide. This behavior can help prevent an inappropriate application of our numerical methods.

We further exhibit the utility and flexibility of our numerical method by performing experiments with dielectric obstacles of different geometries. As seen in Fig. 6.18,

Table 6.2: Absolute error in the 6.3.34 for the reflection and transmission coefficients $|R|$ and $|T|$ calculated from the spline solutions to 6.3.32.

wavenumber k	$\lvert 1-(\lvert R\rvert^2+\lvert T\rvert^2)\rvert$	wavenumber k	$\lvert 1-(\lvert R\rvert^2+\lvert T\rvert^2)\rvert$
0.20	1.02e-07	0.60	3.25e-07
0.25	1.46e-07	0.65	1.93e-06
0.30	1.80e-06	0.70	2.76e-07
0.35	8.31e-07	0.75	4.02e-07
0.40	3.92e-07	0.80	3.09e-07
0.45	1.51e-08	0.85	4.27e-07
0.50	2.54e-07	0.90	5.81e-01
0.55	2.38e-07	0.95	1.25e+00

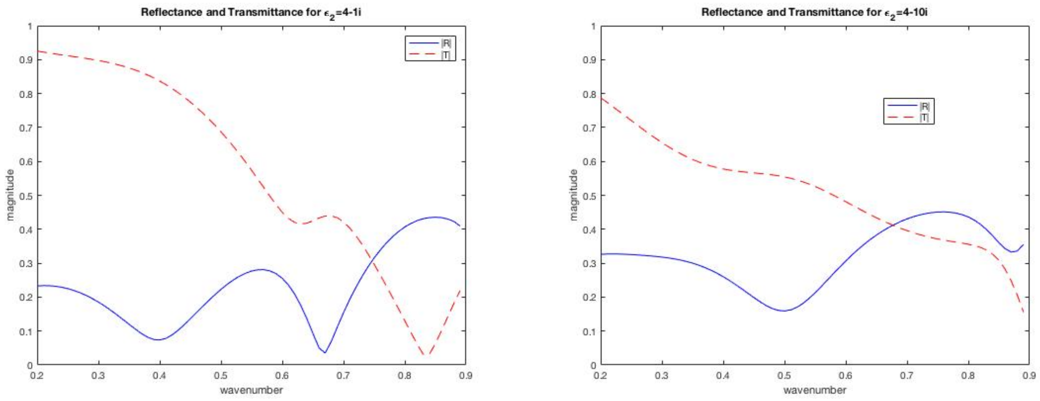


Figure 6.17: The plots of the $|R|$ and $|T|$ computed from the spline solutions in $\mathbb{S}_5^{1*}$ as the wavenumber k varies from 0.2 to 0.9 with lossy dielectrics. On the left, $\epsilon_r = 4-1i$; on the right, $\epsilon_r = 4-10i$.

we first consider dielectrics of relatively simple geometries, one consisting of three thin strips, and one dielectric rod with triangular cross section. The width of the dielectric strips is 1cm, and they are separated by 1cm; the base of the triangle is 4cm long. Table 6.4 shows the accuracy of the spline solutions with respect to 6.3.27 for explict and variational enforcement of the condition for these dielectrics. Within this table, the tables on the left show this error at the midpoint of each edge of the dielectric strips. On the right, the tables show the error at various points spread along the inclined edges of the triangular dielectric. The accuracy of the modified spline smoothness condition surpasses the standard variational enforcement.

Next, the heights of both shapes of dielectrics are allowed to vary, and we compute the reflection and transmission coefficients for these geometries as in Fig. 6.20. Finally, we introduce a more complicated, multilayer dielectric in Fig. 6.21, and, instead of changing the size of the obstruction, we allow the wavenumber to vary from 0 to the cutoff frequency.

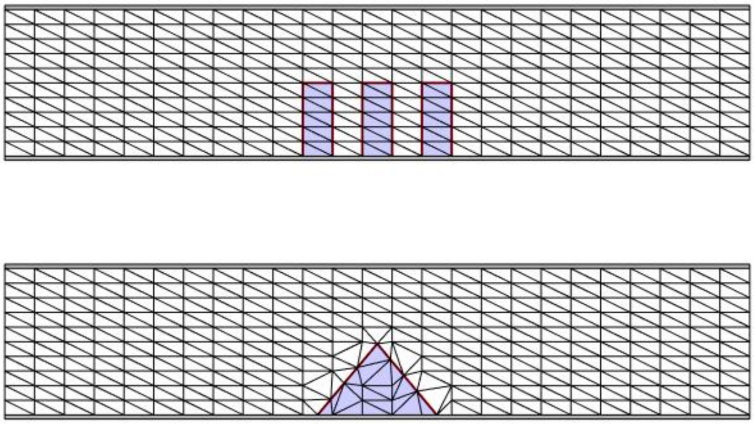


Figure 6.18: Triangulations of waveguide with dielectric obstructions of different geometries. We allow the heights of the obstructions to vary as in the experiment in Jin.

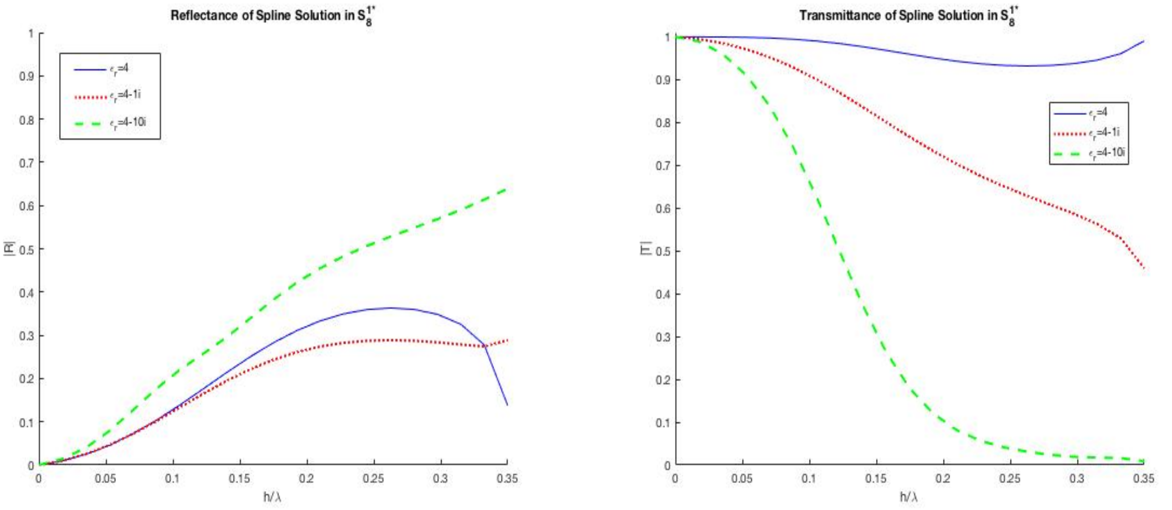


Figure 6.19: The plots of $|R|$ and $|T|$ computed from the spline solution in $\mathbb{S}_8^{1^*}$, calculated as the height of the strip dielectrics from Fig. 6.18 varies from 0 to 3.5.

We observe that the both the geometry and the height of the dielectric clearly affect the portion of the wave's power that is transmitted or reflected. For the fixed wavenumber $k = 2\pi/10$, unlike Jin's experiment, there is no dielectric height at which full reflection occurs. In general, it seems the larger the imaginary part of the medium's relative permittivity, the smaller the transmission coefficient. The inverse, however, does not always hold for the portion of the wave that is reflects.

As before, since we have no analytic solution or existing results with which to compare our spline solutions, we seek to validate our calculations with relation 6.3.35 and 6.3.33.

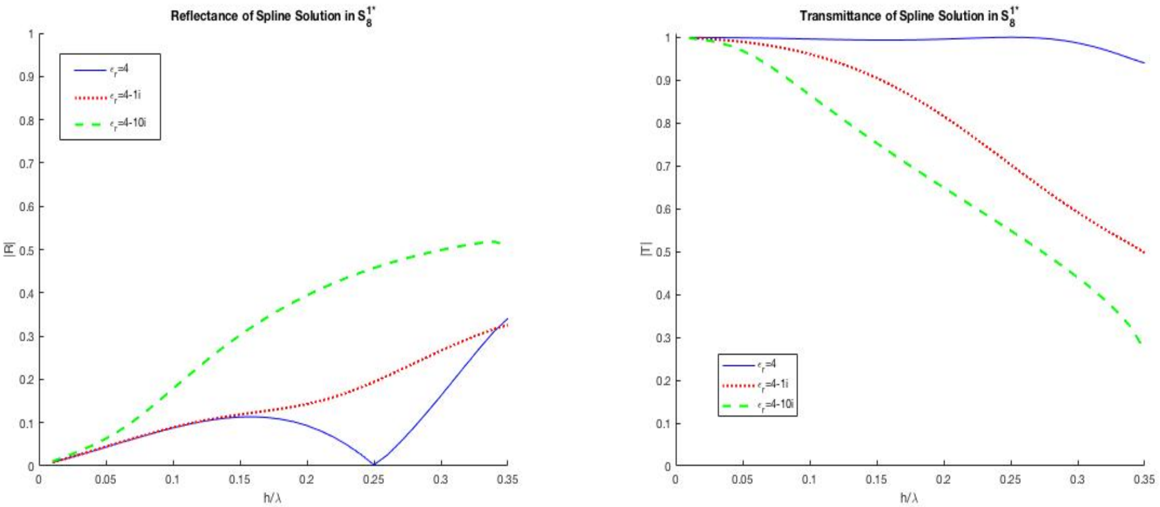


Figure 6.20: The plots of $|R|$ and $|T|$ computed from the spline solution in $\mathbb{S}_8^{1^*}$, calculated as the height of the triangular dielectric in Fig. 6.18 varies from 0 to 3.5.

Table 6.3: The results of 6.3.35 as the heights of the dielectric obstructions seen in Fig. 6.18 vary from 0 to 3.5. The results for the domain with dielectric strips are on the left; the results for the dielectric triangle are on the right. In both cases, the relation is satisfied quite well by the spline solution.

Strip Dielectric

height	$\lvert 1-(R^2+T^2)\rvert$
0.3	3.86e-07
0.7	6.48e-07
1.1	5.56e-08
1.5	7.19e-08
1.9	3.24e-07
2.3	2.69e-08
2.7	6.59e-07
3.1	2.32e-07
3.5	1.58e-06

Triangular Dielectric

height	$\lvert 1-(R^2+T^2)\rvert$
0.3	2.20e-07
0.7	2.26e-07
1.1	3.08e-07
1.5	4.32e-08
1.9	4.13e-07
2.3	6.56e-07
2.7	3.58e-07
3.1	2.32e-07
3.5	8.96e-07

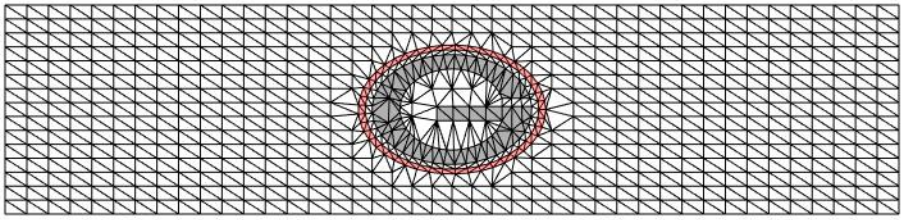


Figure 6.21: Triangulation of waveguide with a complicated, multilayer dielectric obstruction.

Table 6.4: Error in the spline solutions' satisfaction of the interface condition 6.3.27 for various dielectric geometries

Modified Spline Smoothness Condition, $\epsilon_2 = 4$

Three Strip Dielectric

	Left Strip	Center Strip	Right Strip
Left Edge	6.026e-13	3.294e-13	5.837e-13
Top Edge	1.911e-12	2.031e-12	1.911e-12
Right Edge	3.669e-13	5.226e-13	3.300e-13

Triangular Dielectric

Left Edge	Right Edge
2.552e-13	2.442e-13
1.487e-13	1.186e-13
1.216e-13	1.150e-13

Variational Enforcement, $\epsilon_2 = 4$

Three Strip Dielectric

	Left Strip	Center Strip	Right Strip
Left Edge	2.712e-06	3.295e-06	3.630e-06
Top Edge	6.597e-04	1.169e-03	6.597e-04
Right Edge	3.782e-06	3.288e-06	2.771e-06

Triangular Dielectric

Left Edge	Right Edge
6.924e-06	7.769e-05
2.079e-07	2.309e-07
6.735e-05	3.650e-05

Modified Spline Smoothness Condition, $\epsilon_2 = 4 - 1i$

Three Strip Dielectric

	Left Strip	Center Strip	Right Strip
Left Edge	5.142e-13	3.068e-13	5.351e-13
Top Edge	3.572e-12	1.647e-12	3.572e-12
Right Edge	3.964e-13	4.922e-13	3.349e-13

Triangular Dielectric

Left Edge	Right Edge
2.991e-13	2.678e-13
2.067e-13	9.093e-14
1.640e-13	2.352e-13

Variational Enforcement, $\epsilon_2 = 4 - 1i$

Three Strip Dielectric

	Left Strip	Center Strip	Right Strip
Left Edge	2.778e-06	3.300e-06	3.597e-06
Top Edge	7.167e-04	1.221e-03	7.167e-04
Right Edge	3.767e-06	3.454e-06	2.703e-06

Triangular Dielectric

Left Edge	Right Edge
8.015e-06	8.138e-05
2.092e-07	2.548e-07
5.984e-05	4.714e-05

Modified Spline Smoothness Condition, $\epsilon_2 = 4 - 10i$

Three Strip Dielectric

	Left Strip	Center Strip	Right Strip
Left Edge	2.157e-13	1.562e-13	1.657e-13
Top Edge	3.136e-12	1.440e-12	3.136e-12
Right Edge	4.324e-13	1.997e-12	1.683e-13

Triangular Dielectric

Left Edge	Right Edge
2.431e-13	2.359e-13
2.092e-13	1.940e-13
1.573e-13	6.578e-13

Variational Enforcement, $\epsilon_2 = 4 - 10i$

Three Strip Dielectric

	Left Strip	Center Strip	Right Strip
Left Edge	3.595e-06	2.772e-06	4.082e-06
Top Edge	1.472e-03	2.173e-03	1.472e-03
Right Edge	9.070e-06	6.258e-05	4.066e-06

Triangular Dielectric

Left Edge	Right Edge
6.987e-06	4.264e-05
4.402e-07	4.161e-07
3.639e-05	1.309e-04

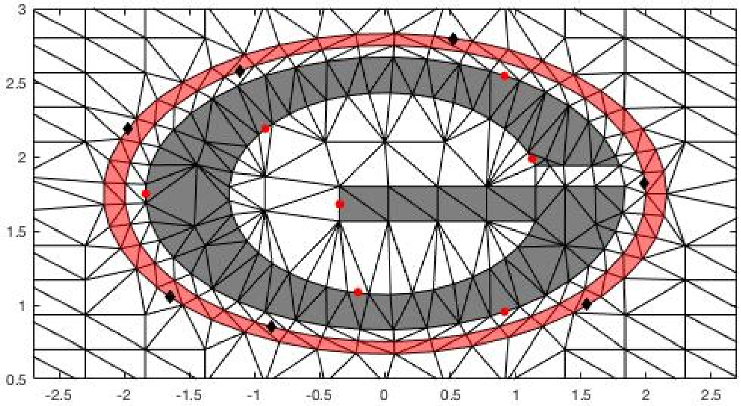


Figure 6.22: A closeup view of the multilayer dielectric. We test the accuracy of the spline solution with respect to continuity condition 6.3.27 at the locations shown. We test the inner dielectric at the locations marked by red dots, and the outer dielectric at the locations marked by black diamonds. The results can be seen in Table 6.6.

Table 6.5: Error in relation 6.3.34 for R and T computed from the spline solution in $S_5^{1^*}(\triangle)$ with dielectric $\epsilon_r = 4$. As in the previous experiment, the relation breaks down as the wavenumber grows past the cutoff frequency.

wavenumber	$\|1-(\|R\|^2+\|T\|^2)\|$
0.20	6.105e-06
0.25	1.758e-06
0.30	5.750e-07
0.35	1.070e-06
0.40	2.956e-06
0.45	1.694e-06
0.50	5.656e-06
0.55	1.063e-06

wavenumber	$\|1-(\|R\|^2+\|T\|^2)\|$
0.60	1.756e-06
0.65	4.148e-06
0.70	5.142e-06
0.75	2.247e-05
0.80	2.361e-05
0.85	4.198e-06
0.90	1.769e-05
0.95	1.870e-03

Table 6.6: Comparison of the spline modified smoothness condition to variational enforcement of the relation 6.3.27 for the multilayer dielectric shown in Fig. 6.21.

Multilayer Dielectric: Inner $\epsilon_r = 4$, Outer $\epsilon_r = 2$

Modified Smoothness Condition

Inner G Dielectric	Outer Dielectric
2.379e-11	2.379e-11
5.221e-12	5.221e-12
9.603e-12	9.603e-12
1.022e-11	3.410e-11
7.636e-12	2.924e-12
4.135e-11	3.329e-11
1.033e-11	6.712e-11

Variational Enforcement

Inner G Dielectric	Outer Dielectric
4.563e-03	4.563e-03
1.456e-03	1.456e-03
1.309e-02	1.309e-02
2.311e-02	2.917e-03
4.438e-05	1.082e-03
4.279e-03	5.733e-03
3.510e-05	5.604e-04

Multilayer Dielectric: Inner $\epsilon_r = 4 - 2i$, Outer $\epsilon_r = 2 - 1i$

Modified Smoothness Condition

Inner G Dielectric	Outer Dielectric
1.092e-11	1.092e-11
4.179e-12	4.179e-12
7.631e-12	7.631e-12
8.234e-12	1.973e-11
6.552e-12	4.124e-12
3.785e-11	2.251e-11
1.001e-11	2.445e-11

Variational Enforcement

Inner G Dielectric	Outer Dielectric
5.419e-03	5.419e-03
1.193e-03	1.193e-03
8.705e-03	8.705e-03
2.269e-02	1.653e-03
1.249e-04	1.003e-03
4.478e-03	4.453e-03
1.483e-04	4.166e-04

Multilayer Dielectric: Inner $\epsilon_r = 4 - 10i$, Outer $\epsilon_r = 2 - 5i$

Modified Smoothness Condition

Inner G Dielectric	Outer Dielectric
5.356e-12	5.356e-12
1.390e-12	1.390e-12
2.422e-12	2.422e-12
5.676e-12	6.724e-12
5.962e-12	2.400e-12
2.024e-11	4.706e-12
8.863e-12	1.041e-11

Variational Enforcement

Inner G Dielectric	Outer Dielectric
7.204e-03	7.204e-03
2.867e-04	2.867e-04
1.815e-03	1.815e-03
2.118e-02	1.317e-04
5.534e-04	4.225e-04
3.023e-03	1.218e-03
6.946e-04	2.604e-04

6.4 Wave Equation with Time-Periodic Source Terms

Next we extend the above study to situations in which the governing physics is time-periodic rather than strictly time-harmonic. In this setting, we expand the known functions $\tilde{f}(\mathbf{x}, t)$ and $\tilde{g}(\mathbf{x}, t)$ in their Fourier series to have

$$\begin{aligned} \tilde{f} &= \sum_{j\in\mathbb{Z}} f_j(\mathbf{x}) \exp(i\omega_j t), \quad \mathbf{x} \in \Omega \\ \tilde{g}(\mathbf{x}, t) &= \sum_{j\in\mathbb{Z}} g_j(\mathbf{x}) \exp(i\omega_j t), \quad \mathbf{x} \in \partial\Omega. \end{aligned}$$

Then our solution $\tilde{u}(\mathbf{x}, t)$ can be expressed as

$$\tilde{u}(\mathbf{x}, t) = \sum_{j\in\mathbb{Z}} u_j(\mathbf{x}) \exp(i\omega_j t), \quad \forall \mathbf{x} \in \Omega,$$

and by matching the Fourier coefficients, we have the Helmholtz boundary value problem

$$\begin{aligned} \Delta u_j(\mathbf{x}) + \frac{(\omega_j)^2}{c^2} u_j(\mathbf{x}) &= f_j(\mathbf{x}), \quad \mathbf{x} \in \Omega \\ \alpha \frac{\partial}{\partial n} u_j(\mathbf{x}) + \beta u_j(\mathbf{x}) &= g_j(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega \end{aligned} \tag{6.4.1}$$

for each $k \in \mathbb{Z}$.

We now describe a numerical scheme under the assumption that the source term and boundary conditions are band-limited. Let ω_{max} be the maximum frequency of interest. We shall use bivariate spline space $S_d^1(\triangle)$ to approximate u_j, where $\triangle$ is a triangulation of Ω. Then we sample the source $\tilde{f}(\mathbf{x}, t)$ and boundary function $\tilde{g}(\mathbf{x}, t)$ at times $t_j = j/N$, $h = 0, 1, ..., N-1$, where N is chosen according to the Nyquist sampling rate so that $N \geq 2\omega_{max}$. For use with the fast Fourier Transform (FFT), we choose $N = 2^j$ for some $j \in \mathbb{N}$ in practice[5].

We compute the discrete Fourier transform (FFT) of the time series corresponding to each domain point of $S_d^1(\triangle)$, and determine the frequencies which contribute to the spectrum at a magnitude greater than a given tolerance tol. For each such $\omega_j \leq \omega_{max}$, we solve (6.4.1) as in the previous sections. Exploiting the symmetry of the FFT of real time signal, we have $\omega_j = \overline{\omega_{N-j}}$. Finally, we apply the inverse FFT at each domain point to recover our time-domain solution.

Example 6.4.1. First, we solve a homogeneous wave equation over the unit hexagonal domain as in Example 5.1.1, scaled so that $\mu_0\epsilon_0 = 1$. The exact solution is given by

$$u(\mathbf{x}, t) = \sum_{n=1}^{3} \sin(5n\pi t)\big(\cos(5n\pi x) + \cos(5n\pi y)\big),$$

We apply Dirichlet boundary conditions, and solve in the space S_{10}^1 over a triangulation with $|h| = 0.1$. The time evolution of the approximate and exact wave at $(x, y) = (0, 0)$ is shown in Figure 6.23, as well as the approximate and exact wave over the entire domain at time $t = 1.64$. The maximum pointwise error, taken over all time in the period, is $8.8154e - 6$ which is an exellent approximation to the given exact solution.

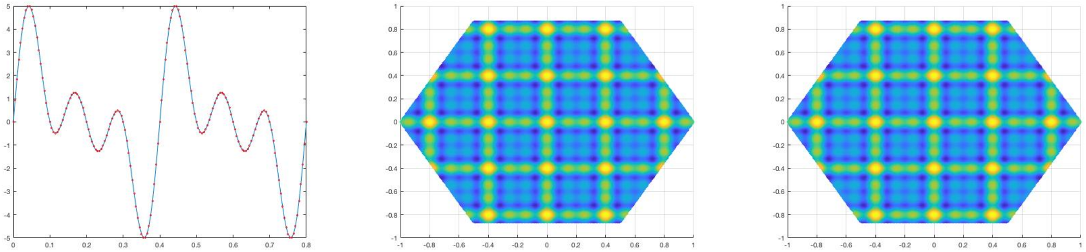


Figure 6.23: Time evolution of the height of the center point of the wave and snapshot of wave at $t = 1.64$. The center point $(0, 0)$ of the spline solution is given by the blue line and exact solution by red points, left; spline wave at $t = 1.64$ center; exact wave at same time shown right.

Our numerical experiments show that this FFT approach works well work a variety of homogeneous and inhomogeneous wave equations arising from periodic source terms and boundary conditions. We test our scheme in this final example by solving the wave equation with an exact solution that is not explicitly a sum of sinusoidal functions in time.

Example 6.4.2. We seek a spline solution $u_s \in S_5^1$ to the inhomogeneous wave equation with exact solution

$$u(\mathbf{x}, t) = \sin(x + y)e^{1/(t^2-2t)}$$

which is time-periodic with $T = 2$. Here we solve over the square domain $\Omega := [0,1] \times [0,1]$, and use Dirichlet boundary conditions. We run the experiment 3 times, using a triangulation with $|h| = 0.1$, and sampling the source functions at increasingly fine time intervals. The results of are summarized in Table 6.7; the errors reported are the maximum pointwise error taken over all time in $t = [0,2]$.

Table 6.7: Spline solutions to time-periodic wave equation based on FFT.

Length of signal	Sampling Freq.	Max Physical Freq.	Max err
32	16	8	4.5822e-01
64	32	16	4.6042e-02
128	64	32	2.3745e-04

In Fig. 6.24 we display the height of spline solution at a spatial location, say the 50^{th} domain point of our triangulation $(x, y) = (.28, .64)$ over the time period $t = [0,2]$. By sampling at a rate of 64 herz, we are able to generate a spline wave whose time evolution is indistinguishable from the exact solution.

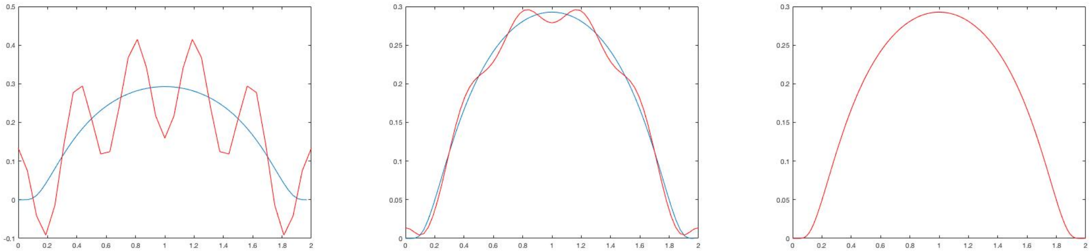


Figure 6.24: Time evolution of point on time-periodic wave for exact and spline solution generated by various sampling frequencies. The height of the point $(.28, .64)$ of the spline solution (red line) and exact solution (blue curve). Spline wave with maximum frequency component $\omega_{max} = 8$ left, spline wave with $\omega_{max} = 16$ center, and spline wave with $\omega_{max} = 32$ (right).

References

[1] *Wireless power transfer*, ANSYS Application Brief (2012).

[2] Gerard Awanou, Ming-Jun Lai, and Paul Wenston, *The multivariate spline method for scattered data fitting and numerical solutions of partial differential equations*, Wavelets and splines: Athens (2005), 24–74.

[3] ———, *The multivariate spline method for scattered data fitting and numerical solutions of partial differential equations*, Wavelets and splines: Athens (2005), 24–74.

[4] CR Boucher, Zehao Li, CI Ahheng, JD Albrecht, and LR Ram-Mohan, *Hermite finite elements for high accuracy electromagnetic field calculations: A case study of homogeneous and inhomogeneous waveguides*, Journal of Applied Physics **119** (2016), no. 14, 143106.

[5] Ronald Newbold Bracewell and Ronald N Bracewell, *The fourier transform and its applications*, vol. 31999, McGraw-Hill New York, 1986.

[6] Bree Ettinger, *Bivariate splines for ozone concentration predictions*, Ph.D. thesis, uga, 2009.

[7] Bree Ettinger, Serge Guillas, and Ming-Jun Lai, *Bivariate splines for ozone concentration forecasting*, Environmetrics **23** (2012), no. 4, 317–328.

[8] Xiaobing Feng and Haijun Wu, *Discontinuous galerkin methods for the helmholtz equation with large wave number*, SIAM Journal on Numerical Analysis **47** (2009), no. 4, 2872–2896.

[9] DO Forfar and CMath FIMA, *James clerk maxwell: Maker of waves*, Scotland's Mathematical Heritage: Napier to Clerk Maxwell, Edinburgh, UK (1995).

[10] Roland Griesmaier and Peter Monk, *Error analysis for a hybridizable discontinuous galerkin method for the helmholtz equation*, Journal of Scientific Computing **49** (2011), no. 3, 291–310.

[11] David J Griffiths, *Introduction to electrodynamics*, 2005.

[12] Juan B Gutierrez, Ming-Jun Lai, and George Slavov, *Bivariate spline solution of time dependent nonlinear pde for a population density over irregular domains*, Mathematical biosciences **270** (2015), 263–277.

[13] Hermann A Haus and James R Melcher, *Electromagnetic fields and energy (massachusetts institute of technology: Mit opencourseware)*, 1989.

[14] Qianying Hong, *Bivariate splines applied to variational model for image processing.*

[15] Stephen C Jardin, *A triangular finite element with first-derivative continuity applied to fusion mhd applications*, Journal of Computational Physics **200** (2004), no. 1, 133–152.

[16] Nédélec Jean-Claude, *elec. mixed finite elements in r3*, Numer. Math **35** (1980), 315–341.

[17] Bo-nan Jiang, *The least-squares finite element method: theory and applications in computational fluid dynamics and electromagnetics*, Springer Science & Business Media, 1998.

[18] Bo-Nan Jiang, Jie Wu, and Louis A Povinelli, *The origin of spurious solutions in computational electromagnetics*, Journal of computational physics **125** (1996), no. 1, 104–123.

[19] Jian-Ming Jin, *The finite element method in electromagnetics*, John Wiley & Sons, 2015.

[20] CS Jog and Arup Nandy, *Mixed finite elements for electromagnetic analysis*, Computers & Mathematics with Applications **68** (2014), no. 8, 887–902.

[21] Fumio Kikuchi, *Theoretical analysis of nedelec's edge elements*, Japan journal of industrial and applied mathematics **18** (2001), no. 2, 321.

[22] Ming-Jun Lai and Larry L Schumaker, *Spline functions on triangulations*, vol. 110, Cambridge University Press, 2007.

[23] Carlos A Leal-Sevillano, Jorge A Ruiz-Cruz, José R Montejo-Garai, and Jesús M Rebollar, *Rigorous analysis of the parallel plate waveguide: From the transverse electromagnetic mode to the surface plasmon polariton*, Radio Science **47** (2012), no. 6.

[24] DA Lowther and EM Freeman, *The application of the research work of james clerk maxwell in electromagnetics to industrial frequency problems*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **366** (2008), no. 1871, 1807–1820.

[25] Xiao Lu, Ping Wang, Dusit Niyato, Dong In Kim, and Zhu Han, *Wireless charging technologies: Fundamentals, standards, and network applications*, IEEE Communications Surveys & Tutorials **18** (2016), no. 2, 1413–1452.

[26] Jens Markus Melenk and Stefan Sauter, *Wavenumber explicit convergence analysis for galerkin discretizations of the helmholtz equation*, SIAM Journal on Numerical Analysis **49** (2011), no. 3, 1210–1243.

[27] Leopold Matamba Messi, *Theoretical and numerical approximation of the rudin-osher-fatemi model for image denoising in the continuous setting*, Ph.D. thesis, University of Georgia, 2012.

[28] Lin Mu, Junping Wang, Xiu Ye, and Shan Zhao, *A numerical study on the weak galerkin method for the helmholtz equation*, Communications in Computational Physics **15** (2014), no. 5, 1461–1479.

[29] Gerrit Mur, *Edge elements, their advantages and their disadvantages*, IEEE transactions on magnetics **30** (1994), no. 5, 3552–3557.

[30] ______, *The fallacy of edge elements*, IEEE Transactions on Magnetics **34** (1998), no. 5, 3244–3247.

[31] Gerrit Mur and Ioan E Lager, *On the causes of spurious solutions in electromagnetics*, Electromagnetics **22** (2002), no. 4, 357–367.

[32] David M Pozar, *Microwave engineering*, John Wiley & Sons, 2009.

[33] Talal Rahman and Jan Valdman, *Fast matlab assembly of fem stiffness-and mass matrices in 2d and 3d: nodal elements*, Applied Mathematics and Computation - AMC **219** (2013).

[34] Larry Schumaker, *Spline functions: basic theory*, Cambridge University Press, 2007.

[35] Dipak L Sengupta and Tapan K Sarkar, *Maxwell, hertz, the maxwellians, and the early history of electromagnetic waves*, IEEE Antennas and Propagation Magazine **45** (2003), no. 2, 13–19.

[36] George Petrov Slavov, *Bivariate spline solution to a class of reaction-diffusion equations*, Ph.D. thesis, uga, 2016.

[37] Stanimir S Valtchev, Elena N Baikova, and Luis R Jorge, *Electromagnetic field as the wireless transporter of energy*, Facta universitatis-series: Electronics and Energetics **25** (2012), no. 3, 171–181.

[38] Rikard Vinge, *Wireless energy transfer by resonant inductive coupling*, Master's thesis, Chalmers University of Technology (2015).