Numerical Solution of Ordinary and Partial Differential Equation using Frames

Mathematics Writing Sample
Research Report

Saad N Khan

August 31, 2020

Abstract

In this report we use Fourier frames to numerically solve ordinary differential equations (ODE) and partial differential equations (PDEs) on irregular bounded domains in two spatial dimensions. The numerical scheme we implement approximates the true solution using truncated Fourier frames and least squares minimization. Our primary focus is on PDEs whose differential operator is elliptic, linear, second order and constant coefficient. Based on the numerical result we see that our method is spectrally accurate when applied to ODEs. For PDEs in two dimension we see much better results for convex domains with infinity smooth boundary. We also outline how the numerical method can be applied to solve the Laplace eigenvalue problem on an irregular domain. Based on our results we are able to approximate the first few eigenvalues to a high order of accuracy. The third Chapter of this report uses the variational framework to analyse the numerical method. In one dimension the variational form corresponding to our numerical method has a unique solution and our numerical approximation is the best approximation from the space spanned by finite Fourier frames. Due to the bilinear form of the variational formulation not being coercive in higher dimension we modify our bilinear form and carry out a similar error analysis. Our main result with the modified bilinear form is an error estimate similar to Céa lemma where the constant in the inequality depends on the Poincaré constant, Sobolev and trace embedding constants and the dimension of the approximating subspace. In Appendix A we outline how to extend the numerical method to time dependent PDEs such as the advection equation, diffusion equation and discuss some of the issues that arise.

Background and Motivation

Elliptic differential operator are one of the most commonly encountered and studied operators in PDE theory with the Laplacian being the prototypical example. While several theoretical estimates for the solution of elliptic operator on

bounded domains exist there is no closed form formula. As a result several numerical methods have been developed to provide accurate and stable solutions. Finite difference and spectral methods are the most popular method for solving PDEs numerically on a square or rectangular domain. Based on the behaviour of the PDE one can construct a stable, accurate and to as high an order a finite difference method as desired. While in some instances finite difference schemes can be applied to solve PDE on arbitrary domains they are not well suited for it. Spectral methods require a basis for the function space in which the solution exist. The basis depends on the domain. There is no set algorithm to come up with a basis for a function space on a arbitrary domain. Additionally, a basis that works for one domain does not work for another.

The most popular method for solving PDEs on irregular domains is finite element methods. Finite element methods decompose the domain into a set of smaller domain called elements. Then using appropriate shape functions a system of equations is created to solve the weak or variational form associated to the PDE. The shape functions are chosen depending on the required regularity of the solution. For example piecewise linear functions are sufficient for second order elliptic operator however fourth order elliptic operator would require more smoothness thus a piecewise quadratic function is required. Implementing finite element method numerically can be complicated. There are several well known programs that implement finite element methods such as DUNE and FreeFEM. These programs are black box solver that require a few commands detailing the weak form, domain and shape functions to implement the finite element method.

In this report we present a numerical method that can solve PDEs on irregular domains and is much simpler to implement. The numerical method can be thought of as a finite element method with two elements the domain and its boundary. The shape functions is not a basis but rather a frame for our function space. Frames satisfy weaker assumptions and are less restrictive then bases. By uniformly sampling points inside the domain and on its boundary we construct a system of two sets of equations one for the domain and another for its boundary. We solve the linear system using least squares. As such the method is quite straightforward and easy to implement.

Outline

This report consists of three Chapter. As frames are central to our numerical method we begin by outlining frames, their properties and numerical approximation using frames. In the second chapter we outline the numerical method to solve ODEs and elliptic PDEs on irregular domain and present numerical results from several examples. Figures showing the irregular regions on which the PDEs were solved is provided in Appendix B. We then outline how the numerical method can be applied to solve the Laplace eigenvalue problem and provide a couple of examples. In Chapter 3 we outline the analysis of our numerical method using the variational framework. In Appendix A we outline

how the numerical method can be extended to time dependent PDEs such as advection and diffusion equation.

Notation

Before we begin we outline the notation used in this report. We will denote vectors in $\mathbb{C}^d$ and matrices in $\mathbb{C}^{M\times N}$ using bold letters.We will denote the j-th term of a vector $\boldsymbol{x} \in \mathbb{C}^d$ by x_j and the i-th row and j-th column of a matrix $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ by $\boldsymbol{A}_{ij}$. We will denote the zero vector and matrix whose elements are all zero by $\mathbf{0}$, where the distinction will be obvious from context. $\boldsymbol{A}^*$ will represent the complex conjugate of the matrix $\boldsymbol{A}$ and $\boldsymbol{A}^\dagger$ is the pseudo-inverse of a non-square matrix $\boldsymbol{A}$. Any bounded domain Ω can be rescaled to lie within the unit hypercube, as a result unless otherwise stated Ω will be an open simply connected subset of $[-1,1]^d$ with boundary $\partial\Omega$. In this report we will work with separable Hilbert spaces which we will denote by $\mathcal{H}$. The dual space of a separable Hilbert space $\mathcal{H}$ will be denoted by $\mathcal{H}'$. We will always use the notation $\mathcal{I}$ and $\mathcal{I}_N$ to denote a countable and finite index set, respectively. Different subscripts will be used to denote different finite index sets.

Chapter 1

Frames and Numerical Frame Approximation

Frames are an expansive topic that have been used in a variety of fields such as estimating signals via compressed sensing [5], signal noise reduction [9] and finite quantum mechanics [8] to name a few. We will work exclusively with Fourier frames in this report. The focus of this chapter is to introduce Fourier frames and contrast their properties with Fourier basis in solving differential equations. The main difference is that frame representation are not unique and systems formed using truncated Fourier frames are highly ill-conditioned. Despite, this ill-conditioning we can solve the linear system accurately.

1.1 Fourier Basis and Spectral Approximation

We begin this section by stating the definition for a basis.

Definition 1.1.1. *(Definition 3.1.1 of [6]) Let X be a Banach space. A sequence $\{\phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}} \subset X$ is a (Schauder) basis for X if, for each $f \in X$ there exist unique scaler coefficients $\{c_{\boldsymbol{n}}(f)\}_{\boldsymbol{n}\in\mathcal{I}}$ such that,*

$$f = \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}}(f)\phi_{\boldsymbol{n}}. \tag{1.1}$$

Next we define the concept of dual basis. There is an analogous concept to dual basis for frames known as dual frames, which will play an important role in this report.

Definition 1.1.2. *(Theorem 3.2.2 of [6]) Let $\{\phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ be a basis for the Hilbert space $\mathcal{H}$. Then there exists a unique family $\{\Phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ in $\mathcal{H}$ for which*

$$\sum_{\boldsymbol{n}\in\mathcal{I}} \langle f, \Phi_{\boldsymbol{n}}\rangle \phi_{\boldsymbol{n}}, \quad \forall f \in \mathcal{H}, \tag{1.2}$$

$\{\Phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ is a basis for $\mathcal{H}$, $\{\phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ and $\{\Phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ form a bi-orthogonal system and $\{\Phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ is called the dual basis.

The dual basis associated to an orthogonal basis for a Hilbert space $\mathcal{H}$ is itself. The dual basis is always unique in the case of Hilbert spaces, this follows from every Hilbert space being reflexive. The dual frame is important for a different reason as will be explained later in the chapter. Hilbert spaces have several well known desirable properties as a result of the inner product structure. We now state a few properties of Hilbert spaces and briefly touch on their importance.

Theorem 1.1.1. *(Theorem 3.4.2 of [6]) Let $\{\phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ be an orthonormal system for a separable Hilbert space $\mathcal{H}$. Then the following are equivalent*

i $\{\phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ *is an orthonormal basis.*

ii $\overline{span\{\phi_{\boldsymbol{n}}\}}_{\boldsymbol{n}\in\mathcal{I}} = \mathcal{H}$.

iii ***Completeness:*** *If $\langle f, \phi_{\boldsymbol{n}}\rangle = 0$ for all $\boldsymbol{n}\in\mathcal{I}$, then $f = 0$.*

iv ***Parseval's Identity:*** $\|f\|_H = \sum_{\boldsymbol{n}\in\mathcal{I}} |\langle f, \phi_{\boldsymbol{n}}\rangle|^2, \qquad \forall f \in \mathcal{H}$.

v For each $f\in\mathcal{H}$, $f = \sum_{\boldsymbol{n}\in\mathcal{I}}\langle f, \phi_{\boldsymbol{n}}\rangle\phi_{\boldsymbol{n}}$, where the term on the right converge unconditionally in norm. Where unconditional refers to the sequence converging irrespective of how the terms are arranged.

Completeness guarantees uniqueness of the coefficients $\{c_{\boldsymbol{n}}(f)\}_{\boldsymbol{n}\in\mathcal{I}}$, where $\mathcal{I}$ is an infinite index set. We now define Fourier basis and state a couple of theorems that are pertinent from the perspective of solving differential equations numerically.

Definition 1.1.3. *The functions $\{e^{i\pi\boldsymbol{n}\cdot\boldsymbol{x}}\}_{\boldsymbol{n}\in\mathbb{Z}^d}$ form an orthonormal basis for $L^2([-1,1]^d)$ known as the Fourier basis. For $f\in L^2([-1,1]^d)$,*

$$\begin{aligned} f &= \sum_{\boldsymbol{n}\in\mathbb{N}^d} \widehat{f(\boldsymbol{n})}\phi_{\boldsymbol{n}}, \\ \text{where} \qquad \widehat{f(\boldsymbol{n})} &= \frac{1}{2^d}\int_{[-1,1]^d} f(\boldsymbol{x})e^{-i\boldsymbol{n}\boldsymbol{x}}dx, \quad \forall \boldsymbol{n}\in\mathbb{N}^d. \end{aligned} \tag{1.3}$$

We now state a theorem that states if a function is sufficiently smooth the truncated Fourier basis representation will be a good approximation to the true function.

Theorem 1.1.2. *(Theorem 3.14 of [19]) Let $\mathbb{T}$ be the unit hypersphere. If $f\in C^k(\mathbb{T}^d)$, then its coefficients $\widehat{f(\boldsymbol{n})}$ decay at least like n^{-kd} as $|n|\longrightarrow\infty$.*

Using the above results we provide a summary of how Fourier basis are used to approximate functions or solution to differential equations. To keep notation simple we consider the case $d = 1$.

Let u be the function that we are interested in approximating or finding its derivative. Provided u is sufficiently smooth, we could find the Fourier coefficients $\{\widehat{f(n)}\}_{n=-N}^{N}$ for some finite N and approximate u using the truncated Fourier basis representation. Given a set of equispaced M points $\{x_j\}$ we can approximate the coefficients as,

$$\widehat{f(n)} \approx \hat{c}_n = h \sum_{j=1}^{M} u(x_j) e^{-i\pi n x_j}, \quad \text{for} \quad n = -N, \cdots, \cdots, N. \tag{1.4}$$

Using the approximated coefficients we can construct the band limited interpolant $p(x)$ which is our approximation of u at the points x_j.

$$\begin{aligned} u(x_j) &\approx p(x) = \sum_{l=-N}^{N} \hat{c}_n e^{i\pi l x_j}, \\ u'(x_j) &\approx p'(x_j). \end{aligned} \tag{1.5}$$

If $u \in C^{\infty}([-1,1])$ then a small N will provide us with a good approximation as the Fourier coefficients will be close to zero for large N. Conversely, if $u \in C^1([-1,1])$ then we would need N to be large to get a good approximation. Approximating a function or its derivative using orthogonal basis comes down to the rate of decay of the basis coefficients, provided the series converges to the true solution.

1.2 Frames

In this section we outline frames and their properties.

Definition 1.2.1. *((2.1) of [2]) A countable family of elements $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ in a Hilbert space $\mathcal{H}$ is a frame for $\mathcal{H}$ if there exist constants $A, B > 0$ such that,*

$$A \left\|f\right\|_H^2 \leq \sum_{\boldsymbol{n}\in\mathcal{I}} |\langle f, \psi_{\boldsymbol{n}}\rangle|^2 \leq B \left\|f\right\|_H^2, \qquad \forall f \in \mathcal{H}. \tag{1.6}$$

The frame condition is a relaxation of Parseval's identity and provides an equivalence relation between the ℓ^2 norm of the coefficients and the norm of f. The next definition is non-standard however it will be used extensively in this report.

Definition 1.2.2. *Let $\boldsymbol{\psi} = \{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}} \subset \mathcal{H}$ be a frame. we refer to the representation of f in $\mathcal{H}$ by $\boldsymbol{\psi}$ as,*

$$f = \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}} \psi_{\boldsymbol{n}}. \tag{1.7}$$

We refer to a finite or truncated frame representation of f in $\mathcal{H}$ by $\boldsymbol{\psi}$ if the sum in series above is over a finite index set. Where it is understood that the series convergence is with respect to the norm.

We will usually use the short-hand, representation of f in $\mathcal{H}$. A necessary condition in order for each element of $\mathcal{H}$ to have a representation in $\boldsymbol{\psi} = \{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ is $\overline{span\{\boldsymbol{\psi}\}} = \mathcal{H}$.

Theorem 1.2.1. *Let $\boldsymbol{\psi} = \{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ be a frame for $\mathcal{H}$ then,*

$$\overline{span\{\psi_{\boldsymbol{n}}\}}_{\boldsymbol{n}\in\mathcal{I}} = \mathcal{H}. \tag{1.8}$$

Proof. Assume $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ is a frame for $\mathcal{H}$ and $\overline{span\{\psi_{\boldsymbol{n}}\}}_{\boldsymbol{n}\in\mathcal{I}} \neq \mathcal{H}$. Then

$$\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}} = \boldsymbol{\psi} \subseteq \mathcal{H}. \tag{1.9}$$

It is a well known result that if $\mathcal{V} \subset \mathcal{H}$ then $span\{\mathcal{V}\}$ is dense in $\mathcal{H}$ if and only if $\{\mathcal{V}\}^{\perp} = 0$. As a result there exist a $f \in \{\boldsymbol{\psi}\}^{\perp}$ such that $f \neq 0$. Thus

$$A\,\|f\|_H \leq \sum_{\boldsymbol{n}\in\mathcal{I}} |\langle f, \psi_{\boldsymbol{n}}\rangle| = 0. \tag{1.10}$$

This implies $A = 0$. This is a contradiction. □

Theorem 1.2.1 states that the frame condition ensures each element in $\mathcal{H}$ can be represented by $\boldsymbol{\psi}$. However, the uniqueness of coefficients is not guaranteed as we will see later in this section. We outline a few definition regarding different frame properties.

Definition 1.2.3. ***Frame Properties:***

i A frame is tight if we can choose A=B as frame bounds.

ii A frame is said to be exact if removing one element of $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ results in $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ not being a frame any more.

iii A frame is called redundant/over-complete if it is not a basis for $\mathcal{H}$.

iv A frame is said to be linearly independent if every finite subset of $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ is linearly independent.

We now define the analogous concept of dual basis for frames. Unless the frame is exact the dual frame will not be unique in contrast to dual basis.

Definition 1.2.4. *((2.8) of [2]) Let $\boldsymbol{\psi} = \{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ be a frame. A dual frame $\boldsymbol{\Psi} = \{\Psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ of $\boldsymbol{\psi}$ is a frame for $\mathcal{H}$ if $\forall\ f \in \mathcal{H}$,*

$$f = \sum_{\boldsymbol{n}\in\mathcal{I}} \langle f, \Psi_{\boldsymbol{n}}\rangle \psi_{\boldsymbol{n}} = \sum_{\boldsymbol{n}\in\mathcal{I}} \langle f, \psi_{\boldsymbol{n}}\rangle \Psi_{\boldsymbol{n}}. \tag{1.11}$$

Next we define a few key operator which will be key in stating several theorems.

Definition 1.2.5. *((5.1)-(5.4) of [6]) Define the synthesis operator $\mathcal{T}$, analysis operator $\mathcal{T}^*$ and frame operator $\mathcal{T}\mathcal{T}^*$ as follow,*

$$\begin{aligned}
&\mathcal{T} : \ell^2(\mathbb{N}) \longrightarrow \mathcal{H}, && \mathcal{T}^* : \mathcal{H} \longrightarrow \ell^2(\mathbb{N}), && \mathcal{T}\mathcal{T}^* = \mathcal{S} : \mathcal{H} \longrightarrow \mathcal{H},\\
&\mathcal{T} : \{c_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}} \longrightarrow \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}}\psi_{\boldsymbol{n}}, && \mathcal{T}^* : \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}}\Phi_{\boldsymbol{n}} \longrightarrow \{c_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}, && \mathcal{S} : f \longrightarrow \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}}\psi_{\boldsymbol{n}}.
\end{aligned} \tag{1.12}$$

1.2.1 Frame Representation

We are now in a position to define what is meant by frames being a generalization of basis. Each element of $\mathcal{H}$ can be represented using the canonical dual frame. Additionally amongst all representations there exist a representation whose coefficients have minimal norm. If one were to consider the minimal norm representation only, each element in $\mathcal{H}$ could be represented uniquely. The first theorem shows that the canonical dual frame can be used to express every element of $\mathcal{H}$.

Theorem 1.2.2. *(Theorem 5.1.6 of [6]) Let $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ be a frame with frame operator $\mathcal{S}$. Then*

$$f = \sum_{\boldsymbol{n}\in\mathcal{I}} \langle f, \mathcal{S}^{-1}\psi_{\boldsymbol{n}}\rangle \psi_{\boldsymbol{n}}, \qquad \forall f \in \mathcal{H}. \tag{1.13}$$

The series converges unconditionally for all $f \in \mathcal{H}$.

Even though we have a frame representation using the canonical dual frame it might not be unique. In general if the frame is over-complete then the frame coefficients are not unique.

Theorem 1.2.3. *(Lemma 6.3.1 of [6]) Let $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ be an over-complete frame. Then there exist a frame $\{\gamma_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}} \neq \{\mathcal{S}^{-1}\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ for which*

$$f = \sum_{\boldsymbol{n}\in\mathcal{I}} \langle f, \gamma_{\boldsymbol{n}}\rangle \psi_{\boldsymbol{n}}, \qquad \forall f \in \mathcal{H}. \tag{1.14}$$

The next result shows that if $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ is an over-complete frame then the frame coefficients $\{\langle f, \mathcal{S}^{-1}\psi_{\boldsymbol{n}}\rangle\}_{\boldsymbol{n}\in\mathcal{I}}$ have minimal ℓ^2 norm among all other frame coefficients $\{c_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$. This is one of the main reason why the canonical dual frame is so important. When we refer to frame coefficients we allude to the minimal norm canonical dual frame coefficients.

Theorem 1.2.4. *(Lemma 5.4.2 of [6]) Let $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ be a frame for $\mathcal{H}$ and let $f \in \mathcal{H}$. If f has a representation $f = \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}}\psi_{\boldsymbol{n}}$ for some coefficients $\{c_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$, then*

$$\sum_{\boldsymbol{n}\in\mathcal{I}} |c_{\boldsymbol{n}}|^2 = \sum_{\boldsymbol{n}\in\mathcal{I}} |\langle f, \mathcal{S}^{-1}\psi_{\boldsymbol{n}}\rangle|^2 + \sum_{\boldsymbol{n}\in\mathcal{I}} |c_{\boldsymbol{n}} - \langle f, \mathcal{S}^{-1}\psi_{\boldsymbol{n}}\rangle|^2. \tag{1.15}$$

The next theorem states that if a frame is tight and linearly independent, then it is self dual. We will utilize Fourier frames which are tight and linearly independent. Frames being self dual will be helpful in approximating the frame coefficients.

Theorem 1.2.5. *(Corollary 5.1.7 of [6]) Let $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ be a tight frame for $\mathcal{H}$ with frame bounds A, then the canonical dual $\{\Psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ is given by*

$$\Psi_{\boldsymbol{n}} = \frac{1}{A}\psi_{\boldsymbol{n}}, \qquad \forall \boldsymbol{n} \in \mathcal{I}. \tag{1.16}$$

1.2.2 Fourier Frames

We now formally define Fourier frames. Our main goal is to show that $\{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\}_{\boldsymbol{n}\in\mathbb{Z}^d}$ is a frame for $L^2(\Omega)$ where $\Omega \subset [-1,1]^d$. In order do so we state a theorem which we will use.

Theorem 1.2.6. *(Corollary 5.2.3 of [6]) Let $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}}$ be a frame for $\mathcal{H}$ with frame bounds A, B and $\mathcal{P}$ denote an orthogonal projection of $\mathcal{H}$ onto a closed subspace $\mathcal{V}$. Then $\{\mathcal{P}\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$ is a frame for $\mathcal{V}$ with frame bounds A, B.*

Proposition 1.2.1. *Let Ω be compactly embedded in $[-1,1]^d$. The set $\mathcal{F} = \{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\mathcal{X}_\Omega\}_{\boldsymbol{n}\in\mathbb{N}^d}$ is a tight, linearly independent frame for $L^2(\Omega)$ with frame bounds $A = 1$. Where $\mathcal{X}$ is the indicator/characteristic function. We refer to $\mathcal{F}$ as the Fourier frame.*

Proof. $\{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\}_{\boldsymbol{n}\in\mathbb{N}^d}$ is an orthonormal basis it satisfies Parseval's identity. Thus, $\{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\}_{\boldsymbol{n}\in\mathbb{N}^d}$ is a tight frame for $[-1,1]^d$ with frame bounds $A = 1 = B$.
$\mathcal{P} = \mathcal{X}_\Omega(\boldsymbol{x})$ is an orthogonal projection,

$$\mathcal{P}^2 = \mathcal{X}_\Omega(\boldsymbol{x})\mathcal{X}_\Omega(\boldsymbol{x}) = \mathcal{X}_\Omega(\boldsymbol{x}) = \mathcal{P}. \tag{1.17}$$

By theorem 1.2.6 $\{\mathcal{P}e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\}_{\boldsymbol{n}\in\mathbb{N}^d}$ is a frame for $L^2(\Omega)$ with frame bounds $A = 1 = B$. Thus $\mathcal{F}$ is a tight linearly independent frame for $L^2(\Omega)$. □

Restricting the Fourier basis on the unit hypercube to a compactly embedded domain Ω is one way to generate a frame. Under some condition the collection of functions $\{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}} g_{\boldsymbol{n}}(\boldsymbol{x})\}$ is a frame for $L^2(\Omega)$ where $g_{\boldsymbol{n}} \in L^2(\mathbb{R})$. A detailed description is provided in [7].

1.3 Numerical Frame Approximation

From the previous section we have that Fourier frames are self dual and as a result the corresponding frame coefficients have minimum ℓ_2 norm. This property of the Fourier frame coefficients provides us with psudo-uniqueness that we will use in finding the Fourier frame coefficients numerically. We will be interested in finding sequences of Fourier frame coefficients that have minimum ℓ_2 norm and the behaviour of the individual coefficients is not of interest to us. This is in contrast to when approximating with Fourier basis where the approximation depended on the decay rate of the basis coefficient.

In order to approximate numerically we will work with truncated Fourier frame $\mathcal{F}_N = \{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\}_{\boldsymbol{n}\in\mathcal{I}_N}$, where $\boldsymbol{x} \in \Omega \subset [-1,1]^d$ and $\mathcal{I}_N$ is a finite index set of cardinality N. The linear systems that we construct to approximate solution of PDEs are highly ill-conditioned. In general truncated frames exhibit highly ill-conditioned systems. In this section we state results that show despite being severely ill-conditioned we can recover accurate approximation of the true Fourier frame coefficients. This section will outline ill-conditioning in the case of approximating a function using Fourier frames.

We extend the definition of the previous section for finite dimensional frames.

Definition 1.3.1. *Let $\mathcal{H}_N$ be a closed finite dimensional subspace of a Hilbert space $\mathcal{H}$. A finite collections $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}_N}$ is a frame for $\mathcal{H}_N$ if there exist constants $A_N, B_N > 0$ such that,*

$$A_N \|f\|_H \leq \sum_{\boldsymbol{n}\in\mathcal{I}_N} |\langle f, \psi_{\boldsymbol{n}}\rangle|^2 \leq B_N \|f\|_H, \qquad \forall f \in \mathcal{H}_N. \tag{1.18}$$

In particular $\mathcal{F}_N$ is a frame for span$\{\mathcal{F}_N\}$.

We approximate $f \in \mathcal{H}$ by $\mathcal{P}_N f \in \mathcal{H}_N$, where $\mathcal{P}_N f = \sum_{\boldsymbol{n}\in\mathcal{I}_N} c_{\boldsymbol{n}}\psi_{\boldsymbol{n}}$ where $c_{\boldsymbol{n}} \in \mathbb{C}$ for $\boldsymbol{n} \in \mathcal{I}_N$ and $\mathcal{I}_N$ has cardinality N. We solve for the finite sequence $\{c_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}_N}$ as follow.

$$\begin{aligned} f \approx \mathcal{P}_N f &= \sum_{\boldsymbol{n}\in\mathcal{I}_N} c_{\boldsymbol{n}}\psi_{\boldsymbol{n}}, \\ \langle \mathcal{P}_N f, \psi_{\boldsymbol{m}}\rangle &= \langle f, \psi_{\boldsymbol{m}}\rangle, \\ \Big\langle \sum_{\boldsymbol{n}\in\mathcal{I}_N} c_{\boldsymbol{n}}\psi_{\boldsymbol{n}}, \psi_{\boldsymbol{m}}\Big\rangle &= \langle f, \psi_{\boldsymbol{m}}\rangle. \end{aligned} \tag{1.19}$$

$$\sum_{\boldsymbol{n}\in\mathcal{I}_N} c_{\boldsymbol{n}}\langle \psi_{\boldsymbol{n}}, \psi_{\boldsymbol{m}}\rangle = \langle f, \psi_{\boldsymbol{m}}\rangle, \quad \forall \boldsymbol{m} \in \mathcal{I}_N. \tag{1.20}$$

We can rewrite the above system as,

$$\boldsymbol{G}_N \boldsymbol{c} = \boldsymbol{b}, \qquad \boldsymbol{G}_N \in \mathbb{C}^{N\times N},\ \boldsymbol{c}, \boldsymbol{b} \in \mathbb{C}^N. \tag{1.21}$$

We refer to the matrix $\boldsymbol{G}_N$ as the Gram matrix. In order to obtain the approximation $\mathcal{P}_N f$ we need to solve the above system for $\boldsymbol{c}$. The next few results state that the Gram matrix corresponding to linearly independent frame are necessarily ill-conditioned.

Theorem 1.3.1. *(Lemma 4.1 of [2]) The truncated Gram matrix $\boldsymbol{G}_N$ of a linearly independent frame $\Psi = \{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}_N}$ is invertible, $\|\boldsymbol{G}_N^{-1}\|^{-1} = A_N$ and $\|\boldsymbol{G}_N\| = B_N$, where A_N and B_N are truncated frame bounds. In particular the condition number of G_N is,*

$$\kappa(\boldsymbol{G}_N) = \|\boldsymbol{G}_N\|\|\boldsymbol{G}_N^{-1}\| = \frac{B_N}{A_N}. \tag{1.22}$$

This shows that the condition number of $\boldsymbol{G}_N$ depends on how small A_N and how large B_N are. The next theorem outline the behaviour of A_N and B_N as N increases.

Theorem 1.3.2. *(Lemma 4.2 of [2]) Let $\Psi = \{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}_N}$ be a linearly independent frame then,*

1. *The sequence $\{A_N\}_{N\in\mathbb{N}}$ and $\{B_N\}_{N\in\mathbb{N}}$ are monotonically non increasing and non-decreasing, respectively.*
2. *$B_N \leq B,\ \forall N$ and $B_N \to B$ as $N \to \infty$.*

3. $\inf_N A_N > 0$.

The above theorems shows that in general Gram matrices of linearly independent frames are ill-conditioned. We are interested in the behaviour of the condition number of the Gram matrix associated to the linearly independent Fourier frames that we stated in section 1.2.2.

Theorem 1.3.3. *(Proposition 4.5 of [2]) Let* $\Omega \subset [-1,1]^d$ *be a compact Lipschitz domain and* $\mathcal{F} = \{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\}_{\boldsymbol{n}\in\mathbb{Z}^d}$*, where* $\boldsymbol{x} \in \Omega$*. Then the condition number* $\kappa(\boldsymbol{G}_N)$ *grows super algebraically fast in* N*.*

From the above result we have that Gram matrices corresponding to Fourier frames are highly ill-conditioned. This suggest that we will be unable to approximate the Fourier frame coefficients. To show this we recall the following derivation from [2] showing that the frame coefficients can grow arbitrarily fast when solving 1.21.

$$\begin{aligned} \|\boldsymbol{c}\| = \|\boldsymbol{G}_N^{-1}\boldsymbol{b}\| &\leq \|\boldsymbol{G}_N^{-1}\|\|\boldsymbol{b}\|, && \text{Cauchy-Schwarz,} \\ &= \frac{\|\boldsymbol{b}\|}{A_N} = \frac{\sqrt{\sum_{\boldsymbol{n}\in\mathcal{I}_N} |\langle f, \psi_{\boldsymbol{n}}\rangle|^2}}{A_N}, && \text{Thm 1.3.1,} \\ &\leq \frac{B_N}{A_N}\|f\|, && \text{Frame conditon (1.18).} \end{aligned} \tag{1.23}$$

As a result of the ill-conditioning we use singular value decomposition (SVD) regularization to approximate the true Fourier frame coefficients.

1.3.1 SVD Regularization

We now outline the regularization method as outlined in [2]. Regularize 1.21 by picking a tolerance threshold $0 < \epsilon << 1$. Consider the SVD of $\boldsymbol{G}_N = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^*$. We set the singular values $\sigma_j \leq \epsilon$ equal to zero. We solve the following alternative to 1.21.

$$\begin{aligned} &\boldsymbol{G}_N^{\epsilon}\boldsymbol{c}^{\epsilon} = \boldsymbol{b}, \\ &\boldsymbol{G}_N^{\epsilon} = \boldsymbol{V}\boldsymbol{\Sigma}^{\epsilon}\boldsymbol{V}^*, \qquad \boldsymbol{\Sigma}^{\epsilon} = \begin{cases} \sigma_{ii}, & \sigma_{ii} \geq \epsilon \\ 0 & \sigma_{ii} < \epsilon. \end{cases} \end{aligned} \tag{1.24}$$

Note in (1.24) we are computing $\boldsymbol{c}^{\epsilon}$ which are an approximation to the true frame coefficients $\boldsymbol{c}$ in (1.21). Let us denote the approximation of f using the regularized SVD as $\mathcal{P}_N^{\epsilon} f$, i.e.,

$$\mathcal{P}_N^{\epsilon} f = \sum_{\boldsymbol{n}\in\mathcal{I}_N} c_{\boldsymbol{n}}^{\epsilon}\psi_{\boldsymbol{n}}. \tag{1.25}$$

The next result shows that our approximation $\boldsymbol{c}^{\epsilon}$ of the true frame coefficients converges to within $\sqrt{\epsilon}$ of the true frame coefficients.

Theorem 1.3.4. *(Theorem 5.4 of [2]) The coefficients $\boldsymbol{c}^\epsilon$ computed using the regularized SVD satisfy,*

$$\|\boldsymbol{c}^\epsilon\| \leq \inf\{\frac{1}{\sqrt{\epsilon}}\|f - \mathcal{T}_N \boldsymbol{z}\| + \|\boldsymbol{z}\| \ : \ \boldsymbol{z} \in \mathbb{C}^N\}. \tag{1.26}$$

Moreover if $\boldsymbol{c} \in \ell_2(\mathbb{C})$ are the frame coefficients and $\tilde{\boldsymbol{c}}^\epsilon$ be the extension of $\boldsymbol{c}^\epsilon$ by zeros then,

$$\|\boldsymbol{c} - \tilde{\boldsymbol{c}}^\epsilon\| \leq \big(1 + \sqrt{\frac{B}{\epsilon}}\big)\sqrt{\sum_{\boldsymbol{n} \in \mathbb{Z}^d \setminus \mathcal{I}_N} |c_{\boldsymbol{n}}|^2} + \sqrt{\frac{\epsilon}{A}}\|\boldsymbol{c}\|. \tag{1.27}$$

In (1.27) as N increases the sum of the frame coefficients in the square root terms approaches zero. Thus as N increases the error is bounded by $\sqrt{\epsilon}\|\boldsymbol{c}\|$. Thus provided the ℓ_2 norm of the frame coefficients is small our approximation is accurate to within $\sqrt{\epsilon}$. We now state results regarding the convergence of $\mathcal{P}_N^\epsilon f$ to f.

Theorem 1.3.5. *(Lemma 5.3 of [2]) The truncated SVD projection $\mathcal{P}_N^\epsilon$ satisfies,*

$$\|f - \mathcal{P}_N^\epsilon f\| \leq \inf\{\|f - \mathcal{T}_N \boldsymbol{z}\| + \sqrt{\epsilon}\|\boldsymbol{z}\| \ : \ \boldsymbol{z} \in \mathbb{C}^N\}. \tag{1.28}$$

Furthermore if $\boldsymbol{c} \in \ell_2(\mathbb{C})$ are the frame coefficients of f then by letting $\boldsymbol{z} = \boldsymbol{c}$ we have,

$$\limsup_{N \to \infty} \|f - \mathcal{P}_N^\epsilon f\| \leq \sqrt{\epsilon}\|\boldsymbol{c}\| \leq \sqrt{\frac{\epsilon}{A}}\|f\|. \tag{1.29}$$

The above theorem shows that our approximation is of the order $\sqrt{\epsilon}$ provided f can be approximated by a sequence $\boldsymbol{z} \in \mathbb{C}^N$ that has small ℓ_2 norm. Both theorems require there being a sequence in $\boldsymbol{z} \in \mathbb{C}^N$ of small norm for a good approximation. We now state a results that relates the regularity of f to the existence and behaviour of Fourier frame coefficients $\boldsymbol{c} \in \mathbb{C}^N$ that approximate f. Furthermore the results provide us with an upper bound for the ℓ_2 norm of the Fourier frame coefficients.

Proposition 1.3.1. *(Theorem 5.7 of [2]) Let $\Omega \subset [-1,1]^d$ be a compact, Lipschitz domain and let $\mathcal{F}$ be the Fourier frame. If $f \in H^k(\Omega)$ then there exist a $\boldsymbol{c} \in \ell_2(\mathbb{C})$ s.t,*

$$\begin{aligned} \|f - \mathcal{P}_N \boldsymbol{c}\| &\leq \frac{C}{N^k}\|f\|_{H^k([-1,1]^d)}, \\ \|\boldsymbol{c}\| &\leq C\|f\|_{H^k([-1,1]^d)}, \end{aligned} \tag{1.30}$$

and for the regularized projection,

$$\|f - \mathcal{P}_N^\epsilon \boldsymbol{c}\| \leq C(N^{-k} + \sqrt{\epsilon})\|f\|_{H^k([-1,1]^d)}. \tag{1.31}$$

We end this chapter by stating that we can improve upon our approximation by oversampling. The oversampling analogue of (1.21) is the following,

$$\begin{aligned} \boldsymbol{G}_{M,N} &= \{\langle \psi_{\boldsymbol{n}}, \psi_{\boldsymbol{m}} \rangle\}_{\boldsymbol{m} \in \mathcal{I}_M, \boldsymbol{n} \in \mathcal{I}_N} \in \mathbb{C}^{M \times N}, \quad M >> N, \\ \boldsymbol{b} &= \{\langle f, \psi_{\boldsymbol{m}} \rangle\}_{\boldsymbol{m} \in \mathcal{I}_M} \in \mathbb{C}^M, \\ \boldsymbol{G}_{M,N}\boldsymbol{c} &= \boldsymbol{b}. \end{aligned} \tag{1.32}$$

The oversampled Gram matrix is also ill-conditioned as a result we will consider the SVD regularized version for the oversampled system as well.

$$\boldsymbol{G}^{\epsilon}_{M,N}\boldsymbol{c} = \boldsymbol{b}. \tag{1.33}$$

In [3] it is shown that by oversampling the results of Theorem 1.3.5 can be improved from $\mathcal{O}(\sqrt{\epsilon})$ to $\mathcal{O}(\epsilon)$. Due to the improvement in accuracy we will also incorporate oversampling and SVD regularization in our numerical method.

Chapter 2

Numerical Method and Examples

In this chapter we outline the numerical method to solve time independent second order constant coefficient elliptic ODEs and PDEs with Dirichlet and Neumann boundary conditions. The numerical method constructs a linear system by substituting the finite Fourier frame representation into the PDE and evaluating the resulting equations at a set of discrete finite points in the domain and on the boundary. We know from the previous chapter that Fourier frames coefficients have minimum ℓ_2 norm due to Fourier frames being self dual and discrete system formed using finite frames suffer from ill-conditioning. Similar to the outlined work in the previous chapter we use least squares with oversampling and SVD regularization to solve our linear system. We end this chapter by providing a brief outline on how to extend the numerical method to the Laplace eigenvalue problem.

Before we begin we state the definition of the Hadamard product and a result regarding the solution of the least square minimization problem. The Hadamard product will help in simplifying notation later on.

Definition 2.0.1. *Let $\boldsymbol{A}$, $\boldsymbol{B}$ be matrices of the same dimensions and $\boldsymbol{f}$, $\boldsymbol{g}$ be vectors of the same dimension. Then the Hadamard product is defined as,*

$$\begin{aligned} (\boldsymbol{A} \circ \boldsymbol{B})_{ij} &= \boldsymbol{A}_{ij}\boldsymbol{B}_{ij}, \\ (\boldsymbol{f} \circ \boldsymbol{g})_i &= f_i g_i. \end{aligned} \tag{2.1}$$

As we will employ least square minimization to solve for the Fourier frame coefficients, for completeness we state the following result that outlines the necessary and sufficient conditions for the existence and uniqueness of the least square minimization problem.

Theorem 2.0.1. *(Theorem 11.1 of [17]) Let $\boldsymbol{A} \in \mathbb{C}^{M\times N}(M >> N)$ and $\boldsymbol{b} \in \mathbb{C}^M$ be given. A vector $\boldsymbol{x} \in \mathbb{C}^N$ minimizes the residual norm $\|\boldsymbol{r}\|_2 = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2$*

if and only if $\boldsymbol{r} \perp range(\boldsymbol{A})$*, that is*

$$\boldsymbol{A}^* \boldsymbol{r} = \boldsymbol{0}, \tag{2.2}$$

or equivalently,

$$\boldsymbol{A}^* \boldsymbol{A} \boldsymbol{x} = \boldsymbol{A}^* \boldsymbol{b}. \tag{2.3}$$

The $N \times N$ *system of equations (2.3) is nonsingular if and only if* $\boldsymbol{A}$ *has full rank. Consequently the solution* $\boldsymbol{x}$ *is unique if and only if* $\boldsymbol{A}$ *has full rank.*

2.1 Numerical Method

We are interested in solving the following second order constant coefficient elliptic PDE,

$$\begin{aligned} (\boldsymbol{p} \circ \nabla) \cdot (\nabla u) + \boldsymbol{q} \cdot \nabla u + ru = f, \quad & \boldsymbol{x} \in \Omega \subset [-1,1]^d, \\ u = g, \quad & \boldsymbol{x} \in \partial\Omega, \end{aligned} \tag{2.4}$$

where $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^d$ and $r \in \mathbb{R}$. We will approximate the true solution u from the function space spanned by the truncated Fourier frame $\mathcal{F}_N = \{e^{i\pi \boldsymbol{n} \cdot \boldsymbol{x}}\}_{\substack{|n_l| \le N \\ 1 \le l \le d}}$, i.e.

$$u(\boldsymbol{x}) = \sum_{\substack{|n_l| \le N \\ 1 \le l \le d}} c_{\boldsymbol{n}} e^{i\pi \boldsymbol{n} \cdot \boldsymbol{x}}. \tag{2.5}$$

Substituting expression (2.5) into (2.4) we get,

$$\begin{aligned} \sum_{\substack{|n_l| \le N \\ 1 \le l \le d}} c_{\boldsymbol{n}} \Big[-\pi^2 (\boldsymbol{p} \circ \boldsymbol{n}) \cdot \boldsymbol{n} + i\pi \boldsymbol{q} \cdot \boldsymbol{n} + r \Big] e^{i\pi \boldsymbol{n} \cdot \boldsymbol{x}} = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \\ \sum_{\substack{|n_l| \le N \\ 1 \le l \le d}} c_{\boldsymbol{n}} e^{i\pi \boldsymbol{n} \cdot \boldsymbol{x}} = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{2.6}$$

We now outline our disretization strategy. Let $X = \{\boldsymbol{x}_j\}_{j \in \mathcal{I}_{M_{cube}}}$ be the set of points of an equispaced uniform finite mesh of the unit hyper cube $[-1,1]^d$. We define Ω_{M_i} to be the set of points in X that are also in Ω. We denote the cardinality of Ω_{M_i} by M_i, i.e.

$$\Omega_{M_i} = X \cap \Omega = \{\boldsymbol{x}_j \in X \cap \Omega \ : \ 1 \le j \le M_i\}. \tag{2.7}$$

We now construct a finite set of boundary points. Let $\partial\Omega(\boldsymbol{x}(t))$ be a parametrization of the boundary, where $0 < t \le 1$. Let $\{t_k\}_{k=1}^{M_b}$ be a finite uniform discretization of the unit interval $(0,1]$. Using the discretization we define our finite set of boundary points $\partial\Omega_{M_b}$ as,

$$\partial\Omega_{M_b} = \{\boldsymbol{x}_k \ : \ \boldsymbol{x}_k = \partial\Omega(\boldsymbol{x}(t_k)), \ 1 \le k \le M_b\}. \tag{2.8}$$

In order to over sample we require the sum of the number of points in Ω_{M_i} and Ω_{M_b} to be greater then the number of Fourier frame modes, i.e $M_i+M_b = M >> (2N+1)^d$. We will refer to the total number of points M as the discretization points. We will refer to the ratio of discretization points M to the number of Fourier frame modes as the sampling rate. If the number of total Fourier frames modes is $\bar{N}$ then a sampling rate of 5 implies $M = 5\bar{N}$. Now we evaluate (2.6) at the discretized interior and boundary points to construct a system of M equations.

$$\begin{aligned}\sum_{\substack{|n_l|\leq N\\ 1\leq l\leq d}} c_{\boldsymbol{n}}\Big[-\pi^2(\boldsymbol{p}\circ\boldsymbol{n})\cdot\boldsymbol{n}+i\pi\boldsymbol{q}\cdot\boldsymbol{n}+r\Big]e^{i\pi\boldsymbol{n}\cdot\boldsymbol{x}_j} &= f(\boldsymbol{x}_j),\quad \boldsymbol{x}_j\in\Omega_{M_i},\\ \sum_{\substack{|n_l|\leq N\\ 1\leq l\leq d}} c_{\boldsymbol{n}}e^{i\pi\boldsymbol{n}\cdot\boldsymbol{x}_k} &= g(\boldsymbol{x}_k),\quad \boldsymbol{x}_k\in\partial\Omega_{M_b}.\end{aligned} \tag{2.9}$$

Rewriting the above in matrix form we have,

$$\begin{aligned}&\boldsymbol{A}\boldsymbol{c}=\boldsymbol{F},\\ \text{where}\quad &\boldsymbol{A}=\begin{bmatrix}\boldsymbol{A}^{int}\\ \boldsymbol{A}^{bdy}\end{bmatrix},\qquad \boldsymbol{F}=\begin{bmatrix}\boldsymbol{f}\\ \boldsymbol{g}\end{bmatrix},\\ &\boldsymbol{A}\in\mathbb{C}^{M\times(2N+1)^d},\quad \boldsymbol{A}^{int}\in\mathbb{C}^{M_i\times(2N+1)^d},\boldsymbol{A}^{bdy}\in\mathbb{C}^{M_b\times(2N+1)^d},\\ &\boldsymbol{F}\in\mathbb{C}^{M},\quad \boldsymbol{f}\in\mathbb{C}^{M_i},\ \boldsymbol{g}\in\mathbb{C}^{M_b}.\end{aligned} \tag{2.10}$$

Our goal is to solve for the unknown Fourier frame coefficients $\boldsymbol{c}\in\mathbb{C}^{(2N+1)^d}$, which we do using least square minimization.

$$\boldsymbol{c}=\underset{\tilde{\boldsymbol{c}}\in\mathbb{C}^{(2N+1)^d}}{\operatorname{argmin}}\|\boldsymbol{A}\tilde{\boldsymbol{c}}-\boldsymbol{F}\|_2. \tag{2.11}$$

From Theorem 2.0.1 we know the solution of the above least square minimization problem is,

$$\begin{aligned}\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{c}&=\boldsymbol{A}^*\boldsymbol{F},\\ \boldsymbol{c}&=(\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\boldsymbol{F}.\end{aligned} \tag{2.12}$$

From the previous chapter we know that systems formed from truncated frames suffer from ill-conditioning. Analogous to the previous chapter, we employ SVD regularization to approximate the Fourier frame coefficients (2.12) due to the ill-conditioning. Let ϵ be a tolerance threshold, then we solve the following alternative to (2.11).

$$\begin{aligned}\boldsymbol{c}^\epsilon&=\underset{\tilde{\boldsymbol{c}}\in\mathbb{C}^{(2N+1)^d}}{\operatorname{argmin}}\|\boldsymbol{A}_\epsilon\tilde{\boldsymbol{c}}-\boldsymbol{F}\|_2,\\ \boldsymbol{c}^\epsilon&=(\boldsymbol{A}_\epsilon^*\boldsymbol{A}_\epsilon)^{-1}\boldsymbol{A}_\epsilon^*\boldsymbol{F},\\ \boldsymbol{c}^\epsilon&=\boldsymbol{V}\boldsymbol{\Sigma}_\epsilon^\dagger\boldsymbol{U}^*\boldsymbol{F},\end{aligned} \tag{2.13}$$

$$\text{where}\quad \boldsymbol{A}_\epsilon=\boldsymbol{U}\boldsymbol{\Sigma}_\epsilon\boldsymbol{V}^*,\ \boldsymbol{\Sigma}_\epsilon=\begin{cases}\sigma_{ii} & \sigma_{ii}\geq\epsilon\\ 0 & \sigma_{ii}<\epsilon.\end{cases}$$

By substituting our Fourier frame coefficients $\boldsymbol{c}^{\epsilon}$ into (2.5) we get our numerical approximation to (2.4).

2.1.1 Neumann Boundary Condition

We now show the implementation of the numerical scheme for Neumann boundary conditions. We are interested in solving the following PDE with Neumann boundary condition,

$$\begin{aligned}(\boldsymbol{p}\circ\nabla)\cdot(\nabla u)+\boldsymbol{q}\cdot\nabla u+ru=f, &\quad \boldsymbol{x}\in\Omega\subset[-1,1]^d,\\ \frac{\partial u}{\partial \boldsymbol{n}}=g, &\quad \boldsymbol{x}\in\partial\Omega.\end{aligned} \tag{2.14}$$

Where $\boldsymbol{n}(\boldsymbol{x})=(n_1(\boldsymbol{x}),\cdots,n_d(\boldsymbol{x}))$ is the unit normal vector, $\boldsymbol{p},\boldsymbol{q}\in\mathbb{R}^d$ and $r\in\mathbb{R}$. Rewriting the boundary condition using the normal vector we have,

$$\frac{\partial u}{\partial \boldsymbol{n}}=\nabla u\cdot\boldsymbol{n}=\sum_{j=1}^{d}n_j(\boldsymbol{x})\frac{\partial u}{\partial x_j}. \tag{2.15}$$

As before we substitute our Fourier frame representation (2.5) into the above equation and evaluate it at the set of discretized boundary points in $\partial\Omega_{M_b}$,

$$\sum_{\substack{|m_l|\le N\\ 1\le l\le d}} c_{\boldsymbol{m}}(i\pi\boldsymbol{n}(\boldsymbol{x}_k)\cdot\boldsymbol{m})e^{i\pi\boldsymbol{m}\cdot\boldsymbol{x}_k}=g(\boldsymbol{x}_k),\quad \boldsymbol{x}_k\in\partial\Omega_{M_b}. \tag{2.16}$$

Note in the above equation we are using $\boldsymbol{m}$ as the finite index as $\boldsymbol{n}$ represents the unit normal vector. We have M_b equations corresponding to the boundary. The equations corresponding to the interior are the same as in (2.9). We can now form the system (2.10) using (2.16) for the boundary equations. Least square minimization and SVD regularization is identical to the Dirichlet boundary case, as a result we do not repeat the details here.

2.2 Numerical Examples

In this section we outline numerical results from implementing the numerical method to seven examples. Our first two examples are of ordinary differential equation (ODEs) with Dirichlet and mixed boundary conditions. In order to show how the numerical scheme outlined in the previous section is implemented we provide the details for the first ODE example. Next we solve two elliptic PDEs on a circular domain. We provide a brief outline of the numerical method for the third example. This is primarily because this will be our first PDE example. Example five and six outline numerical results from solving the same Poisson PDE on an irregular convex and non-convex domain with continuous boundary, respectively. We refer to the domains of example five and six as the convex and non-convex ice cream cone domain, respectively. The final example

solves a second order elliptic PDE on a different non-convex domain with continuous boundary. We refer to this domain as the cloud shaped domain. While we define each domain at the beginning of each example, for simplicity we have provided in Appendix B figures showing the domain.

In order to compute the numerical error where the true solution is unknown, we will use a reference solution. The reference solution is a numerical solution computed using a much larger number of Fourier frame modes and discretization points then the rest of the numerical solutions. The exact number of Fourier frame modes and discretization nodes used to compute the reference solution and numerical solution is stated for each example. For all of the examples in this section N will always represent the number of Fourier frame modes and M will represent the number of discretization points.

The numerical error is computed using the ℓ_2 grid function norm of the difference between the numerical solution and the true solution if available, or the reference solution. Let $e(\boldsymbol{x}_j)$ be the error between a computed numerical solution and the true or reference solution at the point $\boldsymbol{x}_j \in \Omega \subset [-1,1]^d$ where $1 \leq j \leq M_g$. The numerical error E corresponding to that numerical solution using the grid function norm is computed as follow,

$$E = \frac{1}{M_g^{\frac{d}{2}}} \sqrt{\sum_{j=1}^{M_g} |e(\boldsymbol{x}_j)|^2}. \tag{2.17}$$

2.2.1 ODE With Dirichlet Boundary Conditions

Our first example is the following ODE with Dirichlet boundary conditions,

$$\begin{aligned} u''(x) + u'(x) - 2u(x) &= \cos(x) - 3\sin(x), \quad x \in \Omega = (-0.5, 0.5), \\ u(-0.5) &= e^{-0.5} + \sin(-0.5), \\ u(0.5) &= e^{0.5} + \sin(0.5). \end{aligned} \tag{2.18}$$

The true solution to the above problem is,

$$u(x) = e^x + \sin(x). \tag{2.19}$$

For the one dimensional ODE case our set of boundary points $\partial\Omega_{M_b}$ consists of two points $x_1 = -0.5$ and $x_M = 0.5$. Let Ω_{M-2} be a finite uniform discretization of length δ_x of Ω , i.e.

$$\begin{aligned} \Omega_{M-2} &= \{x_k \ : \ x_k \in \Omega, \ x_{k+1} - x_k = \delta_x, \ 2 \leq k \leq M-1\}, \\ x_1 &= -0.5, \quad x_M = 0.5. \end{aligned} \tag{2.20}$$

For a fixed $N \in \mathbb{N}$ our truncated frame approximation is,

$$u_N(x) = \sum_{|n| \leq N} c_n e^{i\pi n x}. \tag{2.21}$$

Substituting the above expression in to (2.18) we get,

$$\sum_{|n|\leq N} c_n(-\pi^2 n^2 + i\pi n - 2)e^{i\pi n x_k} = \cos(x_k) - 3\sin(x_k), \quad 2 \leq k \leq M-1,$$

$$\sum_{|n|\leq N} c_n e^{-0.5 i\pi n} = e^{-0.5} + \sin(-0.5),$$

$$\sum_{|n|\leq N} c_n e^{0.5 i\pi n} = e^{0.5} + \sin(0.5). \tag{2.22}$$

In order to write the above M equations compactly let us introduce the following notation,

$$\begin{aligned} \psi_{k,n} &= e^{i\pi n x_k}, \quad |n| \leq N,\ 1 \leq k \leq M, \\ d_n &= -\pi^2 n^2 + i\pi n - 2, \quad |n| \leq N, \\ f(x_k) &= \cos(x_k) - 3\sin(x_k), \quad 2 \leq k \leq M-1, \\ g(x_k) &= e^{x_k} + \sin(x_k), \quad k = 1, M. \end{aligned} \tag{2.23}$$

Using the above notation we can write (2.22) in matrix form as follow,

$$\begin{bmatrix} d_{-N}\psi_{2,-N} & d_{-N+1}\psi_{2,-N+1} & \cdots & d_N\psi_{2,N} \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ d_{-N}\psi_{M-1,-N} & d_{-N+1}\psi_{M-1,-N+1} & \cdots & d_N\psi_{M-1,N} \\ \psi_{1,-N} & \psi_{1,-N+1} & \cdots & \psi_{1,N} \\ \psi_{M,-N} & \psi_{M,-N+1} & \cdots & \psi_{M,N} \end{bmatrix} \begin{bmatrix} c_{-N} \\ \vdots \\ \vdots \\ \vdots \\ c_N \end{bmatrix} = \begin{bmatrix} f(x_2) \\ \vdots \\ \vdots \\ \vdots \\ f(x_{M-1}) \\ g(x_1) \\ g(x_M) \end{bmatrix}. \tag{2.24}$$

We write the above system in matrix form and carry out the SVD regularization. For completeness we outline the derivation,

$$\begin{aligned} \boldsymbol{A}\boldsymbol{c} &= \boldsymbol{F}, \\ \boldsymbol{c} &= (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\boldsymbol{F}, \\ \boldsymbol{c} &= (\boldsymbol{V}\boldsymbol{\Sigma}^*\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*)^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^*\boldsymbol{U}^*\boldsymbol{F}, \\ \boldsymbol{c}^\epsilon &= (\boldsymbol{V}\boldsymbol{\Sigma}_\epsilon^*\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{\Sigma}_\epsilon\boldsymbol{V}^*)^{-1}\boldsymbol{V}\boldsymbol{\Sigma}_\epsilon^*\boldsymbol{U}^*\boldsymbol{F}, \quad (0 = \sigma_{ii} < \epsilon), \\ \boldsymbol{c}^\epsilon &= (\boldsymbol{V}\boldsymbol{\Sigma}_\epsilon^*\boldsymbol{\Sigma}_\epsilon\boldsymbol{V}^*)^{-1}\boldsymbol{V}\boldsymbol{\Sigma}_\epsilon^*\boldsymbol{U}^*\boldsymbol{F}, \qquad \boldsymbol{U} \text{ is unitary}, \\ \boldsymbol{c}^\epsilon &= \boldsymbol{V}(\boldsymbol{\Sigma}_\epsilon^*\boldsymbol{\Sigma}_\epsilon)^{-1}\boldsymbol{V}^*\boldsymbol{V}\boldsymbol{\Sigma}_\epsilon^*\boldsymbol{U}^*\boldsymbol{F}, \\ \boldsymbol{c}^\epsilon &= \boldsymbol{V}\boldsymbol{\Sigma}_\epsilon^\dagger\boldsymbol{U}^*\boldsymbol{F}, \qquad \boldsymbol{V} \text{ is unitary}. \end{aligned} \tag{2.25}$$

Our numerical solution is thus,

$$u_N(x) = \sum_{|n|\leq N} c_n^\epsilon e^{i\pi n x}. \tag{2.26}$$

Figure 2.1 shows the numerical results of implementing the numerical method with different sampling strategies. The numerical results were computed using a regularization parameter of $\epsilon = 10^{-5}$. The first plot shows the true and numerical solution computed using 71 Fourier frame modes and 350 discretization points. The second plot shows how the error behaves for increasing number of Fourier frame modes for three different sampling rates. As the number of modes increase the error decays which is what we expect. The numerical results computed using a higher sampling rate are much more accurate. With a sampling rate of 10 i.e. $(M = 10N)$ and $N = 31$ we have an accuracy of approximately 10^{-11}. Even without oversampling $(M = N)$ our numerical solution is accurate to within 10^{-6}. The numerical error is approximately the same for the two oversampling cases $M = 3N$ and $M = 10N$. This indicates that a sampling rate of $M = 3N$ is sufficient for this problem. In this example oversampling provides an increase in accuracy of approximately 10^{-4}.

Numerical Results: ODE with Dirichlet BC

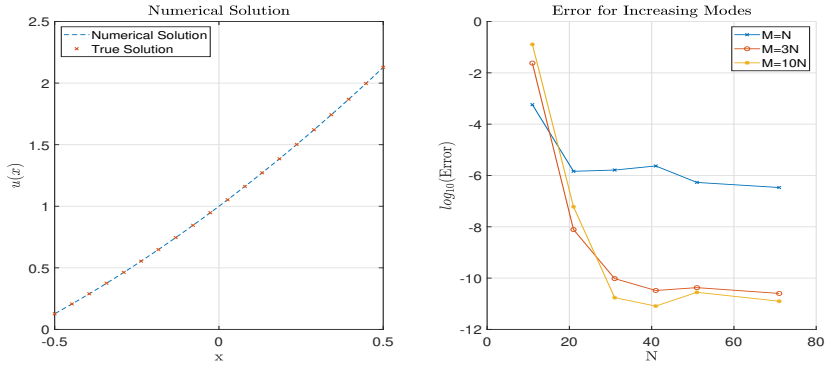

Figure 2.1: The first plot shows the true and numerical solution. The second plot shows the error as we increase the number of Fourier frame modes (N) for different sampling rates.

2.2.2 ODE With Mixed Boundary Conditions

Our second example is the following ODE with mixed boundary condition,

$$\begin{aligned} u''(x) - 2u(x) &= 2e^x - xe^x, \quad x \in \Omega = (-0.5, 0.5), \\ u'(-0.5) &= 0.5e^{-0.5}, \\ u(0.5) &= 0.5e^{0.5}. \end{aligned} \tag{2.27}$$

The true solution to the above ODE is,

$$u(x) = xe^x. \tag{2.28}$$

The above problem is similar to the previous ODE except for the boundary conditions. For completeness we state the equations relating to the boundary

for this ODE as they are different then the previous example. To that end let $x_1 = -0.5$ and $x_M = 0.5$ be as the previous example. By substituting the truncated Fourier frame (2.21) into the boundary conditions of (2.27) and evaluating the resulting equation at the boundary points we have the following two equations.,

$$\sum_{|n|\leq N} c_n(i\pi n)e^{i\pi n x_1} = 0.5e^{-0.5},$$
$$\sum_{|n|\leq N} c_n e^{i\pi n x_M} = 0.5e^{0.5}. \tag{2.29}$$

Using notation (2.23) from the previous example we can write the above two equations in the following matrix form.

$$\begin{bmatrix} -i\pi N\psi_{1,-N} & i\pi(-N+1)\psi_{1,-N+1} & \cdots & i\pi N\psi_{1,N} \\ \psi_{M,-N} & \psi_{M,-N+1} & \cdots & \psi_{M,N} \end{bmatrix} \boldsymbol{c} = \begin{bmatrix} 0.5e^{-0.5} \\ 0.5e^{0.5} \end{bmatrix}. \tag{2.30}$$

Numerical Results: ODE with Mixed BC

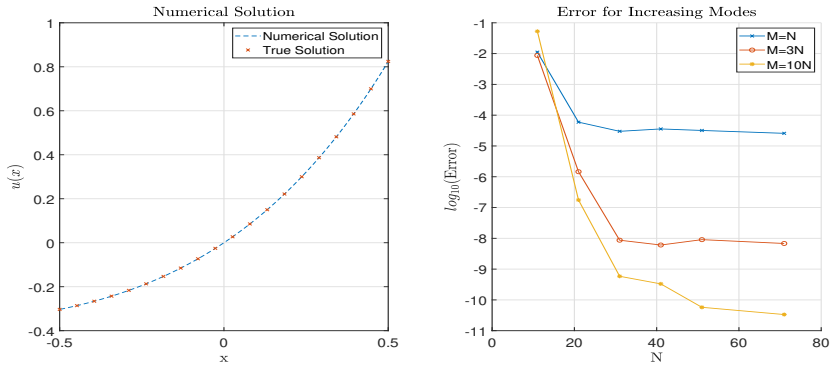

Figure 2.2: The first plot shows the true and numerical solution. The second plot shows the error for increasing N for different sampling rates.

Figure 2.2 shows the numerical results of implementing the numerical method with different sampling strategies. The numerical results were computed using a regularization parameter of $\epsilon = 10^{-5}$. The first plot shows the true and numerical solution that was computed using 71 Fourier frame modes and 350 discretization points. The second plot shows how the error decays for three different sampling rates $M = N$, $M = 3N$ and $M = 10N$. As the number of Fourier frame modes N increases the difference between the accuracy of the numerical solution obtained using different sampling rates is more pronounced. The numerical solution computed with $N = 71$ and without oversampling is accurate to within 10^{-4}. The numerical solution computed without oversampling is less accurate then the previous example. With a sampling rate of 10 and $N = 71$ our numerical solution is accurate to within 10^{-10}. This result

is similar to the previous example. The numerical solution computed using 71 Fourier frame modes and a sampling rate of 3 is accurate to within 10^{-8}. This is in contrast to the previous example where both sampling rates $M = 3N$ and $M = 10N$ were almost equally accurate. In this example using the sampling rate $M = 3N$ and $M = 10N$ provides an increase in accuracy of at least 10^{-3} and 10^{-5}, respectively.

The two example show the our numerical method provides accurate numerical solution with 31 Fourier frame modes and a sufficient sampling rate.

2.2.3 Elliptic PDEs on Circular Domain

Our next example is the following PDE on a circle of radius 0.5.

$$\begin{aligned}
\Omega &= \{(x,y) \in \mathbb{R}^2 \ : \ x^2 + y^2 < 0.25\}, \\
\partial\Omega &= \{(x,y) \in \mathbb{R}^2 \ : \ x^2 + y^2 = 0.25\}, \\
\Delta u(x,y) &= 16\pi \cos(4\pi(x^2+y^2)) - 64\pi^2(x^2+y^2)\sin(4\pi(x^2+y^2)), \quad (x,y) \in \Omega, \\
u(x,y) &= 0, \quad (x,y) \in \partial\Omega.
\end{aligned} \tag{2.31}$$

The true solution to the above PDE is,

$$u(x,y) = \sin(4\pi(x^2+y^2)), \quad (x,y) \in \Omega. \tag{2.32}$$

Our truncated frame approximation of the true solution is,

$$u_N(x,y) = \sum_{|p|\leq N} \sum_{|q|\leq N} c_{p,q} e^{i\pi(px+qy)}, \quad (x,y) \in \Omega. \tag{2.33}$$

We now outline the discretization strategy. Let $X = \{(x_{j_1}, y_{j_2})\}_{(j_1,j_2)\in\mathcal{I}_{M_d}}$ be a uniform equally spaced tensor grid of the unit square centred at 0, where $\mathcal{I}_{M_d}$ is a finite index set. Our set of interior points Ω_{M_i} consists of points of X that are in Ω.

$$\Omega_{M_i} = \{(x_{j_1}, y_{j_2}) \in X : x_{j_1}^2 + y_{j_2}^2 < .25\}. \tag{2.34}$$

In order to construct our set of boundary points $\partial\Omega_{M_b}$ we use polar coordinates to parametrize the boundary. Let $\theta \in (0, 2\pi]$, then $(x,y) \in \partial\Omega$ can be defined as $x = 0.5\cos(\theta)$ and $y = 0.5\sin(\theta)$. Now let $\{\theta_k\}_{k=1}^{M_b}$ be a finite uniform discretization of $(0, 2\pi]$. Then our set of boundary points is defined as,

$$\partial\Omega_{M_b} = \{(x_k, y_k) \ : \ x_k = 0.5\cos(\theta_k), \ y_k = 0.5\sin(\theta_k) \ , 1 \leq k \leq M_b\}. \tag{2.35}$$

As before we require the number of discretized points to be greater then the number of Fourier frame modes, i.e. $M_b + M_i = M >> (2N+1)^2$. For simplicity of notation let us define a few terms.

$$\begin{aligned}
\psi_{p,q}(x_{j_1}, y_{j_2}) &= e^{i\pi(px_{j_1} + qy_{j_2})}, \\
f(x_{j_1}, y_{j_2}) &= 16\pi\cos(4\pi(x_{j_1}^2 + y_{j_2}^2)) - 64\pi^2(x_{j_1}^2 + y_{j_2}^2)\sin(4\pi(x_{j_1}^2 + y_{j_2}^2)).
\end{aligned} \tag{2.36}$$

Now substituting (2.33) into (2.31) and evaluating it at the M_i interior points and M_b boundary points, we have the following system of equations.

$$\begin{aligned}\sum_{|p|\le N}\sum_{|q|\le N} -c_{p,q}\pi^2(p^2+q^2)\psi_{p,q}(x_{j_1},y_{j_2}) = f(x_{j_1},y_{j_2}), \quad (x_{j_1},y_{j_2})\in\Omega_{M_i},\\ \sum_{|p|\le N}\sum_{|q|\le N} c_{p,q}\psi_{p,q}(x_k,y_k) = 0, \quad (x_k,y_k)\in\partial\Omega_{M_b}.\end{aligned} \tag{2.37}$$

Rewriting the above equations in matrix form we have,

$$\begin{aligned}\boldsymbol{A}\boldsymbol{c}=\boldsymbol{F}, \qquad \boldsymbol{A}=\begin{bmatrix}\boldsymbol{A}_{int}\\ \boldsymbol{A}_{bd}\end{bmatrix}, \quad \boldsymbol{F}=\begin{bmatrix}\boldsymbol{f}\\ \boldsymbol{0}\end{bmatrix},\\ \boldsymbol{A}_{int}\in\mathbb{C}^{M_i\times(2N+1)^2}, \ \boldsymbol{A}_{bd}\in\mathbb{C}^{M_b\times(2N+1)^2}, \ \boldsymbol{f}\in\mathbb{R}^{M_i}, \ \boldsymbol{c}\in\mathbb{C}^{(2N+1)^2},\\ (\boldsymbol{A}_{int})_{j,k} = -\pi^2(p^2+q^2)\psi_{p,q}(x_{j_1},y_{j_2}), \quad -N\le p,q\le N, \ (x_{j_1},y_{j_2})\in\Omega_{M_i},\\ (\boldsymbol{A}_{bd})_{j,k} = \psi_{p,q}(x_j,y_j), \quad -N\le p,q\le N, \ (x_j,y_j)\in\partial\Omega_{M_b},\\ f_j = f(x_{j_1},y_{j_2}), \quad (x_{j_1},y_{j_2})\in\Omega_{M_i}.\end{aligned} \tag{2.38}$$

As before, we approximate the Fourier frame coefficients $\boldsymbol{c}$ by solving the regularized least square minimization problem. Substituting the approximated frame coefficients into the truncated frame representation provides us with our numerical approximation.

$$\begin{aligned}u_N(x,y) = \sum_{|p|\le N}\sum_{|q|\le N} c^{\epsilon}_{p,q}e^{i\pi(px+qy)}, \quad (x,y)\in\Omega,\\ \boldsymbol{c}^{\epsilon} = \boldsymbol{V}\boldsymbol{\Sigma}^{\dagger}_{\epsilon}\boldsymbol{U}^{*}\boldsymbol{b}.\end{aligned} \tag{2.39}$$

Numerical Result: Example 3 on Circular Domain

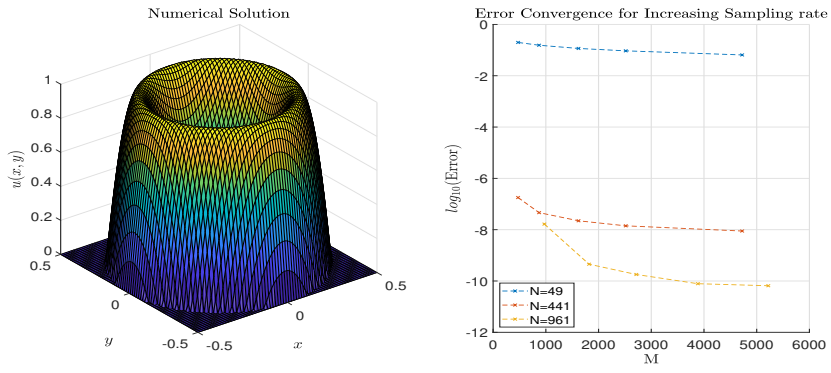

Figure 2.3: The first plot shows the numerical solution. The second plot shows how the error behaves for a given N as M increases.

Figure 2.3 shows the numerical results from implementing the above outlined method. The numerical results were computed using a regularization parameter of $\epsilon = 10^{-4}$. The first plot shows the numerical solution computed using 961 Fourier frame modes, 3720 Interior points and 1500 boundary points. The second plot shows how the error behaves as we increase the sampling rate for three fixed number of Fourier frame modes $N = 49$, $N = 441$ and $N = 961$. From the plot we see that with 49 Fourier frame modes even having a sampling rate of 90 ($M = 90N$) only provides an accuracy of 10^{-1}. In order to approximate the true solution we need to consider a sufficient number of Fourier frame modes. The numerical solution computed using 441 Fourier frame modes and discretization points is accurate to within 10^{-7}. With 441 Fourier Modes oversampling provides an increase in accuracy of approximately 10^{-1}. With 961 Fourier frame modes our numerical solution is approximately accurate to 10^{-8} without oversampling and to 10^{-10} with oversampling. For $N = 961$ oversampling provides an increase in accuracy of 10^{-2}.

The fourth example is the following PDE defined on the same circular domain of radius 0.5 as in the previous example (2.31),

$$\begin{aligned} 2\frac{\partial^2 u}{\partial x^2} + 4\frac{\partial^2 u}{\partial y^2} - 2\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} &= 4x - 2y - 12, \quad (x, y) \in \Omega, \\ u(x, y) &= 0, \qquad (x, y) \in \partial\Omega. \end{aligned} \tag{2.40}$$

The true solution to the above equation is,

$$u(x, y) = \frac{1}{4} - (x^2 + y^2). \tag{2.41}$$

Numerical Results: Example 4 on Circular Domain

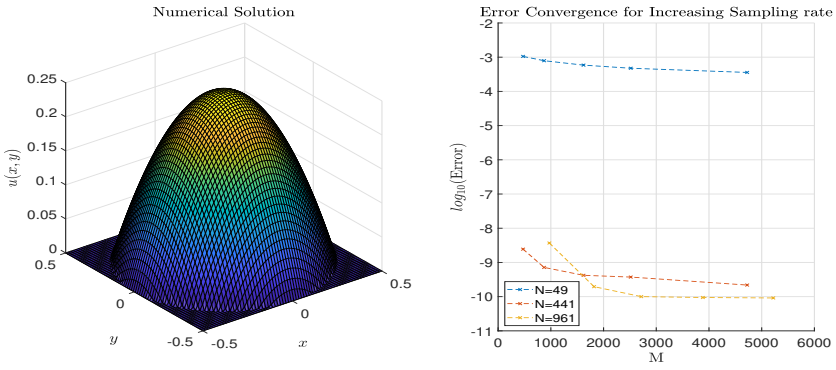

Figure 2.4: The first plot shows the numerical solution. The second plot shows how the error behaves for a given N as M increases, i.e. error behaviour as we sample more for a given N.

Figure 2.4 shows the numerical results from implementing the numerical method. The numerical results were computed using a regularization parame-

ter of $\epsilon = 10^{-4}$. The first plot shows the numerical solution computed using $N = 961$ and $M = 5220$. The second plot show how the error behaves as we oversample for $N = 49$, $N = 441$ and $N = 961$. With 49 Fourier frame modes our numerical solution is accurate to within 10^{-3}. Using 441 Fourier frame modes and discretization points provides an accuracy of 10^{-8} and oversampling provides an increase in accuracy of approximately 10^{-1}. For 961 Fourier frame modes our solution is approximately accurate to $10^{-8.5}$ without oversampling and to 10^{-10} with oversampling.

The numerical results from example three and four indicate that our numerical method approximates the true solution on a circular domain accurately provided we consider a sufficient number of Fourier frame modes. Oversampling does not provide as large an increase in accuracy as the ODE case (example 1 and 2).

2.2.4 Poisson PDE on Icecream Cone Domains

In our next couple of examples we solve the Poisson PDE on two irregular domains. The first domain is a convex domain consisting of a semicircle and an inverted triangle. The second domain is non-convex and consists of three semicircles, a rectangle and an inverted triangle. We will refer to these domains as the convex ice cream cone domain and non-convex ice cream cone domain. A figure of each domain is provided in Appendix B. The circular domain considered in the previous subsection is convex with an infinitely differentiable boundary. Both the convex and non-convex ice cream cone domains have a continuous boundary. The aim of the next two examples is to study how well our numerical method preforms when we reduce the smoothness properties of the domain on which the PDE is defined.

We begin by defining the convex ice cream cone domain.

$$\begin{aligned}\Omega =&\{(x,y)\in\mathbb{R}^2 : y < \sqrt{0.25-x^2},\ -0.5<x<0.5\}\cup\\ &\{(x,y)\in\mathbb{R}^2 : y > -x-0.5,\ y<0,\ -0.5<x<0\}\cup\\ &\{(x,y)\in\mathbb{R}^2 : y > x-0.5,\ y<0,\ 0<x<0.5\},\\ \partial\Omega =&\ \overline{\Omega}\backslash\Omega.\end{aligned} \tag{2.42}$$

We are interested in solving the following Poisson PDE on the convex ice cream cone domain,

$$\begin{aligned}-\Delta u(x,y) &= 5, \quad (x,y)\in\Omega,\\ u(x,y) &= 0, \quad (x,y)\in\partial\Omega.\end{aligned} \tag{2.43}$$

Figure 2.5 shows the numerical results from implementing the numerical method. The numerical results were computed using a regularization parameter of $\epsilon = 10^{-4}$. The first plot of Figure 2.5 shows the reference solution computed using $2,401$ Fourier frame modes, $11,704$ interior points and $1,979$ boundary points. The second plot shows the error as we increase the number of Fourier frame modes for a sampling rate within a given range. The range for the two sampling rates are $1.5N \leq M \leq 2N$ and $6.4N \leq M \leq 6.9N$. The ratio of boundary

points (M_b) to total number of discretization points (M) in the computation of each numerical solution for both sampling rates is between 0.15 and 0.17. The accuracy of the numerical solution increases with the number of Fourier frame modes. The numerical solution with 961 Fourier frame modes and 1695 discretization points (sampling rate of 1.7) is accurate to within $10^{-6.5}$. For the same number of modes $N = 961$ and $6,347$ discretization points (sampling rate of 6.6) we have almost the same numerical accuracy. In this example oversampling provides a marginal increase in numerical accuracy. This is in contrast to example 3 and example 4 where with $N = 961$, oversampling provided an increase in accuracy of order 10^{-2} and $10^{-1.5}$, respectively.

Poisson PDE on Convex Ice-Cream Cone Domain

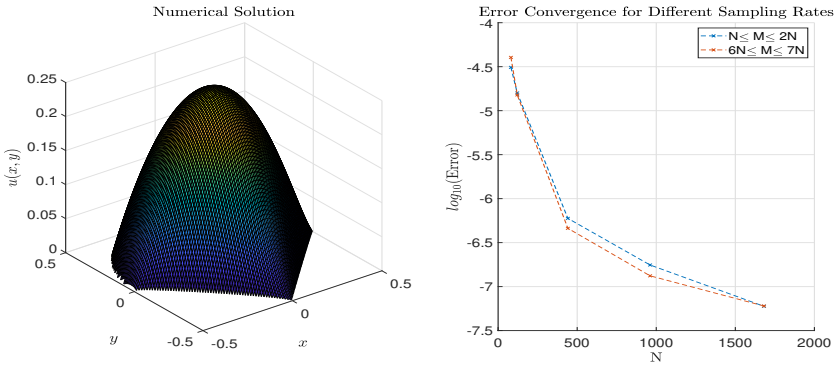

Figure 2.5: The first plot shows the numerical solution. The second plot shows how the error behaves as we consider more Fourier frame modes N with a given sampling rate.

The next example is of a Poisson PDE on the non-convex ice cream cone domain. We being by defining the domain Ω and its boundary.

$$\begin{aligned}\Omega =&\{(x,y)\in\mathbb{R}^2 : (x+0.25)^2+y^2<(0.25)^2,\ -0.5<x<-0.25,\ 0<y<0.25\}\cup\\ &\{(x,y)\in\mathbb{R}^2 : x^2+(y-0.25)^2<(0.25)^2,\ -0.25<x<0.25,\ 0.25<y<0.5\}\cup\\ &\{(x,y)\in\mathbb{R}^2 : (x-0.25)^2+y^2<(0.25)^2,\ 0.25<x<0.5,\ 0<y<0.25\}\cup\\ &\{(x,y)\in\mathbb{R}^2 : -0.25<x<0.25,\ 0<y<0.25\}\cup\\ &\{(x,y)\in\mathbb{R}^2 : y>-x-0.5,\ -0.5<x<0,\ 0<y<-0.5\}\cup\\ &\{(x,y)\in\mathbb{R}^2 : y>x-0.5,\ 0<x<0.5,\ 0<y<-0.5\},\\ \partial\Omega =&\ \overline{\Omega}\backslash\Omega.\end{aligned} \tag{2.44}$$

We will solve the following Poisson PDE on the non-convex ice cream cone domain,

$$\begin{aligned}-\Delta u(x,y) &= 5, \quad (x,y)\in\Omega,\\ u(x,y) &= 0, \quad (x,y)\in\partial\Omega.\end{aligned} \tag{2.45}$$

Figure 2.6 shows the numerical results. The numerical results were computed using a regularization parameter of $\epsilon = 10^{-4}$. The first plot shows the reference solution computed using $2,401$ Fourier frame modes, $11,860$ interior points and $1,999$ boundary points. The second plot shows the error as we increase the number of Fourier frame modes for a sampling rate within a given range. The range for the sampling rates is $1.6N \leq M \leq 1.7N$ and $6N \leq M \leq 6.1N$. The ratio of boundary points (M_b) to total number of discretization points (M) in the computation of each numerical solution for both sampling rates is between 0.150 and 0.181. The numerical solution with 961 Fourier frame modes and 1623 discretization points is accurate to within 10^{-5}. Oversampling with $N = 961$ provides marginally worse results. The numerical result for this example are less accurate then the convex ice cream cone domain example by approximately $10^{-1.5}$.

Poisson PDE on Non-Convex Ice Cream Cone

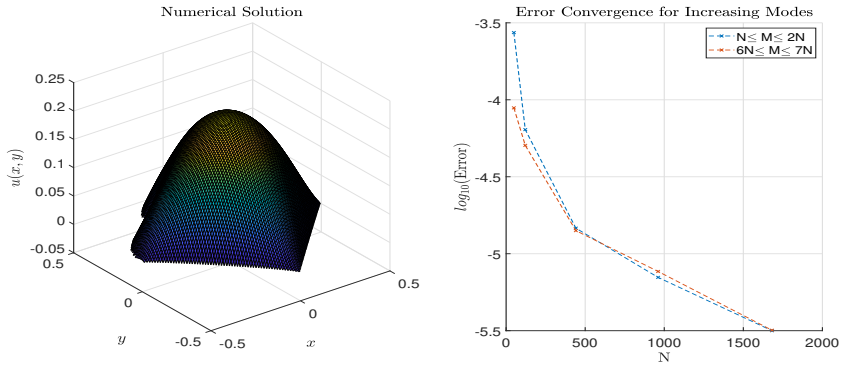

Figure 2.6: The first plot shows the numerical solution. The second plot shows how the error behaves as we consider larger N for a sampling rate within a given range.

Example five and six illustrate that our numerical method provides more accurate results for convex domains. Furthermore, the difference in the accuracy between example 4 (circular domain) and example 5 (convex ice cream cone domain) show that the regularity of the boundary of the domain also affects the accuracy of the numerical solution.

2.2.5 PDE on Cloud Domain

Our last example is of an irregular non-convex domain with continuous boundary which we refer to as the cloud domain. The domain is defined as follow.

$$\begin{aligned}\Omega^+ =&\{(x,y) \in \mathbb{R}^2 : (x+0.25)^2 + y^2 < (0.25)^2,\ -0.5 < x < -0.25,\ 0 < y < 0.25\} \cup \\ &\{(x,y) \in \mathbb{R}^2 : x^2 + (y-0.25)^2 < (0.25)^2,\ -0.25 < x < 0.25,\ 0.25 < y < 0.5\} \cup \\ &\{(x,y) \in \mathbb{R}^2 : (x-0.25)^2 + y^2 < (0.25)^2,\ 0.25 < x < 0.5,\ 0 < y < 0.25\},\end{aligned}$$

$$\begin{aligned}\Omega^- =&\{(x,-y) \ : \ (x,y)\in\Omega^+\},\\ \Omega =&\Omega^+\cup\Omega^-\cup\{(x,y)\in\mathbb{R}^2 : -0.25<x<0.25,\ -0.25<y<0.25\},\\ \partial\Omega =&\overline{\Omega}\backslash\Omega.\end{aligned} \tag{2.46}$$

We will solve the following PDE on the cloud domain,

$$\begin{aligned} u(x,y)+\Delta u(x,y) &= -10^3 xy, \qquad (x,y)\in\Omega,\\ u(x,y) &= 0, \qquad\quad (x,y)\in\partial\Omega. \end{aligned} \tag{2.47}$$

Numerical Results: Example 7 on Cloud Domain

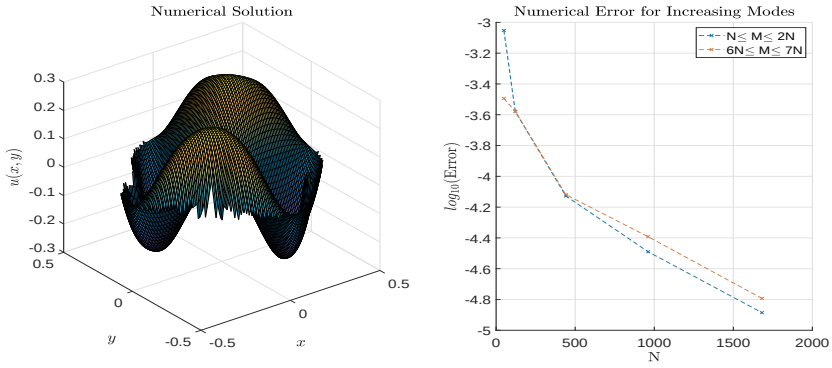

Figure 2.7: The first plot shows the numerical solution. The second plot shows how the error behaves as we consider larger N for a sampling rate within a given range.

Figure 2.7 shows the numerical results. The numerical results were computed using a regularization parameter of $\epsilon = 10^{-4}$. The first plot shows the reference solution computed using $2,401$ Fourier frame modes, $11,712$ interior points and $1,438$ boundary points. The second plot shows the error as we increase the number of Fourier frame modes for a sampling rate within a given range. The range for the sampling rates is $1.5N \leq M \leq 1.9N$ and $6.8N \leq M < 7N$. The ratio of boundary points (M_b) to total number of discretization points (M) in the computation of each numerical solution for both sampling rates is between 0.15 and 0.20. Our results are similar to the non-convex ice cream cone domain. With 961 Fourier frame modes and 1450 discretization points our solution is accurate to within 10^{-4}. Oversampling does not provide any worthwhile increase in accuracy.

2.3 Numerical Method for Laplace Eigenvalue Problem

In this section we briefly outline how the numerical method of section 2.1 can be modified to compute the eigenvalues of the Laplace operator on a bounded

irregular domain Ω. Similar to section 2.1, we will construct a linear system by substituting the truncated Fourier frame representation into the Laplace eigenvalue problem and evaluate the resulting equations at points in the domain Ω and on the boundary $\partial\Omega$. We will then solve for the eigenvalues of the resulting system of equations using the MATLAB routine "eig". We end this section by providing numerical results from applying the numerical method to the Laplace eigenvalue problem on the unit interval and a circle of radius 0.5. We provide the details for the implementation of the numerical method for the first example.

2.3.1 LEP and Numerical Method

The Laplace eigenvalue problem (LEP) is to find $\lambda \in \mathbb{C}$ and $u \in H_0^1(\Omega)$ that solves,

$$\begin{aligned} -\Delta u &= \lambda u, \qquad \boldsymbol{x} \in \Omega \subset [-1,1]^d, \\ u &= 0, \qquad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{2.48}$$

The pair (λ, u) that satisfy (2.48) are known as the eigenvalue and eigenfunction of the Laplace operator.

We begin by stating a couple of results. The first result ensures there exist a solution pair (λ, u) to the LEP.

Theorem 2.3.1. *(Theorem 8.37 of [18]) There exist a countable sequence* $\{(\lambda_n, u_n)\}_{n\in\mathbb{N}}$ *of eigenvalue and eigenfunction pair which satisfies the Laplace eigenvalue problem.*

The next result states a few properties of the eigenvalues that satisfy (2.48). These properties can be used to check if the numerical scheme computes the eigenvalues with the correct behaviour. Furthermore, our numerical scheme will require the eigenvalues to be strictly positive.

Theorem 2.3.2. *(Section 6.5, Theorem 1 of [10]) Let* $\{\lambda_n\}_{n\in\mathbb{N}}$ *be a sequence of eigenvalues of the Laplace eigenvalue problem such that* $|\lambda_i| \leq |\lambda_j|$ *for* $i < j$. *Then,*

i λ_i *is real for* $i \in \mathbb{N}$,

ii $\lambda_1 > 0$,

iii $\lim_{n\to\infty} \lambda_n = \infty$.

Now we outline our numerical scheme to compute the eigenvalues $\boldsymbol{\lambda} \in \mathbb{R}^{(2N+1)^d}$, where we will assume the eigenvalues are arranged in increasing order, i.e. $|\lambda_i| \leq |\lambda_j|$ for $1 \leq i \leq j \leq (2N+1)^d$. Consider the truncated Fourier frame representation of $u(\boldsymbol{x})$,

$$u(\boldsymbol{x}) = \sum_{\substack{|n_l| \leq N \\ 1 \leq l \leq d}} c_{\boldsymbol{n}} e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}. \tag{2.49}$$

Let Ω_{M_i} be the set containing points of an equispaced finite mesh that are in Ω and let $\partial\Omega_{M_b}$ be the finite set of boundary points as defined in (2.7) and (2.8), respectively. We substitute our finite frame representation (2.49) into (2.48) and evaluate the resulting equations at the points in Ω_{M_i} and $\partial\Omega_{M_b}$. We have the following M_i equations corresponding to the points in the domain Ω and M_b equations corresponding to the points on the boundary $\partial\Omega$.

$$\begin{aligned}\sum_{\substack{|n_l|\le N\\ 1\le l\le d}} c_{\boldsymbol{n}}\pi^2(\boldsymbol{n}\cdot\boldsymbol{n})e^{i\pi\boldsymbol{n}\cdot\boldsymbol{x}_j} &= \lambda \sum_{\substack{|n_l|\le N\\ 1\le l\le d}} c_{\boldsymbol{n}}e^{i\pi\boldsymbol{n}\cdot\boldsymbol{x}_j}, \qquad \boldsymbol{x}_j\in\Omega_{M_i},\\ \sum_{\substack{|n_l|\le N\\ 1\le l\le d}} c_{\boldsymbol{n}}e^{i\pi\boldsymbol{n}\cdot\boldsymbol{x}_k} &= 0, \qquad \boldsymbol{x}_k\in\partial\Omega_{M_b}.\end{aligned} \tag{2.50}$$

Writing the above system in matrix form we have,

$$\begin{aligned}&\boldsymbol{A}\boldsymbol{c} = \lambda\boldsymbol{B}\boldsymbol{c},\\ \text{where}\quad &\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}^{int}\\ \boldsymbol{A}^{bdy}\end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix}\boldsymbol{B}^{int}\\ \boldsymbol{B}^{bdy}\end{bmatrix},\\ &\boldsymbol{A}\in\mathbb{C}^{M\times(2N+1)^d},\\ &\boldsymbol{A}^{int}\in\mathbb{C}^{M_i\times(2N+1)^d}, \quad \boldsymbol{A}^{bdy}\in\mathbb{C}^{M_b\times(2N+1)^d},\\ &\boldsymbol{B}^{int}\in\mathbb{C}^{M_i\times(2N+1)^d}, \quad \boldsymbol{B}^{bdy}\in\boldsymbol{0}^{M_b\times(2N+1)^d}.\end{aligned} \tag{2.51}$$

In order to solve for the eigenvalues $\boldsymbol{\lambda}$, we convert the above $M\times(2N+1)^d$ linear system into a square $(2N+1)^d\times(2N+1)^d$ system as follow,

$$\begin{aligned}\boldsymbol{A}\boldsymbol{c} &= \lambda\boldsymbol{B}\boldsymbol{c},\\ \boldsymbol{c} &= \lambda\boldsymbol{A}^\dagger\boldsymbol{B}\boldsymbol{c},\\ \frac{1}{\lambda}\boldsymbol{c} &= \boldsymbol{A}^\dagger\boldsymbol{B}\boldsymbol{c}, \qquad (\lambda>0),\\ \boldsymbol{Z}\boldsymbol{c} &= \tilde{\lambda}\boldsymbol{c},\\ \text{where}\quad \boldsymbol{Z} &= \boldsymbol{A}^\dagger\boldsymbol{B}, \quad \tilde{\lambda} = \frac{1}{\lambda}.\end{aligned} \tag{2.52}$$

In the third line we have used the fact that the eigenvalues are strictly positive. Similar to the previous section we use SVD regularization as the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are ill-conditioned. Let $\boldsymbol{A} = \boldsymbol{U}_A(\boldsymbol{\Sigma}_A)_\epsilon\boldsymbol{V}_A^*$ and $\boldsymbol{B} = \boldsymbol{U}_B(\boldsymbol{\Sigma}_B)_\epsilon\boldsymbol{V}_B^*$ be the regularized SVD of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. Then,

$$\begin{aligned}\boldsymbol{Z}_\epsilon\boldsymbol{c}_\epsilon &= \tilde{\lambda}_\epsilon\boldsymbol{c}_\epsilon,\\ \text{where}\quad \boldsymbol{Z}_\epsilon &= \boldsymbol{A}_\epsilon^\dagger\boldsymbol{B}_\epsilon,\\ &= \boldsymbol{V}_A(\boldsymbol{\Sigma}_A)_\epsilon^\dagger\boldsymbol{U}_A^*\boldsymbol{U}_B(\boldsymbol{\Sigma}_B)_\epsilon\boldsymbol{V}_B^*.\end{aligned} \tag{2.53}$$

$\boldsymbol{Z}_\epsilon$ is a square matrix with $(2N+1)^d$ rows and columns. Now we numerically approximate $\tilde{\boldsymbol{\lambda}}_\epsilon\in\mathbb{R}^{(2N+1)^d}$, the eigenvalues of $\boldsymbol{Z}_\epsilon$. By taking the reciprocal of

the eigenvalues of $\boldsymbol{Z}_\epsilon$ we have $\boldsymbol{\lambda}_\epsilon$, our approximation to the eigenvalues of the LEP. The subscript ϵ denotes the dependence on the regularization parameter.

We now provide numerical results from two different numerical implementations of the above method. The first method which we refer to as the Regularized-SVD method, computes the eigenvalues of $\boldsymbol{Z}_\epsilon$ (2.53) using the MATLAB command "eig" and takes the reciprocal of the eigenvalues. The second method which we refer to as the Non-SVD Regularized method, computes the matrix $\boldsymbol{Z} = \boldsymbol{A}^\dagger \boldsymbol{B}$ using the MATLAB command "\". The MATLAB routine "\" implements its own pre-conditioning and regularization method. We then compute the eigenvalues of $\boldsymbol{Z}$ using the MATLAB command "eig" and take the reciprocal of the eigenvalues. In order to compute the error between the true and computed eigenvalue we use the ℓ_∞ norm. If λ_{true} and λ_{num} are the true and computed eigenvalue then the error (E) is computed as,

$$E = |\lambda_{num} - \lambda_{true}|. \tag{2.54}$$

2.3.2 LEP on Unit Interval

Our first example is the following LEP on the unit interval,

$$\begin{aligned} -u''(x) &= \lambda u(x), \qquad x \in \Omega = (0,1), \\ u(0) &= 0 = u(1). \end{aligned} \tag{2.55}$$

The true eigenvalue and eigenfunction pair to the above LEP are $\{(n^2\pi^2, \sin(n\pi x))\}_{n\in\mathbb{N}}$.

Our set of boundary points $\partial\Omega_{M_b}$ consists of two points $x_1 = 0$ and $x_M = 1$. For the interior points, let Ω_{M-2} be a uniform discretization of length δ_x of the unit interval.

$$\begin{aligned} \Omega_{M-2} &= \{x_k \ : \ x_k \in \Omega, \ x_{k+1} - x_k = \delta_x, \ 2 \le k \le M-1\}, \\ \partial\Omega_{M_b} &= \{x_1, x_M \ : \ x_1 = 0, \ \ x_M = 1\}. \end{aligned} \tag{2.56}$$

For $N \in \mathbb{N}$ our truncated Fourier frame is,

$$u_N(x) = \sum_{|n|\le N} c_n e^{i\pi n x}. \tag{2.57}$$

By substituting the frame representation into the LEP and and evaluating the resulting equations on the set of boundary and interior points we have the following system,

$$\begin{aligned} \boldsymbol{A}\boldsymbol{c} &= \lambda \boldsymbol{B}\boldsymbol{c}, \\ \text{where} \qquad \boldsymbol{A} &= \begin{bmatrix} \boldsymbol{A}^{int} \\ \boldsymbol{A}^{bdy} \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}^{int} \\ \boldsymbol{B}^{bdy} \end{bmatrix}, \\ \boldsymbol{A}^{int}_{j,k} &= k^2\pi^2 e^{ik\pi x_j}, \qquad x_j \in \Omega_{M_i}, \\ \boldsymbol{B}^{int}_{j,k} &= e^{ik\pi x_j}, \qquad x_j \in \Omega_{M_i}, \\ \boldsymbol{A}^{bdy}_{j,k} &= e^{ik\pi x_j}, \qquad x_j \in \partial\Omega_{M_b}, \\ \boldsymbol{B}^{bdy}_{j,k} &= 0, \qquad x_j \in \partial\Omega_{M_b}. \end{aligned} \tag{2.58}$$

Rewriting (2.58) as a square system of $2N+1$ equations and carrying out SVD regularization we have,

$$\begin{aligned} \boldsymbol{Z}_\epsilon \boldsymbol{c}_\epsilon &= \tilde{\lambda}_\epsilon \boldsymbol{c}_\epsilon, \\ \text{where} \quad \boldsymbol{Z}_\epsilon &= \boldsymbol{A}_\epsilon^\dagger \boldsymbol{B}_\epsilon, \\ &= \boldsymbol{V}_A (\boldsymbol{\Sigma}_A)_\epsilon^\dagger \boldsymbol{U}_A^* \boldsymbol{U}_B (\boldsymbol{\Sigma}_B)_\epsilon \boldsymbol{V}_B^*. \end{aligned} \tag{2.59}$$

Numerical Error for Regularized SVD Eigenvalues

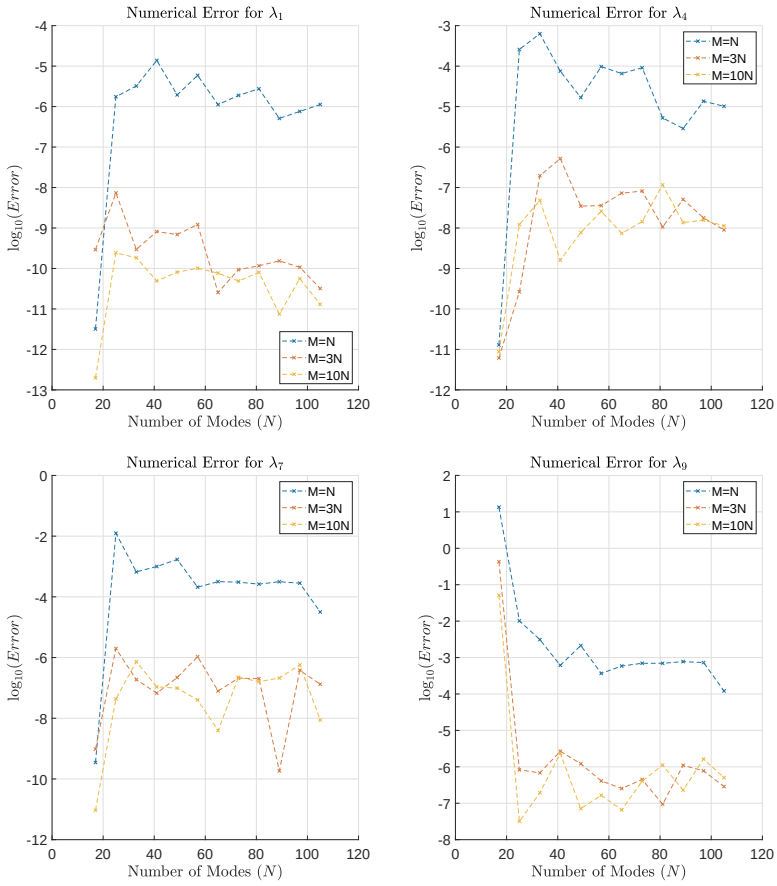

Figure 2.8: Each plot shows the convergence of an eigenvalue for different sampling strategies. Oversampling provides more accurate results for all four eigenvalues.

Figure 2.8 shows the numerical result from computing the reciprocal of the eigenvalues of $\boldsymbol{Z}_\epsilon$ computed using the MATLAB command "eig". The numerical results were computed using a regularization parameter of $\epsilon = 10^{-5}$. The

computed first eigenvalues λ_1 is accurate to within 10^{-6} without oversampling and to 10^{-10} with oversampling. The computed fourth eigenvalues λ_4 is accurate to within 10^{-5} without oversampling and to 10^{-8} with oversampling. The computed ninth eigenvalues λ_9 is accurate to within 10^{-4} without oversampling and to 10^{-6} with oversampling. Oversampling provides an increase an accuracy in the computation of all three eigenvalues λ_1, λ_4 and λ_9. Our numerical approximation is less accurate for higher eigenvalues. With 105 Fourier frame modes and 1050 discretization points the numerical approximation of the first eigenvalue is more accurate then the numerical approximation of the ninth eigenvalue by approximately 10^{-5}.

Numerical Error for Non-SVD Regularized Eigenvalues

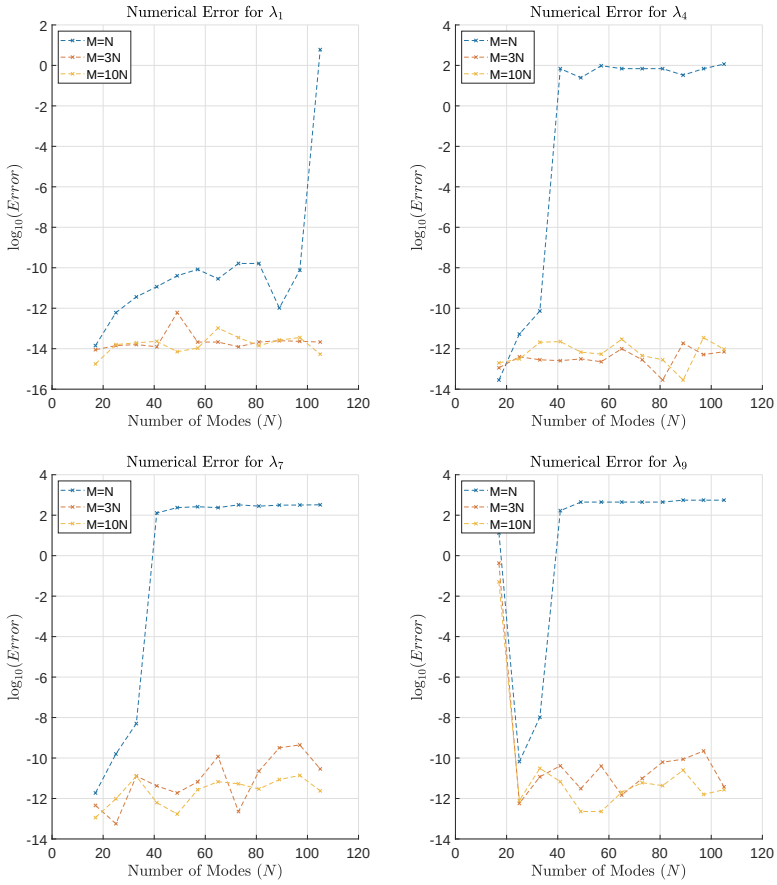

Figure 2.9: Each plot shows the convergence of an eigenvalue for different sampling strategies.

Figure 2.9 shows the numerical results from using the Non-SVD Regular-

ization method to compute the eigenvalues. From the results we can see that with oversampling all four eigenvalues λ_1, λ_4, λ_7 and λ_9 are accurate to within 10^{-11}. Without oversampling ($M = N$) our computed eigenvalues do not converge to the true eigenvalues. This is in contrast to the eigenvalues computed using Regularized SVD method.

2.3.3 LEP on Circle

Our second example is the following LEP on a circular domain of radius 0.5.

$$\begin{aligned} -\Delta u(x,y) &= \lambda u(x,y), \qquad (x,y) \in \Omega = \{x^2 + y^2 < 0.25\}, \\ u(x,y) &= 0, \qquad\qquad\quad (x,y) \in \partial\Omega = \{x^2 + y^2 = 0.25\}. \end{aligned} \tag{2.60}$$

For $n \in \mathbb{N}$, let $J_n(x)$ denote the Bessel function of the first kind and $j_{n,k}$ denote the k-th root of $J_n(x)$ in increasing order. The eigenvalues that satisfy (2.60) are $\lambda_{n,k} = 4j_{n,k}^2$. In order to compute the true eigenvalues we have used the MATLAB function Bessel Zero Solver [14] to compute the roots of the Bessel function of the first kind.

Numerical Error for Eigenvalues on Circular Domain

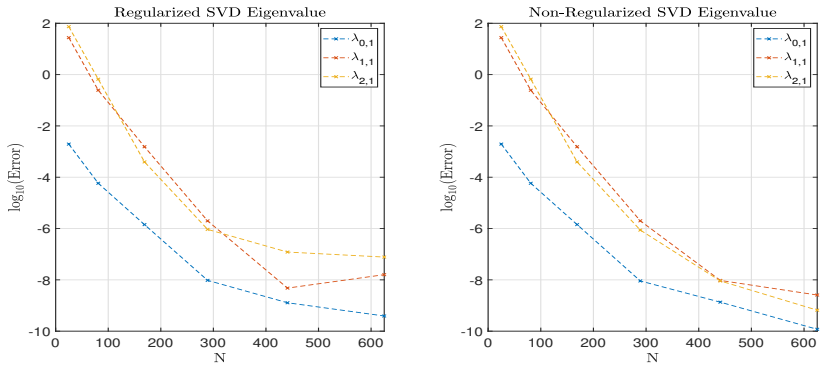

Figure 2.10: The first plot shows the numerical results for the Regularized SVD method. The second plot shows numerical result from the Non-SVD Regularized method.

Figure 2.10 shows the numerical results from the Regularized SVD and Non-SVD Regularized method in computing the three eigenvalues $\lambda_{0,1}$, $\lambda_{1,1}$ and $\lambda_{2,1}$. The numerical results in the first plot were computed using a regularization parameter of 10^{-5}. The number of interior and boundary points used in the computation of each numerical solution for both plots is $5,224$ and $1,500$, respectively. With 625 Fourier frame modes the numerical eigenvalues $\lambda_{0,1}$ and $\lambda_{1,1}$ are approximately accurate to 10^{-10} and 10^{-8} for both methods. The eigenvalue $\lambda_{2,1}$ computed using the Non-SVD Regularized method is approximately accurate to 10^{-9}, whereas the same eigenvalue computed using the Regularized

SVD method is accurate to within 10^{-7}. Similar to the previous example, higher eigenvalues are more accurate for the Non-SVD regularized method.

Chapter 3

Numerical Analysis for Poisson PDE

In this chapter we use the variational framework to study the numerical method applied to the Poisson PDE on a bouded domain. From the previous chapter we saw that our numerical method was spectrally accuracy for the ODE case and for the two dimensional case the rate of convergence depended on the domain and smoothness of the boundary. The main result that we will work towards in this chapter shows that our numerical approximation from a closed subspace of $H^2(\Omega)$ to the true weak solution of the Poisson PDE is the best approximation up to a constant. The constant depends on the dimension of the closed subspace, embedding constants from the several embedding theorems and the domain Ω via the Poincaré inequality.

We begin this chapter by providing an outline of results from Sobolev space and elliptic PDE regularity theory that we will be using in this report. In the next section we state the variational formulation and provide an example of how the variational framework is applied to the Poisson PDE. Next using the Lax-Milgram theorem we show that the variational formulation corresponding to our numerical method has a unique solution for the one dimensional case. For dimension greater then or equal to two the bilinear form corresponding to our numerical method is not coercive. As a result we consider a modified variational formulation and carry out a similar analysis. In the last section we outline several results that lead towards our error estimate for the modified variational formulation.

3.1 Sobolev Spaces and Elliptic Regularity

In this section we provide a short and concise summary of the background material that we will need in our analysis of the numerical method - outlined in Chapter 2 - applied to the Poisson PDE.

We begin with stating the definition of Hölder continuity.

Definition 3.1.1. *Let $\boldsymbol{x}_0$ be a point in $\mathbb{R}^d$ and f a function defined on a bounded set D containing $\boldsymbol{x}_0$.*

1. *For $0 < \mu < 1$, we say that f is Hölder continuous with exponent μ at $\boldsymbol{x}_0$ if the quantity,*
$$[f]_\mu = \sup_{\boldsymbol{x} \in D} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{x}_0)|}{|\boldsymbol{x} - \boldsymbol{x}_0|^\mu} < \infty. \tag{3.1}$$
If (3.1) is finite with $\mu = 1$ then f is said to be Lipschitz continuous.

2. *We say f is uniformly Hölder continuous with exponent μ in D if,*
$$[f]_\mu = \sup_{\substack{\boldsymbol{x}, \boldsymbol{y} \in D \\ \boldsymbol{x} \neq \boldsymbol{y}}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|^\mu} < \infty. \tag{3.2}$$
A function f that is uniformly Hölder continuous on D with exponent μ will be referred to as μ continuous.

Hölder continuity provides a way to differentiate between the smoothness of two continuous functions, furthermore the exponent can be thought of as fractional differentiability. Using the above definition we state the space of Hölder continuous function $C^{k,\mu}(\Omega)$. Let Ω be an open subset of $\mathbb{R}^d$, then

$$C^{k,\mu}(\Omega) = \left\{ f : \Omega \to \mathbb{R} \ : \ \frac{\partial^k f(\boldsymbol{x})}{\partial x_j^k} \text{is } \mu \text{ continuous}, \ 1 \leq j \leq d \right\}. \tag{3.3}$$

We now work towards providing a formal definition of the regularity of the boundary of the domain Ω.

Definition 3.1.2. *Let $\zeta : \mathbb{R}^{d-1} \to \mathbb{R}$ be such that,*

$$\Omega = \{ \boldsymbol{x} \in \mathbb{R}^{d-1} \ : \ x_d < \zeta(\boldsymbol{x}'), \ \forall \ \boldsymbol{x}' = (x_1, \cdots, \cdots, x_{d-1}) \in \mathbb{R}^{d-1} \}. \tag{3.4}$$

If ζ is in $C^{k,\mu}(\mathbb{R}^{d-1})$ we say that Ω is a $C^{k,\mu}$ hypograph.

Definition 3.1.3. *(Definition 3.28 of [12]) The open set Ω is a $C^{k,\mu}$ domain or of class $C^{k,\mu}$ if its boundary $\partial\Omega$ is compact and if there exist finite families $\{W_j\}$ and $\{\Omega_j\}$ having the following properties:*

1. *The family $\{W_j\}$ is a finite open cover of $\partial\Omega$, i.e. each W_j is an open subset of $\mathbb{R}^d$ such that $\partial\Omega \subset \cup_j W_j$.*

2. *Each Ω_j can be transformed to a $C^{k,\mu}$ hypograph by a rigid motion, i.e. by a rotation plus a translation.*

3. *The set Ω satisfies $W_j \cap \Omega = W_j \cap \Omega_j$ for each j.*

A $C^{0,1}$ domain is referred to as a Lipschitz domain. At this point we introduce multi-index notation. The multi index notation makes stating certain terms and formulas much easier and compact.

Definition 3.1.4. *A multi-index $\boldsymbol{\alpha}$ is an n-tuple of non-negative integers, α_j. The length of $\boldsymbol{\alpha}$ is given by*

$$|\boldsymbol{\alpha}| = \sum_{j=1}^{n} \alpha_j. \tag{3.5}$$

Given a vector $\boldsymbol{x}$ in $\mathbb{R}^n$ we define

$$\boldsymbol{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_n^{\alpha_n}. \tag{3.6}$$

For ϕ in $C^\infty(\Omega)$ we define the partial derivative as follow

$$D^{\boldsymbol{\alpha}}\phi := \left(\frac{\partial}{\partial x_1}^{\alpha_1}\right) \cdots \left(\frac{\partial}{\partial x_n}^{\alpha_n}\right)\phi. \tag{3.7}$$

We now define a few function spaces that will enable us to develop Sobolev space theory.

Definition 3.1.5. *Let Ω be an open bounded subset of $\mathbb{R}^n$ then,*

1. *$C_c^\infty(\Omega)$ is the set of $C^\infty(\Omega)$ functions with compact support.*
2. *$L^p(\Omega) := \left\{f : f : \Omega \longrightarrow \mathbb{C},\ \int_\Omega |f|^p\ dx < \infty\right\}$ for $p \in \mathbb{N}$.*
3. *$L^1_{loc}(\Omega) := \{f\ :\ f \in L^1(K),\ \forall$ compact $K \subset int(\Omega)\}$.*

Definition 3.1.6. *(Definitioin 1.2.4 of [16]) A given function f in $L^1_{loc}(\Omega)$ has a weak derivative, $D_w^{\boldsymbol{\alpha}} f$, provided there exist a g in $L^1_{loc}(\Omega)$ such that*

$$\int_\Omega g\phi\ dx = (-1)^{|\boldsymbol{\alpha}|} \int_\Omega f D^{\boldsymbol{\alpha}}\phi\ dx, \qquad \forall \phi \in C_c^\infty. \tag{3.8}$$

If such a g exist we define $D_w^{\boldsymbol{\alpha}} f = g$.

We are now in a position to define Sobolev spaces. Sobolev spaces are ideal for the study of PDEs as even if there are no classical or strong solution $u \in C^{k,\mu}(\Omega)$ to a PDE, there might exist a solution u in the weak sense to the PDE in some Sobolev space. Sobolev spaces are central to our analysis.

Definition 3.1.7. *Let Ω be an open bounded subset of $\mathbb{R}^n$. Given positive integers m and p we define,*

1. *$W^{m,p}(\Omega) \equiv \{u \in L^p(\Omega)\ :\ D^{\boldsymbol{\alpha}} u \in L^p(\Omega)$ for $0 \le |\boldsymbol{\alpha}| \le m\}$.*
2. *$W_0^{m,p}(\Omega) \equiv \{$Closure of $C_0^\infty(\Omega)$ in the space $W^{m,p}(\Omega)\}$.*

The above spaces are endowed the following norm,

$$\|u\|_{W^{k,p}(\Omega)} = \left(\sum_{0 \le |\boldsymbol{\alpha}| \le k} \|D^{\boldsymbol{\alpha}} u\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}}, \qquad \text{for } 1 \le p < \infty. \tag{3.9}$$

Theorem 3.1.1. *(Theorem 3.2 of [1]) $W^{m,p}(\Omega)$ is a Banach space.*

The spaces $W_0^{m,2}(\Omega)$ and $W^{m,2}(\Omega)$ are Hilbert spaces. As a result we will use the following notation $H_0^m(\Omega)$ and $H^m(\Omega)$, respectively. We now state the Sobolev embedding theorem.

Theorem 3.1.2. *(Corollary 9.15 of [4]) Let $\Omega \subset \mathbb{R}^n$ be an open $C^{1,0}$ domain with bounded $\partial\Omega$. Let $m \geq 1$ be an integer and $p \in [1, +\infty)$. We have*

$$\begin{aligned} &W^{m,p}(\Omega) \subset L^q(\Omega), \text{ where } \frac{1}{q} = \frac{1}{p} - \frac{m}{n} && \text{if } \frac{1}{p} - \frac{m}{n} > 0, \\ &W^{m,p}(\Omega) \subset L^q(\Omega) \ \forall q \in [p, +\infty), && \text{if } \frac{1}{p} - \frac{m}{n} = 0, \\ &W^{m,p}(\Omega) \subset L^q(\Omega), && \text{if } \frac{1}{p} - \frac{m}{n} < 0, \end{aligned} \tag{3.10}$$

and all these injections are continuous.

The Sobolev embedding theorem provides us with a way to bound the $L^2(\Omega)$ norm in terms of the $H^2(\Omega)$ norm. We now state Friedrichs's inequality which provides an estimate for the lower bound of the derivatives of a function.

Theorem 3.1.3. *(Theorem 30.2 of [15]) Let Ω be a bounded subset of $\mathbb{R}^n$ with diameter d and u in $W_0^{k,2}(\Omega)$. Then*

$$\|u\|_{L^p(\Omega)} \leq d^k \Bigg(\sum_{|\boldsymbol{\alpha}|=k} \|D^{\boldsymbol{\alpha}} u\|_{L^p(\Omega)}^p \Bigg)^{\frac{1}{p}}. \tag{3.11}$$

We will be interested in the following versions of Friedrichs's inequality.

$$\begin{aligned} &\|u\|_{L^2(\Omega)}^2 \leq C_\Omega \|\nabla u\|_{L^2(\Omega)}^2, && \forall\ u \in H_0^1(\Omega), \\ &\|u\|_{L^2(\Omega)}^2 \leq C_\Omega \|D^2 u\|_{L^2(\Omega)}^2, && \forall\ u \in H_0^2(\Omega). \end{aligned} \tag{3.12}$$

The first inequality in (3.12) is known as the Poincaré inequality. The constant C_Ω depends on the domain. The optimal constant C_Ω will be referred to as the Poincaré constant. The next result relates the Poincaré constant to the eigenvalues of the Laplace operator on the domain Ω.

Corollary 3.1.1. *(Section 6.5, Theorem 2 of [10]) Let λ_1 be the smallest eigenvalue of the Laplace operator on Ω (by Thm 2.3.2 $\lambda_1 > 0$). For any $u \in H_0^1(\Omega)$,*

$$\|u\|_{L^2(\Omega)}^2 \leq \lambda_1^{-1} \|\nabla u\|_{L^2(\Omega)}^2, \qquad \forall\ u \in H_0^1(\Omega). \tag{3.13}$$

Furthermore λ_1^{-1} is the optimal constant.

In the rest of the report we will use $C_\Omega = \lambda_1^{-1}$ to denote the optimal Poincaré constant. We now define fractional Sobolev spaces.

Definition 3.1.8. *(2.1 of [13]) Let Ω be an open subset of $\mathbb{R}^n$ and $s \in (0,1)$. For $p \in [1,\infty)$ we define $W^{s,p}(\Omega)$ as follow*

$$W^{s,p}(\Omega) := \left\{ u \in L^p(\Omega) \ : \ \frac{|u(\boldsymbol{x}) - u(\boldsymbol{y})|^p}{|\boldsymbol{x}-\boldsymbol{y}|^{\frac{n}{p}+s}} \in L^p(\Omega \times \Omega) \right\}. \tag{3.14}$$

$W^{s,p}(\Omega)$ is a Banach space endowed with the norm,

$$\|u\|_{W^{s,p}(\Omega)} = \left(\int_\Omega |u|^p \ dx + \int_\Omega \int_\Omega \frac{|u(\boldsymbol{x}) - u(\boldsymbol{y})|^p}{|\boldsymbol{x}-\boldsymbol{y}|^{n+sp}} \ dxdy \right)^{\frac{1}{p}}. \tag{3.15}$$

Now we extend the definition of fractional Sobolev spaces for $s \geq 1$.

Definition 3.1.9. *(2.10 of [13]) Let Ω be an open subset of $\mathbb{R}^n$ and for $s > 1$ let $s = m + \sigma$ where m is an integer and $\sigma \in (0,1)$. Then*

$$W^{s,p}(\Omega) = \{ u \in W^{m,p}(\Omega) \ : \ D^{\boldsymbol{\alpha}}_w u \in W^{\sigma,p}(\Omega) \text{ for any } \boldsymbol{\alpha} \text{ s.t } |\boldsymbol{\alpha}| = m \}. \tag{3.16}$$

$W^{s,p}(\Omega)$ is a Banach space endowed with the norm,

$$\|u\|_{W^{s,p}(\Omega)} = \left(\|u\|^p_{W^{m,p}(\Omega)} + \sum_{|\boldsymbol{\alpha}|=m} \|D^{\boldsymbol{\alpha}}_w u\|^p_{W^{\sigma,p}(\Omega)} \right)^{\frac{1}{p}}. \tag{3.17}$$

Now we state an embedding result for fractional Sobolev spaces.

Theorem 3.1.4. *(Corollary 2.3 of [13]) Let $p \in [1,\infty)$, $s' > s > 1$ and Ω be an open subset of $\mathbb{R}^n$ of class $C^{0,1}$ then,*

$$W^{s',p}(\Omega) \subseteq W^{s,p}(\Omega). \tag{3.18}$$

We now move onto defining the trace operator and some trace embedding theorems.

Theorem 3.1.5. *(Section 5.5, Thm 1 of [10]) Let Ω be a bounded subset of $\mathbb{R}^n$ with $C^1(\Omega)$ boundary. Then there exist a linear bounded operator*

$$\gamma : H^1(\Omega) \longrightarrow L^2(\partial\Omega), \tag{3.19}$$

such that for each u in $H^1(\Omega)$,

1. $\gamma u = u_{|\partial\Omega}, \qquad$ *for* $u \in H^1(\Omega) \cap C(\bar{\Omega})$.
2. $\|\gamma u\|_{L^2(\partial\Omega)} \leq C \|u\|_{H^1(\Omega)}$.

Definition 3.1.10. *We refer to γ in the above definition as the trace operator.*

We now provide a general trace embedding result for a domain of class $C^{k,\mu}$.

Theorem 3.1.6. *(Theorem 3.37 of [12]) If Ω is a $C^{k-1,1}$ domain and if $\frac{1}{2} < s \leq k$, then γ has a unique extension to a bounded linear operator*

$$\gamma : H^s(\Omega) \to H^{s-\frac{1}{2}}(\partial\Omega), \tag{3.20}$$

and this extension has a continuous right inverse.

For a Lipschitz domain ($C^{0,1}$) the previous theorem implies $\frac{1}{2} < s \leq 1$. Our next result improves upon Theorem 3.1.6.

Theorem 3.1.7. *(Theorem 3.38 of [12]) If Ω is a Lipschitz domain then the trace operator operator in (3.20) is bounded for $\frac{1}{2} < s < \frac{3}{2}$.*

We end this section by stating the existence, uniqueness and regularity results regarding elliptic PDEs. We provide the results in the most general form. The first result states the conditions for the existence and uniqueness of a strong or classical solutions $u \in C^{k,\mu}(\Omega)$.

Definition 3.1.11. *Let Ω be a bounded subset of $\mathbb{R}^d$, a_{jk}, b_j and c be functions in $L^\infty(\Omega)$, then the operator*

$$\mathcal{L} = -\sum_{j=1}^{d}\sum_{k=1}^{d} \partial_j(a_{jk}\partial_k) + \sum_{j=1}^{d} \partial_j(b_j) + c, \tag{3.21}$$

is

1. ***elliptic*** *if the matrix (a_{jk}) is positive definite.*
2. ***Uniformely elliptic*** *if there exist a constant $\theta > 0$ such that,*

$$\sum_{j=1}^{d}\sum_{k=1}^{d} a_{jk}(\boldsymbol{x})\xi_j\xi_k \geq \theta|\boldsymbol{\xi}|^2, \qquad \forall \boldsymbol{x} \in \Omega,\ \forall \boldsymbol{\xi} \in \mathbb{R}^d. \tag{3.22}$$

Theorem 3.1.8. *(Theorem 6.14 of [18]) Consider the following elliptic PDE,*

$$\begin{aligned} \mathcal{L}u &= f, \qquad \boldsymbol{x} \in \Omega, \\ u &= \phi, \qquad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{3.23}$$

Let $\mathcal{L}$ be strictly elliptic in a bounded domain Ω with $c \leq 0$ and let f and the coefficients of $\mathcal{L}$ belong to $C^\mu(\overline{\Omega})$. Suppose that Ω is a $C^{2,\mu}$ domain and that $\phi \in C^{2,\mu}(\overline{\Omega})$. Then (3.23) has a unique solution in $C^{2,\mu}(\overline{\Omega})$.

Our next result states the conditions for the existence and uniqueness of a weak or generalized solution $u \in W^{m,p}(\Omega)$.

Theorem 3.1.9. *(Theorem 8.3 of [18]) Consider the elliptic operator in divergence form,*

$$\begin{aligned} \sum_{j=1}^{d}\sum_{k=1}^{d} \partial_j(a_{jk}(\boldsymbol{x})\partial_k u + b_j(\boldsymbol{x})u) + \sum_{j=1}^{d} c_j(\boldsymbol{x})\partial_j u + d(\boldsymbol{x})u = g + \sum_{j=1}^{d} \partial_j f,\ \boldsymbol{x} \in \Omega, \\ \gamma u = \phi, \qquad \boldsymbol{x} \in \partial\Omega, \end{aligned} \tag{3.24}$$

where a_{jk}, b_j, c_j and d are assumed to be measurable functions on a domain $\Omega \subset \mathbb{R}^d$. Furthermore assume there exist a θ, β and ζ such that,

$$\begin{aligned} \sum_{j=1}^{d}\sum_{k=1}^{d} a_{jk}(\boldsymbol{x})\xi_j\xi_k \geq \theta|\boldsymbol{\xi}|^2, \qquad \forall \boldsymbol{x} \in \Omega,\ \forall \boldsymbol{\xi} \in \mathbb{R}^d, \\ \sum_{j=1}^{d}\sum_{k=1}^{d} |a_{jk}(\boldsymbol{x})|^2 \leq \beta^2, \\ \frac{1}{\theta^2}\sum_{j=1}^{d}(|b_j(\boldsymbol{x})|^2 + |c_j(\boldsymbol{x})|^2) + \frac{1}{\theta}|d(\boldsymbol{x})| \leq \zeta^2, \\ \int_\Omega dv - \sum_{j=1}^{d} b_j(\boldsymbol{x})\partial_j v \ dx \leq 0,\ \forall v \in C_0^1(\Omega)\ s.t\ v \geq 0. \end{aligned} \tag{3.25}$$

For $\phi \in W^{1,2}(\Omega)$ and $g, \partial_j f \in L^2(\Omega)$ there exist a unique weak or generalized solution of (3.24).

The next result provides the conditions required for the existence of a weak solution in the space $H^2(\Omega)$ along with a global estimate for the solution.

Theorem 3.1.10. *(Theorem 8.12 of [18]) Let Ω be an open bounded subset of $\mathbb{R}^n$ with C^2 boundary ($\partial\Omega$), $a_{j,k}$ and b_j be uniformly Lipschitz continuous in Ω, c_j and d be essentially bounded in Ω, $f \in L^2(\Omega)$ and there exist a $\phi \in W^{2,2}(\Omega)$ for which $u - \phi \in W_0^{1,2}(\Omega)$. Then $u \in W^{2,2}(\Omega)$ and*

$$\|u\|_{H^2(\Omega)} \leq C\big(\|u\|_{L^2(\Omega)} + \|f\|_{L^2(\Omega)}\big). \tag{3.26}$$

Our last result extends the above existence and uniqueness result to convex domains.

Theorem 3.1.11. *(Theorem 3.2.1.2 of [11]) Let Ω be a convex, bounded open subset of $\mathbb{R}^d$, $f \in L^2(\Omega)$ $a_{j,k} \in C^{0,1}(\overline{\Omega})$ and*

$$\begin{aligned} Au = \sum_{j=1}^{d}\sum_{k=1}^{d} \partial_j(a_{jk}(\boldsymbol{x})\partial_k u) = f, \quad \boldsymbol{x} \in \Omega, \\ \gamma u = 0, \quad \boldsymbol{x} \in \partial\Omega, \end{aligned} \tag{3.27}$$

such that A is strongly elliptic. Then there exist a unique $u \in H^2(\Omega)$ that solves (3.27).

3.2 Variational Formulation

In this section we outline the variational framework, also known as the weak formulation. The variational framework is the primary analytical tool used to

carry out error analysis in finite element methods. We will analyse the error of our numerical method using the variational framework. Our numerical method does not satisfy the variational framework completely as a result we will make required adjustments to carry out our error analysis. After providing details on the variational framework we state an example outlining how the variational framework is implemented to study the error of a finite subspace approximation to the Poisson PDE. In this section $\mathcal{H}$ will represent a Hilbert space and $\mathcal{H}'$ will represent the dual space of the Hilbert space $\mathcal{H}$.

We begin by stating a few definitions.

Definition 3.2.1. *Let V be a vector space. A bilinear form, $a(\cdot,\cdot)$ is a mapping from $V \times V \longrightarrow \mathbb{C}$ which is linear in both its arguments i.e. for u, v, w in V and λ, μ in $\mathbb{C}$ we have,*

1. $a(\lambda u + \mu w, v) = \lambda \cdot a(u, v) + \mu \cdot a(w, v)$,
2. $a(u, \lambda v + \mu w) = \lambda^* \cdot a(u, v) + \mu^* \cdot a(w, v)$.

A bilinear form, $a(\cdot,\cdot)$, on a vector space V is symmetric if $a(u, v) = a(v, u)^*$.

The next couple of properties of bilinear forms play a fundamental role in the variational framework.

Definition 3.2.2. *(Definition 2.5.2 of [16]) A bilinear form, $a(\cdot,\cdot)$ on a normed linear space V is said to be,*

1. *Continuous: If there exist a $C < \infty$ such that*
$$a(u, v) \leq C\|u\|_V \|v\|_V, \quad \forall u, v \in V. \tag{3.28}$$
2. *Coercive: If there exist an $\alpha > 0$ such that*
$$a(u, u) \geq \alpha \|u\|_V^2, \quad \forall u \in V. \tag{3.29}$$

We now state the variational problem (**V**) and the corresponding Approximation problem (**A**).

Variational Problem (V): Given a bilinear form, $a(\cdot,\cdot)$, on a Hilbert space $\mathcal{H}$ and a linear functional L in $\mathcal{H}'$, find u in $\mathcal{H}$ such that

$$a(u, v) = L(v), \qquad \forall v \in \mathcal{H}. \tag{3.30}$$

We will be approximating the solution u to the variational problem using a finite dimensional subspace, as a result we are more interested in the finite dimensional analogue of the Variational problem.

Approximation problem (A): Given a bilinear form, $a(\cdot,\cdot)$, on a Hilbert space $\mathcal{H}$, a finite n-dimensional closed subspace $\mathcal{H}_n$ of the Hilbert space $\mathcal{H}$ and a linear functional L in $\mathcal{H}'$, find u_n in $\mathcal{H}_n$ such that

$$a(u_n, v) = L(v), \qquad \forall v \in \mathcal{H}_n. \tag{3.31}$$

Assuming we have a solution $u_n \in \mathcal{H}_n$ to the approximation problem (**A**), we are interested in finding how well it approximates the solution of the variational problem (**V**). We state an important property known as Galerkin orthogonality that says the error is orthogonal to $\mathcal{H}_n$. We will use this result later when computing the error, however it seems appropriate to state it here.

Proposition 3.2.1. *Let u and u_n be solutions to the variational and approximation problem respectively, then*

$$a(u - u_n, v) = 0, \qquad \forall v \in \mathcal{H}_n. \tag{3.32}$$

Proof.

$$\begin{aligned} a(u - u_n, v) &= a(u, v) - a(u_n, v), \\ &= L(v) - L(v), \\ &= 0. \end{aligned} \tag{3.33}$$

□

We now state the Lax-Milgram theorem which states that given certain conditions the variational and approximation problem have a unique solution.

Theorem 3.2.1. *(Theorem 2.7.7 of [16]) Let $\mathcal{H}$ be a Hilbert space, $a(\cdot, \cdot)$ a continuous, coercive bilinear form on $\mathcal{H}$ and L a continuous linear functional in $\mathcal{H}'$. Then there exists a unique u in $\mathcal{H}$ such that*

$$a(u, v) = L(v), \qquad \forall v \in \mathcal{H}. \tag{3.34}$$

Corollary 3.2.1. *If all the conditions of the Lax-Milgram theorem are satisfied then there exist a unique solution to the approximation problem **(A)**, as well.*

A natural question one asks is how well does u_n the solution of the approximation problem (**A**) approximates the solution u of the variational problem (**V**). The next result provides an answer to this question by stating that the solution to the approximation problem **(A)** is the best approximation from the subspace $\mathcal{H}_n$ upto a constant.

Theorem 3.2.2. *(Theorem 2.8.1 of [16]) Let $\mathcal{H}$ be a Hilbert space and $a(\cdot, \cdot)$ a coercive, continuous bilinear form on $\mathcal{H}$. Suppose u and u_n solve the variational problem **(V)** and approximation problem **(A)**, respectively. Then*

$$\|u - u_n\|_{\mathcal{H}} \leq \frac{C}{\alpha} \min_{v \in \mathcal{H}_n} \|u - v\|_{\mathcal{H}}, \tag{3.35}$$

where C and α are constants from the continuity and coercive property of the bilinear form, respectively.

Our last result states that if the bilinear form is symmetric then the above estimate can be improved.

Corollary 3.2.2. *(Theorem 0.3.3 of [16]) Let $a(\cdot,\cdot)$ be a symmetric bilinear form that is coercive. Then $a(\cdot,\cdot)$ is an inner product. The norm induced by the symmetric and coercive bilinear inner product is known as the energy norm.*

$$\|u\|_E := \sqrt{a(u,u)}. \tag{3.36}$$

Then u_n is the best approximation to u from the subspace $\mathcal{H}_n$ in the energy norm.

$$\|u - u_n\|_E \leq \|u - v\|_E, \qquad \forall v \in \mathcal{H}_n. \tag{3.37}$$

3.2.1 Variational Formulation Example

We now provide an example of how the variational formulation is applied to PDEs. We analyse the Poisson PDE with homoegeneous boundary conditions using the variational formulation. This example will serve to illustrate how the results and concepts previously presented in the section are utilized. Furthermore it will also provide a contrast to the error estimate we carry out for our numerical method later on.

Let Ω be a subset of $[-1,1]^d$ of class $C^{1,1}$. Consider the Poisson PDE with Dirichlet boundary conditions on Ω.

$$\begin{aligned} -\Delta u &= f, \qquad \boldsymbol{x} \in \Omega, \\ u &= 0, \qquad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{3.38}$$

We begin by deriving a weak solution for the above Poisson PDE. To do so let us multiply the above equation with a $v \in C_c^\infty(\Omega)$.

$$\int_\Omega -\Delta u v \; dx = \int_\Omega f v \; dx. \tag{3.39}$$

Integrating over Ω using Green's first identity and using the zero boundary condition we have

$$\int_\Omega \nabla u \cdot \nabla v \; dx = \int_\Omega f v \; dx. \tag{3.40}$$

In order to find a u that satisfies the above equation for all v we require $u \in H_0^1(\Omega)$. Thus it seems natural to pick $H_0^1(\Omega)$ as our Hilbert space in which to seek a solution. In order to prove that there exists a $u \in H_0^1(\Omega)$ such that (3.40) holds for all $v \in H_0^1(\Omega)$ we use the variational framework. To that end let us define our bilinear form $a(\cdot,\cdot)$ and linear functional L_f.

$$\begin{aligned} a(u,v) &= \int_\Omega \nabla u \cdot \nabla v \; dx, \\ L_f(v) &= \int_\Omega f v \; dx. \end{aligned} \tag{3.41}$$

We now show that the bilinear form is continuous and coercive. Continuity of the bilinear form is as follow,

$$\begin{aligned} a(u,v) &= \int_\Omega \nabla u \cdot \nabla v \; dx, \\ &\leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}, \qquad \text{(Cauchy-Schwarz)}, \\ &\leq \|u\|_{H_0^1(\Omega)} \|v\|_{H_0^1(\Omega)}. \end{aligned} \tag{3.42}$$

We now show that the bilinear form is coercive.To see this we observe

$$\begin{aligned} \|u\|^2_{H_0^1(\Omega)} &= \|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}, \\ &\leq C_\Omega \|\nabla u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}, \qquad \text{by (3.12)}, \\ \|\nabla u\|^2_{L^2(\Omega)} &\geq \frac{1}{1+C_\Omega} \|u\|^2_{H_0^1(\Omega)}. \end{aligned} \tag{3.43}$$

With the above observation our result follows.

$$\begin{aligned} a(u,u) &= \int_\Omega |\nabla u|^2 \; dx, \\ &= \|\nabla u\|^2_{L^2(\Omega)}, \\ &\geq \frac{1}{1+C_\Omega} \|u\|^2_{H_0^1(\Omega)}. \end{aligned} \tag{3.44}$$

We now show that the linear functional $L_f(v)$ is continuous.

$$\begin{aligned} L_f(v) &= \int_\Omega fv \; dx, \\ &\leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}, \\ &\leq \|f\|_{L^2(\Omega)} \|v\|_{H_0^1(\Omega)}. \end{aligned} \tag{3.45}$$

All requirements of the Lax-Milgram theorem are satisfied. There exist a unique weak solution i.e. there exist $u \in H_0^1(\Omega)$ that satisfies,

$$\int_\Omega \nabla u \cdot \nabla v \; dx = \int_\Omega fv \; dx, \quad \forall v \in H_0^1(\Omega). \tag{3.46}$$

Let $\Psi_N = \text{span}\{\psi_1, \cdots, \psi_N\}$ be a closed subspace of $H_0^1(\Omega)$. As all the conditions of the Lax-Milgram theorem are satisfied there exist a unique $u_N \in \Psi$ such that,

$$\int_\Omega \nabla u_N \cdot \nabla v \; dx = \int_\Omega fv \; dx, \quad \forall v \in \Psi. \tag{3.47}$$

Theorem 3.2.2 provides an error estimate for the approximation u_N.

$$\|u - u_N\|_{H_0^1(\Omega)} \leq \frac{C}{\alpha} \min_{v \in \Psi} \|u - v\|_{H_0^1(\Omega)}. \tag{3.48}$$

3.3 Numerical Analysis

We now begin to carry out our analysis of the numerical method using the variational framework and results stated in the previous two sections. We will analyse the numerical method applied to the Poisson PDE. Let $\Omega \subset [-1,1]^d$ be a convex domain of class $C^{0,1}$, i.e. it has Lipschitz boundary and $f \in L^2(\Omega)$. We are interested in the following homogeneous Poisson PDE,

$$\begin{aligned} \Delta u &= f, \qquad \boldsymbol{x} \in \Omega, \\ u &= 0, \quad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{3.49}$$

We begin by providing an informal argument which shows why it is natural to consider the variational formulation to study our numerical method. The numerical scheme solved for a solution u_N in the space spanned by the truncated Fourier frame $\mathcal{F}_N = \{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\}_{\substack{|n_l|\leq N \\ 1\leq l\leq d}}$ by evaluating the Poisson PDE (3.49) at interior points $\boldsymbol{x}_j \in \Omega_{M_i}$ and boundary points $\boldsymbol{x}_k \in \partial\Omega_{M_b}$. Our numerical solution was,

$$\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{c} = \boldsymbol{A}^*\boldsymbol{F}. \tag{3.50}$$

We are interested in knowing how well our numerical method approximates the true solution from the space $\mathcal{F}_N$ and are not considering the error being introduced from using the SVD regularization. As there are other preconditioning methods that can be used to solve (3.50) that might be more accurate and we want to know how well our method performs in the absence of any numerical implementation issues.

Every $v \in \text{span}\{\mathcal{F}_N\}$ can be represented as,

$$v(\boldsymbol{x}) = \sum_{\substack{|n_l|\leq N \\ 1\leq l\leq d}} d_{\boldsymbol{n}} e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}, \tag{3.51}$$

for some $\boldsymbol{d} \in \mathbb{C}^{(2N+1)^d}$. Let us multiply (3.50) by $\boldsymbol{d}^*$ to get,

$$\begin{aligned} \boldsymbol{d}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{c} &= \boldsymbol{d}^*\boldsymbol{A}^*\boldsymbol{F}, \\ (\boldsymbol{A}\boldsymbol{d})^*\boldsymbol{A}\boldsymbol{c} &= (\boldsymbol{A}\boldsymbol{d})^*\boldsymbol{F}, \\ \sum_{j=1}^{M_i} \Delta u_N(\boldsymbol{x}_j)(\Delta v_N(\boldsymbol{x}_j))^* + \sum_{k=1}^{M_b} u_N(\boldsymbol{x}_k)(v_N(\boldsymbol{x}_k))^* &= \sum_{j=1}^{M_i} f(\boldsymbol{x}_j)(\Delta v_N(\boldsymbol{x}_j))^*. \end{aligned} \tag{3.52}$$

If we were to consider u, v in the infinite dimensional space $\mathcal{F} = \{e^{i\pi \boldsymbol{n}\cdot\boldsymbol{x}}\}_{\boldsymbol{n}\in\mathbb{Z}^d}$ instead of $u_N \in \mathcal{F}_N$ we would have,

$$\sum_{j=1}^{M_i} \Delta u(\boldsymbol{x}_j)(\Delta v(\boldsymbol{x}_j))^* + \sum_{k=1}^{M_b} u(\boldsymbol{x}_k)(v(\boldsymbol{x}_k))^* = \sum_{j=1}^{M_i} f(\boldsymbol{x}_j)(\Delta v(\boldsymbol{x}_j))^*. \tag{3.53}$$

Informally if we were to take the limit as M_i and M_b go to infinity we would have the continuous analogue of (3.53),

$$\int_\Omega \Delta u\overline{\Delta v}\, dx + \int_{\partial\Omega} \gamma u\overline{\gamma v}\, dx = \int_\Omega f\overline{\Delta v}\, dx. \tag{3.54}$$

This shows that our numerical method approximates u such that (3.54) holds for all v. Thus we consider the following bilinear form $a(\cdot,\cdot)$ and continuous linear functional L_f,

$$\begin{aligned} a(u,v) &= \int_\Omega \Delta u \overline{\Delta v}\ dx + \int_{\partial\Omega} \gamma u \overline{\gamma v}\ dx, \\ L_f(v) &= \int_\Omega f\overline{\Delta v}\ dx. \end{aligned} \tag{3.55}$$

We now need to pick a Hilbert space in order to implement the Lax-Milgram theorem and the variational formulation. In order for the bilinear form (3.55) to be defined the minimum level of regularity required is for $u \in H^2(\Omega)$. It follows that we have the following variational formulation corresponding to our numerical method.

$$\begin{aligned} &\mathbf{V}\text{: Find } u \in H^2(\Omega) \text{ s.t } a(u,v) = L_f(v),\ \forall v \in H^2(\Omega), \\ &\mathbf{A}\text{: Find } u_n \in H^2_N(\Omega) \text{ s.t } a(u_n,v) = L_f(v),\ \forall v \in H^2_N(\Omega). \end{aligned} \tag{3.56}$$

In the above formulation the bilinear form $a(\cdot,\cdot)$, linear functional L_f correspond to (3.55) and $H^2_N(\Omega)$ is a closed finite dimensional subspace of $H^2(\Omega)$. We begin our analysis by considering the one dimensional (ODE) case.

3.3.1 ODE Results

For the ODE case ($d = 1$) our bilinear form and linear functional are,

$$\begin{aligned} a(u,v) &= \int_a^b u''\bar{v}''\ dx + u(a)\bar{v}(a) + u(b)\bar{v}(b), \qquad \Omega = (a,b) \subset [-1,1], \\ L_f v &= \int_a^b f v''\ dx. \end{aligned} \tag{3.57}$$

Proposition 3.3.1. *The bilinear form in (3.57) is continuous, i.e. there exist a constant C such that*

$$a(u,v) \le C\|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}. \tag{3.58}$$

Proof. We consider the terms of the bilinear form separately.

$$\begin{aligned} \int_a^b u''\bar{v}''\ dx &\le \int_a^b |u''||\bar{v}''|\ dx, \\ &\le \|u''\|_{L^2(\Omega)}\|v''\|_{L^2(\Omega)}, \\ &\le \|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}. \end{aligned} \tag{3.59}$$

We now find an upper bound for the boundary terms,

$$|u(a)| \le \sup_{x\in\Omega} |u(x)| = \|u\|_{L^\infty(\Omega)}. \tag{3.60}$$

We have the following continuous injection ,

$$W^{m,p}(\Omega) \subset L^{\infty}(\Omega), \quad if \quad \frac{1}{p} - m < 0. \tag{3.61}$$

As a result we have,

$$\begin{aligned} W^{2,2}(\Omega) &\subset L^{\infty}(\Omega), \\ \|u\|_{L^{\infty}(\Omega)} &\leq C\|u\|_{H^2(\Omega)}. \end{aligned} \tag{3.62}$$

Combining (3.62) and (3.60) we have an upper bound for the boundary terms.

$$\begin{aligned} u(a)\bar{v}(a) + u(b)\bar{v}(b) &\leq |u(a)||\bar{v}(a)| + |u(b)||\bar{v}(b)|, \\ &\leq 2\|u\|_{L^{\infty}((\Omega))}\|v\|_{L^{\infty}(\Omega)}, \\ &\leq 2C^2\|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}. \end{aligned} \tag{3.63}$$

Putting (3.63) and (3.59) together we have our desired result.

$$\begin{aligned} \int_a^b u''\bar{v}''\, dx + u(a)\bar{v}(a) + u(b)\bar{v}(b) &\leq \|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)} + 2C^2\|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}, \\ &\leq Q\|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}, \end{aligned} \tag{3.64}$$

where $Q = \max\{1, 2C^2\}$. □

We now show that the bilinear form is also coercive.

Proposition 3.3.2. *The bilinear form $a(u, v)$ in (3.57) is coercive, i.e. there exist a constant C such that,*

$$a(u, u) \geq C\|u\|^2_{H^2(\Omega)}, \quad \forall u \in H^2(\Omega). \tag{3.65}$$

Proof. We begin by stating that the following norm is equivalent to the $\|\cdot\|_{H^2}$ norm.

$$\|u\|^2 = \|u\|^2_{L^2(\Omega)} + \|u''\|^2_{L^2(\Omega)}, \tag{3.66}$$

For this proof we will use (3.66) to indicate the $\|\cdot\|_{H^2}$ norm.

In order to show the above result we will apply the fundamental theorem of calculus and mean value theorem. In order to do so we note that by the Meyers-Serrin theorem $C^{\infty}(\Omega) \cap H^2(\Omega)$ is dense in $H^2(\Omega)$. In order to apply the theorems to a $u \in H^2(\Omega)$ we can consider a sequence of $C^{\infty}(\Omega)$ function $\{u_n\}_{n\in\mathbb{N}}$ that converge to u. We prove the result for a $u_n \in C^{\infty}(\Omega)$ and then take the limit. The change in limits is possible due to the dominated convergence theorem. Due to the density argument we apply the fundamental theorem of calculus and mean value theorem to a $u \in H^2(\Omega)$ directly.

By the mean value theorem there exist $c \in (a, b)$ such that,

$$\begin{aligned} |u'(c)| &= \left|\frac{u(b) - u(a)}{b - a}\right|, \\ &\leq C(|u(b)| + |u(a)|), \qquad C = \frac{1}{|b - a|}, \\ |u'(c)|^2 &\leq 2C^2(|u(b)|^2 + |u(a)|^2). \end{aligned} \tag{3.67}$$

Let $x \in \Omega$, then by the fundamental theorem of calculus we have,

$$\begin{aligned}
|u'(x)| &= \big|u'(c) + \int_c^x u''(t)\ dt\big|, \\
&\le |u'(c)| + \int_c^x |u''(t)|\ dt, \\
|u'(x)|^2 &\le 2\bigg(|u'(c)|^2 + \int_c^x |u''(t)|^2\ dt\bigg), \\
&\le 2\bigg(2C^2(|u(b)|^2 + |u(a)|^2)\bigg) + 2\int_c^x |u''(t)|^2\ dt, \\
&\le \overline{C}\bigg(|u(a)|^2 + |u(b)|^2 + \int_a^b |u''(t)|^2\ dt\bigg), \quad \overline{C} = \max\{4C^2, 2\}, \\
\int_a^b |u'(x)|^2\ dx &\le \overline{C}|b-a|\bigg(|u(a)|^2 + |u(b)|^2 + \int_a^b |u''(t)|^2\ dt\bigg), \\
\int_a^b |u'(x)|^2\ dx &\le K\bigg(|u(a)|^2 + |u(b)|^2 + \int_a^b |u''(t)|^2\ dt\bigg), \quad K = \max\{4C, \frac{2}{C}\}.
\end{aligned} \tag{3.68}$$

The term on the right hand side is the bilinear form $a(u,u)$. We use the fundamental theorem of calculus to compute a lower bound for the left hand side.

$$\begin{aligned}
|u(x)| &= \big|u(a) + \int_a^x u'(t)\ dt\big|, \\
&\le |u(a)| + \int_a^x |u'(t)|\ dt|, \\
|u(x)|^2 &\le 2|u(a)|^2 + 2\int_a^x |u'(t)|^2\ dt|, \\
&\le 2|u(a)|^2 + 2\int_a^b |u'(t)|^2\ dt|, \\
\int_a^b |u(t)|^2\ dt &\le 2|b-a|(|u(a)|)^2 + 2|b-a|\int_a^b |u'(t)|^2\ dt|, \\
&= \frac{2}{C}(|u(a))|^2 + \frac{2}{C}\int_a^b |u'(t)|^2\ dt|.
\end{aligned} \tag{3.69}$$

Using (3.68), (3.69) and a few manipulation we will derive our result. From equation (3.68) we have,

$$\begin{aligned}
\int_a^b |u'(x)|^2\ dx &\le K|u(a)|^2 + K|u(b)|^2 + K\int_a^b |u''(t)|^2\ dt, \\
\frac{2}{C}\int_a^b |u'(x)|^2\ dx &\le \frac{2K}{C}|u(a)|^2 + \frac{2K}{C}|u(b)|^2 + \frac{2K}{C}\int_a^b |u''(t)|^2\ dt, \\
\frac{2}{C}|u(a)|^2 + \frac{2}{C}\int_a^b |u'(x)|^2\ dx &\le \frac{2K+2}{C}|u(a)|^2 + \frac{2K}{C}|u(b)|^2 + \frac{2K}{C}\int_a^b |u''(t)|^2\ dt.
\end{aligned} \tag{3.70}$$

Now adding the integral of the second derivative to both sides we get

$$\begin{aligned}\frac{2}{C}|u(a)|^2+\frac{2}{C}\int_a^b |u'(x)|^2\ dx + \int_a^b |u''(t)|^2\ dt \\ \leq \frac{2K+2}{C}|u(a)|^2 + \frac{2K}{C}|u(b)|^2 + \frac{2K+C}{C}\int_a^b |u''(t)|^2\ dt, \quad (3.71)\\ \leq \overline{K}\bigg(\ |u(a)|^2 + |u(b)|^2 + \int_a^b |u''(t)|^2\ dt\bigg),\end{aligned}$$

where,

$$\overline{K} = \max\{\frac{2K+2}{C}, \frac{2K+C}{C}\}. \quad (3.72)$$

Rearranging (3.71) we have,

$$\begin{aligned}\overline{K}\bigg(\ |u(a)|^2 + |u(b)|^2 + \int_a^b |u''(t)|^2\ dt\bigg) \geq \int_a^b |u''(t)|^2\ dt + \\ \frac{2}{C}|u(a)|^2 + \frac{2}{C}\int_a^b |u'(x)|^2\ dx, \\ \geq \int_a^b |u(t)|^2\ dt + \int_a^b |u''(t)|^2\ dt, \\ |u(a)|^2 + |u(b)|^2 + \int_a^b |u''(t)|^2\ dt \geq \frac{1}{\overline{K}}\bigg(\int_a^b |u(t)|^2\ dt + \int_a^b |u''(t)|^2\ dt\bigg). \quad (3.73)\end{aligned}$$

From the above equation our result follows.

$$\begin{aligned}a(u,u) = |u(a)|^2 + |u(b)|^2 + \int_a^b |u''(t)|^2\ dt, \\ \geq \frac{1}{\overline{K}}\|u\|^2_{H^2(\Omega)}, \quad (3.74)\\ \text{where} \qquad \overline{K} = \max\{\frac{2K+2}{C}, \frac{2K+C}{C}\}.\end{aligned}$$

□

We now show that linear functional L_f in (3.57) is continuous.

Proposition 3.3.3. *The linear functional $L_f(v)$ in (3.57) is continuous.*

Proof. Using Cauchy-Schwarz we have,

$$\begin{aligned}L_f(v) = \int_\Omega fv''\ dx, \\ \leq \|f\|_{L^2(\Omega)}\|v''\|_{L^2(\Omega)}, \quad (3.75)\\ \leq \|f\|_{L^2(\Omega)}\|v\|_{H^2(\Omega)}, \\ L_f(v) \leq C\|v\|_{H^2(\Omega)}.\end{aligned}$$

□

All the conditions for the Lax-Milgram theorem are satisfied. Thus there exist a unique $u \in H^2(\Omega)$ such that,

$$a(u,v) = L_f(v),\ \forall v \in H^2(\Omega). \tag{3.76}$$

Now the set of truncated Fourier frames $\mathcal{F}_N = \text{span}\{e^{i\pi xn}\}_{|n|\leq N}$ is a closed subspace of $H^2(\Omega)$. As a result by the Lax-Milgram theorem we have unique $u_N \in \mathcal{F}_N$ such that,

$$a(u_N,v) = L_f(v),\ \forall v \in \mathcal{F}_N. \tag{3.77}$$

By Theorem 3.2.2 we know that our approximation u_N is the best approximation of u from $\mathcal{F}_N$ up to a constant i.e,

$$\|u-u_N\|_{H^2(\Omega)} \leq \overline{K}Q\|u-v\|_{H^2(\Omega)}, \quad \forall v \in \mathcal{F}_N. \tag{3.78}$$

We now show that the bilinear form in (3.55) for $d \geq 2$ is not coercive.

Proposition 3.3.4. *Let $\Omega \subset (-1,1)^2$. The bilinear defined on $H^2(\Omega)$ by,*

$$a(u,v) = \int_\Omega \Delta u \Delta \bar{v}\ dx + \int_{\partial\Omega} \gamma u \gamma \bar{v}\ dx. \tag{3.79}$$

is not coercive.

Proof. We prove by contradiction. Suppose $\forall u \in H^2(\Omega)\ \exists$ a $C \in \mathbb{R}$ s.t,

$$a(u,u) \geq C\|u\|^2_{H^2(\Omega)}. \tag{3.80}$$

Consider the following function ,

$$u_n(x,y) = \frac{\sinh(n\pi x)\sin(n\pi y)}{\sinh(n\pi)}, \qquad (x,y) \in (0,1)\times(0,1), \quad n \in \mathbb{N}. \tag{3.81}$$

The function u_n satisfy Laplace equation with non zero boundary condition $\forall n \in \mathbb{N}$, i.e.,

$$\begin{aligned} \Delta u_n &= 0, \qquad (x,y) \in (0,1)\times(0,1), \\ u_n(0,y) &= u_n(x,0) = u_n(x,1) = 0, \\ u_n(1,y) &= \sin(n\pi). \end{aligned} \tag{3.82}$$

Since, $\|u_n\|_{H^1(\Omega)} \leq \|u_n\|_{H^2(\Omega)}$ we will show that by picking n large enough there cannot be a $C > 0$ s.t $a(u_n,u_n) \geq C\|u_n\|_{H^1(\Omega)}$. To that end we have,

$$\begin{aligned} a(u_n,u_n) &= \frac{1}{2} - \frac{\sin(2\pi n)}{4\pi n}, \\ \|u_n\|^2_{L^2(\Omega)} &= \frac{(2\pi n - \sin(2\pi n))(\cosh(\pi n) - \pi n\,\text{csch}(\pi n))}{8\pi^2 n^2}, \\ \|\partial_x u_n\|^2_{L^2(\Omega)} &= \frac{(2\pi n - \sin(2\pi n))(\cosh(\pi n) + \pi n\,\text{csch}(\pi n))}{8}, \\ \|\partial_y u_n\|^2_{L^2(\Omega)} &= \frac{(2\pi n + \sin(2\pi n))(\cosh(\pi n) - \pi n\,\text{csch}(\pi n))}{8}. \end{aligned} \tag{3.83}$$

For any fixed $C > 0,\ \exists N \in \mathbb{R}$ s.t for $n > N$,

$$C\|u_n\|^2_{H^1(\Omega)} \geq \frac{1}{2} > a(u_n, u_n). \tag{3.84}$$

□

Due to the bilinear form not being coercive for $d \geq 2$ we cannot apply the Lax-Milgram theorem and error analysis presented in section 3.2 to our variational formulation (3.56).

3.4 Modified Variational Formulation

In this section we modify the variational formulation (3.56) presented in the previous section and derive a similar, albeit weaker error estimate as Theorem 3.2.2. In order to derive this estimate we will need to assume higher regularity on the boundary of the domain Ω. In this section we will take Ω be an open subset of $[-1, 1]^d$ of class $C^{2,1}$, unless otherwise stated. Similar to the previous section we carry out our analysis for the following Poisson PDE,

$$\begin{aligned} \Delta u &= f, \qquad \boldsymbol{x} \in \Omega, \\ u &= 0, \quad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{3.85}$$

Our modified variational formulation is the following,

$$\begin{aligned} &\mathbf{V}\text{: Find } u \in H^2(\Omega) \text{ s.t } a_\lambda(u, v) = L_f(v),\ \forall v \in H^2(\Omega), \\ &\mathbf{A}\text{: Find } u_n \in H^2_N(\Omega) \text{ s.t } a(u_n, v)_\lambda = L_f(v),\ \forall v \in H^2_N(\Omega), \end{aligned} \tag{3.86}$$

where $H^2_N(\Omega)$ is a closed finite dimensional subspace of $H^2(\Omega)$ and

$$\begin{aligned} a_\lambda(u, v) = &\int_\Omega \Delta u \Delta \bar{v}\ dx + \lambda_\Omega \Bigg[\int_{\partial\Omega} \gamma u \gamma \bar{v}\ dx + \\ &\int_{\partial\Omega}\int_{\partial\Omega} \frac{(\gamma u(\boldsymbol{x}) - \gamma u(\boldsymbol{y}))\overline{(\gamma v(\boldsymbol{x}) - \gamma v(\boldsymbol{y}))}}{|\boldsymbol{x} - \boldsymbol{y}|^{d+1}}\ dxdy \Bigg], \qquad \lambda_\Omega \in \mathbb{R}, \\ L_f(v) = &\int_\Omega f \Delta \bar{v}\ dx. \end{aligned} \tag{3.87}$$

The constant λ_Ω depends on the domain Ω. Later on in the section we state the conditions that the constant λ_Ω will need to satisfy. The term in the bracket of the bilinear form is the inner product associated to $H^{\frac{1}{2}}(\Omega)$. We can thus write the bilinear form as,

$$\begin{aligned} a_\lambda(u, v) &= \int_\Omega \Delta u \Delta \bar{v}\ dx + \lambda_\Omega \langle \gamma u, \gamma v \rangle_{H^{\frac{1}{2}}(\partial\Omega)}, \\ a_\lambda(u, v) &= \langle \Delta u, \Delta v \rangle_{L^2(\Omega)} + \lambda_\Omega \langle \gamma u, \gamma v \rangle_{H^{\frac{1}{2}}(\partial\Omega)}. \end{aligned} \tag{3.88}$$

The Lax-Milgram theorem states the existence of a unique solution u to a variational problem $(\mathbf{V})$ provided the associated bilinear form is coercive and continuous. As our bilinear form in not coercive we will use the existence of a weak solution of the Poisson PDE (3.85) to show that variational problem in (3.86) has a solution.

Proposition 3.4.1. *There exist a $u \in H^2(\Omega)$ that solves the variational problem $(\boldsymbol{V})$ in (3.86).*

Proof. By Theorem 3.1.10 there exist a weak solution $u \in H^2(\Omega)$ to the Poisson PDE (3.85). As a result for any $h \in C_c^\infty(\Omega)$ we have,

$$\int_\Omega \Delta u h \, dx = \int_\Omega f h \, dx, \quad \forall h \in C_c^\infty(\Omega). \tag{3.89}$$

Now pick a $v \in H^2(\Omega)$ then there exist a $g \in L^2(\Omega)$ such that $g = \Delta \bar{v}$ in the weak sense. Since the set of $C_c^\infty(\Omega)$ functions is dense in $L^2(\Omega)$, there exist a sequence $\{h_n\}_{n\in\mathbb{N}}$, where $h_n \in C_c^\infty(\Omega)$ such that $h_n \to g$. Thus,

$$\begin{aligned}
\int_\Omega \Delta u h_n \, dx &= \int_\Omega f h_n \, dx, \quad \forall n \in \mathbb{N}, \\
\lim_{n\to\infty} \int_\Omega \Delta u h_n \, dx &= \lim_{n\to\infty} \int_\Omega f h_n \, dx, \\
\int_\Omega \lim_{n\to\infty} \Delta u h_n \, dx &= \int_\Omega \lim_{n\to\infty} f h_n \, dx, \qquad \text{by DCT}, \\
\int_\Omega \Delta u g \, dx &= \int_\Omega f g \, dx, \\
\int_\Omega \Delta u \Delta \bar{v} \, dx &= \int_\Omega f \Delta \bar{v} \, dx, \quad \forall v \in H^2(\Omega).
\end{aligned} \tag{3.90}$$

In the third line we have used the dominated convergence theorem. Finally since u is a weak solution to the Poisson PDE (3.85) we have that $\gamma u = 0$. As a result,

$$\langle \gamma u, \gamma v \rangle_{H^{\frac{1}{2}}(\partial\Omega)} = \langle 0, \gamma v \rangle_{H^{\frac{1}{2}}(\partial\Omega)} = 0. \tag{3.91}$$

Combining (3.90) and (3.91) we have,

$$\begin{aligned}
\int_\Omega \Delta u \Delta \bar{v} \, dx + 0 &= \int_\Omega f \Delta \bar{v} \, dx, \quad \forall v \in H^2(\Omega), \\
\int_\Omega \Delta u \Delta \bar{v} \, dx + \lambda_\Omega \langle \gamma u, \gamma v \rangle_{H^{\frac{1}{2}}(\partial\Omega)} &= \int_\Omega f \Delta \bar{v} \, dx, \quad \forall v \in H^2(\Omega), \\
a_\lambda(u, v) &= L_f(v), \quad \forall v \in H^2(\Omega).
\end{aligned} \tag{3.92}$$

□

The next result states the conditions under which the approximation problem $(\mathbf{A})$ has a solution.

Proposition 3.4.2. *Let $H^2_N(\Omega)$ be a finite dimensional closed subspace of $H^2(\Omega)$ such that for each $h \in H^2_N(\Omega)$, $\gamma h = 0$. Then there exist a $u_N \in H^2_N(\Omega)$ that solves the approximation problem (**A**).*

Proof. By Proposition 3.4.1 there exist a $u \in H^2(\Omega)$ that solves the variational problem (**V**). Define $u_N = P_N u$, where P_N is the projection operator. By Proposition 3.4.1 we have,

$$\begin{aligned} a_\lambda(u,v) &= L_f(v), \quad \forall v \in H^2(\Omega), \\ \langle \Delta u, \Delta v\rangle_{L^2(\Omega)} + \lambda_\Omega \langle \gamma u, \gamma v\rangle_{H^{\frac{1}{2}}(\partial\Omega)} &= \langle f, \Delta v\rangle_{L^2(\Omega)}, \\ \langle \Delta u, \Delta v\rangle_{L^2(\Omega)} &= \langle f, \Delta v\rangle_{L^2(\Omega)}, \quad \gamma u = 0, \\ \langle P_N \Delta u, \Delta v\rangle_{L^2(\Omega)} &= \langle P_N f, \Delta v\rangle_{L^2(\Omega)}, \\ \langle P_N^2 \Delta u, \Delta v\rangle_{L^2(\Omega)} &= \langle P_N f, \Delta v\rangle_{L^2(\Omega)}, \quad P_N^2 = P_N, \\ \langle P_N \Delta u, P_N \Delta v\rangle_{L^2(\Omega)} &= \langle f, P_N \Delta v\rangle_{L^2(\Omega)}, \quad P_N = P_N^*, \\ \langle \Delta u_N, \Delta v_N\rangle_{L^2(\Omega)} &= \langle f, \Delta v_N\rangle_{L^2(\Omega)}. \end{aligned} \tag{3.93}$$

Since $\gamma u_N = 0$,

$$\begin{aligned} \langle \Delta u_N, \Delta v_N\rangle_{L^2(\Omega)} + 0 &= \langle f, \Delta v_N\rangle_{L^2(\Omega)}, \\ \langle \Delta u_N, \Delta v_N\rangle_{L^2(\Omega)} + \lambda_\Omega \langle \gamma u_N, \gamma v_N\rangle_{H^{\frac{1}{2}}(\partial\Omega)} &= \langle f, \Delta v_N\rangle_{L^2(\Omega)}, \\ a_\lambda(u_N, v_N) &= L_f(v_N), \quad \forall v_N \in H^2_N(\Omega). \end{aligned} \tag{3.94}$$

□

Corollary 3.4.1. *Let $H^2_N(\Omega)$ be a finite dimensional closed subspace of $H^2(\Omega)$ such that for each $h \in H^2_N(\Omega)$, $\gamma h = 0$. Let $u \in H^2(\Omega)$ and $u_N \in H^2_N(\Omega)$ be the solution of the variational problem (**V**) and approximation problem (**A**), respectively. Then,*

$$a(u - u_N, v) = 0, \quad \forall v \in H^2_N(\Omega). \tag{3.95}$$

Next we show that the modified bilinear form $a_\lambda(u,v)$ is continuous.

Proposition 3.4.3. *The bilinear form $a_\lambda(u,u)$ is continuous for $v \in H^2(\Omega)$.*

Proof. We consider the terms of the bilinear form separately. Applying Cauchy-Schwarz inequality to the first term we have,

$$\begin{aligned} \int_\Omega \Delta u \Delta \bar{v}\, dx &\le \|\Delta u\|_{L^2(\Omega)} \|\Delta v\|_{L^2(\Omega)}, \\ &\le \|u\|_{H^2(\Omega)} \|v\|_{H^2(\Omega)}. \end{aligned} \tag{3.96}$$

Now for the second term we have,

$$\begin{aligned} \langle \gamma u, \gamma v\rangle_{H^{\frac{1}{2}}(\partial\Omega)} &\le \|\gamma u\|_{H^{\frac{1}{2}}(\partial\Omega)} \|\gamma v\|_{H^{\frac{1}{2}}(\partial\Omega)}, \quad \text{by Cauchy-Schwarz}, \\ &\le K_T \|u\|_{H^1(\Omega)} K_T \|v\|_{H^1(\Omega)}, \qquad \text{Thm 3.1.6}, \\ &\le K_T^2 \|u\|_{H^2(\Omega)} \|v\|_{H^2(\Omega)}. \end{aligned} \tag{3.97}$$

In (3.97) K_T is the embedding constant from Theorem 3.1.6. Combining (3.96) and (3.97) we have,

$$\begin{aligned}\int_\Omega \Delta u \Delta v \; dx + \lambda_\Omega \langle \gamma u, \gamma v \rangle_{H^{\frac{1}{2}}(\partial\Omega)} &\leq \|u\|_{H^2(\Omega)} \|v\|_{H^2(\Omega)} + \lambda_\Omega K_T^2 \|u\|_{H^2(\Omega)} \|v\|_{H^2(\Omega)}, \\ &\leq K_c \|u\|_{H^2(\Omega)} \|v\|_{H^2(\Omega)},\end{aligned} \tag{3.98}$$

where $K_c = \max\{1, \lambda_\Omega K_T^2\}$. □

We now work towards proving a weaker coercivity condition for the modified bilinear form $a_\lambda(u, v)$ in (3.87). In our next results we use the fact the $H^2(\Omega)$ is a Hilbert space and $H_0^2(\Omega)$ is a closed subspace of $H^2(\Omega)$, to write each $u \in H^2(\Omega)$ as $u = w - z$, where $z \in H_0^2(\Omega)$ and w is in the orthogonal complement of $H_0^2(\Omega)$.

Proposition 3.4.4. *Let $u \in H^2(\Omega)$ then there exist a $w \in H^2(\Omega)$ and $z \in H_0^2(\Omega)$ s.t $u = w - z$ in the sense of distributions. Furthermore, we have*

$$\begin{aligned}\int_\Omega \Delta u v \; dx &= \int_\Omega -\Delta z v \; dx, \quad \forall v \in C_c^\infty(\Omega), \\ \int_\Omega \Delta w v \; dx &= 0, \quad \forall v \in C_c^\infty(\Omega), \\ \gamma u &= \gamma w, \quad \gamma z = 0.\end{aligned} \tag{3.99}$$

Proof. Let $u \in H^2(\Omega)$, $\gamma u = g \in H^1(\Omega)$ and $\Delta u = f \in L^2(\Omega)$ in the weak sense. By Theorem 3.1.9 there exist unique weak solutions to each of the following Poisson PDEs,

$$(1) \begin{cases} -\Delta z = f & \boldsymbol{x} \in \Omega, \\ z = 0, & \boldsymbol{x} \in \partial\Omega. \end{cases}, \; (2) \begin{cases} \Delta w = 0 & \boldsymbol{x} \in \Omega, \\ w = g, & \boldsymbol{x} \in \partial\Omega. \end{cases}, \; (3) \begin{cases} \Delta u = f & \boldsymbol{x} \in \Omega, \\ u = g, & \boldsymbol{x} \in \partial\Omega. \end{cases} \tag{3.100}$$

Since the solutions are unique it follows that,

$$\begin{aligned}\int_\Omega \Delta(-z + w) v \; dx &= \int_\Omega -\Delta z v \; dx, \quad \forall v \in C_c^\infty(\Omega), \\ &= \int_\Omega f v \; dx, \quad \forall v \in C_c^\infty(\Omega), \\ &= \int_\Omega \Delta u v, \quad \forall v \in C_c^\infty(\Omega).\end{aligned} \tag{3.101}$$

Thus $u = w - z$ in the sense of distributions. Furthermore,

$$\gamma w = g = \gamma u, \quad \gamma z = 0. \tag{3.102}$$

□

We note that the above result can be generalized to strong solutions provided we assume sufficient smoothness of the function f and the boundary to guarantee a unique strong solution in $C^{k,\mu}(\Omega)$ for the three PDEs in (3.100). The next result provides an lower bound for the first term of the bilinear form (3.87).

Proposition 3.4.5. *Let Ω be a bounded subset of $\mathbb{R}^n$ of class $C^{2,1}$, $f \in L^2(\Omega)$ and $z \in H^2(\Omega) \cap H_0^1(\Omega)$ be the unique weak solution to the following PDE,*

$$\begin{aligned} -\Delta z &= f, \qquad \boldsymbol{x} \in \Omega, \\ z &= 0, \quad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{3.103}$$

Then,

$$\frac{1}{C_\Omega}\|z\|_{L^2(\Omega)} \leq \|\Delta z\|_{L^2(\Omega)}, \tag{3.104}$$

where C_Ω is the Poincaré constant.

Proof. For $f \in L^2(\Omega)$ by definition we have,

$$\begin{aligned} \|f\|_{L^2(\Omega)} &= \sup_{g \in L^2(\Omega)} \frac{\langle f, g\rangle_{L^2(\Omega)}}{\|g\|_{L^2(\Omega)}}, \\ &\geq \frac{\langle f, z\rangle_{L^2(\Omega)}}{\|z\|_{L^2(\Omega)}}, \qquad z \in L^2(\Omega). \end{aligned} \tag{3.105}$$

Since z is a weak solution of (3.103) we have,

$$\int_\Omega fv \; dx = \int_\Omega -\Delta z v \; dx, \quad \forall v \in C_c^\infty. \tag{3.106}$$

For $h_n \in C_c^\infty$, let $\{h_n\}_{n\in\mathbb{N}}$ be a sequence that converges to $\bar{z} \in L^2(\Omega)$. Thus we have,

$$\begin{aligned} \int_\Omega f h_n \; dx &= \int_\Omega -\Delta z h_n \; dx, \quad \forall n \in \mathbb{N}, \\ \int_\Omega f h_n \; dx &= \int_\Omega \nabla z \cdot \nabla h_n \; dx, \quad \forall n \in \mathbb{N}, \\ \lim_{n\to\infty} \int_\Omega f h_n \; dx &= \lim_{n\to\infty} \int_\Omega \nabla z \cdot \nabla h_n \; dx, \\ \int_\Omega \lim_{n\to\infty} f h_n \; dx &= \int_\Omega \lim_{n\to\infty} \nabla z \cdot \nabla h_n \; dx, \quad \text{by DCT}, \\ \int_\Omega f \bar{z} \; dx &= \int_\Omega \nabla z \cdot \nabla \bar{z} \; dx. \\ \int_\Omega f \bar{z} \; dx &= \|\nabla z\|_{L^2(\Omega)}^2. \end{aligned} \tag{3.107}$$

In the second line we have used Greens first identity and in the third line we have used the dominated convergence theorem. Using the Poincaré Inequality we have,

$$\begin{aligned} \frac{\langle f, z\rangle_{L^2(\Omega)}}{\|z\|_{L^2(\Omega)}} &= \frac{\|\nabla z\|_{L^2(\Omega)}^2}{\|z\|_{L^2(\Omega)}}, \\ &\geq \frac{\|z\|_{L^2(\Omega)}^2}{C_\Omega \|z\|_{L^2(\Omega)}}, \\ &= \frac{1}{C_\Omega}\|z\|_{L^2(\Omega)}. \end{aligned} \tag{3.108}$$

Since $f = -\Delta z$ in the weak sense, combining (3.105) and (3.108) we have,

$$\frac{1}{C_\Omega}\|z\|_{L^2(\Omega)} \leq \|\Delta z\|_{L^2(\Omega)}. \tag{3.109}$$

□

We are now in a position to state our weaker coercivity assumption.

Proposition 3.4.6. *For a fixed domain $\Omega \subset [-1,1]^d$ of class $C^{k,\mu}$ let C_Ω and K_T be the Poincaré constant and the trace embedding constant from Theorem 3.1.6, respectively. If*

$$\lambda_\Omega \geq \frac{2K_T^2}{C_\Omega^2}, \tag{3.110}$$

then we have,

$$a_\lambda(u,u) \geq \frac{1}{C_\Omega}\|u\|^2_{L^2(\Omega)}, \qquad \forall u \in H^2(\Omega). \tag{3.111}$$

Proof. The bilinear form is defined as,

$$\begin{aligned} a_\lambda(u,u) &= \int_\Omega |\Delta u|^2 \, dx + \lambda_\Omega \langle \gamma u, \gamma u \rangle_{H^{\frac{1}{2}}(\partial\Omega)}, \\ &= \int_\Omega |\Delta u|^2 \, dx + \lambda_\Omega \|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}. \end{aligned} \tag{3.112}$$

From Proposition 3.4.3 we have that $\Delta u = -\Delta z$ in the weak sense. Thus,

$$a_\lambda(u,u) = \int_\Omega |\Delta z|^2 \, dx + \lambda_\Omega \|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}. \tag{3.113}$$

By Proposition 3.4.4 and Proposition 3.4.5 we have,

$$\begin{aligned} a_\lambda(u,u) &\geq \frac{1}{C_\Omega^2}\|z\|^2_{L^2(\Omega)} + \lambda_\Omega \|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}, && \text{Prop 3.4.5,} \\ &= \frac{1}{C_\Omega^2}\|w-u\|^2_{L^2(\Omega)} + \lambda_\Omega \|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}, && \text{Prop 3.4.4,} \\ &\frac{1}{C_\Omega^2}\big(\|u\|^2_{L^2(\Omega)} + \|w\|^2_{L^2(\Omega)} - 2\langle u, w\rangle\big) + \lambda_\Omega \|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}. \end{aligned} \tag{3.114}$$

Applying Cauchy-Schwarz in reverse gives us,

$$\begin{aligned} a_\lambda(u,u) &\geq \frac{1}{C_\Omega^2}\big(\|u\|^2_{L^2(\Omega)} + \|w\|^2_{L^2(\Omega)} - 2\|u\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}\big) + \lambda_\Omega \|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}, \\ &\geq \frac{1}{C_\Omega^2}\big(\|u\|^2_{L^2(\Omega)} + \|w\|^2_{L^2(\Omega)} - 2\|u\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}\big) + \lambda_\Omega \|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}. \end{aligned} \tag{3.115}$$

By Theorem 3.1.6 we have,

$$\begin{aligned} a_\lambda(u,u) &\geq \frac{1}{C_\Omega^2}\big(\|u\|^2_{L^2(\Omega)} + \|w\|^2_{L^2(\Omega)} - 2K_T^2\|\gamma u\|_{H^{\frac{1}{2}}(\partial\Omega)}\|\gamma w\|_{H^{\frac{1}{2}}(\partial\Omega)}\big) + \lambda_\Omega\|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}, \\ &\geq \frac{1}{C_\Omega^2}\big(\|u\|^2_{L^2(\Omega)} - 2K_T^2\|\gamma u\|_{H^{\frac{1}{2}}(\partial\Omega)}\|\gamma w\|_{H^{\frac{1}{2}}(\partial\Omega)}\big) + \lambda_\Omega\|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}. \end{aligned} \tag{3.116}$$

By Proposition 3.4.4 $\gamma u = \gamma w$. Thus,

$$\begin{aligned} a_\lambda(u,u) &\geq \frac{1}{C_\Omega^2}\big(\|u\|^2_{L^2(\Omega)} - 2K_T^2\|\gamma u\|_{H^{\frac{1}{2}}(\partial\Omega)}\|\gamma u\|_{H^{\frac{1}{2}}(\partial\Omega)}\big) + \lambda_\Omega\|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}, \\ &\geq \frac{1}{C_\Omega^2}\|u\|^2_{L^2(\Omega)} - 2\frac{K_T^2}{C_\Omega^2}\|\gamma u\|_{H^{\frac{1}{2}}(\partial\Omega)} + \lambda_\Omega\|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}, \\ &\geq \frac{1}{C_\Omega^2}\|u\|^2_{L^2(\Omega)} + \left(\lambda_\Omega - 2\frac{K_T^2}{C_\Omega^2}\right)\|\gamma u\|^2_{H^{\frac{1}{2}}(\partial\Omega)}. \end{aligned} \tag{3.117}$$

For $\lambda_\Omega \geq 2\frac{K_T^2}{C_\Omega^2}$,

$$a_\lambda(u,u) \geq \frac{1}{C_\Omega^2}\|u\|^2_{L^2(\Omega)}. \tag{3.118}$$

□

Given a domain Ω we can always find a $\lambda_\Omega \geq 2\frac{K_T^2}{C_\Omega^2}$. With the above weaker coercivity condition we can find a similar error estimate as Theorem 3.2.2 for our modified variational problem.

Proposition 3.4.7. *Let $H^2_m(\Omega)$ be a closed finite dimensional subspace of $H^2(\Omega)$ such that for each $h \in H^2_m(\Omega)$, $\gamma h = 0$. Let $u \in H^2(\Omega)$ and $u_m \in H^2_m(\Omega)$ be the solution of the variational problem (**V**) and approximation problem (**A**) (3.86), respectively. Furthermore, for the given domain Ω of class $C^{2,1}$ let,*

$$\lambda_\Omega \geq 2\frac{K_T^2}{C_\Omega^2}, \tag{3.119}$$

where C_Ω is the Poincaré constant and K_T is the embedding constant from Theorem 3.1.6. Then,

$$\|u - u_m\|_{L^2(\Omega)} \leq C\|u - v\|_{H^2(\Omega)}, \qquad \forall v \in H^2_m(\Omega). \tag{3.120}$$

Proof. From Proposition 3.4.6, we have,

$$C_\Omega^2 a_\lambda(u,u) \geq \|u\|^2_{L^2(\Omega)}. \tag{3.121}$$

Using the above result we have,

$$\begin{aligned}
\|u-u_m\|^2_{L^2(\Omega)} &\leq C^2_\Omega a_\lambda(u-u_m, u-u_m),\\
&= C^2_\Omega a_\lambda(u-u_m, u-v+v-u_m),\\
&= C^2_\Omega a_\lambda(u-u_m, u-v) + a(u-u_m, v-u_m),\\
&= C^2_\Omega a_\lambda(u-u_m, u-v), && \text{Corollary 3.4.1,}\\
&\leq C^2_\Omega K_c \|u-u_m\|_{H^2(\Omega)} \|u-v\|_{H^2(\Omega)}, && \text{Prop 3.4.3,}\\
&\leq C^2_\Omega K_c \|u-v+v-u_m\|_{H^2(\Omega)} \|u-v\|_{H^2(\Omega)}, \quad v \in H^2_m(\Omega)\\
&\leq C^2_\Omega K_c \bigg(\|u-v\|_{H^2(\Omega)} + \|v-u_m\|_{H^2(\Omega)} \bigg) \|u-v\|_{H^2(\Omega)}.
\end{aligned} \tag{3.122}$$

Consider the term,

$$C(H^2_m) := \sup_{v \in H^2_m(\Omega)} \frac{\|v\|_{H^2(\Omega)}}{\|v\|_{L^2(\Omega)}}. \tag{3.123}$$

Using the above term we have,

$$\|v-u_m\|_{H^2(\Omega)} \leq C(H^2_m) \|v-u_m\|_{L^2(\Omega)}. \tag{3.124}$$

Thus we have,

$$\begin{aligned}
\|u-u_m\|^2_{L^2(\Omega)} &\leq C^2_\Omega K_c \bigg(\|u-v\|_{H^2(\Omega)} + C(H^2_m) \|v-u_m\|_{L^2(\Omega)} \bigg) \|u-v\|_{H^2(\Omega)},\\
&\leq C^2_\Omega K_c \|u-v\|^2_{H^2(\Omega)} + C^2_\Omega K_c C(H^2_m) \|v-u_m\|_{L^2(\Omega)} \|u-v\|_{H^2(\Omega)}.
\end{aligned} \tag{3.125}$$

Using Young's Inequality we get,

$$\begin{aligned}
\|u-u_m\|^2_{L^2(\Omega)} &\leq C^2_\Omega K_c \|u-v\|^2_{H^2(\Omega)} + \frac{\|v-u_m\|^2_{L^2(\Omega)}}{2} + \frac{C^4_\Omega K^2_c C^2(H^2_m) \|u-v\|^2_{H^2(\Omega)}}{2},\\
&\leq \bigg(C^2_\Omega K_c + \frac{C^4_\Omega K^2_c C^2(H^2_m)}{2} \bigg) \|u-v\|^2_{H^2(\Omega)} + \frac{\|v-u_m\|^2_{L^2(\Omega)}}{2},\\
&\leq \bigg(C^2_\Omega K_c + \frac{C^4_\Omega K^2_c C^2(H^2_m)}{2} \bigg) \|u-v\|^2_{H^2(\Omega)} + \frac{\|v-u+u-u_m\|^2_{L^2(\Omega)}}{2},\\
&\leq \bigg(C^2_\Omega K_c + \frac{C^4_\Omega K^2_c C^2(H^2_m)}{2} \bigg) \|u-v\|^2_{H^2(\Omega)} + \frac{\|u-v\|^2_{L^2(\Omega)}}{2} +\\
&\frac{\|u-u_m\|^2_{L^2(\Omega)}}{2}.
\end{aligned} \tag{3.126}$$

By the Sobolev embedding theorem we have $H^2(\Omega) \hookrightarrow L^2(\Omega)$. Let S be the embedding constant i.e, $\|u\|^2_{L^2(\Omega)} \leq S\|u\|^2_{H^2(\Omega)}$.

$$\|u-u_m\|^2_{L^2(\Omega)} \leq \bigg(\frac{S}{2} + C^2_\Omega K_c + \frac{C^4_\Omega K^2_c C^2(H^2_m)}{2} \bigg) \|u-v\|^2_{H^2(\Omega)} + \frac{\|u-u_m\|^2_{L^2(\Omega)}}{2}. \tag{3.127}$$

Rearranging we get,

$$\|u - u_m\|^2_{L^2(\Omega)} \leq 2\left(\frac{S}{2} + C^2_\Omega K_c + \frac{C^4_\Omega K^2_c C^2(H^2_m)}{2}\right)\|u - v\|^2_{H^2(\Omega)}. \tag{3.128}$$

Thus we have,

$$\|u - u_m\|_{L^2(\Omega)} \leq \left(S + 2C^2_\Omega K_c + C^4_\Omega K^2_c C^2(H^2_m)\right)\|u - v\|^2_{H^2(\Omega)}, \quad \forall v \in H^2_m(\Omega). \tag{3.129}$$

□

Chapter 4

Conclusion and Future Work

In this report we applied a collocation method to solve ODEs, elliptic PDEs and Laplace eigenvalue problem on irregular domains using Fourier frames. Based on results we can see that the accuracy of our numerical method depends on the domain and the smoothness of the boundary. For PDEs the numerical solution was accurate to within 10^{-10} on the circular domain (convex with infinitely smooth boundary), 10^{-7} on convex ice cream cone domain (continuous boundary) and 10^{-5} on the non-convex ice cream cone domain that also has a continuous boundary. Oversampling did not provide any substantial increase in accuracy for PDEs on domains with piecewise smooth boundary. Furthermore, we required a larger number of Fourier frame modes when solving PDEs on domain with piecewise smooth boundary. For the circular domain example oversampling provided an increase in accuracy of 10^{-2}. The numerical results of section 2.3 show that with oversampling we were able to approximate the first few eigenvalues corresponding to the Laplace eigenvalues problem up to an accuracy of at least 10^{-8}.

In Chapter 3 we utilized the variational framework to analyse the numerical method. We were able to apply the variational framework to our numerical method for the one dimensional case. Due to a lack of coercivity of the bilinear form in higher dimension we modified our variational formulation. For the modified variational formulation we provided a weaker error estimate then Theorem 3.2.2.

4.1 Future Work

When applying our numerical method we used SVD regularization in order to deal with the ill-conditioning. In the future a main area of study is to examine the method if we were to apply some other regularization method to solve the ill-conditioned linear system. As an example utilizing the Matlab backslash

operator which implements QR factorization with some conditioning algorithm to solve the linear system provides better results in a some instances such as the Laplace eigenvalue problems in section 2.3 and the ODE examples in section 2.2.1.

Secondly, in the future we would like to extend this method to time dependent PDEs such as the advection or diffusion equation. We can extend this method by using an implicit time stepping scheme such as backward differentiation formulas and solving a linear system using least square at each time step. In the Appendix we provide one example each for the advection equation and diffusion equation with one spatial dimension. The advection equation example in the appendix shows first order convergence for the backward Euler which is what we expect. However, our application of second and third order backward differentiation formula schemes do not provide the correct rate of convergence. For example, BDF2 applied to the advection equation provides a convergence rate of approximately 1.6, when we are expecting second order convergence. We received similar results when implementing BDF2 to solve the diffusion equation in 2D on a circular domain. When implementing BDF3 we observe several instances of reducing the time step leading to worse results for the advection equation. Solving time dependent PDEs numerically is a very broad topic and generally numerical schemes that are able to capture the behaviour of the PDE perform better. For example, forward time differencing is much more accurate and common for the advection equation whereas centred time differencing schemes are more preferred for the diffusion equation.

Appendix A

Numerical Method Extension for Time Dependent PDEs

We outline very briefly how the numerical method can be extended to solve the advection equation and diffusion equation numerically.

A.1 Advection Equation

We solve the following advection equation with one spatial dimension,

$$\begin{aligned} u_t + cu_x = 0, \quad t > 0,\ x \in \left(-\frac{1}{2}, \frac{1}{2}\right),\ c \in \mathbb{R}, \\ u(0,x) = \sin(x), \\ u(t,-0.5) = \sin(-0.5 - ct), \end{aligned} \tag{A.1}$$

The true solution to the above equation is,

$$u(t,x) = \sin(x - ct). \tag{A.2}$$

We refer to c as the wave speed. We will consider the above problem with $c = 4$ and $c = 20$. The case of $c = 4$ and $c = 20$ will be referred to as low and high wave speed, respectively. Generally when solving the transport or advection equation only the initial condition is assumed to be given. The problem we are solving provides us with a boundary conditions at the left end for each time as well. Having one boundary condition is reasonable as we can think of it as the flow into the domain at each time step. Without having a boundary conditions at each time step our numerical solution becomes less accurate with time. If we were to consider having the boundary value at both ends for all time our method provides more accurate and better convergence results.

We will utilize the implicit backward Euler as our time stepping scheme. Let δ_t be our time step. We can think of u_x as the forcing term for a time dependent ODE i.e.,

$$\begin{aligned} u_t &= -cu_x, \\ \frac{U^{n+1} - U^n}{\delta_t} &= -cU_x^{n+1}, \\ U^{n+1} + c\delta_t U_x^{n+1} &= U^n. \end{aligned} \tag{A.3}$$

Now we substitute the frame representation (2.21) into (A.3) to get,

$$\begin{aligned} \sum_{|k|\leq N} C_k^{n+1} e^{i\pi kx} + c \sum_{|k|\leq N} C_k^{n+1} ik\pi e^{i\pi kx} &= U^n, \\ \sum_{|k|\leq N} C_k^{n+1} e^{-0.5i\pi k} &= \sin(-0.5 - c\delta_t(n+1)). \end{aligned} \tag{A.4}$$

Note in the above equation C_k^{n+1} are the frame coefficient of the function $u(\delta_t(n+1), x)$. We are using capital "C" for frame coefficients this time as the wave speed is denoted by "c". The "n+1" superscript denotes the time index. As before let Ω_{M-1} be a finite uniform discretization of length δ_x of Ω , i.e.

$$\begin{aligned} \Omega_{M-1} &= \{x_k \ : \ x_k \in \Omega, \ x_{k+1} - x_k = \delta_x, \ 1 \leq k \leq M-1\}, \\ x_M &= -0.5. \end{aligned} \tag{A.5}$$

Evaluating (A.4) at the discretization points we have,

$$\begin{aligned} \sum_{|k|\leq N} (1 + cik\pi) C_k^{n+1} e^{i\pi kx} &= U^n, \\ \sum_{|k|\leq N} C_k^{n+1} e^{-0.5i\pi k} &= \sin(-0.5 - c\delta_t(n+1)). \end{aligned} \tag{A.6}$$

Writing the above N equations in a linear system we have,

$$\begin{aligned} \boldsymbol{A}\boldsymbol{C}^{n+1} &= \boldsymbol{F}^n, \\ \boldsymbol{A}_{p,q} &= 1 + ciq\pi e^{i\pi q x_p}, \quad 1 \leq p \leq M-1, \ -N \leq q \leq N, \\ \boldsymbol{A}_{M,q} &= e^{-0.5iq\pi}, \quad -N \leq q \leq N. \end{aligned} \tag{A.7}$$

We solve the above system using least squares with SVD regularization. Thus at each time step we solve a least square solve to march forward in time. Figure A.1 shows the numerical convergence result of computing the solution at time $t = 1$ i.e. we are computing $u(1, x)$ and then computing the error by using the true solution. Both plots of figure A.1 were obtained using 250 discretization points and total of 25 Fourier frame modes ($N = 25$). Our results are much better both in terms of convergence and accuracy when the wave speed is low $c = 4$. Our method is unable to handle a very high wave speed such as $c = 1000$ even if we increase the numbers of mode and/or consider a finer spatial mesh. This holds true even if we consider both boundary points. We have first order

convergence for $c = 4$ as expected. However for a wave speed of 20 we can see that our convergence results are not accurate.

Convergence result for different c at $T = 1$

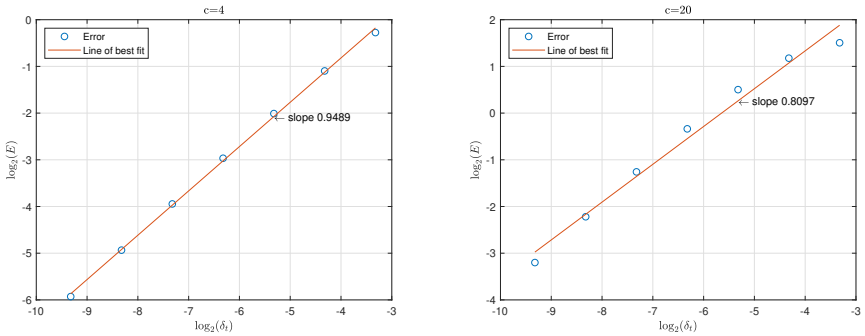

Figure A.1: Error convergence for advection equation for wave speed $c = 4$ and $c = 20$.

The above results computed the error between the computed solution and true solution at $T = 1$. Now we consider a final time of $T = 2$ and see how the error behaves at each time step. We only consider the low wave speed case $c = 4$. The numerical results are obtained using 250 equally spaced spatial points and total of 25 Fourier frame modes. Figure A.2 shows the results.

Error at Each Time Step

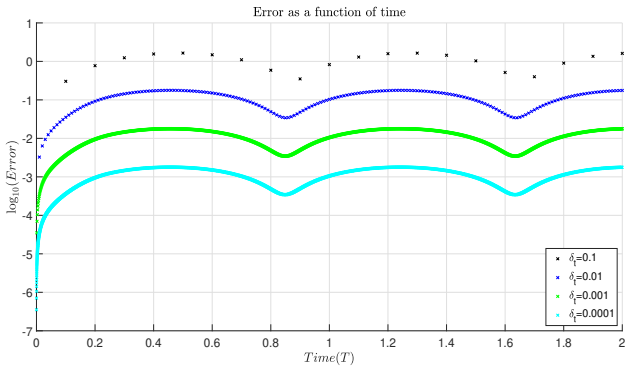

Figure A.2: The error over time for different time step using backward Euler.

Our results for low wave speed show the correct convergence and are fairly accurate. With a step size of 0.0001 our error is approximately 10^{-3}. If we used a higher order time stepping method such as BDF3 we would expect much higher accuracy.

A.2 Diffusion Equation

We now solve the 1D diffusion equation with a piecewise smooth initial condition.

$$\begin{aligned} u_t &= u_{xx}, \qquad x \in \Omega = \Big(-\frac{1}{2}, \frac{1}{2} \Big), \\ u(t,-0.5) &= 0, \quad u(t,0.5) = 0, \end{aligned} \tag{A.8}$$

$$u(0,x) = \begin{cases} 4x+2 & -0.5 \le x \le -0.25, \\ -4x & -0.25 \le x \le 0, \\ 40x & 0 \le x \le .25, \\ -40x+20 & 0.25 \le x \le .5. \end{cases} \tag{A.9}$$

For this example let us use the second order backward differentiation formula 2 (BDF2) for our time stepping.

$$\begin{aligned} U^{n+2} - \frac{4}{3}U^{n+1} + \frac{1}{3}U^{n} &= \frac{2}{3}U_{xx}^{n+2}, \\ U^{n+2} - \frac{2}{3}U_{xx}^{n+2} &= \frac{4}{3}U^{n+1} - \frac{1}{3}U^{n}. \end{aligned} \tag{A.10}$$

Similar to the previous example we substitute the frame representation (2.21), and evaluate it at a set of discrete interior and boundary points to get,

$$\begin{aligned} &\sum_{|k|\le M} (1+k^2\pi^2) C_k^{n+2} e^{i\pi k x_j} = \frac{4}{3}U^{n+1} - \frac{1}{3}U^{n}, \quad 1 \le j \le N-2 \\ &\sum_{|k|\le M} C_k^{n+2} e^{-0.5 i\pi k} = 0, \\ &\sum_{|k|\le M} C_k^{n+2} e^{0.5 i\pi k} = 0. \end{aligned} \tag{A.11}$$

Numerical Solution for Diffusion Equation

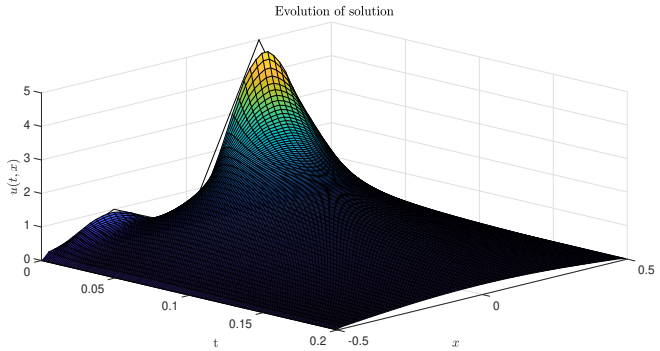

Figure A.3: Evolution of numerical solution of heat equation.

Rewriting (A.11) as $\boldsymbol{A}\boldsymbol{c}^{n+2} = \boldsymbol{F}^{n+1,n}$ and solving for the Fourier frame coefficient at each time step we march forward in time. Figure A.3 show our numerical solution computed using a time step of $\delta_t = 0.001$. At each time step we used 11 Fourier frame modes and 25 discretization points. In order to implement the BDF2 scheme we used backward Euler to compute U^1. The numerical solution shows that our solution behaves as would expect. The solution becomes smooth instantaneously and decays over time.

Appendix B

Domain Shapes

Convex Ice Cream Cone Domain

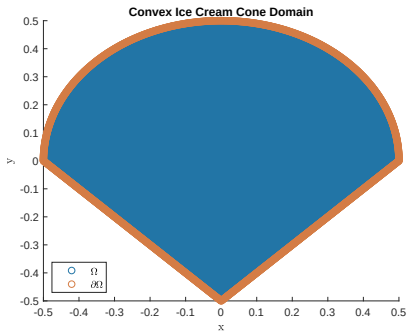

Figure B.1

Non-Convex Ice Cream Cone Domain

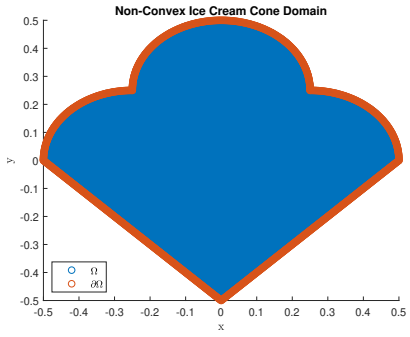

Figure B.2

Cloud domain

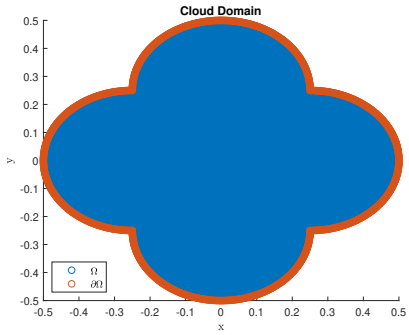

Figure B.3

Bibliography

[1] Robert A.Adams. *Sobolev Spaces*. Academic Press, 1975.

[2] Ben Adcock and Daan Huybrechs. Frames and numerical approximation, 2016.

[3] Ben Adcock and Daan Huybrechs. Frames and numerical approximation ii: generalized sampling, 2018.

[4] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, 2010.

[5] Peter Casazza and Gitta Kutyniok. *Finite Frames: Theory and Applications*. 01 2012.

[6] Ole Christensen. *An Introduction to Frames and Riesz Basis*. Birkhauser, 1966.

[7] Ole Christensen and Elnaz Osgooei. On frame properties for fourier-like systems. *Journal of Approximation Theory*, 172:47 – 57, 2013.

[8] Nicolae Cotfas and Jean-Pierre Gazeau. Finite tight frames and some applications. *Journal of Physics A Mathematical and Theoretical*, 43, 03 2008.

[9] Ingrid. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[10] Lawrence Evans. *Partial Differential Equations*. American Mathematical Society, 1998.

[11] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman Publishing, 1985.

[12] William McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, 2000.

[13] Eleonora Di Nezza, Giampiero Palatucci, and Enrico Valdinoci. Hitchhiker's guide to the fractional sobolev spaces, 2011.

[14] Jason Nicholson. Bessel zero solver, 2020. The MathWorks, Natick, MA, USA.

[15] K. Rektorys. *Variational Methods in Mathematics, Science and Engineering*. Springer, 2001.

[16] S Brenner L. Scott. *The mathematical Theory of Finite Element Methods*. Springer, 2008.

[17] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997.

[18] D. Gilbarg N. Trudinger. *Elliptic Partial Differential Equaions of Second Order*. Springer-Verlag, 1970.

[19] Maria Pereyra Lesley Ward. *Harmonic Analysis From Fourier to Wavelets*. American Mathematical Society, 2012.