ABSTRACT

Title of dissertation:	MESHLESS COLLOCATION METHODS FOR THE NUMERICAL SOLUTION OF ELLIPTIC BOUNDARY VALUED PROBLEMS AND THE ROTATIONAL SHALLOW WATER EQUATIONS ON THE SPHERE
	Christopher D. Blakely, Doctor of Philosophy, 2009
Dissertation directed by:	Professor John Osborn Department of Mathematics Professor Ferdinand Baer Department of Atmospheric and Oceanic Science

This dissertation thesis has three main goals: 1) To explore the anatomy of meshless collocation approximation methods that have recently gained attention in the numerical analysis community; 2) Numerically demonstrate why the meshless collocation method should clearly become an attractive alternative to standard finite-element methods due to the simplicity of its implementation and its high-order convergence properties; 3) Propose a meshless collocation method for large scale computational geophysical fluid dynamics models.

We provide numerical verification and validation of the meshless collocation scheme applied to the rotational shallow-water equations on the sphere and demonstrate computationally that the proposed model can compete with existing high performance methods for approximating the shallow-water equations such as the SEAM (spectral-element atmospheric model) developed at NCAR. A detailed anal-

ysis of the parallel implementation of the model, along with the introduction of parallel algorithmic routines for the high-performance simulation of the model will be given. We analyze the programming and computational aspects of the model using Fortran 90 and the message passing interface (mpi) library along with software and hardware specifications and performance tests. Details from many aspects of the implementation in regards to performance, optimization, and stabilization will be given.

In order to verify the mathematical correctness of the algorithms presented and to validate the performance of the meshless collocation shallow-water model, we conclude the thesis with numerical experiments on some standardized test cases for the shallow-water equations on the sphere using the proposed method.

MESHLESS COLLOCATION METHODS FOR THE NUMERICAL SOLUTION OF ELLIPTIC BOUNDARY VALUED PROBLEMS THE ROTATIONAL SHALLOW WATER EQUATIONS ON THE SPHERE

by

Christopher D. Blakely

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor John E. Osborn, Chair
Professor Ferdinand Baer, Co-Chair
Dr. Michael Fox-Rabinovitz
Professor Konstantina Trivisa
Professor James Drake

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Historical Context and Motivations

The meshless collocation method for approximating solutions to partial differential equations (PDEs) is a recent and fast growing research area that spans many different fields in applied mathematics, science and engineering. As the name suggests, it deals with computing the numerical solution of a PDE on a set of given data locations that are scattered throughout the domain of interest, all without the use of a computational mesh which are traditionally required in standard spectral, finite element, and finite difference methods. Instead, the method uses collocation in the sense that it poses the approximation to be satisfied exactly on a set of collocation points, or nodes, in the domain of interest and on the boundary.

Historically, collocation methods for the solution of PDEs have been centered mathematically on a fairly simple approach to approximation. The methods typically choose a finite dimensional approximation space of candidate solutions along with a finite collection of points in the domain and boundary (called *collocation points* or *node*) and aim at selecting the solution that satisfies the given PDE exactly at those points. A popular choice in the literature for the finite dimensional approximation space of candidate solutions are (orthogonal) polynomials up to a certain degree. While these methods have generally shown great success using such polynomial spaces as Chebychev or Legendre polynomials (see Boyd [13]), adapting them to nontrivial domains which are either nonconvex, nonsmooth, or both has been a limiting drawback to the method due to the fact that the collocation points are required to be Gaussian-type quadrature points. While the method has been shown to converge ([13]), this type of collocation limits the method geometrically.

1.1.1 Why Meshless Collocation?

Freedom to choose the collocation points is of course a highly desirable feature in collocation and hasn't been available until the recent mathematical innovations of meshless collocation. In chapter 2, we investigate these innovations, which include the so-called *Native Space* theory of symmetric positive definite kernels, developed in the seminal paper by Schaback [44], and their connection with reproducing kernel Hilbert spaces. We will try to demonstrate why and how the theory of native spaces work in the context of scattered data approximation and, more generally, collocation for PDEs.

As we will show, this approach to collocation enables the choice of a collection of collocation points that can be a rather general set of scattered data points satisfying mild restrictions to ensure convergence. The finite dimensional space is the span of functions of the form $\Psi(\cdot, \mathbf{x}_1), \ldots, \Psi(\cdot, \mathbf{x}_N)$ where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are the collocation nodes and $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ will be chosen to be a compactly supported symmetric positive definite kernel which satisfies high-order smoothness and algebraic spectral decay.

So why would one want to use meshless collocation for numerically solving PDEs in the first place? Aren't finite element and difference methods (FEM, FDM) powerful enough? After all, the theory of both methods reach back through the 1950s, with thousands of research papers and a slew of books on the subject. Furthermore, new innovations for FE solution and refinement techniques are making way into many commercial software packages (FEMlab, FEMpack,etc.) for tackling large-scale/high-performance problems in all engineering fields, atmospheric science, physics, and so on.

Motivations for employing meshless collocation might stem from the desire to approximate solutions to PDEs where boundaries to the domain of interest have either very complicated geometric structures or even move in time dependent problems. As we hope to demonstrate in this thesis, another desirable feature of meshless collocation is its approximation robustness and fast convergence rates for smooth problems, as well as overall ease of implementation. One of the major advantages of meshless collocation over traditional and generalized finite element methods (see [7] for example) that we stress to communicate in this thesis is that meshless collocation does not require numerical quadrature for integration since integration is not performed in the collocation method.

A grand challenge in the current trend of generalized finite element methods comes in selecting optimal quadrature points and weights for the numerical integration over supports of the underlying basis functions. The introduction of quadrature then of course leads to additional numerical errors which can propagate throughout the global approximation. The fact that no integration is done in collocation greatly simplifies the implementation of the method and allows much freedom in selecting the basis kernels and the placement of collocation nodes. Furthermore, since no triangularization or rectangularization of the domain is required, collocation can be used on much more general smooth manifolds without the problem of domain discretization errors.

The final goal for this thesis will be to demonstrate numerically how and why meshless collocation can compete with standard computational methods for large-scale numerical solutions to nonlinear PDEs. For this, we tackle the problem of applying meshless collocation to a large-scale geophysical dynamics model.

Two fundamental problems in numerically simulating Global Climate Models (GCM) which have challenged researchers during the past few decades are the long-term stability requirements in the underlying numerical approximations of the climate model along with the sensitivity to small changes in regional scales of the model. Most GCMs in the past have had difficulty being able to simulate regional meteorological phenomena such as tropical storms, which play an important part in the latitudinal transfer of energy and momentum. This is partly due to the fact that climate models have traditionally employed spectral methods using spherical harmonics which are global and require excessively large resolution in order to inherit any properties which can be used to study regional scale phenomenon (see Baer et al. [6]). This is a massive computational burden since the increase in resolution must be done globally due to the nature of the numerical method. In order to tackle these two problems in a computationally efficient manner, the high-order convergence and approximation properties of global spectral methods would ultimately need to be coupled with the ability to locally refine approximation results of the model in certain regions of the global domain.

In the final part of the thesis, we propose, implement, and experiment with a meshless collocation paradigm for use in the parallel high-performance simulation of nonlinear geophysical models while keeping these two problems that have challenged researchers in computational geophysics and climatology in mind. Due to its simplicity while still retaining some of the nonlinear challenges of larger, more complex geophysical models, our geophysical model of choice will be the rotational shallow-water equations on the sphere.

Ultimately, we wish to show that our proposed meshless collocation method can compete with the many models and benchmarks published in today's high-performance scientific computing industries, and can also provide an attractive alternative to standard finite-element techniques for regional and global modeling of geophysical fluid dynamics.

1.2 Thesis Outline

This dissertation thesis has three main goals: 1) To explore the anatomy of meshless collocation approximation methods that have recently gained attention in the numerical analysis community; 2) Numerically demonstrate why the meshless collocation method should clearly become an attractive alternative to standard finite-element methods due to the simplicity of its implementation and its high-order convergence properties; 3) Propose a meshless collocation method for large scale computational geophysical fluid dynamics models.

In chapter 2 of the thesis, we give a tour of some of the theoretical highlights of meshless collocation that we hope will demonstrate the robust approximation power of the method. Much of the theory of meshless collocation that is featured in this thesis was developed by Schaback and Wendland the past decade in a combination of papers. We summarize the important aspects of the theory and give detailed proofs for the main results.

In the second part of chapter 2, we will explore implementational issues of meshless collocation and give a suite of numerical experiments for the approximation method which will aim at verifying and validating many of the theoretical claims and results discussed in the first part of the chapter. We hope to demonstrate the computational robustness of meshless collocation for different types of domains which are convex, nonconvex, smooth, and nonsmooth where we focus primarily on determining numerical convergence rates for the method while using different collocation node distributions.

In the final part of the second chapter, we provide a computational comparison between the meshless collocation method and the standard finite-element method with piecewise linear elements on a few different test problems to assess the differences in numerical convergence and stability issues. As already mentioned, one of the major advantages of meshless collocation over traditional and generalized finite element methods that we will continue to highlight in this thesis is that meshless collocation does not require numerical quadrature for integration since integration is not performed in the collocation method. Clearly, this is one desirable feature since it renders implementational issues much easier to handle computationally.

In an effort to demonstrate that meshless collocation can compete with spectral/finite-element methods in regards to numerical robustness and computational speed for solving large-scale problems in geophysical dynamics problems on the sphere, in the last two chapters of the thesis, we introduce a new meshless collocation method for solving the rotational shallow-water equations on the sphere. The method we propose is quite versatile in that it can be constructed using any set of collocation points in the domain. Being roughly based on the theory of pseudospectral approximation, this meshless collocation method seeks to approximate pointwise values and their derivatives by using a differentiation matrix construction similar to pseudospectral methods. One of the advantages that we attempt to demonstrate is that the method is not only fast and easy to implement, but also that the matrices resulting from the linear systems in conjuction with the semi-implicit time stepping scheme that we propose are symmetric and positive definite. This enables fast parallel conjugate gradient solvers for obtaining the solution at each collocation node.

We begin chapter 3 by first discussing the geophysical model that we will use, namely the rotational shallow-water equations on the sphere, and then give its discretization on the so-called cubed-sphere followed by a semi-implicit time stepping scheme to integrate the geophysical model in time. Lastly, we discuss in detail the construction of the meshless collocation approach using compactly supported radial basis functions. We show how to apply the method at each semi-implicit time step to yield a symmetric positive definite system that can be solved using conjugate gradient.

In the final part of the thesis, we provide numerical verification and validation of the meshless collocation scheme applied to the rotational shallow-water equations on the sphere introduced in chapter 3. Our goal is to show computationally that this proposed model can compete with existing high performance methods for approximating the shallow-water equations such as the SEAM (spectral-element atmospheric model) developed at NCAR all while highlighting the advantages and disadvantages of the method. A detailed analysis of the parallel implementation of the model, along with the introduction of parallel algorithmic routines for the high-performance simulation of the model will be given. We analyze the programming and computational aspects of the model using Fortran 90 and the message passing interface (mpi) library along with software and hardware specifications and performance tests. Details from many aspects of the implementation in regards to performance, optimization, and stabilization will be given.

A theoretical discussion of regional modeling in geophysical systems will then introduce the final numerical experiments section of the thesis. In order to verify the mathematical correctness of the algorithms presented and to validate the performance of the meshless collocation shallow-water model, we conclude the thesis with numerical experiments on some standardized test cases for the shallow-water equations on the sphere using the proposed method. These test cases, introduced by Williamson et al. in [71], are now considered the standard test suite for analyzing the performance of newly proposed numerical schemes for the spherical SWEs.

We end the thesis with a conclusion summarizing the numerical and theoretical efforts of this project including a discussion on some of the advantages and disadvantages that we have experienced using meshless collocation as a method for solving large-scale PDEs. Directions into further ongoing developments and future research in the field of meshless collocation will also be given.

Chapter 2

The Meshless Collocation Method

2.1 Introduction

The meshless collocation method for approximating solutions to partial differential equations (PDEs) is a recent and fast growing research area that spans many different fields in applied mathematics, science and engineering. As the name suggests, it deals with computing the numerical solution of a PDE on a set of given data locations that are scattered throughout the domain of interest, all without the use of a computational mesh as in the classical finite element and difference methods. The method uses collocation in the sense that it poses the approximation to be satisfied exactly on a set of collocation points, or nodes, in the domain of interest and on the boundary.

Motivations for employing meshless collocation might stem from the desire to approximate solutions to PDEs where boundaries to the domain of interest have either very complicated geometric structures or even move in time dependent problems. As we hope to demonstrate, we take interest in meshless collocation namely due to its approximation robustness and ease of implementation, and also to the fact that approximation refinement with collocation is rather simple to implement as we will see; simply increase the number of collocation nodes in the domain, no need to refine a mesh.

The theory behind the meshless collocation method that we use in this thesis actually stems from the theory of scattered data approximation (an excellent resource to the field of scattered data approximation is provided by H. Wendland in [66].) Scattered data approximation heavily relies on the theory of *reproducing kernel Hilbert spaces* and so-called *native spaces* of positive definite functions. The notion of native spaces of positive definite functions was initially introduced by Schaback in [46] to provide a framework for the approximation of functions u from a certain Hilbert space $\mathcal{H}$. The approximation was assumed to be done on scattered samples of the function u and thus an interpolant was created out of positive definite kernels to interpolate the scattered data. The idea of the native spaces of the kernel was to show equivalence with the Hilbert space in regards to the norm of the Hilbert space. This way, analysis of error estimates and stability could be done directly in the native space.

In this chapter, we wish to investigate the symmetric meshless collocation method as an attractive alternative to standard finite element methods by highlighting some of the main features which makes this collocation method such a numerical success. In the first part of the chapter, we review results from reproducing kernel Hilbert spaces and the notion of native spaces for positive definite functions and discuss their properties which provide much of the theoretical backbone of meshless collocation. We then give a fairly detailed analysis of the main fundamental error estimate for scattered data approximation in native spaces which is the underlying prerequisite to understanding the success of meshless collocation. The third part of the chapter then discusses the symmetric meshless collocation method for boundary-valued elliptic partial differential using the theory derived from scattered data approximation in native spaces of positive definite kernels. Deriving the main error estimate for meshless collocation will then conclude the theoretical portion of the chapter. For the final part of the chapter, we take an in-depth numerical tour of symmetric meshless collocation which will aim at providing further insight into the numerical robustness of the collocation method. We will give an explicit example of how the collocation method is constructed computationally and applied to a simple elliptic PDE. The numerical convergence rate for a few example problems will then be given along with a section dedicated to comparing the performance of the finite element method and the meshless collocation method for a couple elliptic problems on both smooth and nonsmooth nonconvex domains. We conclude the numerical section with a study on the numerical stability of the collocation method.

2.2 Theory of Meshless Collocation

2.2.1 Preliminaries

Before we begin the discussing the tools needed for the meshless collocation method, we must mention some notation that will be used for the domain Ω on which we work. Unless otherwise stated, we will define Ω to be an open bounded connected set in $\mathbb{R}^2$ (which we will frequently call a *domain*) and assume in addition that Ω satisfies an interior cone condition and has a Lipschitz boundary $\partial\Omega$. We will use the following definition of the cone condition.

Definition The set Ω satisfies an interior cone condition if there exists an angle

$\theta \in (0, \pi/2)$ and a radius $r > 0$ such that for every $\mathbf{x} \in \Omega$ a unit vector $\xi(\mathbf{x})$ exists such that the cone

$$C(\mathbf{x}, \xi(\mathbf{x}), \theta, r) := \{\mathbf{x} + \lambda \mathbf{y} : \mathbf{y} \in \mathbb{R}^2, \|\mathbf{y}\|_2 = 1, \mathbf{y}^T \xi(\mathbf{x}) \geq \cos(\theta), \lambda \in [0, r]\} \quad (2.1)$$

is contained in Ω.

As we will see, the cone condition property will become important in the analysis of the meshless collocation method as it allows for different approximation results to hold which we will discuss later in the thesis.

Fundamental to the concept of the meshless collocation method is the set of collocation nodes on which approximations are computed. To this end, we define a set of N *pairwise distinct points* (or simply *nodes* as we will sometimes call them) as $\mathcal{X}_N = \{\mathbf{x}_1^N, \ldots, \mathbf{x}_N^N\} \subseteq \Omega$, where the superscript on each point $\mathbf{x}_j^N \in \mathcal{X}_N$ denotes the dependence on N. For a given integer N, we associate with the set $\mathcal{X}_N$ a measure $h_{\mathcal{X}_N, \Omega}$ defined by

$$h_{\mathcal{X}_N, \Omega} = \sup_{\mathbf{x} \in \Omega} \min_{\mathbf{x}_j^N \in \mathcal{X}_N} \|\mathbf{x} - \mathbf{x}_j^N\|_2. \quad (2.2)$$

We will call this the *point saturation measure* or *fill distance* of $\mathcal{X}_N$ and it can be interpreted as follows: for any $\mathbf{x} \in \Omega$, there is a node $\mathbf{x}_j^N \in \mathcal{X}_N$ within a distance at most $h_{\mathcal{X}_N, \Omega}$.

It is natural for convergence of approximations to require that the distribution of points $\mathcal{X}_N$ has the property that as $N \to \infty$, then $h_{\mathcal{X}_N, \Omega} \to 0$. In other words the distribution of points in $\mathcal{X}_N \subseteq \Omega$ should "cover" Ω rather well. In order to help us in determining if the domain Ω is "covered" sufficiently well, we introduce another useful quantity associated with the point distribution $\mathcal{X}_N$. This is the *separation* *distance* defined as

$$q_{\mathcal{X}_N} := \frac{1}{2} \min_{j \neq k} \|\mathbf{x}_j^N - \mathbf{x}_k^N\|_2, \tag{2.3}$$

which is half the shortest distance between any two distinct points in $\mathcal{X}_N$. Using $q_{\mathcal{X}_N}$ and $h_{\mathcal{X}_N,\Omega}$, we can define the notion of *quasi-uniform* for a family of sets $\mathcal{X}_N$.

Definition We say that the family of sets $\mathcal{X}_N \subseteq \Omega$ is *quasi-uniform* with respect to a constant $c > 0$ if

$$q_{\mathcal{X}_N} \leq h_{\mathcal{X}_N,\Omega} \leq c q_{\mathcal{X}_N} \tag{2.4}$$

for any $N > 1$.

The quasi-uniform property for the family of sets $\mathcal{X}_N$ ensures that the separation distance $q_{\mathcal{X}_N}$ and the fill distance $h_{\mathcal{X}_N,\Omega}$ are equivalent in that their ratio $h_{\mathcal{X}_N,\Omega}/q_{\mathcal{X}_N}$ can be bounded below by 1 and above by a constant c for any $N > 1$.

To see the importance of the quasi-uniform property, let us first consider a uniform family of centers $\mathcal{X}_N$ in the following example. Let $\Omega = (0,1)^2$ and define $\mathcal{X}_N := (1/N)\mathbb{Z}^2 \cap \Omega$ for $N \geq 2$ (for example $\mathcal{X}_3 := \{(1/3,1/3),(1/3,2/3),(2/3,1/3),(2/3,2/3)\}$. Then the separation distance is $q_{\mathcal{X}_N} = \frac{1}{2N}$ while the fill distance is $h_{\mathcal{X}_N,\Omega} = \frac{\sqrt{2}}{N}$ and thus the quasi-uniform constant is $c = 2\sqrt{2}$. Obviously in the uniform case, the nodes in $\mathcal{X}_N$ "cover" Ω arbitrarily well in that no point $\mathbf{x} \in (0,1)^2$ is greater than a distance of $\sqrt{2} q_{\mathcal{X}_N}$ to its nearest neighboring node in $\mathcal{X}_N$. In the quasi-uniform case, we would thus like for the constant c to be as close to $2\sqrt{2}$ as possible; the closer c is to $2\sqrt{2}$, the closer $\mathcal{X}_N$ is to being purely uniform. This ensures that the nodes in $\mathcal{X}_N$ are not too close together and at the same time "cover" Ω rather well as $N \to \infty$.

We note that the left-hand side of the inequality in (2.4) is not only a restriction on $\mathcal{X}_N$, but also on Ω. There are instances in which the left-hand inequality does not hold true for an open bounded set Ω and a given $\mathcal{X}_N$. However, if Ω satisfies the interior cone condition with radius $r > 0$ and $q_{\mathcal{X}_N} < r$, then it can be shown that $q_{\mathcal{X}_N} \leq h_{\mathcal{X}_N,\Omega}$ holds (see page 232 of Wendland [66]). Another instance in which the inequality holds is in the case Ω is convex.

Without any chance of misunderstanding, we will often use throughout the remainder of this thesis an abbreviated notation for the point saturation and separation distances by suppressing the notation of the dependence on $\mathcal{X}_N$ and Ω by simply letting $h := h_{\mathcal{X}_N,\Omega}$ and $q := q_{\mathcal{X}_N}$. We will also occasionally denote the set of nodes $\mathcal{X}_N$ by $\mathcal{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

2.2.2 Radial Functions

Before we define the notion of reproducing kernels and native spaces, we discuss radial functions which will be used to derive the reproducing kernels we are interested in. In this thesis, we will consider classes of reproducing kernels that are *radial functions.*

Definition A function $\Psi_0 : \mathbb{R}^2 \mapsto \mathbb{R}$ is said to be radial if there exists a function $\psi : [0, \infty) \mapsto \mathbb{R}$ such that $\Psi_0(\mathbf{x}) = \psi(\|\mathbf{x}\|_2)$, for all $\mathbf{x} \in \mathbb{R}^2$.

Specifically, we are interested in radial functions $\Psi_0 \in L^1(\mathbb{R}^2) \cap C(\mathbb{R}^2)$ possessing a multivariate Fourier transform defined as

$$\widehat{\Psi}_0(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} \Psi_0(\mathbf{x}) e^{-i\mathbf{x}\cdot\boldsymbol{\omega}} d\mathbf{x},$$

that decays algebraically of order $s \in \mathbb{R}^+$. This means there exists constants $0 \leq c_1 \leq c_2$ such that

$$c_1(1 + \|\boldsymbol{\omega}\|_2^2)^{-s} \leq \widehat{\Psi}_0(\boldsymbol{\omega}) \leq c_2(1 + \|\boldsymbol{\omega}\|_2^2)^{-s}, \quad \boldsymbol{\omega} \in \mathbb{R}^2 \tag{2.5}$$

for $\|\boldsymbol{\omega}\|_2 \to \infty$. Now if $s > 1$, we see that $\widehat{\Psi}_0 \in L^1(\mathbb{R}^2)$. Thus we consider radial functions $\Psi_0 \in L^1(\mathbb{R}^2) \cap C(\mathbb{R}^2)$ such that $s > 1$. This is an important property which we will need later on.

The final property that we will be interested in for our class of radial functions is the *positive definite* property.

Definition A continuous and even radial function $\Psi_0 : \mathbb{R}^2 \mapsto \mathbb{R}$ is said to be positive definite on $\mathbb{R}^2$ iff for any $N \in \mathbb{N}$, all sets of pairwise distinct nodes $\mathcal{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_N\} \subseteq \mathbb{R}^2$, and all vectors $\alpha \in \mathbb{R}^N/0$, the quadratic form

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \Psi_0(\mathbf{x}_i - \mathbf{x}_j) \tag{2.6}$$

is positive.

It is easy to see that the definition of positive definiteness for a radial function Ψ_0 is equivalent to the requirement that the matrix $\mathcal{A}_{\mathcal{X}}$ with entries $\mathcal{A}_{\mathcal{X}}[i,j] = \Psi_0(\mathbf{x}_i - \mathbf{x}_j)$ is positive definite for any set $\mathcal{X}$ of distinct nodes. This property will become quite useful later when we attempt to solve scattered data interpolation problems.

For the moment, let us assume that such a radial function $\Psi_0 \in L^1(\mathbb{R}^2) \cap C(\mathbb{R}^2)$ discussed in this section exists. Namely, one that satisfies the Fourier decay property

(2.5) with $s > 1$ and the positive definite property (2.6). Later, we will review a class of radial functions called the Wendland functions that satisfy these properties.

We now discuss the notion of reproducing kernels and reproducing kernel Hilbert spaces. We then show how we connect our class of radial functions Ψ_0 with reproducing kernels.

2.2.3 Reproducing Kernel Hilbert Spaces

Reproducing kernel Hilbert spaces which have been thoroughly studied in recent decades since the seminal paper by Aronszajin in ([4]). Most of the recent theory has been studied by Schaback in papers such as [46], [47] and references therein. In this section, we briefly summarize and give some fundamental results from the theory of reproducing kernel Hilbert spaces that we will use throughout the thesis.

We note that the theory of reproducing kernel Hilbert spaces only requires that $\Omega \subseteq \mathbb{R}^2$ be a nonempty set. Of course, this is much more general than what we will need later in the thesis when we try to numerically solve PDEs. So without taking away from the generality of the theory, we will continue to consider open bounded connected sets $\Omega \subset \mathbb{R}^2$ discussed in 2.2.1. We will also only consider real vector spaces of real-value functions in this thesis. We begin with the definition of a *kernel* and a special type of kernel called a *reproducing kernel.*

Definition A function $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ is called a *kernel.*

Definition Let $\mathcal{H}$ be a Hilbert space of functions $f : \Omega \mapsto \mathbb{R}$ with inner product

$(\cdot,\cdot)_{\mathcal{H}}$. A function $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ is called a *reproducing kernel* for $\mathcal{H}$ if

1. $\Psi(\cdot, \mathbf{y}) \in \mathcal{H}$ for all $\mathbf{y} \in \Omega$,
2. $f(\mathbf{y}) = (f, \Psi(\cdot, \mathbf{y}))_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and all $\mathbf{y} \in \Omega$.

From these properties in the definition, we can easily see that the reproducing kernel of a Hilbert space is uniquely determined. Suppose there are two reproducing kernels Ψ_1 and Ψ_2 for $\mathcal{H}$ that satisfy the above properties. Then 2. gives $(f, \Psi_1(\cdot, \mathbf{y}) - \Psi_2(\cdot, \mathbf{y}))_{\mathcal{H}} = 0$ for any $f \in \mathcal{H}$ and $\mathbf{y} \in \Omega$. But if we let $f \equiv \Psi_1(\cdot, \mathbf{y}) - \Psi_2(\cdot, \mathbf{y}) \in \mathcal{H}$ for any fixed $\mathbf{y} \in \Omega$, then $(\Psi_1(\cdot, \mathbf{y}) - \Psi_2(\cdot, \mathbf{y}), \Psi_1(\cdot, \mathbf{y}) - \Psi_2(\cdot, \mathbf{y}))_{\mathcal{H}} = 0$ which implies $\Psi_1(\cdot, \mathbf{y}) = \Psi_2(\cdot, \mathbf{y})$ and we see the uniqueness.

We now consider the dual space $\mathcal{H}'$ of $\mathcal{H}$. It contains all bounded linear functionals λ on $\mathcal{H}$ and is equipped with a dual norm

$$\|\lambda\|_{\mathcal{H}'} := \sup_{f \in \mathcal{H}/\{0\}} \frac{\lambda(f)}{\|f\|_{\mathcal{H}}}, \quad \forall \lambda \in \mathcal{H}'. \tag{2.7}$$

We will typically denote functions from $\mathcal{H}$ using Latin letters $f, g, \ldots$ and functionals from the dual space $\mathcal{H}'$ using Greek letters $\lambda, \delta, \ldots$. One functional that we will use frequently is the point evaluation functional $\delta_{\mathbf{x}} \in \mathcal{H}'$ defined by $\delta_{\mathbf{x}}(f) = f(\mathbf{x})$ for any $f \in \mathcal{H}$ and $\mathbf{x} \in \Omega$. Using the point evaluation functional and Riesz' representation Theorem, we give a characterization of a Hilbert space with a reproducing kernel, originally found in Schaback [46].

Theorem 2.2.1. *Suppose $\mathcal{H}$ is a Hilbert space of functions $f : \Omega \mapsto \mathbb{R}$. Then the following statements are equivalent.*

1. *The point evaluation functions are continuous, i.e., $\delta_{\mathbf{y}} \in \mathcal{H}'$ for all $\mathbf{y} \in \Omega$.*

2. $\mathcal{H}$ has a reproducing kernel.

Proof. Suppose that the point evaluations are continuous. Using Riesz' representation theorem, there exists an element $\Psi_{\mathbf{y}} \in \mathcal{H}$ such that $\delta_{\mathbf{y}}(f) = (f, \Psi_{\mathbf{y}})_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Thus we see that $\Psi(\mathbf{x}, \mathbf{y}) := \Psi_{\mathbf{y}}(\mathbf{x})$ is the reproducing kernel of $\mathcal{H}$. Now suppose that $\mathcal{H}$ has a reproducing kernel Ψ. This means that for any $f \in \mathcal{H}$, $\delta_{\mathbf{y}}(f) = f(\mathbf{y}) = (f, \Psi(\cdot, \mathbf{y}))_{\mathcal{H}}$. Then $|\delta_{\mathbf{y}}(f)| = |(f, \Psi(\cdot, \mathbf{y}))_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|\Psi(\cdot, \mathbf{y})\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$, thus $\delta_{\mathbf{y}} \in \mathcal{H}'$. □

We can also use the Riesz representation Theorem to show the following

Theorem 2.2.2. *Suppose $\mathcal{H}$ is a Hilbert space of functions $f : \Omega \mapsto \mathbb{R}$ with reproducing kernel Ψ. Then we have*

1. $\Psi(\mathbf{x}, \mathbf{y}) = (\Psi(\cdot, \mathbf{x}), \Psi(\cdot, \mathbf{y}))_{\mathcal{H}} = (\delta_{\mathbf{x}}, \delta_{\mathbf{y}})_{\mathcal{H}'}$ *for all* $\mathbf{x}, \mathbf{y} \in \Omega$

2. $\Psi(\mathbf{x}, \mathbf{y}) = \Psi(\mathbf{y}, \mathbf{x})$ *for all* $\mathbf{x}, \mathbf{y} \in \Omega$*.*

Proof. The Riesz' representation $R : \mathcal{H}' \mapsto \mathcal{H}$ for point evaluations is given by $R(\delta_{\mathbf{y}}) = \Psi(\cdot, \mathbf{y})$ due to the reproducing kernel properties. Thus since $\mathcal{H}'$ carries the inner product $(\delta_{\mathbf{x}}, \delta_{\mathbf{y}})_{\mathcal{H}'} = (R(\delta_{\mathbf{x}}), R(\delta_{\mathbf{y}}))_{\mathcal{H}}$, we have

$$(\delta_{\mathbf{x}}, \delta_{\mathbf{y}})_{\mathcal{H}'} = (\Psi(\cdot, \mathbf{x}), \Psi(\cdot, \mathbf{y}))_{\mathcal{H}}.$$

Furthermore, we have $\Psi(\mathbf{x}, \mathbf{y}) = \delta_{\mathbf{x}}(\Psi(\cdot, \mathbf{y})) = (\Psi(\cdot, \mathbf{y}), \Psi(\cdot, \mathbf{x}))_{\mathcal{H}} = (\Psi(\cdot, \mathbf{x}), \Psi(\cdot, \mathbf{y}))_{\mathcal{H}}$, from which we get properties 1. and 2. □

We see that Riesz' representation Theorem in reproducing kernel Hilbert spaces is quite useful since it allows us to represent the kernel by point evaluation functionals in the dual space and vice-versa. We use this in following Theorem from

Schaback [46] that connects some important properties about reproducing kernel Hilbert spaces. The Theorem will be useful throughout the remainder of this thesis as we will be able to deduce many results on native spaces which will be encountered later.

Theorem 2.2.3. *Let $\mathcal{H}$ be a reproducing kernel Hilbert space of functions on Ω with kernel Ψ and let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be any set of N distinct points in Ω. Then the following properties are equivalent.*

1. *The functions $\Psi(\cdot, \mathbf{x}_j)$ for all $\mathbf{x}_j \in \mathcal{X}$ are linearly independent on Ω.*
2. *The matrix $\mathcal{A}_{\mathcal{X}}$ defined by $\mathcal{A}_{\mathcal{X}}[i, j] = \Psi(\mathbf{x}_i, \mathbf{x}_j)$ is symmetric positive definite.*
3. *The point evaluation functionals $\{\delta_{\mathbf{x}_j}\}$ for all $\mathbf{x}_j \in \mathcal{X}$ are linearly independent.*

Proof. The equivalence of properties 2 and 3 follows directly from the Riesz' representation theorem in the reproducing kernel Hilbert space. Indeed, for any set of N pairwise distinct points $\mathcal{X} \subset \Omega$ and $\alpha \in \mathbb{R}^N\{0\}$ we have

$$\begin{aligned}\sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j \alpha_k \Psi(\mathbf{x}_j, \mathbf{x}_k) &= \Big(\sum_{j=1}^{N} \alpha_j \Psi(\cdot, \mathbf{x}_j), \sum_{k=1}^{N} \alpha_k \Psi(\cdot, \mathbf{x}_k)\Big)_{\mathcal{H}} \\ &= \Big(\sum_{j=1}^{N} \alpha_j \delta_{\mathbf{x}_j}, \sum_{k=1}^{N} \alpha_k \delta_{\mathbf{x}_k}\Big)_{\mathcal{H}'} = \|\sum_{j=1}^{N} \alpha_j \delta_{\mathbf{x}_j}\|_{\mathcal{H}'}^2 \geq 0.\end{aligned} \tag{2.8}$$

The last expression is zero only if the point evaluation functionals $\delta_{\mathbf{x}_j}$ are linearly dependent. Thus properties 2 and 3 are equivalent.

We now show that properties 1 and 2 are equivalent. Suppose that the set of functions $\Psi(\cdot, \mathbf{x}_j) \in \mathcal{H}$ for all $\mathbf{x}_j \in \mathcal{X}$ are linearly independent. Then for any

$\alpha \in \mathbb{R}^N$ not all zero, we have by Theorem 2.2.2 that

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(\Psi(\cdot,\mathbf{x}_i),\Psi(\cdot,\mathbf{x}_j))_{\mathcal{H}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\Psi(\mathbf{x}_i,\mathbf{x}_j). \tag{2.9}$$

The left-hand side of this equality is the Gram matrix for the functions $\Psi(\cdot,\mathbf{x}_j)$, for $\mathbf{x}_j \in \mathcal{X}$, with entries $(\Psi(\cdot,\mathbf{x}_i),\Psi(\cdot,\mathbf{x}_j))_{\mathcal{H}}$. So we see that $\Psi(\cdot,\mathbf{x}_1),\ldots,\Psi(\cdot,\mathbf{x}_N)$ are linearly independent if and only if the matrix $\Psi(\mathbf{x}_i,\mathbf{x}_j)$ is symmetric positive definite making the right side of the equality positive. This implies that 1 and 2 are equivalent. □

If a kernel Ψ is a reproducing kernel for the Hilbert space $\mathcal{H}$, then we will say that $\mathcal{H}$ is the *native space* for Ψ, where we denote the native space of Ψ by $\mathcal{N}_\Psi(\Omega)$, namely $\mathcal{H} = \mathcal{N}_\Psi(\Omega)$. We will see this in a more formal manner in the next section on Native spaces for positive definite functions.

2.2.4 Native Spaces of Positive Definite Kernels

Let us consider kernels $\Psi \in C(\Omega \times \Omega)$ which are symmetric, $(\Psi(\mathbf{x},\mathbf{y}) = \Psi(\mathbf{y},\mathbf{x}))$, and positive definite in the sense that for any $N \in \mathbb{N}$, all sets of pairwise distinct nodes $\mathcal{X} = \{\mathbf{x}_1,\ldots\mathbf{x}_N\} \subseteq \mathbb{R}^2$, and all vectors $\alpha \in \mathbb{R}^N/0$, the quadratic form

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\Psi(\mathbf{x}_i,\mathbf{x}_j) \tag{2.10}$$

is positive. We will refer to such kernels as SPD kernels. For example, if we define the kernel $\Psi(\mathbf{x},\mathbf{y}) := \Psi_0(\mathbf{x}-\mathbf{y})$ where Ψ_0 is a positive definite and radial function, then certainly Ψ is a symmetric positive definite kernel. We have thus generalized the notion of positive definite functions to kernels.

From Theorem 2.2.3 in the previous section, we are aware that SPD kernels Ψ appear naturally as the reproducing kernel of Hilbert spaces. However, one does not usually begin with a Hilbert function space and attempt to derive the reproducing kernel. Instead, we are rather interested in constructing native Hilbert spaces from SPD kernels. As we will investigate later in this section, this allows us to build Hilbert spaces of certain degrees of smoothness dependent on the smoothness properties of the kernel Ψ.

We show in this section how to construct native Hilbert spaces on sets $\Omega \subseteq \mathbb{R}^2$ using SPD kernels. Once again, as in the previous section, the theory of native spaces allows Ω to be quite general, as long as it is a measurable set. In fact, this theory can even be generalize to $\Omega \subseteq \mathbb{R}^d$ for any integer choice of d. Later in the thesis however, we will need to restrict ourselves to cases in which Ω is open, bounded, and connected and satisfies a cone condition. But for right now, we can assume $\Omega \subseteq \mathbb{R}^2$ is quite general.

Consider the infinite dimensional real linear space on Ω defined by

$$F_\Psi(\Omega) := \operatorname{span}\{\Psi(\cdot, \mathbf{x}) \ : \ \mathbf{x} \in \Omega\} \tag{2.11}$$

equipped with the bilinear form

$$\Big(\sum_{i=1}^{N} \alpha_i \Psi(\cdot, \mathbf{x}_i), \sum_{j=1}^{M} \beta_j \Psi(\cdot, \mathbf{y}_j)\Big)_\Psi := \sum_{i=1}^{N}\sum_{j=1}^{M} \alpha_i \beta_j \Psi(\mathbf{x}_i, \mathbf{y}_j)$$

for any sets of distinct points $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_M\}$ and integers N,M. (For example, if we take $N = M = 1$ and $\alpha_1 = \beta_1 = 1$, we simply get the formula $(\Psi(\cdot, \mathbf{x}), \Psi(\cdot, \mathbf{y}))_\Psi = \Psi(\mathbf{x}, \mathbf{y}),\ \mathbf{x}, \mathbf{y} \in \Omega$.) In the following Theorem, we show that $F_\Psi(\Omega)$ is a pre-Hilbert space with reproducing kernel Ψ and inner product $(\cdot,\cdot)_\Psi$.

Theorem 2.2.4. *(Wendland [66], Chapter 10) If $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ is a symmetric positive definite kernel, then $(\cdot,\cdot)_\Psi$ defines an inner product on $F_\Psi(\Omega)$. Furthermore, $F_\Psi(\Omega)$ is a pre-Hilbert space with reproducing kernel Ψ.*

Proof. Since Ψ is symmetric, it is clear that $(\cdot,\cdot)_\Psi$ is symmetric and bilinear and thus defines an inner product on the space $F_\Psi(\Omega)$. Choose an arbitrary function $f = \sum_{i=1}^N \alpha_i \Psi(\cdot, \mathbf{x}_i) \not\equiv 0$ from $F_\Psi(\Omega)$. Then the definition of the bilinear form gives

$$(f, f)_\Psi = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \Psi(\mathbf{x}_j, \mathbf{x}_k) > 0 \tag{2.12}$$

by the fact that Ψ is positive definite. Furthermore, we have for any $\mathbf{y} \in \Omega$,

$$(f, \Psi(\cdot, \mathbf{y}))_\Psi = (\sum_{j=1}^N \alpha_j \Psi(\cdot, \mathbf{x}_j), \Psi(\cdot, \mathbf{y}))_\Psi = \sum_{j=1}^N \alpha_j \Psi(\mathbf{y}, \mathbf{x}_j) = f(\mathbf{y}),$$

which establishes that Ψ is the reproducing kernel for $F_\Psi(\Omega)$. □

We now show that the completion of the pre-Hilbert space $F_\Psi(\Omega)$, denoted by $\mathcal{F}_\Psi(\Omega)$, with respect to the inner product is the native space for Ψ. We need to interpret the abstract elements of the completion as functions. Since the point-evaluation functionals are continuous on the pre-Hilbert space $F_\Psi(\Omega)$, their extensions to the completion remain continuous. Thus the point evaluation functionals $\delta_\mathbf{x}$ can be used to define elements of the completion. To see this clearly, let $f \in \mathcal{F}_\Psi(\Omega)$. Since $\mathcal{F}_\Psi(\Omega)$ is the closure of $F_\Psi(\Omega)$, there exists a sequence $f_j \in F_\Psi(\Omega)$ such that $\lim_{j\to\infty} f_j \to f$ in $\mathcal{F}_\Psi(\Omega)$. Now we can define $\delta_\mathbf{x}(f) = \lim_{j\to\infty} \delta_\mathbf{x}(f_j)$ by extending the functional $\delta_\mathbf{x}$ to $\mathcal{F}_\Psi(\Omega)$ by continuity.

Now let R be defined as the mapping

$$R : \mathcal{F}_\Psi(\Omega) \mapsto C(\Omega), \qquad R(f)(\mathbf{x}) := (f, \Psi(\cdot, \mathbf{x}))_\Psi. \tag{2.13}$$

The resulting functions are indeed continuous since

$$|Rf(\mathbf{x}) - Rf(\mathbf{y})| = (f, \Psi(\cdot, \mathbf{x}) - \Psi(\cdot, \mathbf{y}))_\Psi \leq \|f\|_\Psi \|\Psi(\cdot, \mathbf{x}) - \Psi(\cdot, \mathbf{y})\|_\Psi$$

and

$$\|\Psi(\cdot, \mathbf{x}) - \Psi(\cdot, \mathbf{y})\|_\Psi^2 = \Psi(\mathbf{x}, \mathbf{x}) + \Psi(\mathbf{y}, \mathbf{y}) - 2\Psi(\mathbf{x}, \mathbf{y}),$$

which goes to zero as $\mathbf{x} \to \mathbf{y}$ by the continuity of Ψ. Furthermore, we have $R(f)(\mathbf{x}) = (f, \Psi(\cdot, \mathbf{x}))_\Psi = f(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ and all $f \in F_\Psi(\Omega)$. From this we have the following lemma.

Lemma 2.2.1. *The linear mapping* $R : \mathcal{F}_\Psi(\Omega) \mapsto C(\Omega)$ *is injective.*

Proof. $Rf = 0$ for an $f \in \mathcal{F}_\Psi(\Omega)$ would mean that $(f, \Psi(\cdot, \mathbf{x}))_\Psi = 0$ for all $\mathbf{x} \in \Omega$ implying that $f \perp F_\Psi(\Omega)$. But since $\mathcal{F}_\Psi(\Omega)$ is the completion of $F_\Psi(\Omega)$, the only element from $\mathcal{F}_\Psi(\Omega)$ which is perpendicular to $F_\Psi(\Omega)$ is $f = 0$. □

We can now conclude that the native Hilbert space of the positive definite kernel Ψ is indeed the completion of $F_\Psi(\Omega)$, $\mathcal{F}_\Psi(\Omega)$.

Definition The native Hilbert function space corresponding to the symmetric positive definite kernel $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ is defined by

$$\mathcal{N}_\Psi(\Omega) := R(\mathcal{F}_\Psi(\Omega))$$

and carries the inner product

$$(f, g)_{\mathcal{N}_\Psi(\Omega)} := (R^{-1}f, R^{-1}g)_\Psi.$$

We just saw that the space defined by $\mathcal{N}_\Psi(\Omega) := R(\mathcal{F}_\Psi(\Omega))$ is indeed a Hilbert space of continuous functions on Ω with reproducing kernel Ψ. Furthermore, since $\Psi(\cdot, \mathbf{x})$ is an element of $F_\Psi(\Omega)$ for any $\mathbf{x} \in \Omega$, it is unchanged under the mapping R and hence $f(\mathbf{x}) = (R^{-1}f, \Psi(\cdot, \mathbf{x}))_\Psi = (f, \Psi(\cdot, \mathbf{x}))_{\mathcal{N}_\Psi(\Omega)}$ for all $f \in \mathcal{N}_\Psi(\Omega)$ and $\mathbf{x} \in \Omega$.

Theorem 2.2.5. *([66], Chapter 10) Suppose that $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ is a SPD kernel. Then the associated native space $\mathcal{N}_\Psi(\Omega)$ is a Hilbert function space with reproducing kernel Ψ.*

A natural question now is to ask exactly how native spaces of symmetric positive definite kernels Ψ and reproducing kernel Hilbert spaces relate. Namely we want to know that if $\mathcal{H}(\Omega)$ is a reproducing kernel Hilbert space with kernel Ψ, then are the spaces $\mathcal{N}_\Psi(\Omega)$ and $\mathcal{H}(\Omega)$ the same? The next Theorem answers this question.

Theorem 2.2.6. *([66], Chapter 10) Suppose that Ψ is a symmetric positive definite kernel. Suppose further that $\mathcal{H}$ is a Hilbert space of functions $f : \Omega \mapsto \mathbb{R}$ with reproducing kernel Ψ. Then $\mathcal{H}$ is the native space $\mathcal{N}_\Psi(\Omega)$ and the inner products are the same.*

Proof. Let Ψ be the reproducing kernel of $\mathcal{H}$. Then any $f = \sum_{i=1}^N \alpha_i \Psi(\cdot, \mathbf{x}_i) \in F_\Psi$ is obviously in $\mathcal{H}$ and

$$\|f\|_{\mathcal{H}}^2 = \sum_i^N \sum_j^N \alpha_i \alpha_j (\Psi(\cdot, \mathbf{x}_i), \Psi(\cdot, \mathbf{x}_j))_{\mathcal{H}} = \sum_i^N \sum_j^N \alpha_i \alpha_j \Psi(\mathbf{x}_i, \mathbf{x}_j) = \|f\|_{\mathcal{N}_\Psi(\Omega)}^2 \tag{2.14}$$

for some $\alpha \in \mathbb{R}^N/\{0\}$ and set of distinct points $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

To show that $\mathcal{N}_\Psi(\Omega) \subseteq \mathcal{H}$, take any $f \in \mathcal{N}_\Psi(\Omega)$. There exists a sequence $f_n \subseteq F_\Psi$ converging to f in $\mathcal{N}_\Psi(\Omega)$. We see that f is given pointwise by $f(\mathbf{x}) = \lim_{n\to\infty} f_n(\mathbf{x})$ since

$$|f(\mathbf{x}) - f_n(\mathbf{x})| = |(f - f_n, \Psi(\cdot, \mathbf{x}))_{\mathcal{N}_\Psi(\Omega)}| \leq \|f - f_n\|_{\mathcal{N}_\Psi(\Omega)} \|\Psi(\cdot, \mathbf{x})\|_{\mathcal{N}_\Psi(\Omega)}.$$

This means that f_n is also a Cauchy sequence in $\mathcal{H}$ since $\|f_n - f_m\|_{\mathcal{N}_\Psi} \to 0$ implies $\|f_n - f_m\|_{\mathcal{H}} \to 0$ as $n \to m$ by the equivalence of the norms (2.14). Thus in $\mathcal{H}$, f_n converges to some $g \in \mathcal{H}$. But the reproducing property of $\mathcal{H}$ gives

$$|g(\mathbf{x}) - f_n(\mathbf{x})| = |(g - f_n, \Psi(\cdot, \mathbf{x}))_{\mathcal{H}}| \leq \|g - f_n\|_{\mathcal{H}} \|\Psi(\cdot, \mathbf{x})\|_{\mathcal{H}}$$

and so $g(\mathbf{x}) = \lim_{n\to\infty} f_n(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ implying that $f = g \in \mathcal{H}$ and thus $\mathcal{N}_\Psi(\Omega) \subseteq \mathcal{H}$. Now suppose that $\mathcal{N}_\Psi(\Omega) \neq \mathcal{H}$. Since $\mathcal{N}_\Psi(\Omega)$ is closed, there exists some $0 \not\equiv f \in \mathcal{H}$ that is orthogonal to $\mathcal{N}_\Psi(\Omega)$. But this means that $f(\mathbf{x}) = (f, \Psi(\cdot, \mathbf{x}))_{\mathcal{H}} = 0$ for all $\mathbf{x} \in \Omega$ which implies $f \equiv 0$, a contradiction. Thus $\mathcal{N}_\Psi(\Omega) = \mathcal{H}$ and the inner products are the same since the norms are same by the polarization identity. □

We now give a concrete example of a reproducing kernel Hilbert space on $\Omega = \mathbb{R}^2$ by constructing a native space out of the kernel $\Psi(\mathbf{x}, \mathbf{y}) := \Psi_0(\mathbf{x} - \mathbf{y})$ where $\Psi_0 \in L^1(\mathbb{R}^2) \cap C(\mathbb{R}^2)$ is radial and satisfies the decay rate (2.5) with $s > 1$. The following Theorem will allow us to build native spaces on $\mathbb{R}^2$ that are in fact smoothness spaces. We will then see in the next section on restriction and extension how we can generalize this to get native spaces $\mathcal{N}_\Psi(\Omega)$ which are smoothness spaces on domains $\Omega \subset \mathbb{R}^2$ that satisfy certain conditions.

Theorem 2.2.7. *(Wendland [66], Chapter 10) Suppose that $\Psi_0 \in L^1(\mathbb{R}^2) \cap C(\mathbb{R}^2)$ is radial and satisfies the Fourier decay rate (2.5) with $s > 1$. Define*

$$\mathcal{G} := \left\{ f \in L^2(\mathbb{R}^2) \cap C(\mathbb{R}^2) \,:\, \hat{f}/\sqrt{\widehat{\Psi}_0} \in L^2(\mathbb{R}^2) \right\}$$

and equip this space with the bilinear form

$$(f,g)_{\mathcal{G}} := (2\pi)^{-1}(\hat{f}/\sqrt{\widehat{\Psi}_0}, \hat{g}/\sqrt{\widehat{\Psi}_0})_{L^2(\mathbb{R}^2)} = (2\pi)^{-1}\int_{\mathbb{R}^2} \frac{\hat{f}(\boldsymbol{\omega})\overline{\hat{g}(\boldsymbol{\omega})}}{\widehat{\Psi}_0(\boldsymbol{\omega})} d\boldsymbol{\omega}.$$

Then $\mathcal{G}$ is a real Hilbert space with inner product $(\cdot,\cdot)_{\mathcal{G}}$ and reproducing kernel $\Psi(\mathbf{x},\mathbf{y}) := \Psi_0(\mathbf{x}-\mathbf{y})$.

Proof. First of all, we know that $\widehat{\Psi}_0 \in L^1(\mathbb{R}^2)$ since $\widehat{\Psi}_0$ satisfies the decay rate (2.5) with $s > 1$. Hence for any $f \in \mathcal{G}$, $\hat{f} \in L^1(\mathbb{R}^2)$ since

$$\int_{\mathbb{R}^2} |\hat{f}(\boldsymbol{\omega})| d\boldsymbol{\omega} \leq \Big(\int_{\mathbb{R}^2} \frac{|\hat{f}(\boldsymbol{\omega})|^2}{\widehat{\Psi}_0(\boldsymbol{\omega})} d\omega \Big)^{1/2} \Big(\int_{\mathbb{R}^2} \widehat{\Psi}_0(\boldsymbol{\omega}) d\boldsymbol{\omega} \Big)^{1/2}$$

which is finite. Since $f \in L^2(\mathbb{R}^2)$ is continuous and $\hat{f} \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$, by Plancherel's Theorem, f can be recovered pointwise from

$$f(\mathbf{x}) = (2\pi)^{-1} \int_{\mathbb{R}^2} \hat{f}(\boldsymbol{\omega}) e^{i\mathbf{x}\cdot\boldsymbol{\omega}} d\boldsymbol{\omega}, \quad \boldsymbol{\omega} \in \mathbb{R}^2$$

almost everywhere.

We now show that the bilinear form $(\cdot,\cdot)_{\mathcal{G}}$ is real and positive definite, and thus is an inner product on $\mathcal{G}$. Indeed, for any real $f, g \in L^2(\mathbb{R}^2)$, we have $\overline{\hat{f}(\boldsymbol{\omega})} = \hat{f}(-\boldsymbol{\omega})$ where $\boldsymbol{\omega} := (\omega_1, \omega_2) \in \mathbb{R}^2$ which results in

$$\begin{aligned}
\int_{\mathbb{R}^2} \frac{\hat{f}(\boldsymbol{\omega})\overline{\hat{g}(\boldsymbol{\omega})}}{\widehat{\Psi}_0(\boldsymbol{\omega})} d\boldsymbol{\omega} &= \int_{\omega_1>0} \frac{\hat{f}(\boldsymbol{\omega})\overline{\hat{g}(\boldsymbol{\omega})}}{\widehat{\Psi}_0(\boldsymbol{\omega})} + \frac{\hat{f}(-\boldsymbol{\omega})\overline{\hat{g}(-\boldsymbol{\omega})}}{\widehat{\Psi}_0(-\boldsymbol{\omega})} d\boldsymbol{\omega} \\
&= \int_{\omega_1>0} \frac{\hat{f}(\boldsymbol{\omega})\overline{\hat{g}(\boldsymbol{\omega})} + \overline{\hat{f}(\boldsymbol{\omega})}\hat{g}(\boldsymbol{\omega})}{\widehat{\Psi}_0(\boldsymbol{\omega})} d\boldsymbol{\omega} \qquad (2.15) \\
&= 2\int_{\omega_1>0} \frac{\mathcal{R}[\hat{f}(\boldsymbol{\omega})\overline{\hat{g}(\boldsymbol{\omega})}]}{\Psi_0(\boldsymbol{\omega})} d\boldsymbol{\omega},
\end{aligned}$$

which is real. Since $(\cdot,\cdot)_{\mathcal{G}}$ is a weighted $L^2(\mathbb{R}^2)$ inner product with strictly positive weight $\widehat{\Psi}_0(\boldsymbol{\omega})$ for all $\boldsymbol{\omega} \in \mathbb{R}^2$, $(\cdot,\cdot)_{\mathcal{G}}$ is positive definite and thus defines an inner product.

To show that $\mathcal{G}$ is complete, we consider a sequence $\{f_n\}$ that is a Cauchy sequence in $\mathcal{G}$ which means that $\{\hat{f}_n/\sqrt{\widehat{\Psi}_0}\}$ is a Cauchy sequence in $L^2(\mathbb{R}^2)$. Thus there exists a function $\hat{g} \in L^2(\mathbb{R}^2)$ such that $\hat{f}_n/\sqrt{\widehat{\Psi}_0} \to \hat{g}$ in $L^2(\mathbb{R}^2)$. Now since $\hat{g}\sqrt{\widehat{\Psi}_0} \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ by using Schwartz's inequality and the fact that $\hat{g} \in L^2(\mathbb{R}^2)$ and $\widehat{\Psi}_0$ is bounded, we define

$$f(\mathbf{x}) := (2\pi)^{-1} \int_{\mathbb{R}^2} \hat{g}(\boldsymbol{\omega})\sqrt{\widehat{\Psi}_0} e^{i\mathbf{x}\cdot\boldsymbol{\omega}} d\boldsymbol{\omega}, \quad \mathbf{x} \in \mathbb{R}^2,$$

which is continuous, in $L^2(\mathbb{R}^2)$, and satisfies $\hat{f}/\sqrt{\widehat{\Psi}_0} = \hat{g} \in L^2(\mathbb{R}^2)$. It is also real-valued since, by using the inequality of Cauchy-Schwartz

$$|f(\mathbf{x}) - f_n(\mathbf{x})| \leq (2\pi)^{-1} \int_{\mathbb{R}^2} |\hat{g}(\boldsymbol{\omega})\sqrt{\widehat{\Psi}_0} - \hat{f}_n(\boldsymbol{\omega})| d\boldsymbol{\omega} \leq (2\pi)^{-1} \|\hat{g} - \hat{f}_n/\sqrt{\widehat{\Psi}_0}\|_{L^2(\mathbb{R}^2)} \big(\|\widehat{\Psi}_0\|_{L^1(\mathbb{R}^2)}\big)^{1/2}.$$

Hence $f \in \mathcal{G}$. Lastly, to show f_n converges to f in $\mathcal{G}$, we have that

$$\|f - f_n\|_{\mathcal{G}} = (2\pi)^{-1/2} \Big\| \frac{\hat{f}}{\sqrt{\widehat{\Psi}_0}} - \frac{\hat{f}_n}{\sqrt{\widehat{\Psi}_0}} \Big\|_{L^2} = (2\pi)^{-1/2} \Big\| \hat{g} - \frac{\hat{f}_n}{\sqrt{\widehat{\Psi}_0}} \Big\|_{L^2} \to 0$$

for $n \to 0$. Thus $\mathcal{G}$ is complete and consequently a Hilbert space.

It remains to show that $\Psi(\cdot,\cdot) := \Psi_0(\cdot - \cdot)$ is the reproducing kernel for $\mathcal{G}$. First of all, $\Psi_0 \in L^2(\mathbb{R}^2)$ by the decay rate property (2.5) with $s > 1$ and so

$$\|\Psi_0\|_{\mathcal{G}} = (2\pi)^{-1} \|\widehat{\Psi}_0^2/\widehat{\Psi}_0\|_{L^2(\mathbb{R}^2)} = (2\pi)^{-1} \|\widehat{\Psi}_0\|_{L^2(\mathbb{R}^2)} < \infty$$

giving $\Psi_0(\cdot - \mathbf{y}) \in \mathcal{G}$ for any $\mathbf{y} \in \mathbb{R}^2$. The reproduction property follows from using the translation property of the Fourier transform, the definition of the inner product on $\mathcal{G}$, and the fact that f can be recovered pointwise by its inverse Fourier transform.

$$\begin{aligned}(f,\Psi(\cdot,\mathbf{y}))_{\mathcal{G}} := (f,\Psi_0(\cdot-\mathbf{y}))_{\mathcal{G}} &= (2\pi)^{-1}\int_{\mathbb{R}^2}\frac{\hat{f}(\boldsymbol{\omega})\overline{\widehat{\Psi}_0(\boldsymbol{\omega})e^{-i\boldsymbol{\omega}\cdot\mathbf{y}}}}{\widehat{\Psi}_0(\boldsymbol{\omega})}d\boldsymbol{\omega} \\ &= (2\pi)^{-1}\int_{\mathbb{R}^2}\hat{f}(\boldsymbol{\omega})e^{i\boldsymbol{\omega}\cdot\mathbf{y}}d\boldsymbol{\omega} \\ &= f(\mathbf{y}).\end{aligned} \tag{2.16}$$

Thus Ψ is the reproducing kernel for the Hilbert space $\mathcal{G}$. □

In light of this result and Theorem 2.2.6, we know that for this particular kernel Ψ, the native space $\mathcal{N}_\Psi$ is the same as $\mathcal{G}$, and the inner products are equal. We now see that native spaces can be thought of as generalizations of Sobolev spaces. To see this, recall that for $s > 1$ the Sobolev space of order s is defined by

$$H^s(\mathbb{R}^2) = \{f \in L^2(\mathbb{R}^2)\cap C(\mathbb{R}^2) \,:\, \hat{f}(\boldsymbol{\omega})(1+\|\boldsymbol{\omega}\|_2^2)^{s/2} \in L^2(\mathbb{R}^2)\}.$$

It should now be clear that if Ψ_0 satisfies the decay rate (2.5), then its native space is equivalent to the Sobolev space $H^s(\mathbb{R}^2)$ and the norms are equivalent. This means, that we can find finite positive constants c_1, c_2 such that for any function $f \in \mathcal{N}_\Psi(\mathbb{R}^2)\cap H^s(\mathbb{R}^2)$, we have

$$c_1\cdot\|f\|_{H^s} \le \|f\|_{\mathcal{N}_\Psi} \le c_2\cdot\|f\|_{H^s}. \tag{2.17}$$

Corollary 2.2.1. *(Wendland [66], Chapter 10) Suppose that $\Psi_0 \in C(\mathbb{R}^2)\cap L^1(\mathbb{R}^2)$ satisfies the decay rate (2.5) with $s > 1$. Then the native space $\mathcal{N}_\Psi$ for Ψ coincides with the Sobolev space $H^s(\mathbb{R}^2)$ and the native space norm and Sobolev norms are equivalent.*

This Theorem and Corollary clearly demonstrate the necessary conditions for a native space on $\mathbb{R}^2$ to coincide with a Sobolev space on $\mathbb{R}^2$. We will use this result often throughout the thesis.

One last characterization of native spaces that we will need relates the space of continuous functions $C(\Omega)$ and $\mathcal{N}_\Psi(\Omega)$. For any set $\Omega \subset \mathbb{R}^2$, define the space of functionals

$$L(\Omega) := \text{span}\{\delta_{\mathbf{x}} \ : \ \mathbf{x} \in \Omega\} \tag{2.18}$$

and equip it with the inner product

$$(\lambda, \mu)_\Psi := \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \beta_j \Psi(\mathbf{x}_i, \mathbf{y}_j)$$

and norm $\|\lambda\|_\Psi^2 := \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \Psi(\mathbf{x}_i, \mathbf{x}_j)$ where $\lambda = \sum_{i=1}^N \alpha_i \delta_{\mathbf{x}_i} \in L(\Omega)$ and $\mu = \sum_{j=1}^M \beta_j \delta_{\mathbf{y}_j} \in L(\Omega)$ for some vectors $\boldsymbol{\alpha} \in \mathbb{R}^N$, $\boldsymbol{\beta} \in \mathbb{R}^M$, $N, M \in \mathbb{N}$, and distinct points $\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{y}_1, \ldots, \mathbf{y}_M \in \Omega$. By these definitions, there is an obvious one to one relationship between $L(\Omega)$ and $F_\Psi(\Omega)$ given by

$$L(\Omega) \to F_\Psi(\Omega), \quad \lambda \to \lambda_{\mathbf{x}} \Psi(\cdot, \mathbf{x})$$

where $\lambda_{\mathbf{x}}$ denotes λ operating on the $\mathbf{x}$, or second, argument of $\Psi(\cdot, \mathbf{x})$. Furthermore, the norms on $L(\Omega)$ and $F_\Psi(\Omega)$ are the same. This shows in particular that $\lambda(f) = (f, \lambda_{\mathbf{x}} \Psi(\cdot, \mathbf{x}))_\Psi$ for any $\lambda \in L(\Omega)$ and all $f \in F_\Psi(\Omega)$ since

$$\lambda(f) = \sum_{i=1}^N \alpha_i f(\mathbf{x}_i) = \sum_{i=1}^N \alpha_i (f, \Psi(\cdot, \mathbf{x}_i))_\Psi = (f, \sum_{i=1}^N \alpha_i \Psi(\cdot, \mathbf{x}_i))_\Psi = (f, \lambda_{\mathbf{x}} \Psi(\cdot, \mathbf{x}))_\Psi. \tag{2.19}$$

We now use $L(\Omega)$ to characterize the space of functions on which all functionals from $L(\Omega)$ are continuous.

Theorem 2.2.8. *Suppose that $\Psi \in C(\Omega \times \Omega)$ is an SPD kernel. Define the space*

$$\mathcal{G} = \{f \in C(\Omega) : |\lambda(f)| \leq c_f \|\lambda\|_\Psi, \ \forall \lambda \in L(\Omega)\}.$$

Then $\mathcal{G} = \mathcal{N}_\Psi(\Omega)$ *and*

$$\|f\|_{\mathcal{N}_\Psi(\Omega)} = \sup_{0\neq\lambda\in L(\Omega)} \frac{|\lambda(f)|}{\|\lambda\|_\Psi}. \tag{2.20}$$

Proof. Suppose that $f \in \mathcal{N}_\Psi(\Omega)$. To show $f \in \mathcal{G}$, we first note that $f \in C(\Omega)$ by definition of the native space $\mathcal{N}_\Psi(\Omega)$ (see (2.13)). Choose any $\lambda \in L(\Omega)$ (thus $\lambda = \sum_{i=1}^N \alpha_i\delta_{\mathbf{x}_i}$ for some distinct points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \Omega$ and coefficients $\boldsymbol{\alpha} \in \mathbb{R}^N$.) Then by the reproduction formula $f(\mathbf{x}) = (f, \Psi(\cdot, \mathbf{x}))_{\mathcal{N}_\Psi(\Omega)}$ and Schwartz's inequality, we easily see that

$$|\lambda(f)| = |(f, \lambda_{\mathbf{x}}\Psi(\cdot,\mathbf{x}))_{\mathcal{N}_\Psi(\Omega)}| \leq \|f\|_{\mathcal{N}_\Psi(\Omega)}\|\sum_{i=1}^N \alpha_i\Psi(\cdot,\mathbf{x}_i)\|_{\mathcal{N}_\Psi(\Omega)} = c_f\|\lambda\|_\Psi \tag{2.21}$$

and thus $f \in \mathcal{G}$. This also establishes that

$$\sup_{0\neq\lambda\in L(\Omega)} \frac{|\lambda(f)|}{\|\lambda\|_\Psi} \leq \|f\|_{\mathcal{N}_\Psi(\Omega)}.$$

Now we show $f \in \mathcal{G}$ implies that $f \in \mathcal{N}_\Psi(\Omega)$. Let $f \in \mathcal{G}$. For the given f, we define a linear functional

$$F_f : F_\Psi(\Omega) \mapsto \mathbb{R}, \quad \lambda_{\mathbf{x}}\Psi(\cdot,\mathbf{x}) \mapsto \lambda(f)$$

which is continuous by definition of $\mathcal{G}$. This means that since $F_\Psi(\Omega) \subseteq \mathcal{F}_\Psi(\Omega)$, by the Hahn-Banach Theorem, the functional F_f has a continuous extension to $\mathcal{F}_\Psi(\Omega)$. By the Riesz representation theorem, we can find an element $Sf \in \mathcal{F}_\Psi(\Omega)$ such that $F_f(g) = (g, Sf)_\Psi$ for all $g \in \mathcal{F}_\Psi(\Omega)$. To show that $f \in \mathcal{N}_\Psi(\Omega)$ we need to show that $f = R(Sf)$ where R was defined in (2.13). For any $\lambda \in L(\Omega)$, we have

$$\begin{aligned}\lambda(f - R(Sf)) &= \lambda(f) - (Sf, \lambda_{\mathbf{x}}\Psi(\cdot,\mathbf{x}))_\Psi \\ &= \lambda(f) - F_f(\lambda_{\mathbf{x}}\Psi(\cdot,\mathbf{x})) = \lambda(f) - \lambda(f) = 0.\end{aligned} \tag{2.22}$$

In particular, this shows that $f(\mathbf{x}) = R(Sf)(\mathbf{x})$ for any $\mathbf{x} \in \Omega$ (simply take $\lambda = \delta_{\mathbf{x}}$). Thus f is in the native space $\mathcal{N}_\Psi(\Omega)$.

Lastly, since $Sf \in \mathcal{F}_\Psi(\Omega)$ and $\mathcal{F}_\Psi(\Omega)$ is the completion of $F_\Psi(\Omega)$, we can choose a sequence $\lambda_j \in L(\Omega)$ such that $\lambda_{j,\mathbf{x}}\Psi(\cdot,\mathbf{x}) \to Sf \in \mathcal{F}_\Psi(\Omega)$ for $j \to \infty$. Hence $\lambda_j(f) = (Sf, \lambda_{j,\mathbf{x}}\Psi(\cdot,\mathbf{x}))_\Psi \to \|Sf\|_\Psi^2$ and $\|\lambda_j\|_\Psi \to \|Sf\|_\Psi$ for $j \to \infty$. Now we have the bound

$$\sup_{0 \neq \lambda \in L(\Omega)} \frac{|\lambda(f)|}{\|\lambda\|_\Psi} \geq \lim_{j \to \infty} \frac{|\lambda_j(f)|}{\|\lambda_j\|_\Psi} = \frac{\|Sf\|_\Psi^2}{\|Sf\|_\Psi} = \|f\|_{\mathcal{N}_\Psi(\Omega)}, \tag{2.23}$$

where the last equality comes from the fact that $f = R(Sf)$. Using this with $\|f\|_{\mathcal{N}_\Psi(\Omega)} \geq \sup_{0\neq\lambda\in L(\Omega)} \frac{|\lambda(f)|}{\|\lambda\|_\Psi}$ already established, we can conclude that $\|f\|_{\mathcal{N}_\Psi(\Omega)} = \sup_{0\neq\lambda\in L(\Omega)} \frac{|\lambda(f)|}{\|\lambda\|_\Psi}$. □

2.2.5 Restriction and Extension

After having studied native spaces on $\mathbb{R}^2$ and demonstrated their equivalence with Sobolev space in the norm $\|\cdot\|_{\mathcal{N}_\Psi(\mathbb{R}^2)}$, it is natural to investigate the extension and restriction of functions from native spaces. Namely, we want to know if a function $f \in \mathcal{N}_\Psi(\Omega)$ can be naturally extended to a function in $\mathcal{N}_\Psi(\mathbb{R}^2)$. We also want to investigate the restriction $f|_\Omega$ of functions from $\mathcal{N}_\Psi(\mathbb{R}^2)$ and determine if they belong to $\mathcal{N}_\Psi(\Omega)$.

Consider two open sets, Ω_1 and Ω_2, that satisfy $\Omega_1 \subseteq \Omega_2 \subseteq \mathbb{R}^2$ and a set of M distinct points $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ that lies in Ω_2. We denote the restriction of the points $\mathcal{X}$ to Ω_1 as $\mathcal{X}|_{\Omega_1} := \mathcal{X} \cap \Omega_1$. We will label these points as $\mathcal{X}|_{\Omega_1} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Furthermore, let Ψ be a continuous symmetric positive definite kernel defined on

$\mathbb{R}^2 \times \mathbb{R}^2$ and consider the native spaces $\mathcal{N}_\Psi(\Omega_1)$ and $\mathcal{N}_\Psi(\Omega_2)$. We have the following extension Theorem for native spaces.

Theorem 2.2.9. *(Wendland [66], Chapter 10) Any function $f \in \mathcal{N}_\Psi(\Omega_1)$ has a natural extension to a function $Ef \in \mathcal{N}_\Psi(\Omega_2)$ such that $\|Ef\|_{\mathcal{N}_\Psi(\Omega_2)} = \|f\|_{\mathcal{N}_\Psi(\Omega_1)}$.*

Proof. Since $\Omega_1 \subseteq \Omega_2$, we first define a natural extension $e : F_\Psi(\Omega_1) \mapsto F_\Psi(\Omega_2)$ simply by evaluation of a function $f \in F_\Psi(\Omega_1)$ at points in Ω_2. This means that for any $f \in F_\Psi(\Omega_1)$ written as $f(\mathbf{x}) = \sum_{j=1}^N \alpha_j \Psi(\mathbf{x}, \mathbf{x}_j)$ for some $\mathcal{X}|_{\Omega_1} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega_1$ and all $\mathbf{x} \in \Omega_1$, we can define the natural extension as $ef(\mathbf{x}) = \sum_{j=1}^N \alpha_j \Psi(\mathbf{x}, \mathbf{x}_j)$ for any $\mathbf{x} \in \Omega_2$ which is well defined since Ψ is continuous on $\Omega_2 \times \Omega_2$.

Since the norm $\|\cdot\|_{\Psi,\Omega_1}$ of any $f \in F_\Psi(\Omega_1)$ depends only on the points $\mathcal{X}|_{\Omega_1}$ and the coefficients α_j, we have $\|f\|_{\Psi,\Omega_1} = \|ef\|_{\Psi,\Omega_2}$. Hence e is an isometric embedding that has a continuous extension $e : \mathcal{F}_\Psi(\Omega_1) \mapsto \mathcal{F}_\Psi(\Omega_2)$. The extension operator $E : \mathcal{N}_\Psi(\Omega_1) \mapsto \mathcal{N}_\Psi(\Omega_2)$ can then be constructed using the operator R defined in (2.13) as follows. Every $f \in \mathcal{N}_\Psi(\Omega_1)$ has the representation $f(\mathbf{x}) = R_{\Omega_1}(\hat{f})(\mathbf{x})$ with $\hat{f} \in \mathcal{F}_\Psi(\Omega_1)$. For this f and any $\mathbf{x} \in \Omega_2$, we define

$$Ef(\mathbf{x}) = R_{\Omega_2}(e\hat{f})(\mathbf{x}).$$

Thus, for $\mathbf{x} \in \Omega_1$, we have

$$\begin{aligned} R_{\Omega_2}(e\hat{f})(\mathbf{x}) &= (e\hat{f}, \Psi(\cdot, \mathbf{x}))_{\Psi,\Omega_2} = (e\hat{f}, e\Psi|_{\Omega_1}(\cdot, \mathbf{x}))_{\Psi,\Omega_2} \\ &= (\hat{f}, \Psi|_{\Omega_1}(\cdot, \mathbf{x}))_{\Psi,\Omega_1} \end{aligned} \tag{2.24}$$

which shows that $Ef(\mathbf{x}) = f(\mathbf{x})$ for $f \in \mathcal{N}_\Psi(\Omega_1)$ and $\mathbf{x} \in \Omega_1$. Finally, for two functions $f, g \in \mathcal{N}_\Psi(\Omega_1)$, the identities

$$(Ef, Eg)_{\mathcal{N}_\Psi(\Omega_2)} = (e\hat{f}, e\hat{g})_{\Psi,\Omega_2} = (\hat{f}, \hat{g})_{\Psi,\Omega_1} = (f, g)_{\mathcal{N}_\Psi(\Omega_1)}$$

show that E is isometric and that $\|Ef\|_{\mathcal{N}_\Psi(\Omega_2)} = \|f\|_{\mathcal{N}_\Psi(\Omega_1)}$. □

We now show that the restriction of a function $f \in \mathcal{N}_\Psi(\Omega_2)$ to Ω_1 is in $\mathcal{N}_\Psi(\Omega_1)$. Firstly, for any pair of domains Ω_1 and Ω_2 such that $\Omega_1 \subseteq \Omega_2 \subseteq \mathbb{R}^2$, we obviously have that $L(\Omega_1) \subseteq L(\Omega_2)$ where the space L was defined in (2.18).

Now consider any $f \in \mathcal{N}_\Psi(\Omega_2)$ and denote the restriction of f to Ω_1 by $f|_{\Omega_1}$. To show that $f|_{\Omega_1} \in \mathcal{N}_\Psi(\Omega_1)$, we utilize Theorem 2.2.8 to show that if there exists a constant c_f such that $|\lambda(f|_{\Omega_1})| \leq c_f \|\lambda\|_{\Psi,\Omega_1}$ for all $\lambda \in L(\Omega_1)$, then $f|_{\Omega_1} \in \mathcal{N}_\Psi(\Omega_1)$. Indeed, choose any $\sum_{i=1}^N \alpha_i \delta_{\mathbf{x}_i} = \lambda \in L(\Omega_1) \subseteq L(\Omega_2)$, then since $f \in \mathcal{N}_\Psi(\Omega_2)$, there exists a constant c_f such that

$$|\lambda(f|_{\Omega_1})| = |\sum_{i=1}^N \alpha_i f(\mathbf{x}_i)| = |(f, \sum_{i=1}^N \alpha_i \Psi(\cdot, \mathbf{x}_i))_{\mathcal{N}_\Psi(\Omega_2)}| \leq \|f\|_{\mathcal{N}_\Psi(\Omega_2)} \|\lambda\|_{\Psi,\Omega_2} = c_f \|\lambda\|_{\Psi,\Omega_1},$$

where the last equality comes from the fact that $\|\lambda\|_{\Psi,\Omega_1} = \|\lambda\|_{\Psi,\Omega_2}$ since the the norms only depend on the points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \Omega_1$ and the coefficients $\boldsymbol{\alpha} \in \mathbb{R}^N$. Thus $f|_{\Omega_1} \in \mathcal{N}_\Psi(\Omega_1)$. Finally, we have again using Theorem 2.2.8,

$$\begin{aligned} \|f|_{\Omega_1}\|_{\mathcal{N}_\Psi(\Omega_1)} &= \sup_{0 \neq \lambda \in L(\Omega_1)} \frac{|\lambda(f)|}{\|\lambda\|_{\Psi,\Omega_1}} = \sup_{0 \neq \lambda \in L(\Omega_1)} \frac{|\lambda(f)|}{\|\lambda\|_{\Psi,\Omega_2}} \\ &\leq \sup_{0 \neq \lambda \in L(\Omega_2)} \frac{|\lambda(f)|}{\|\lambda\|_{\Psi,\Omega_2}} = \|f\|_{\mathcal{N}_\Psi(\Omega_2)}. \end{aligned} \tag{2.25}$$

This finishes the proof of the following theorem.

Theorem 2.2.10. *The restriction $f|_{\Omega_1}$ of a function $f \in \mathcal{N}_\Psi(\Omega_2)$ is contained in $\mathcal{N}_\Psi(\Omega_1)$, with a norm that is less than or equal to the norm of f.*

With the extension and restriction of functions from native spaces well defined now, we can couple these results with Sobolev spaces. As shown in the previous section, we already know that for a positive definite radial function $\Psi_0 \in C(\mathbb{R}^2) \cap$

$L^1(\mathbb{R}^2)$ that satisfies the Fourier decay rate (2.5) of order $s > 1$, then the native space $\mathcal{N}_\Psi(\mathbb{R}^2)$ coincides with the Sobolev space $H^s(\mathbb{R}^2)$ and they have equivalent norms. Now let $\Omega \subset \mathbb{R}^2$ be a domain with Lipschitz boundary $\partial\Omega$. Recall that the Sobolev space $H^k(\Omega)$, $k \in \mathbb{N}$, for a measurable set Ω is defined as the set of all functions $f \in L^2(\Omega)$ such that their weak derivatives of order $|\alpha| = \alpha_1 + \alpha_2 \leq k$ are in $L^2(\Omega)$. The norm on $H^k(\Omega)$ is defined by $\|f\|^2_{H^k(\Omega)} = \sum_{|\alpha|\leq k} \|D^\alpha f\|^2_{L^2(\Omega)}$. Then we can establish the following result.

Corollary 2.2.2. *(Wendland [62], Chapter 10) Suppose that $\Psi_0 \in C(\mathbb{R}^2) \cap L^1(\mathbb{R}^2)$ is a positive definite function that has a Fourier transform that satisfies (2.5) with $s > 1$, $s \in \mathbb{N}$. Suppose further that $\Omega \subset \mathbb{R}^2$ has a Lipschitz boundary. Then $\mathcal{N}_\Psi(\Omega) = H^s(\Omega)$ with equivalent norms.*

Proof. Any $f \in \mathcal{N}_\Psi(\Omega)$ has an extension $Ef \in \mathcal{N}_\Psi(\mathbb{R}^2) = H^s(\mathbb{R}^2)$. The restriction of Ef to Ω, denoted by $Ef|_\Omega$, satisfies $Ef|_\Omega = f \in H^s(\Omega)$ and so $\|f\|_{H^s(\Omega)} \leq \|Ef\|_{H^s(\mathbb{R}^2)} \leq c_1\|Ef\|_{\mathcal{N}_\Psi(\mathbb{R}^2)} = c_1\|f\|_{\mathcal{N}_\Psi(\Omega)}$ for some constant $c_1 > 0$.

Now for $\Omega \subset \mathbb{R}^2$ with Lipschitz boundary $\partial\Omega$, we use the well known result that for a function $f \in H^s(\Omega)$, there exists an extension $\tilde{E}f \in H^s(\mathbb{R}^2) = \mathcal{N}_\Psi(\mathbb{R}^2)$ satisfying $\|\tilde{E}f\|_{H^s(\mathbb{R}^2)} \leq c_2\|f\|_{H^s(\Omega)}$ (cf. Grisvard [30]) for some constant $C_2 > 0$ depending on the domain Ω. Since $\tilde{E}f \in \mathcal{N}_\Psi(\mathbb{R}^2)$, the Native space restriction satisfies $\tilde{E}f|_\Omega \in \mathcal{N}_\Psi(\Omega)$ and $f = \tilde{E}f|_\Omega$ giving

$$\|f\|_{\mathcal{N}_\Psi(\Omega)} \leq \|\tilde{E}f\|_{\mathcal{N}_\Psi(\mathbb{R}^2)} \leq c_2\|\tilde{E}f\|_{H^s(\mathbb{R}^2)} \leq c_2\|f\|_{H^s(\Omega)}.$$

□

We now turn to the problem of approximation in native spaces.

2.2.6 Approximation in Native Spaces

In this section, properties of positive definite kernels $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ and their associated native spaces $\mathcal{N}_\Psi(\Omega)$ that we have developed in the previous subsections are now put to use in the problem of approximating scattered data. Scattered data approximation, as we will see, is the prerequisite to the meshless collocation method introduced later in this chapter. We note again, that this theory will work in $\mathbb{R}^d$ for any integer d, however we restrict ourselves to $\mathbb{R}^2$ since our numerical experiments will take place in $\mathbb{R}^2$

For a given SPD kernel $\Psi \in C^{2k}(\Omega \times \Omega)$, let $\Lambda_N = \{\lambda_1, \lambda_2, \ldots, \lambda_N\} \in \mathcal{N}_\Psi(\Omega)'$ be a set of N linear bounded functionals on the native space $\mathcal{N}_\Psi(\Omega)$. Using the notation $\lambda_{j,2}\Psi(\cdot, \mathbf{x})$ to mean the j-th functional λ_j acting on the second argument of the kernel Ψ, we define the finite dimensional spaces

$$\begin{aligned} \mathcal{N}'_{\Psi,N} &= \operatorname{span}\{\lambda_j \in \Lambda_N,\ 1 \le j \le N\} \\ \mathcal{N}_{\Psi,N} &= \operatorname{span}\{\lambda_{j,2}\,\Psi(\cdot, \mathbf{x}),\ \ 1 \le j \le N\}. \end{aligned} \tag{2.26}$$

In the scattered data interpolation problem, we take the functionals $\lambda_j \in \Lambda_N$ to be the point evaluation functionals $\lambda_j := \delta_{\mathbf{x}_j},\ \mathbf{x}_j \in \mathcal{X}$. The scattered data interpolation problem can be formulated as follows. For a given set of nodes $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega$ and data $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)$ from a continuous function $f \in C(\Omega)$, we find an element $I_\mathcal{X} f \in \mathcal{N}_{\Psi,N} := \operatorname{span}\{\Psi(\cdot, \mathbf{x}_1), \ldots, \Psi(\cdot, \mathbf{x}_N)\}$ such that

$$I_\mathcal{X} f(\mathbf{x}_i) = f(\mathbf{x}_i), \quad \forall \mathbf{x}_i \in \mathcal{X}. \tag{2.27}$$

This amounts to finding a vector of coefficients $\boldsymbol{\alpha} \in \mathbb{R}^N$ such that $I_\mathcal{X} f(\mathbf{x}_i) = \sum_{j=1}^N \alpha_j \Psi(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i)$ for all $\mathbf{x}_i \in \mathcal{X}$. If we define the vector $\mathbf{f} := (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))^T$, then the interpolation problem can be written in matrix-vector notation as: find $\boldsymbol{\alpha} \in \mathbb{R}^N$ such that

$$\mathcal{A}_{\mathcal{X}}\boldsymbol{\alpha} = \mathbf{f} \tag{2.28}$$

where $\mathcal{A}_{\mathcal{X}}[i,j] = \Psi(\mathbf{x}_i, \mathbf{x}_j)$. Clearly, since Ψ is SPD and since the points in $\mathcal{X}$ are distinct, the matrix $\mathcal{A}_{\mathcal{X}}$ is symmetric and positive definite (recall Theorem 2.2.3). We thus have unique solvability for the vector $\boldsymbol{\alpha} \in \mathbb{R}^N$. We can equally express this interpolation problem using the functionals $\lambda_j := \delta_{\mathbf{x}_j} \in \mathcal{N}'_{\Psi,N}$. Find $I_{\mathcal{X}}f \in \mathcal{N}_{\Psi,N}$ such that

$$\lambda_j(I_{\mathcal{X}}f) = \lambda_j(f), \;\; 1 \leq j \leq N. \tag{2.29}$$

We now derive an orthogonality property of the interpolant $I_{\mathcal{X}}f$ in the native space. If the scattered data $(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))$ is in fact sampled from some function $f \in \mathcal{N}_\Psi$ on the set of points $\mathcal{X}$, we can show that $f - I_{\mathcal{X}}f \in \mathcal{N}_\Psi(\Omega)$ and $I_{\mathcal{X}}f \in \mathcal{N}_{\Psi,N}$ are mutually orthogonal with respect to the native space inner product.

Lemma 2.2.2. *Suppose that $f \in \mathcal{N}_\Psi(\Omega)$ and $I_{\mathcal{X}}f \in \mathcal{N}_{\Psi,N}$ is the unique interpolant that satisfies $I_{\mathcal{X}}f(\mathbf{x}_i) = f(\mathbf{x}_i)$ for all $\mathbf{x}_i \in \mathcal{X}$. Then we have*

$$(f - I_{\mathcal{X}}f, s)_{\mathcal{N}_\Psi(\Omega)} = 0$$

for all $s \in \mathcal{N}_{\Psi,N}$. In particular, this gives

$$(f - I_{\mathcal{X}}f, I_{\mathcal{X}}f)_{\mathcal{N}_\Psi(\Omega)} = 0.$$

Proof. Any $s \in \mathcal{N}_{\Psi,N}$ can be written in the form $s = \sum_{j=1}^N \alpha_j \Psi(\cdot, \mathbf{x}_j)$ for some vector of coefficients $\boldsymbol{\alpha} \in \mathbb{R}^N$. The result follows easily using the reproduction property of the kernel Ψ and that $I_{\mathcal{X}}f(\mathbf{x}_j) = f(\mathbf{x}_j)$ for all $\mathbf{x}_j \in \mathcal{X}$ by the interpolation property of $I_{\mathcal{X}}$, giving

$$\begin{aligned}(f - I_{\mathcal{X}}f, s)_{\mathcal{N}_\Psi(\Omega)} &= (f - I_{\mathcal{X}}f, \sum_{j=1}^{N} \alpha_j \Psi(\cdot, \mathbf{x}_j))_{\mathcal{N}_\Psi(\Omega)} \\ &= \sum_{j=1}^{N} \alpha_j f(\mathbf{x}_j) - \sum_{j=1}^{N} \alpha_j I_{\mathcal{X}}f(\mathbf{x}_j) = \sum_{j=1}^{N} \alpha_j (f(\mathbf{x}_j) - I_{\mathcal{X}}f(\mathbf{x}_j)) = 0.\end{aligned} \tag{2.30}$$

Since $I_{\mathcal{X}}f \in \mathcal{N}_{\Psi,N}$, in particular we have

$$(f - I_{\mathcal{X}}f, I_{\mathcal{X}}f)_{\mathcal{N}_\Psi(\Omega)} = 0. \tag{2.31}$$

□

We can now apply this Lemma to show that $I_{\mathcal{X}}f$ can also be characterized as the orthogonal projection of $f \in \mathcal{N}_\Psi(\Omega)$ onto $\mathcal{N}_{\Psi,N}$.

Theorem 2.2.11. *Suppose $\Psi \in C^{2k}(\Omega \times \Omega)$ is SPD and that $f \in \mathcal{N}_\Psi(\Omega)$ is known only at $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. The the interpolant $I_{\mathcal{X}}f$ is the best approximation to f from $\mathcal{N}_{\Psi,N}$ with respect to the native space norm*

$$\|I_{\mathcal{X}}f - f\|_{\mathcal{N}_\Psi(\Omega)} \leq \|f - s\|_{\mathcal{N}_\Psi(\Omega)}$$

for all $s \in \mathcal{N}_{\Psi,N}$. Hence $I_{\mathcal{X}}f$ is the orthogonal projection of f onto $\mathcal{N}_{\Psi,N}$.

Proof. By the previous Lemma, we know that $(f - I_{\mathcal{X}}f, s)_{\mathcal{N}_\Psi(\Omega)} = 0$ for all $s \in \mathcal{N}_{\Psi,N}$. But this is the characterization of the best approximation in a Hilbert space. □

By the orthogonality of $f - I_{\mathcal{X}}f$ and $I_{\mathcal{X}}f$ we also have the following consequence using the Pythagorean theorem. This result will be used in error bounds for interpolation with positive definite kernels as well as in the analysis of solving elliptic PDEs with such kernels.

Corollary 2.2.3. *We have the estimates* $\|I_{\mathcal{X}}f\|_{\mathcal{N}_\Psi(\Omega)} \le \|f\|_{\mathcal{N}_\Psi(\Omega)}$ *and* $\|f - I_{\mathcal{X}}f\|_{\mathcal{N}_\Psi(\Omega)} \le \|f\|_{\mathcal{N}_\Psi(\Omega)}$.

2.2.7 Lagrangian interpolant form

In order to derive the error estimate, it will be helpful to rewrite the interpolant $I_{\mathcal{X}}f$ in the so-called *Lagrangian interpolant* form. This is done as follows. For $j = 1, \ldots, N$, define the vector $\boldsymbol{\alpha}^{(j)} \in \mathbb{R}^N$ that satisfies

$$\mathcal{A}\boldsymbol{\alpha}^{(j)} = \mathbf{e}^{(j)}, \tag{2.32}$$

where $\mathbf{e}^{(j)} \in \mathbb{R}^N$ is the j-th unit vector in $\mathbb{R}^N$ and $\mathcal{A} := \mathcal{A}_{\mathcal{X}}$ is the interpolation matrix defined earlier. The system is uniquely solvable for each $j = 1, \ldots, N$ due to positive definiteness of $\mathcal{A}$. The resulting functions defined by

$$\tilde{v}_j(\cdot) := \sum_{i=1}^{N} \alpha_i^{(j)} \Psi(\cdot, \mathbf{x}_i) \tag{2.33}$$

satisfy $\tilde{v}_j(\mathbf{x}_k) = \delta_{k,j}$ where $\delta_{k,j} = 1$ if $k = j$ and 0 otherwise. The $\tilde{v}_j$ functions will be called the *Lagrangian interpolant* functions and clearly belong to the space $\mathcal{N}_{\Psi,N}$. Due to the property that $\tilde{v}_j(\mathbf{x}_k) = \delta_{k,j}$, any function $f \in \mathcal{N}_\Psi(\Omega)$ has the unique interpolant

$$I_{\mathcal{X}}f(\mathbf{x}) = \sum_{j=1}^{N} f(\mathbf{x}_j)\tilde{v}_j(\mathbf{x}). \tag{2.34}$$

and clearly $I_{\mathcal{X}}f(\mathbf{x}_i) = f(\mathbf{x}_i)$ for any $\mathbf{x}_i \in \mathcal{X}$.

Later in the chapter, it will be useful to write the Lagrangian interpolant functions evaluated at any point $\mathbf{x} \in \Omega$ in vector form. For any given $\mathbf{x} \in \Omega$, if we define the vector $\mathbf{R}(\mathbf{x}) := \mathbf{R}_{\mathcal{X}}(\mathbf{x}) = (\Psi(\mathbf{x}, \mathbf{x}_1), \ldots, \Psi(\mathbf{x}, \mathbf{x}_N))^T$, then we can write $\tilde{\mathbf{v}}(\mathbf{x}) = (\tilde{v}_1(\mathbf{x}), \ldots, \tilde{v}_N(\mathbf{x}))^T$ as

$$\tilde{\mathbf{v}}(\mathbf{x}) = \mathcal{A}^{-1}\mathbf{R}(\mathbf{x}). \tag{2.35}$$

Thus the solution vector $\tilde{\mathbf{v}}(\mathbf{x})$ satisfies $\Psi(\mathbf{x}, \mathbf{x}_j) = \sum_{i=1}^{N} \tilde{v}_i(\mathbf{x})\Psi(\mathbf{x}_i, \mathbf{x}_j)$ for any $\mathbf{x}_j \in \mathcal{X}$.

Now since $\tilde{v}_j \in C^k(\Omega)$ for each $j = 1, \ldots, N$ due to the fact that $\Psi \in C^{2k}(\Omega \times \Omega)$, we can apply the differentiation operator D^α element wise to both sides of equation (2.35) giving

$$D^\alpha \tilde{\mathbf{v}}(\mathbf{x}) = \mathcal{A}^{-1} D^\alpha \mathbf{R}(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{2.36}$$

where $D^\alpha \mathbf{R}(\mathbf{x}) = (D_1^\alpha \Psi(\mathbf{x}, \mathbf{x}_1), \ldots, D_1^\alpha \Psi(\mathbf{x}, \mathbf{x}_N))^T$. We also we see that $D^\alpha I_{\mathcal{X}} f(\mathbf{x}) = \sum_{j=1}^{N} f(\mathbf{x}_j) D^\alpha \tilde{v}_j(\mathbf{x})$ for any $\mathbf{x} \in \Omega$. Our interest now is in finding out how close the approximate $D^\alpha I_{\mathcal{X}} f(\mathbf{x})$ is to $D^\alpha f$ for $f \in \mathcal{N}_\Psi(\Omega)$. This is the subject of the next section.

2.2.8 Error estimate for derivatives of scattered data interpolation

We now discuss the subject of bounding the interpolation error which has been an area of active research the past decade for scattered data interpolation. We are concerned with estimating the difference between derivatives of an unknown function f coming from the native Hilbert space $\mathcal{N}_\Psi(\Omega)$ and derivatives of its interpolant $I_{\mathcal{X}} f$

by bounding the difference in terms of the point saturation measure h for the set $\mathcal{X} \subseteq \Omega$ defined in (2.2). In particular, we derive an error estimate for positive definite kernel interpolation of the following form

$$\|D^\alpha(f - I_{\mathcal{X}} f)\|_{L^\infty(\Omega)} \leq Ch^{k-|\alpha|}\|f\|_{\mathcal{N}_\Psi(\Omega)}, \tag{2.37}$$

where the interpolant $I_{\mathcal{X}} f$ was defined in (2.27). Before deriving the error estimate, we first need a smoothness result for the native space $\mathcal{N}_\Psi(\Omega)$ where $\Psi \in C^{2k}(\Omega \times \Omega)$. Namely, we want to show that $\mathcal{N}_\Psi(\Omega) \subseteq C^k(\Omega)$. For this, we first need the following Lemma.

Lemma 2.2.3. *Suppose $\Psi \in C^{2k}(\Omega \times \Omega)$ is a symmetric positive definite kernel on a domain $\Omega \subset \mathbb{R}^2$. Then Ψ is k-times continuously differentiable with respect to the second argument and for any $\mathbf{x} \in \Omega$ and $|\alpha| := \alpha_1 + \alpha_2 \leq k$, the function $D_2^\alpha \Psi(\cdot, \mathbf{x})$ is in $\mathcal{N}_\Psi(\Omega)$.*

Proof. For any $f \in F_\Psi(\Omega)$, we define the functional $\lambda_n \in F_\Psi(\Omega)'$ by

$$\lambda_n(f) = \Delta_{\alpha,1/n} f(\mathbf{x}) := \Delta^1_{\alpha_1,1/n}\Delta^2_{\alpha_2,1/n} f(\mathbf{x})$$

where $\Delta^i_{\alpha_i,1/n}$ is the forward difference operator on the $i - th$ variable of order α_i given by

$$\Delta_{k,h} f(x) := \sum_{j=0}^{k} (-1)^{k-j} \left(\begin{array}{c} k \\ \\ j \end{array} \right) f(x + jh),$$

for $0 < h \leq 1$. This gives

$$\lim_{n\to\infty} (1/n)^{-|\alpha|} \Delta_{\alpha,1/n} g(\mathbf{x}) = D^\alpha g(\mathbf{x}).$$

for any $g \in C^k(\Omega)$. Now since $f \in F_\Psi(\Omega)$, we have the representation

$$\begin{aligned}\lambda_n(f) &= (f, \lambda_{n,\mathbf{x}}\Psi(\cdot,\mathbf{x}))_\Psi \\ &= (f, (1/n)^{-|\alpha|}\Delta_{\alpha,1/n,\mathbf{x}}\Psi(\cdot,\mathbf{x}))_\Psi = (f, \phi_n)_\Psi\end{aligned} \tag{2.38}$$

where we have defined $\phi_n := (1/n)^{-|\alpha|}\Delta_{\alpha,1/n,2}\Psi(\cdot,\mathbf{x})$. Now since $\Psi \in C^{2k}(\Omega \times \Omega)$, clearly we have that $\lim_{n\to\infty}\phi_n = \lim_{n\to\infty}(1/n)^{-|\alpha|}\Delta_{\alpha,1/n,2}\Psi(\cdot,\mathbf{x}) = D_2^\alpha\Psi(\cdot,\mathbf{x})$. Secondly, ϕ_n is a Cauchy sequence in $F_\Psi(\Omega)$. To see this, by the reproducing property of Ψ and the fact that Ψ is k-times differentiable in both variables, we compute

$$\begin{aligned}\lim_{n,m\to\infty}(\phi_n,\phi_m)_\Psi &= \lim_{n,m\to\infty}(1/n)^{-|\alpha|}\Delta_{\alpha,1/n,1}(1/m)^{-|\alpha|}\Delta_{\alpha,1/m,2}\Psi(\mathbf{x},\mathbf{x}) \\ &= D_1^\alpha D_2^\alpha\Psi(\mathbf{x},\mathbf{x}) =: c.\end{aligned} \tag{2.39}$$

Now we have

$$\lim_{n,m\to\infty}\|\phi_n - \phi_m\|_\Psi^2 = \lim_{n,m\to\infty}\left(\|\phi_n\|_\Psi^2 + \|\phi_m\|_\Psi^2 - 2(\phi_n,\phi_m)_\Psi\right) \to c + c - 2c = 0. \tag{2.40}$$

Since ϕ_n is Cauchy, we can find a $\phi \in \mathcal{F}_\Psi(\Omega)$ with $\|\phi - \phi_n\|_\Psi \to 0$ as $n \to \infty$. Using this ϕ and the definition of the mapping R defined in 2.13, we have

$$\begin{aligned}R(\phi(\mathbf{y})) = (\phi, \Psi(\cdot,\mathbf{y}))_\Psi &= \lim_{n\to\infty}(\phi_n, \Psi(\cdot,\mathbf{y}))_\Psi \\ &= \lim_{n\to\infty}\phi_n(\mathbf{y}) = D_2^\alpha\Psi(\mathbf{y},\mathbf{x}).\end{aligned} \tag{2.41}$$

This shows that $D_2^\alpha\Psi(\cdot,\mathbf{x})$ belongs to $\mathcal{N}_\Psi(\Omega)$. □

We now know that if $\Psi \in C^{2k}(\Omega\times\Omega)$, then $D_2^\alpha\Psi(\cdot,\mathbf{x})$ is in the native space $\mathcal{N}_\Psi(\Omega)$. Next, we apply this to functions, namely, if $f \in \mathcal{N}_\Psi(\Omega)$ then $f \in C^k(\Omega)$.

Theorem 2.2.12. *Suppose that $\Psi \in C^{2k}(\Omega\times\Omega)$ is an SPD kernel on a domain $\Omega \subset \mathbb{R}^2$. Then $\mathcal{N}_\Psi(\Omega) \subset C^k(\Omega)$ and for any $f \in \mathcal{N}_\Psi(\Omega)$, every $\alpha \in \mathbb{N}_0^2$ with $|\alpha| := \alpha_1 + \alpha_2 \le k$, and any $\mathbf{x} \in \Omega$, we have the representation*

$$D^\alpha f(\mathbf{x}) = (f, D_2^\alpha\Psi(\cdot,\mathbf{x}))_{\mathcal{N}_\Psi(\Omega)}. \tag{2.42}$$

Proof. We show by induction on $|\alpha|$. For $|\alpha| = 0$, we already know f is continuous and can be represented in the form (2.42). Now for $|\alpha| > 0$, we can assume that the representation (2.42) holds for some $k > |\beta| > 0$ where $\beta = (\alpha_1 - 1, \alpha_2)$. Denoting e_1 as $(1, 0) \in \mathbb{R}^2$, we have for any $\mathbf{x} \in \Omega$

$$\begin{aligned} D^\alpha f(\mathbf{x}) &= \lim_{n\to\infty} (1/n)^{-1} \big(D^\beta f(\mathbf{x} + e_1/n) - D^\beta f(\mathbf{x}) \big) \\ &= \lim_{n\to\infty} (f, (1/n)^{-1} \big(D_2^\beta \Psi(\cdot, \mathbf{x} + e_1/n) - D_2^\beta \Psi(\cdot, \mathbf{x}) \big))_{\mathcal{N}_\Psi(\Omega)} \qquad (2.43) \\ &= (f, D_2^\alpha \Psi(\cdot, \mathbf{x}))_{\mathcal{N}_\Psi(\Omega)} \end{aligned}$$

where we applied from the proof of Lemma 2.2.3 that the sequence $(1/n)^{-1} \big(D_2^\beta \Psi(\cdot, \mathbf{x} + e_1/n)$ is Cauchy in $\mathcal{N}_\Psi(\Omega)$ and converges to $D_2^\alpha \Psi(\cdot, \mathbf{x})$. Thus $D^\alpha f(\mathbf{x})$ exists. Now we must show that the function $D^\alpha f$ is indeed continuous. We have for any $f \in \mathcal{N}_\Psi(\Omega)$,

$$\begin{aligned} |D^\alpha (f(\mathbf{x}) - f(\mathbf{y}))| &= |(f, D_2^\alpha \big(\Psi(\cdot, \mathbf{x}) - \Psi(\cdot, \mathbf{y}) \big))_{\mathcal{N}_\Psi(\Omega)}| \\ &\leq \|f\|_{\mathcal{N}_\Psi(\Omega)} \|D_2^\alpha \Psi(\cdot, \mathbf{x}) - D_2^\alpha \Psi(\cdot, \mathbf{y})\|_{\mathcal{N}_\Psi}. \end{aligned} \qquad (2.44)$$

Calculating $\|D_2^\alpha \Psi(\cdot, \mathbf{x}) - D_2^\alpha \Psi(\cdot, \mathbf{y})\|_{\mathcal{N}_\Psi}$, we have

$$\begin{aligned} &\|D_2^\alpha \Psi(\cdot, \mathbf{x}) - D_2^\alpha \Psi(\cdot, \mathbf{y})\|^2_{\mathcal{N}_\Psi(\Omega)} \\ &= (D_2^\alpha \Psi(\cdot, \mathbf{x}), D_2^\alpha \Psi(\cdot, \mathbf{x}))_{\mathcal{N}_\Psi(\Omega)} - 2(D_2^\alpha \Psi(\cdot, \mathbf{x}), D_2^\alpha \Psi(\cdot, \mathbf{y}))_{\mathcal{N}_\Psi(\Omega)} + (D_2^\alpha \Psi(\cdot, \mathbf{y}), D_2^\alpha \Psi(\cdot, \mathbf{y}))_{\mathcal{N}_\Psi(\Omega)} \\ &= D_1^\alpha D_2^\alpha \Psi(\mathbf{x}, \mathbf{x}) + D_1^\alpha D_2^\alpha \Psi(\mathbf{y}, \mathbf{y}) - 2 D_1^\alpha D_2^\alpha \Psi(\mathbf{x}, \mathbf{y}). \end{aligned}$$

$$(2.45)$$

Since $\Psi \in C^{2k}(\Omega \times \Omega)$, as $\mathbf{x} \to \mathbf{y}$ we see that $\|D_2^\alpha \Psi(\cdot, \mathbf{x}) - D_2^\alpha \Psi(\cdot, \mathbf{y})\|_{\mathcal{N}_\Psi} \to 0$. Thus $D^\alpha f$ is continuous. □

The second step in obtaining error estimates for interpolation is to introduce the so-called *power function*, which is an important concept in kernel-based approximation.

Definition Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subseteq \Omega$ be pairwise distinct on a domain $\Omega \subset \mathbb{R}^2$ and $\Psi \in C^{2k}(\Omega \times \Omega)$ a symmetric positive definite kernel. Then for any $\mathbf{x} \in \Omega$ and $\alpha \in \mathbb{N}_0^2$ with $|\alpha| := \alpha_1 + \alpha_2 \leq k$ the *power function* with respect to $\mathcal{X}$, Ψ, and α is defined by

$$\left[P_{\Psi,\mathcal{X}}^{(\alpha)}(\mathbf{x})\right]^2 := D_1^\alpha D_2^\alpha \Psi(\mathbf{x},\mathbf{x}) - 2\sum_{i=1}^N D^\alpha \tilde{v}_i(\mathbf{x}) D_1^\alpha \Psi(\mathbf{x},\mathbf{x}_i) + \sum_{i,j=1}^N D^\alpha \tilde{v}_i(\mathbf{x}) D^\alpha \tilde{v}_j(\mathbf{x}) \Psi(\mathbf{x}_i,\mathbf{x}_j). \tag{2.46}$$

By using the native space norm, we can also write the power function as

$$\left[P_{\Psi,\mathcal{X}}^{(\alpha)}(\mathbf{x})\right]^2 = \|D_2^\alpha \Psi(\cdot,\mathbf{x}) - \sum_{i=1}^N D^\alpha \tilde{v}_i(\mathbf{x}) \Psi(\cdot,\mathbf{x}_i)\|_{\mathcal{N}_\Psi(\Omega)}^2. \tag{2.47}$$

In this form, we see that the power function measures how well the finite summation of derivatives of the Lagrangian interpolant functions, $D^\alpha \tilde{v}_i(\mathbf{x})$, approximate $D_2^\alpha \Psi(\cdot,\mathbf{x})$ in the native space. One would imagine that the approximation gets better as the number of points in $\mathcal{X}$ increases. We will see that this is indeed the case by showing that we can in fact bound the power function by a constant times $h^{k-|\alpha|}$.

It will be useful in the proceeding analysis to define the power function in terms of a quadratic form. Let v be any vector in $\mathbb{R}^N$. For fixed $\mathbf{x}, \mathcal{X}, \Psi$, and $\alpha \in \mathbb{N}_0^2$ with $|\alpha| \leq k$ we define the quadratic form $\mathcal{Q} : \mathbb{R}^N \mapsto \mathbb{R}$ as

$$\mathcal{Q}(v) = D_1^\alpha D_2^\alpha \Psi(\mathbf{x},\mathbf{x}) - 2\sum_{i=1}^N v_i D^\alpha \Psi(\mathbf{x},\mathbf{x}_i) + \sum_{i,j=1}^N v_i v_j \Psi(\mathbf{x}_i,\mathbf{x}_j), \quad v \in \mathbb{R}^N. \tag{2.48}$$

It is easy to see that $\mathcal{Q}(D^\alpha \tilde{v}(\mathbf{x})) = \left[P_{\Psi,\mathcal{X}}^{(\alpha)}(\mathbf{x})\right]^2$. In fact, we now show that the vector $D^\alpha \tilde{v}(\mathbf{x})$ for any fixed $\mathbf{x} \in \Omega$ *minimizes* the quadratic form, namely $\mathcal{Q}(D^\alpha \tilde{v}(\mathbf{x})) \leq \mathcal{Q}(v)$ for all $v \in \mathbb{R}^N$. We will use this result in obtaining the error estimate (2.37).

Theorem 2.2.13. *Suppose that $\Psi \in C^{2k}(\Omega \times \Omega)$. Then for any $\mathbf{x} \in \Omega$ and $\alpha \in \mathbb{N}_0^2$ with $|\alpha| \leq k$ the quadratic function $\mathcal{Q} : \mathbb{R}^N \mapsto \mathbb{R}$ defined in (2.48) obtains a minimum in $\mathbb{R}^N$ given by the vector $D^\alpha \tilde{v}(\mathbf{x})$. Thus we have*

$$\mathcal{Q}(D^\alpha \tilde{v}(\mathbf{x})) \leq \mathcal{Q}(v).$$

for all $v \in \mathbb{R}^N$.

Proof. Choose any $\mathbf{x} \in \Omega$. First we write (2.48) in matrix-vector form as

$$\mathcal{Q}(v) = v^T \mathcal{A} v - 2v^T D^\alpha \mathbf{R}(\mathbf{x}) + D_1^\alpha D_2^\alpha \Psi(\mathbf{x}, \mathbf{x})$$

using the matrix $\mathcal{A}$ and vector $\mathbf{R}(\mathbf{x}) = (\Psi(\mathbf{x}, \mathbf{x}_1), \ldots, \Psi(\mathbf{x}, \mathbf{x}_N))^T$ defined previously. Differentiation of the quadratic function $\mathcal{Q}$ with respect to $v \in \mathbb{R}^N$ and setting to 0, we get

$$\mathcal{Q}'(v) = \mathcal{A}v - D^\alpha \mathbf{R}(\mathbf{x}) = 0.$$

Since $\mathcal{A}$ is symmetric positive definite due to Ψ being SPD, the solution to this system is uniquely solvable and is given by $v_0 = \mathcal{A}^{-1} D^\alpha \mathbf{R}(\mathbf{x})$. By equation (2.36), this implies $v_0 = D^\alpha \tilde{\mathbf{v}}(\mathbf{x})$. Thus $D^\alpha \tilde{\mathbf{v}}(\mathbf{x}) \in \mathbb{R}^N$ minimizes $\mathcal{Q}$ on $\mathbb{R}^N$. □

Now that we know the vector $v = D^\alpha \tilde{\mathbf{v}}(\mathbf{x})$ for any fixed $\mathbf{x} \in \Omega$ minimizes the quadratic function $\mathcal{Q}$, our final step in obtaining the error estimate (2.37) is to find a vector $\tilde{u} \in \mathbb{R}^N$ such that we can bound $\mathcal{Q}(\tilde{u})$ by $h^{k-|\alpha|}$ times some constant $C > 0$ that will depend on the domain Ω, the points $\mathcal{X}$, and the kernel Ψ, giving $\mathcal{Q}(\tilde{u}) \leq Ch^{k-|\alpha|}$. So what kind of $\tilde{u} \in \mathbb{R}^N$ will accomplish this? It turns out that $\tilde{u}$ is a vector that satisfies a local polynomial reproduction, as discussed in Appendix C. Only here, the vector needs to satisfy local reproduction of derivatives of polynomials. The following Theorem gives existence of local polynomial reproduction for derivatives of polynomials.

Theorem 2.2.14. *Suppose that $\Omega \subset \mathbb{R}^2$ is open and bounded and satisfies an interior cone condition with radius $r > 0$ and angle $\theta \in (0, \pi/2)$. Let $m \in \mathbb{N}_0$ and $\alpha \in \mathbb{N}_0^2$ be given such that $m \geq |\alpha|$. Then there exists constants $h_0, c_1^{(\alpha)}, c_2^{(\alpha)} > 0$ such that for all $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega$ with saturation measure $h := h_{\mathcal{X},\Omega} \leq h_0$ and for every $\mathbf{x} \in \Omega$, there exists numbers $\tilde{u}_1^{(\alpha)}(\mathbf{x}), \ldots, \tilde{u}_N^{(\alpha)}(\mathbf{x})$ that satisfy*

$$\begin{cases} 1) \sum_{i=1}^N \tilde{u}_i^{(\alpha)}(\mathbf{x}) p(\mathbf{x}_j) = D^{(\alpha)} p(\mathbf{x}) \quad \forall p \in \mathcal{P}_m^2 \\ 2) \sum_{i=1}^N |\tilde{u}_i^{(\alpha)}(\mathbf{x})| \leq c_1^{(\alpha)} h^{-|\alpha|}, \\ 3) \tilde{u}_i^{(\alpha)}(\mathbf{x}) = 0, \quad if \ \|\mathbf{x}_i - \mathbf{x}\|_2 > c_2^{(\alpha)} h \end{cases} \tag{2.49}$$

Proof. See proof of Theorem B.6.1 in Appendix C, section B.6. □

We now have all the tools necessary to prove the main error estimate for scattered data interpolation. (We note again that this theory will work for $\Omega \subset \mathbb{R}^d$ for any integer d, as long as the correct assumptions are made on Ω according to the Theorem. However for simplicity we restrict ourselves to $\mathbb{R}^2$ since our numerical experiments later will take place in $\mathbb{R}^2$.)

Theorem 2.2.15. *(Wendland [66], Chapter 11) Let $\Omega \subset \mathbb{R}^2$ be open and bounded, satisfying an interior cone condition. Suppose that $\Psi \in C^{2k}(\Omega \times \Omega)$ is a symmetric positive definite kernel and denote the interpolant to any $f \in \mathcal{N}_\Psi(\Omega)$ on a set of distinct points $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ as $I_\mathcal{X} f$. Fix $\alpha \in \mathbb{N}_0^2$ with $|\alpha| \leq k$. Then there exists constants $h_0, C > 0$ such that the following error estimate holds for any $\mathbf{x} \in \Omega$*

and any set $\mathcal{X} \subseteq \Omega$ *with* $h := h_{\mathcal{X},\Omega} \leq h_0$

$$|D^\alpha(f(\mathbf{x}) - I_\mathcal{X} f(\mathbf{x}))| \leq C C_\Psi(\mathbf{x})^{1/2} h^{k-|\alpha|} \|f\|_{\mathcal{N}_\Psi(\Omega)}, \quad \mathbf{x} \in \Omega. \tag{2.50}$$

The number $C_\Psi(\mathbf{x})$ *is given by*

$$C_\Psi(\mathbf{x}) := \max_{|\beta|+|\rho|=2k} \max_{\mathbf{z},\mathbf{y} \in \Omega \cap B(\mathbf{x}, c_2^{(\alpha)} h)} |D_1^\beta D_2^\rho \Psi(\mathbf{z}, \mathbf{y})|, \tag{2.51}$$

and the constant C *is independent of* $\mathbf{x}$, f *and* Ψ.

Proof. Choose any $\mathbf{x} \in \Omega$. We first show that

$$|D^\alpha(f(\mathbf{x}) - I_\mathcal{X} f(\mathbf{x}))| \leq P^{(\alpha)}_{\Psi,\mathcal{X}}(\mathbf{x}) \|f\|_{\mathcal{N}_\Psi(\Omega)}$$

where $P^{(\alpha)}_{\Psi,\mathcal{X}}(\mathbf{x})$ is the power function defined in (2.46). To this end, using the reproducing properties of the kernel Ψ and the definition of the interpolants $\tilde{v}_j(\mathbf{x})$, we can write

$$\begin{aligned} D^\alpha I_\mathcal{X} f(\mathbf{x}) &= \sum_{i=1}^N f(\mathbf{x}_j) D^\alpha \tilde{v}_j(\mathbf{x}) \\ &= \sum_{i=1}^N D^\alpha \tilde{v}_j(\mathbf{x}) \big[(f, \Psi(\cdot, \mathbf{x}_j))_{\mathcal{N}_\Psi(\Omega)}\big] \\ &= \big(f, \sum_{i=1}^N D^\alpha \tilde{v}_j(\mathbf{x}) \Psi(\cdot, \mathbf{x}_j)\big)_{\mathcal{N}_\Psi(\Omega)}. \end{aligned} \tag{2.52}$$

In a similar manner, invoking Theorem 2.2.12, we have

$$D^\alpha f(\mathbf{x}) = (f, D_2^\alpha \Psi(\cdot, \mathbf{x}))_{\mathcal{N}_\Psi(\Omega)}. \tag{2.53}$$

Subtracting this from the last equality in (2.52), we get

$$\begin{aligned} |D^\alpha(f(\mathbf{x}) - I_\mathcal{X} f(\mathbf{x}))| &= \Big|\Big(f, D^\alpha \Psi(\cdot, \mathbf{x}) - \sum_{i=1}^N D^\alpha \tilde{v}_j(\mathbf{x}) \Psi(\cdot, \mathbf{x}_j)\Big)_{\mathcal{N}_\Psi(\Omega)}\Big| \\ &\leq P^{(\alpha)}_{\Psi,\mathcal{X}}(\mathbf{x}) \|f\|_{\mathcal{N}_\Psi(\Omega)}, \end{aligned} \tag{2.54}$$

where we applied the Cauchy-Schwarz inequality and the definition of the power function. Now we seek to bound the power function $P^{(\alpha)}_{\Psi,\mathcal{X}}(\mathbf{x})$ in terms of h. In order to do this, we will be applying some Taylor expansions along with Theorem 2.2.14. The Taylor expansions are given as follows. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$, we will denote a point on the line between $\mathbf{x}$ and $\mathbf{y}$ by ξ. Fix $\rho \in \mathbb{N}_0^2$ with $|\rho| < k$. The first type of Taylor expansion with respect to the second argument of $D_1^\rho \Psi(\mathbf{x}, \mathbf{y})$ that we use is given by

$$D_1^\rho \Psi(\mathbf{x}, \mathbf{y}) = \sum_{|\beta|<2k-|\rho|} \frac{D_1^\rho D_2^\beta \Psi(\mathbf{x}, \mathbf{x})}{\beta!} (\mathbf{y} - \mathbf{x})^\beta + R(\mathbf{x}, \mathbf{y}, \rho) \tag{2.55}$$

with remainder term

$$R(\mathbf{x}, \mathbf{y}, \rho) = \sum_{|\beta|=2k-|\rho|} \frac{D_1^\rho D_2^\beta \Psi(\mathbf{x}, \xi)}{\beta!} (\mathbf{y} - \mathbf{x})^\beta.$$

The second type of Taylor expansion we will use is with respect to the second argument of $D_2^\rho \Psi(\mathbf{x}, \mathbf{y})$ and is given as

$$D_2^\rho \Psi(\mathbf{x}, \mathbf{y}) = \sum_{|\beta|<2k-|\rho|} \frac{D_2^{\rho+\beta} \Psi(\mathbf{x}, \mathbf{x})}{\beta!} (\mathbf{y} - \mathbf{x})^\beta + S(\mathbf{x}, \mathbf{y}, \rho), \tag{2.56}$$

with remainder term

$$S(\mathbf{x}, \mathbf{y}, \rho) = \sum_{|\beta|=2k-|\rho|} \frac{D_2^{\beta+\rho} \Psi(\mathbf{x}, \xi)}{\beta!} (\mathbf{y} - \mathbf{x})^\beta.$$

Now we can bound the power function. Let $v := \tilde{u}^{(\alpha)}(\mathbf{x}) \in \mathbb{R}^N$ where the vector $\tilde{u}^{(\alpha)}(\mathbf{x}) = (\tilde{u}_1^{(\alpha)}(\mathbf{x}), \ldots, \tilde{u}_N^{(\alpha)}(\mathbf{x})) \in \mathbb{R}^N$ satisfies the conditions of Theorem 2.2.14. Inserting this in the quadratic form

$$\mathcal{Q}(v) = D_1^\alpha D_2^\alpha \Psi(\mathbf{x}, \mathbf{x}) - 2 \sum_{k=1}^{N} v_j D_1^\alpha \Psi(\mathbf{x}, \mathbf{x}_j) + \sum_{i,j=1}^{N} v_i v_j \Psi(\mathbf{x}_i, \mathbf{x}_j), \tag{2.57}$$

we apply the first Taylor expansion on the second and third terms of the above equation to get

$$\begin{aligned}\mathcal{Q}(v) =&D_1^\alpha D_2^\alpha \Psi(\mathbf{x},\mathbf{x})\\ &-2\sum_{j=1}^N v_j\Big(\sum_{|\beta|<2k-|\alpha|}\frac{D_1^\alpha D_2^\beta\Psi(\mathbf{x},\mathbf{x})}{\beta!}(\mathbf{x}_j-\mathbf{x})^\beta+R(\mathbf{x},\mathbf{x}_j,\alpha)\Big)\\ &+\sum_{i,j}^N v_iv_j\Big(\sum_{|\beta|<2k}\frac{D_2^\beta\Psi(\mathbf{x}_i,\mathbf{x}_i)}{\beta!}(\mathbf{x}_j-\mathbf{x}_i)^\beta+R(\mathbf{x}_i,\mathbf{x}_j,0)\Big).\end{aligned}\tag{2.58}$$

Now we envoke Theorem 2.2.14 with $m \geq 2k-|\alpha|$ at the point $\mathbf{x}$. We apply the reproduction property 1) of Theorem 2.2.14 to all the nonzero terms. Noticing that $\sum_{|\beta|<2k-|\alpha|} D_1^\alpha D_2^\beta\Psi(\mathbf{x},\mathbf{x})(\beta!)^{-1}(\mathbf{x}_j-\mathbf{x})^\beta$ is a polynomial in P_m^2 with respect to $\mathbf{x}_j$, we define $p(\mathbf{z}) := \sum_{|\beta|<2k-|\alpha|} D_1^\alpha D_2^\beta\Psi(\mathbf{x},\mathbf{x})(\beta!)^{-1}(\mathbf{z}-\mathbf{x})^\beta$ and then apply the reproduction property 1) to get

$$\begin{aligned}&\sum_{j=1}^N v_j(\mathbf{x})p(\mathbf{x}_j) = D_{\mathbf{z}}^\alpha p|_{\mathbf{x}}\\ &= D^\alpha\Big(\sum_{|\beta|<2k-|\alpha|}\frac{D_1^\alpha D_2^\beta\Psi(\mathbf{x},\mathbf{x})}{\beta!}(\mathbf{x}-\mathbf{x})^\beta\Big)\\ &= D_1^\alpha D_2^\alpha\Psi(\mathbf{x},\mathbf{x}),\end{aligned}\tag{2.59}$$

since all terms cancel except when $|\beta| = 0$. For the third term in (2.58), we apply the reproducing property on $\sum_{|\beta|<2k} D_2^\beta\Psi(\mathbf{x}_i,\mathbf{x}_i)(\beta!)^{-1}(\mathbf{x}_j-\mathbf{x}_i)^\beta$ while fixing the point $\mathbf{x}_i$. Defining $p(\cdot) := \sum_{|\beta|<2k} D_2^\beta\Psi(\mathbf{x}_i,\mathbf{x}_i)(\beta!)^{-1}(\cdot-\mathbf{x}_i)^\beta$, a polynomial in P_m^2, we get

$$\begin{aligned}\sum_{j=1}^{N} v_j p(\mathbf{x}_j) &= D^{\alpha}\Big(\sum_{|\beta|<2k} \frac{D_2^{\beta}\Psi(\mathbf{x}_i,\mathbf{x}_i)}{\beta!}(\mathbf{x}-\mathbf{x}_i)^{\beta}\Big) \\ &= \sum_{|\beta|<2k} \frac{D_2^{\beta}\Psi(\mathbf{x}_i,\mathbf{x}_i)}{\beta!} D^{\alpha}(\mathbf{x}-\mathbf{x}_i)^{\beta} \\ &= \sum_{|\alpha|\le|\beta|<2k} \frac{D_2^{\beta}\Psi(\mathbf{x}_i,\mathbf{x}_i)}{\beta!}(\mathbf{x}-\mathbf{x}_i)^{\beta-\alpha} \\ &= \sum_{|\beta|<2k-|\alpha|} \frac{D_2^{\beta+\alpha}\Psi(\mathbf{x}_i,\mathbf{x}_i)}{\beta!}(\mathbf{x}-\mathbf{x}_i)^{\beta}\end{aligned} \tag{2.60}$$

where the third line comes about by re-indexing the summation after noticing that $|\alpha|$ terms vanish from the original summation. Now, applying the second Taylor expansion (2.56), we see that

$$\sum_{|\beta|<2k-|\alpha|} \frac{D_2^{\beta+\alpha}\Psi(\mathbf{x}_i,\mathbf{x}_i)}{\beta!}(\mathbf{x}-\mathbf{x}_i)^{\beta} = D_2^{\alpha}\Psi(\mathbf{x}_i,\mathbf{x}) - S(\mathbf{x}_i,\mathbf{x},\alpha).$$

Using these three computations in (2.58) and collecting like terms on the remainders R, we get

$$\begin{aligned}\mathcal{Q}(v) =& D_1^{\alpha}D_2^{\alpha}\Psi(\mathbf{x},\mathbf{x}) - 2D_1^{\alpha}D_2^{\alpha}\Psi(\mathbf{x},\mathbf{x}) - 2\sum_{j=1}^{N} v_j R(\mathbf{x},\mathbf{x}_j,\alpha) \\ &+ \sum_{i=1}^{N} v_i \sum_{|\beta|<2k-|\alpha|} \frac{D_2^{\beta+\alpha}\Psi(\mathbf{x}_i,\mathbf{x}_i)}{\beta!}(\mathbf{x}-\mathbf{x}_i)^{\beta} + \sum_{i,j=1}^{N} v_i v_j R(\mathbf{x}_i,\mathbf{x}_j,0) \\ &= -D_1^{\alpha}D_2^{\alpha}\Psi(\mathbf{x},\mathbf{x}) - \sum_{j=1}^{N} v_j \Big(2R(\mathbf{x},\mathbf{x}_j,\alpha) - \sum_{i=1}^{N} v_i R(\mathbf{x}_i,\mathbf{x}_j,0)\Big) \\ &+ \sum_{i=1}^{N} v_i \Big[D_2^{\alpha}\Psi(\mathbf{x}_i,\mathbf{x}) - S(\mathbf{x}_i,\mathbf{x},\alpha)\Big].\end{aligned} \tag{2.61}$$

Finally, by symmetry $D_2^{\alpha}\Psi(\mathbf{x}_i,\mathbf{x}) = D_1^{\alpha}\Psi(\mathbf{x},\mathbf{x}_i)$, and so we again apply the first Taylor expansion formula on $D_1^{\alpha}\Psi(\mathbf{x},\mathbf{x}_i)$ giving

$$D_1^{\alpha}\Psi(\mathbf{x},\mathbf{x}_i) = \sum_{|\beta|<2k-|\alpha|} \frac{D_1^{\alpha}D_2^{\beta}\Psi(\mathbf{x},\mathbf{x})}{\beta!}(\mathbf{x}_i-\mathbf{x})^{\beta} + R(\mathbf{x},\mathbf{x}_i,\alpha),$$

and then apply the reproduction formula on the polynomial $p(\mathbf{z}) = \sum_{|\beta|<2k-|\alpha|} \frac{D_1^{\alpha} D_2^{\beta} \Psi(\mathbf{x},\mathbf{x})}{\beta!}(\mathbf{z}-\mathbf{x})^{\beta}$ of order $2k - |\alpha|$ which leads to $D_1^{\alpha} D_2^{\alpha} \Psi(\mathbf{x}, \mathbf{x})$ as in (2.59). This leads to the expression for $\mathcal{Q}(v)$ in terms of the Taylor remainders. Collecting like terms above, we have

$$\mathcal{Q}(v) = -\sum_{j=1}^{N} v_j \Big(R(\mathbf{x}, \mathbf{x}_j, \alpha) + S(\mathbf{x}_j, \mathbf{x}_i, \alpha) - \sum_{i=1}^{N} v_i R(\mathbf{x}_i, \mathbf{x}_j, 0) \Big) \tag{2.62}$$

Now using property 3) of Theorem 2.2.14, only the points $\mathbf{x}_j \in \mathcal{X}$ that satisfy $\|\mathbf{x}_j - \mathbf{x}\|_2 \le c_2^{(\alpha)} h$ are nonzero for $v_j := \tilde{u}_j^{(\alpha)}(\mathbf{x})$. Furthermore, we have $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \le 2c_2^{(\alpha)} h$ for any other $\mathbf{x}_i$ in the support of $\tilde{u}_j^{(\alpha)}(\mathbf{x})$. We know from property 2) of Theorem (2.2.14) that $\sum_{j=1}^{N} |\tilde{u}_j^{(\alpha)}(\mathbf{x})| \le c_1^{(\alpha)} h^{-|\alpha|}$ which gives the bounds $R(\mathbf{x}, \mathbf{x}_i, \alpha) \le C C_{\Psi}(\mathbf{x}) h^{2k-|\alpha|}$ and similarly $S(\mathbf{x}_i, \mathbf{x}, \alpha) \le C C_{\Psi}(\mathbf{x}) h^{2k-|\alpha|}$ where $C_{\Psi}(\mathbf{x})$ was defined in (2.51). Lastly, we can bound the last term by $\sum_{i=1}^{N} v_i R(\mathbf{x}_i, \mathbf{x}_j, 0) \le \sum_{i=1} |v_i| \sum_{i=1} |R(\mathbf{x}_i, \mathbf{x}_j, 0)| \le c_1^{(\alpha)} h^{-\alpha} C_{\Psi}(\mathbf{x}) h^{2k} \le C C_{\Psi}(\mathbf{x}) h^{2k-|\alpha|}$ using the bound on the $\tilde{u}_j^{(\alpha)}(\mathbf{x})$ vector. This gives substituting into (2.62)

$$\mathcal{Q}(v) \le -\sum_{j=1}^{N} v_j (C C_{\Psi}(\mathbf{x}) h^{2k-|\alpha|}) \le C h^{-|\alpha|} C_{\Psi}(\mathbf{x}) h^{2k-|\alpha|} \tag{2.63}$$

and thus the final desired bound $P_{\Psi,x}^{(\alpha)}(x) \le C C_{\Psi}(\mathbf{x})^{1/2} h^{k-|\alpha|}$. □

Further analysis on the kernel Ψ and bounds on its corresponding power function $P_{\Psi,x}^{(\alpha)}(x)$ need to be done in order to achieve error bounds in other norms, such as the standard Sobolev norm $\|\cdot\|_{H^s(\Omega)}$ for $s > k + 1$. However, this analysis is out of the scope of this thesis and the interested reader is referred to [66], [44], [41], where error estimates in Sobolev spaces of scattered data interpolation with SPD kernels have been derived. For our purposes, the pointwise error bound derived in the previous Theorem is sufficient in the context of meshless collocation for numerically solving elliptic partial differential equations, which is the subject of the next section.

2.3 Meshless Collocation Method for Elliptic PDEs in Native Spaces

We now discuss the topic of meshless collocation for numerically solving elliptic boundary valued PDEs. We apply the theory of reproducing kernel Hilbert spaces and scattered data approximation introduced in the previous section along with a couple of additional tools to construct a collocation method which is completely meshless in the sense that only a set of scattered distinct points in the domain of interest Ω and on the boundary $\partial\Omega$ is needed to construct the approximation. The approach for meshless collocation that we introduce in this section is a kernel-based symmetric collocation method which turns out to be closely related to the well-known Hermite-Birkoff generalized interpolation problem (see Narcowich and Ward [42]). The analysis and notation in this section will be loosely based on analysis and notation of Wendland [66], Chapter 16.

Let $\Omega \subset \mathbb{R}^2$ be an open, bounded and connected set with Lipschitz boundary $\partial\Omega$. We will consider elliptic boundary-valued PDEs on Ω of the form

$$\begin{aligned} Lu &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{2.64}$$

where $f \in C(\Omega)$ and $g \in C(\partial\Omega)$ are given. The operator $L : C^2(\Omega) \mapsto C(\Omega)$ is an elliptic differential operator

$$Lu(\mathbf{x}) = \sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\big(a_{i,j}(\mathbf{x})\frac{\partial u}{\partial x_j}\big) + \sum_{j=1}^{2} b_j(\mathbf{x})\frac{\partial u}{\partial x_j}(\mathbf{x}) + u(\mathbf{x}), \tag{2.65}$$

where $a_{i,j} \in C^1(\Omega)$ and the matrix with entries $a_{i,j}(\mathbf{x})$ for any $\mathbf{x} \in \Omega$ is positive definite.

The symmetric meshless collocation method for approximating solutions to (2.64) begins by introducing the native space framework from section 2.2.4. Let $\Psi_0 \in C^{2k}(\mathbb{R}^2)$ for $k \geq 2$ be a radial and positive definite function. We define a symmetric positive definite kernel $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ by $\Psi(\mathbf{x}, \mathbf{y}) := \Psi_0(\mathbf{x} - \mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in \Omega$. As usual, we will denote $\mathcal{N}_\Psi(\Omega)$ the native space of the kernel Ψ on Ω. By construction, since the kernel Ψ is bounded and continuous on $\Omega \times \Omega$, we clearly have that $\mathcal{N}_\Psi(\Omega) \subset C(\overline{\Omega})$ where $\overline{\Omega}$ is the closure of Ω. In particular, the trace $u|_{\partial\Omega}$ of any function $u \in \mathcal{N}_\Psi(\Omega)$ is well-defined and in $C(\partial\Omega)$.

We assume (2.64) is well-posed in the sense that there exists a unique solution $u \in \mathcal{N}_\Psi(\Omega)$ satisfying both equations in (2.64). At this point, we will not discuss the necessary conditions on f and g that imply the existence and uniqueness of a solution $u \in \mathcal{N}_\Psi(\Omega)$. For right now we will only assume they are continuous.

To obtain an approximation, we introduce sets of distinct points $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \subseteq \overline{\Omega}$ where $\mathcal{X}_1 := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq \Omega$ and $\mathcal{X}_2 := \{\mathbf{x}_{n+1}, \ldots \mathbf{x}_N\} \subseteq \partial\Omega$ for a total of N distinct points. Each set of points is equipped with a point saturation measure defined by

$$\begin{aligned} h_{\Omega,\mathcal{X}_1} &= \sup_{\mathbf{x}\in\Omega} \min_{\mathbf{x}_i\in\mathcal{X}_1} \|\mathbf{x} - \mathbf{x}_i\|_2 \\ h_{\partial\Omega,\mathcal{X}_2} &= \sup_{\mathbf{x}\in\partial\Omega} \min_{\mathbf{x}_i\in\mathcal{X}_2} \|\mathbf{x} - \mathbf{x}_i\|_2, \end{aligned} \tag{2.66}$$

although we will often use the abbreviated notation $h_1 := h_{\Omega,\mathcal{X}_1}$ and $h_2 := h_{\partial\Omega,\mathcal{X}_2}$. Using h_1 and h_2, we can also define the saturation measure on all of $\mathcal{X}$ and $\overline{\Omega}$ as $h := \max(h_1, h_2)$. We will call these sets of distinct points in $\overline{\Omega}$ the *collocation nodes.*

With respect to $\mathcal{X}_1$ and $\mathcal{X}_2$, we define the functionals

$$\lambda_j = \begin{cases} \delta_{\mathbf{x}_j} \circ L, & 1 \leq j \leq n, \\ \delta_{\mathbf{x}_j}, & n+1 \leq j \leq N. \end{cases} \tag{2.67}$$

The action of the functional λ_j on the SPD kernel Ψ is defined as $\lambda_{1,j}\Psi(\mathbf{x},\mathbf{y})$ for the first argument of the kernel and $\lambda_{2,j}\Psi(\mathbf{x},\mathbf{y})$ for the second argument. For example, if $1 \leq j \leq n$, then by definition $\lambda_{2,j}\Psi(\mathbf{x},\mathbf{y}) := (\delta_{\mathbf{x}_j} \circ L)_2\Psi(\mathbf{x},\mathbf{y}) = L_2\Psi(\mathbf{x},\mathbf{x}_j)$, $\mathbf{x}_j \in \mathcal{X}_1$. Similarly, if $n+1 \leq j \leq N$, then $\lambda_{1,j}\Psi(\mathbf{x},\mathbf{y}) := (\delta_{\mathbf{x}_j})_1\Psi(\mathbf{x},\mathbf{y}) = \Psi(\mathbf{x}_j,\mathbf{y})$, $\mathbf{x}_j \in \mathcal{X}_2$.

For $1 \leq j \leq n$, since $\Psi \in C^{2k}(\Omega \times \Omega)$ it is easy to see that the functionals $\lambda_j := (\delta_{\mathbf{x}_j} \circ L)$ are in the dual space. Indeed, for any $u \in \mathcal{N}_\Psi(\Omega)$, using Theorem 2.2.12, we have

$$|\lambda_j(u)| = |Lu(\mathbf{x}_j)| = |(u, L_2\Psi(\cdot,\mathbf{x}_j))_{\mathcal{N}_\Psi(\Omega)}| \leq \|u\|_{\mathcal{N}_\Psi(\Omega)}\|L_2\Psi(\cdot,\mathbf{x}_j)\|_{\mathcal{N}_\Psi(\Omega)} \tag{2.68}$$

which is bounded by the fact that $L_2\Psi(\cdot,\mathbf{x}_j) \in \mathcal{N}_\Psi(\Omega)$. Similarly, for $n+1 \leq j \leq N$, since $u|_{\partial\Omega} \in C(\partial\Omega)$ for any $u \in \mathcal{N}_\Psi(\Omega)$, the functionals $\lambda_j := \delta_{\mathbf{x}_j}$ for $\mathbf{x}_j \in \mathcal{X}_2$ are in $\mathcal{N}_\Psi(\Omega)'$. Indeed, we have $|\lambda_j(u)| = |u(\mathbf{x}_j)| = |(u, \Psi(\cdot,\mathbf{x}_j))_{\mathcal{N}_\Psi(\Omega)}| \leq \|u\|_{\mathcal{N}_\Psi(\Omega)}\|\Psi(\cdot,\mathbf{x}_j)\|_{\mathcal{N}_\Psi(\Omega)} = \sqrt{\Psi_0(0)}\|u\|_{\mathcal{N}_\Psi(\Omega)}$.

With the functionals $\Lambda_N = \{\lambda_1, \ldots, \lambda_N\}$ in the dual space $\mathcal{N}_\Psi(\Omega)'$, we can form the finite dimensional spaces

$$\begin{aligned}
\mathcal{N}_{\Psi,N} &:= \operatorname{span}\{\lambda_{2,1}\Psi(\cdot,\mathbf{x}),\ldots,\lambda_{2,N}\Psi(\cdot,\mathbf{x})\} \subset \mathcal{N}_\Psi(\Omega) \\
\mathcal{N}'_{\Psi,N} &:= \operatorname{span}\{\lambda_1,\ldots,\lambda_N\} \subset \mathcal{N}_\Psi(\Omega)'.
\end{aligned} \tag{2.69}$$

The meshless collocation method can now be formulated with respect to these spaces. Assuming that $\{\lambda_1,\ldots,\lambda_N\}$ are linearly independent, we seek an approximation $u_{h,\Lambda}$ to the exact solution $u \in \mathcal{N}_\Psi(\Omega)$ of (2.64) of the form

$$\begin{aligned}
u_{h,\Lambda}(\mathbf{x}) &= \sum_{i=1}^{N} \alpha_i \lambda_{2,i}\Psi(\mathbf{x},\mathbf{y}) \\
&= \sum_{i=1}^{n} \alpha_i L_2\Psi(\mathbf{x},\mathbf{x}_i) + \sum_{i=n+1}^{N} \alpha_i \Psi(\mathbf{x},\mathbf{x}_i),
\end{aligned} \tag{2.70}$$

which satisfies

$$\begin{aligned}
Lu_{h,\Lambda}(\mathbf{x}_j) &= f(\mathbf{x}_j), \quad 1 \le j \le n, \\
u_{h,\Lambda}(\mathbf{x}_j) &= g(\mathbf{x}_j), \quad n+1 \le j \le N.
\end{aligned} \tag{2.71}$$

In matrix-vector form, this amounts to finding a vector of coefficients $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_N) \in \mathbb{R}^N$ such that

$$\begin{pmatrix} A & C \\ C^T & D \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} f|_{\mathcal{X}_1} \\ g|_{\mathcal{X}_2} \end{pmatrix}, \tag{2.72}$$

where the block matrices are defined by

$$\begin{aligned}
A_{i,j} &= L_1 L_2 \Psi(\mathbf{x}_i,\mathbf{x}_j),\ 1 \le i,j \le n, \\
C_{i,j} &= L_2\Psi(\mathbf{x}_i,\mathbf{x}_j),\ n+1 \le i \le N, 1 \le j \le n, \\
D_{i,j} &= \Psi(\mathbf{x}_i,\mathbf{x}_j),\ n+1 \le i,j \le N.
\end{aligned} \tag{2.73}$$

Using shorthand notation, this system can be written as $\mathcal{A}\boldsymbol{\alpha} = \mathbf{f}$ where $\mathcal{A}$ is the matrix with entries $\mathcal{A}[i,j] = \lambda_{1,i}\lambda_{2,j}\Psi(\mathbf{x},\mathbf{y})$ for $1 \le i,j \le N$ and $\mathbf{f} = (f(\mathbf{x}_1),\ldots,f(\mathbf{x}_n),g(\mathbf{x}_{n+1}),\ldots,$

As can be deduced from equations (2.71), the method is indeed of collocation type due to the requirement that our approximation $u_{h,\Lambda}$ to the solution u be satisfied pointwise on the sets of distinct collocation nodes $\mathcal{X}_1$ and $\mathcal{X}_2$. The method is also symmetric due to the obvious symmetry of the collocation matrix $\mathcal{A}$.

The first step in our analysis of symmetric meshless collocation is to show that the matrix in (2.72) is positive definite, thus implying that $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)$ exists and is unique, resulting in a unique approximation $u_{h,\Lambda}$ to u. To do this, we require that the set of functionals $\Lambda_N = \{\lambda_1, \dots, \lambda_N\}$ in the dual space $\mathcal{N}_\Psi(\Omega)'$ are linearly independent. Given that the set of functionals Λ_N are linearly independent, we can conclude that the matrix $\mathcal{A}$ is indeed symmetric positive definite.

Theorem 2.3.1. *(Wendland [68], Chapter 16) Suppose that $\Psi \in C^{2k}(\Omega \times \Omega)$ is a positive definite reproducing kernel with native Hilbert space $\mathcal{N}_\Psi(\Omega)$. If the functionals in Λ_N defined at (2.67) are linearly independent, then the matrix defined by $\mathcal{A}[i,j] = (\lambda_{1,i}\lambda_{2,j}\Psi(\mathbf{x},\mathbf{y}))$, $1 \le i,j \le N$, is symmetric positive definite.*

Proof. Let $\lambda = \sum_{i=1}^N \alpha_i \lambda_i$ with $\alpha_i \in \mathbb{R}$ be given. Then we have using the definition of the inner product on the dual space $\mathcal{N}_\Psi(\Omega)'$

$$\sum_{i,j=1}^{N} \alpha_i \alpha_j \lambda_{1,i}\lambda_{2,j}\Psi(\mathbf{x},\mathbf{y}) = (\lambda,\lambda)_{\mathcal{N}_\Psi(\Omega)'} = \|\lambda\|^2_{\mathcal{N}_\Psi(\Omega)'} \tag{2.74}$$

which is nonnegative and is 0 only if $\boldsymbol{\alpha} = 0$. But the set $\Lambda_N = \{\lambda_1, \dots, \lambda_N\}$ is linearly independent, thus $\sum_{i,j=1}^N \alpha_i\alpha_j\lambda_{1,i}\lambda_{2,j}\Psi(\mathbf{x},\mathbf{y}) = \|\lambda\|^2_{\mathcal{N}_\Psi(\Omega)'} > 0$, and so the matrix $\mathcal{A}$ is positive definite. Symmetry comes directly from the symmetry of Ψ. □

Now that we know the matrix $\mathcal{A}$ is symmetric positive definite, there exists a unique solution $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ to the system (2.72), and thus a unique approximation $u_{h,\Lambda}$ to u. We would now like to know how "close" the approximation $u_{h,\Lambda}$ is to u and if $u_{h,\Lambda}$ converges to u as $h \to 0$ in the sense that $\|u - u_{h,\Lambda}\|_{L^\infty(\Omega)} \to 0$ as $h \to 0$. In order to do this, we will need a few additional tools for the error analysis. We will frequently make use of the so called *modified* kernel defined as follows.

Definition For an SPD kernel $\Psi \in C^{2k}(\Omega \times \Omega)$ and differential operator $L : C^2(\Omega) \mapsto C(\Omega)$, the *modified* kernel with respect to L will be defined as

$$\Psi_L(\mathbf{x}, \mathbf{y}) := (\delta_{\mathbf{x}} \circ L)_1 (\delta_{\mathbf{y}} \circ L)_2 \Psi(\mathbf{u}, \mathbf{v}), \quad \mathbf{x}, \mathbf{y} \in \Omega. \tag{2.75}$$

Since Ψ is an SPD kernel on $\Omega \times \Omega$ it is clear that if

$$\{\delta_{\mathbf{x}} \circ L \ : \ \mathbf{x} \in \Omega\} \text{ is linearly independent over } \mathcal{N}_\Psi(\Omega), \tag{2.76}$$

then Ψ_L is also an SPD kernel. (Recall that a set of linear functionals $\lambda_1, \ldots, \lambda - N \in \mathcal{N}_\Psi(\Omega)'$ is linear independent on $\mathcal{N}_\Psi(\Omega)$ if $\sum_{i=1}^N \alpha_i \lambda_i(f) = 0$ implies that $\boldsymbol{\alpha} = 0$. In other words, $\sum_{i=1}^N \alpha_i \lambda_i \Psi(\cdot, \mathbf{x}) = 0$ implies $\boldsymbol{\alpha} = 0$).

To see this, let $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be any set of distinct points in Ω and $\boldsymbol{\alpha} \in \mathbb{R}^N$ be given. Then we have

$$\begin{aligned} \sum_{i,j}^N \alpha_i \alpha_j \Psi_L(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i,j}^N \alpha_i \alpha_j (\delta_{\mathbf{x}_i} \circ L)_1 (\delta_{\mathbf{x}_j} \circ L)_2 \Psi(\mathbf{x}, \mathbf{y}) \\ &= \| \sum_{i=1}^N \alpha_i (\delta_{\mathbf{x}_i} \circ L) \|^2_{\mathcal{N}_\Psi(\Omega)'} \end{aligned} \tag{2.77}$$

which is nonnegative and is 0 only if $\boldsymbol{\alpha} = 0$.

Now since Ψ_L is an SPD kernel, we can consider its native Hilbert space $\mathcal{N}_{\Psi_L}(\Omega)$ with reproducing kernel Ψ_L. Since $\Psi \in C^{2k}(\Omega \times \Omega)$ and $\mathcal{N}_\Psi(\Omega) \subset C^k(\Omega)$, clearly we must have $\Psi_L \in C^{2k-4}(\Omega \times \Omega)$ and $\mathcal{N}_{\Psi_L}(\Omega) \subset C^{k-2}(\Omega)$ for $k \geq 2$. Our next Theorem enlightens the connection between the kernels Ψ and Ψ_L.

Theorem 2.3.2. *(Wendland [66], Chapter 16) Suppose that $\Psi \in C^{2k}(\Omega \times \Omega)$ is a positive definite kernel and that $L : \mathcal{N}_\Psi(\Omega) \mapsto C(\Omega)$ satisfies (2.76). Then $L(\mathcal{N}_\Psi(\Omega)) = \mathcal{N}_{\Psi_L}(\Omega)$, and the following mappings*

$$\begin{aligned} L : \mathcal{N}_\Psi(\Omega) \mapsto \mathcal{N}_{\Psi_L}(\Omega), \quad & f \mapsto Lf \\ \mathcal{N}_{\Psi_L}(\Omega)' \mapsto \mathcal{N}_\Psi(\Omega)', \quad & \lambda \mapsto \lambda \circ L \end{aligned} \tag{2.78}$$

are isometric isomorphisms. In particular, if $f \in \mathcal{N}_\Psi(\Omega)$ and $\lambda \in \mathcal{N}_{\Psi_L}(\Omega)'$, then $Lf \in \mathcal{N}_{\Psi_L}(\Omega)$ and $\lambda \circ L \in \mathcal{N}_\Psi(\Omega)'$ with $\|Lf\|_{\mathcal{N}_{\Psi_L}(\Omega)} = \|f\|_{\mathcal{N}_\Psi(\Omega)}$ and $\|\lambda\|_{\mathcal{N}_{\Psi_L}(\Omega)'} = \|\lambda \circ L\|_{\mathcal{N}_\Psi(\Omega)'}$.

Proof. Our goal in the proof will be to show the existence of an isometric isomorphism $\widetilde{T} : \mathcal{N}_{\Psi_L}(\Omega) \mapsto \mathcal{N}_\Psi(\Omega)$ such that $\widetilde{T}^{-1} = L$. We will use the space of functionals $\mathcal{L}^0 := \text{span}\{\delta_{\mathbf{x}},\ \mathbf{x} \in \Omega\}$, introduced previously in (2.18) with slightly different notation, and the space $F_\Psi(\Omega) := \{\lambda_2 \Psi(\cdot, \mathbf{x}),\ \lambda \in \mathcal{L}^0\} \subseteq \mathcal{N}_\Psi(\Omega)$ originally defined in (2.11). When necessary, we will use the notation $\mathcal{L}^0_\Psi$ (resp. $\mathcal{L}^0_{\Psi_L}$) to mean the space of functionals $\mathcal{L}^0$ equipped with the inner product on $\mathcal{N}_\Psi(\Omega)'$ (resp. $\mathcal{N}_{\Psi_L}(\Omega)'$). The proof will be done in three parts.

(1) We begin by introducing the following additional spaces with respect to the operator L:

$$\mathcal{L}^0_L := \{(\lambda \circ L) : \lambda \in \mathcal{L}^0\} \subseteq \mathcal{N}_\Psi(\Omega)',\ \mathcal{F}^0_L := \{(\lambda \circ L)_2 \Psi(\cdot, \mathbf{v})\ :\ \lambda \in \mathcal{L}^0\} \subseteq \mathcal{N}_\Psi(\Omega). \tag{2.79}$$

As with $\mathcal{L}^0$ and $F_\Psi(\Omega)$, there is an obvious one-to-one correspondance between $\mathcal{L}_L^0$ and $\mathcal{F}_L^0$ given by the Riesz mapping restricted to $\mathcal{L}_L^0$

$$R_\Psi|_{\mathcal{L}_L^0} : \mathcal{L}_L^0 \mapsto \mathcal{F}_L^0,\ \lambda \circ L \mapsto (\lambda \circ L)_2 \Psi(\cdot, \mathbf{v}), \tag{2.80}$$

which is an isometric isomorphism due to the norm preserving property of the Riesz representer.

Using the spaces $\mathcal{L}_L^0$ and $\mathcal{F}_L^0$, we now define the following mappings induced by the operator L:

$$\begin{aligned} &T : F_{\Psi_L}(\Omega) \mapsto \mathcal{F}_L^0, \quad \lambda_2 \Psi_L(\cdot, \mathbf{x}) \to (\lambda \circ L)_2 \Psi(\cdot, \mathbf{x}), \\ &T' : \mathcal{L}_{\Psi_L}^0 \to \mathcal{L}_L^0, \quad \lambda \to \lambda \circ L. \end{aligned} \tag{2.81}$$

It is clear that both mappings are one-to-one since a unique $\lambda \in \mathcal{L}^0$ determines a unique element in $\mathcal{F}_L^0$ and $\mathcal{L}_L^0$. Furthermore, for any $\lambda \in \mathcal{L}^0$, we have

$$\|(\lambda \circ L)_2 \Psi(\cdot, \mathbf{v})\|_\Psi^2 = (\lambda \circ L)_1 (\lambda \circ L)_2 \Psi(\mathbf{u}, \mathbf{v}) = \lambda_1 \lambda_2 \Psi_L(\mathbf{u}, \mathbf{v}) = \|\lambda_2 \Psi_L(\cdot, \mathbf{v})\|_{\Psi_L}^2. \tag{2.82}$$

Thus the norms of $(\lambda \circ L)_2 \Psi(\cdot, \mathbf{v}) \in \mathcal{F}_L^0$ and $\lambda_2 \Psi_L(\cdot, \mathbf{v}) \in F_{\Psi_L}(\Omega)$ are the same. A similar result holds for the mapping T'. Thus $T : F_{\Psi_L}(\Omega) \mapsto \mathcal{F}_L^0$ and $T' : \mathcal{L}_{0,\Psi_L} \to \mathcal{L}_L^0$ are isometric isomorphisms.

Lastly, we want to show that the inverse mapping $T^{-1} : \mathcal{F}_L^0 \mapsto F_{\Psi_L}(\Omega)$ coincides with L. Let $f = (\lambda \circ L)_2 \Psi(\cdot, \mathbf{v}) \in \mathcal{F}_L^0$ for any arbitrary $\lambda \in \mathcal{L}^0$. Then $T^{-1} f = \lambda_2 \Psi_L(\cdot, \mathbf{v})$ which leads to

$$\begin{aligned} T^{-1} f(\mathbf{x}) &= \lambda_2 \Psi_L(\mathbf{x}, \mathbf{v}) = \lambda_2 \big((\delta_\mathbf{x} \circ L)_1 L_2\big) \Psi(\mathbf{u}, \mathbf{v}) \\ &= (\delta_\mathbf{x} \circ L)_1 (\lambda \circ L)_2 \Psi(\mathbf{u}, \mathbf{v}) = (\delta_\mathbf{x} \circ L) f(\mathbf{u}) \\ &= L f(\mathbf{x}) \end{aligned} \tag{2.83}$$

for any $\mathbf{x} \in \Omega$. Thus $T^{-1} = L$.

(2) Since the isometric isomorphic mappings T and T' map the dense subspaces of Hilbert spaces into Hilbert spaces, they possess unique isometric extensions given by

$$\begin{aligned} \widetilde{T}' : \mathcal{N}_{\Psi_L}(\Omega)' \mapsto \mathcal{L}_T := \widetilde{T}'(\mathcal{N}_{\Psi_L}(\Omega)') \subseteq \mathcal{N}_{\Psi}(\Omega)' \\ \widetilde{T} : \mathcal{N}_{\Psi_L}(\Omega) \mapsto \mathcal{F}_T := \widetilde{T}(\mathcal{N}_{\Psi_L}(\Omega)) \subseteq \mathcal{N}_{\Psi}(\Omega). \end{aligned} \tag{2.84}$$

We want to show that $\mathcal{L}_T = \mathcal{N}_{\Psi}(\Omega)'$ and $\mathcal{F}_T = \mathcal{N}_{\Psi}(\Omega)$. Indeed, since $\mathcal{N}_{\Psi_L}(\Omega)$ is complete and $\widetilde{T} : \mathcal{N}_{\Psi_L}(\Omega) \mapsto \mathcal{F}_T$ is an isometric isomorphism, the space $\mathcal{F}_T$ must also be complete. But this means $\mathcal{F}_T$ is a reproducing kernel Hilbert space with reproducing kernel Ψ. Thus, by Theorem 2.2.6, since both $\mathcal{F}_T$ and $\mathcal{N}_{\Psi}(\Omega)$ have reproducing kernel Ψ, we must have $\mathcal{F}_T = \mathcal{N}_{\Psi}(\Omega)$. Furthermore, since $\mathcal{F}_T = \mathcal{N}_{\Psi}(\Omega)$ is the image of $\mathcal{N}_{\Psi_L}(\Omega)$ under the mapping $\widetilde{T}$, which is the extension of the mapping $T : F_{\Psi_L}(\Omega) \mapsto \mathcal{F}_L^0$, this implies that $T(F_{\Psi_L}(\Omega)) = \mathcal{F}_L^0$ is dense in $\mathcal{N}_{\Psi}(\Omega)$.

Now to see that $\mathcal{L}_T = \mathcal{N}_{\Psi}(\Omega)'$, for an arbitrary $\lambda \in \mathcal{N}_{\Psi}(\Omega)'$, we show that $\lambda \in \mathcal{L}_T$ as well. Indeed, using the Riesz mapping R_{Ψ} we have $R_{\Psi}(\lambda) = \lambda_2\Psi(\cdot, \mathbf{x}) \in \mathcal{N}_{\Psi}(\Omega) = \mathcal{F}_T$. Since $\lambda_2\Psi(\cdot, \mathbf{x}) \in \mathcal{F}_T$, using the inverse Riesz map $(R_{\Psi}|_{\mathcal{L}_T})^{-1}$, we see that $(R_{\Psi}|_{\mathcal{L}_T})^{-1}(\lambda_2\Psi(\cdot, \mathbf{x})) = \lambda \in \mathcal{L}_T$ and therefore $R_{\Psi}|_{\mathcal{L}_T} = R_{\Psi}$. This implies $\mathcal{L}_T = \mathcal{N}_{\Psi}(\Omega)'$.

(3) Finally, we can use parts (1) and (2) to show that $\widetilde{T}^{-1} = L$ and $\|Lf\|_{\mathcal{N}_{\Psi_L}(\Omega)} = \|f\|_{\mathcal{N}_{\Psi}(\Omega)}$. We first show $\widetilde{T}^{-1} = L$ by using the fact that $\mathcal{F}_L^0$ is dense in $\mathcal{N}_{\Psi}(\Omega)$. Choose any $f \in \mathcal{N}_{\Psi}(\Omega)$. Then by density of $\mathcal{F}_L^0$ in $\mathcal{N}_{\Psi}(\Omega)$, we can choose a sequence $f_n = (\lambda_n \circ L)_2\Psi(\cdot, \mathbf{x}) \in \mathcal{F}_L^0$ with $\lambda_n \in \mathcal{L}^0$ such that $\|f - f_n\|_{\mathcal{N}_{\Psi}(\Omega)} \to 0$ as $n \to \infty$.

Since $\delta_{\mathbf{x}} \circ L \in \mathcal{N}_\Psi(\Omega)'$, this means that

$$|Lf(\mathbf{x}) - Lf_n(\mathbf{x})| \leq \|\delta_{\mathbf{x}} \circ L\|_{\mathcal{N}_\Psi(\Omega)'} \|f - f_n\|_{\mathcal{N}_\Psi(\Omega)} \to 0, \tag{2.85}$$

as $n \to \infty$. On the other hand, since $T^{-1}f_n(\mathbf{x}) = Lf_n(\mathbf{x})$ from part (1), we see that

$$|\widetilde{T}^{-1}f(\mathbf{x}) - Lf_n(\mathbf{x})| = |\widetilde{T}^{-1}f(\mathbf{x}) - T^{-1}f_n(\mathbf{x})| \leq \|\delta_{\mathbf{x}} \circ \widetilde{T}^{-1}\|_{\mathcal{N}_\Psi(\Omega)'} \|f - f_n\|_{\mathcal{N}_\Psi(\Omega)} \to 0. \tag{2.86}$$

Both (2.85) and (2.86) together imply that $L = \widetilde{T}^{-1}$. Thus for every $f \in \mathcal{N}_\Psi(\Omega)$, we have $Lf = \widetilde{T}^{-1}f \in \mathcal{N}_{\Psi_L}(\Omega)$ and since $\widetilde{T}^{-1} : \mathcal{N}_\Psi(\Omega) \mapsto \mathcal{N}_{\Psi_L}(\Omega)$ is an isometric isomorphism, $\|Lf\|_{\mathcal{N}_{\Psi_L}(\Omega)} = \|\widetilde{T}^{-1}f\|_{\mathcal{N}_{\Psi_L}(\Omega)} = \|f\|_{\mathcal{N}_\Psi(\Omega)}$.

Lastly, we show in a similar manner that $\widetilde{T}'(\lambda) = \lambda \circ L$ and $\|\lambda \circ L\|_{\mathcal{N}_\Psi(\Omega)'} = \|\lambda\|_{\mathcal{N}_{\Psi_L}(\Omega)'}$ for all $\lambda \in \mathcal{N}_{\Psi_L}(\Omega)'$. Choose any $\lambda \in \mathcal{N}_{\Psi_L}(\Omega)'$. By density of $\mathcal{L}_{0,\Psi_L}$ in $\mathcal{N}_{\Psi_L}(\Omega)'$, there is a sequence $\{\lambda_n\} \in \mathcal{L}_{0,\Psi_L}$ which converges to λ in $\mathcal{N}_{\Psi_L}(\Omega)'$ as $n \to \infty$. Thus for any $f \in \mathcal{N}_\Psi(\Omega)$, we have $Lf \in \mathcal{N}_{\Psi_L}(\Omega)$ and so

$$|(\lambda \circ L)f - (\lambda_n \circ L)f| \leq \|\lambda - \lambda_n\|_{\mathcal{N}_{\Psi_L}(\Omega)'} \|Lf\|_{\mathcal{N}_{\Psi_L}(\Omega)} \to 0,\ n \to \infty. \tag{2.87}$$

Now by definition we have $T'(\lambda) = \lambda \circ L$ so that

$$\begin{aligned} |\widetilde{T}'(\lambda)(f) - (\lambda_n \circ L)f| &\leq \|\widetilde{T}'(\lambda - \lambda_n)\|_{\mathcal{N}_\Psi(\Omega)'} \|f\|_{\mathcal{N}_\Psi(\Omega)} \\ &= \|\lambda - \lambda_n\|_{\mathcal{N}_\Psi(\Omega)'} \|f\|_{\mathcal{N}_\Psi(\Omega)} \to 0,\ n \to \infty. \end{aligned} \tag{2.88}$$

Both (2.87) and (2.88) imply that $\widetilde{T}'(\lambda) = \lambda \circ L$. Thus for $\lambda \in \mathcal{N}_{\Psi_L}(\Omega)'$, we know that $\lambda \circ L = \widetilde{T}'(\lambda) \in \mathcal{N}_\Psi(\Omega)'$ and $\|\lambda \circ L\|_{\mathcal{N}_\Psi(\Omega)'} = \|\lambda\|_{\mathcal{N}_{\Psi_L}(\Omega)'}$. This ends the proof. □

Now that we have a connection between the kernels Ψ and Ψ_L and their native spaces, we now discuss their power functions which is the next step in obtaining error estimates for the interpolant $u_{h,\Lambda}$. We will use a slightly different notation than in the previous section which will allow for more general power functions. Let $\lambda \in \mathcal{N}_\Psi(\Omega)'$ be any functional on $\mathcal{N}_\Psi(\Omega)$ and let $\Lambda_N = \{\lambda_1, \dots, \lambda_N\}$ be any set of N functionals on $\mathcal{N}_\Psi(\Omega)$ and define $\mathcal{N}'_{\Psi,N} := \text{span}\{\Lambda_N\} \subset \mathcal{N}_\Psi(\Omega)'$. Then the power function with respect to Ψ and Λ_N is defined by

$$P_{\Psi,\Lambda_N}(\lambda) := \inf_{\mu \in \mathcal{N}'_{\Psi,N}} \|\lambda - \mu\|_{\mathcal{N}_\Psi(\Omega)'}, \tag{2.89}$$

where we recall the norm on $\mathcal{N}_\Psi(\Omega)'$

$$\sqrt{(\lambda, \lambda)_{\mathcal{N}_\Psi(\Omega)'}} := \sqrt{\lambda_1 \lambda_2 \Psi(\mathbf{x}, \mathbf{y})}.$$

For example, if $\lambda = \delta_{\mathbf{x}}$, for any $\mathbf{x} \in \Omega$, and $\mu = \sum_{i=1}^N \beta_i \delta_{\mathbf{x}_i}$ for certain coefficients $\boldsymbol{\beta} \in \mathbb{R}^N$, then it is easy to see that (2.89) is equivalent to the definition of the power function given in (2.46) with $\alpha = (0,0)$. This more generalized definition of the power function describes how well the functional $\lambda \in \mathcal{N}_\Psi(\Omega)'$ can be approximated by functionals from $\mathcal{N}'_{\Psi,N}$.

We can now prove the following properties of this power function which will then be used in obtaining error estimates for the symmetric meshless collocation method. The first is a generalized best approximation result.

Theorem 2.3.3. *Let $u \in \mathcal{N}_\Psi(\Omega)$ and $\mathcal{N}'_{\Psi,N} := span\{\Lambda_N\} = span\{\lambda_1, \dots, \lambda_N\} \subset \mathcal{N}_\Psi(\Omega)'$ and suppose that $u_{h,\Lambda} \in \mathcal{N}_{\Psi,N}$ satisfies $\lambda_j(u_{h,\Lambda}) = \lambda_j(u)$ for all $\lambda_j \in \Lambda_N$. Then for any $\lambda \in \mathcal{N}_\Psi(\Omega)'$ we have the bound*

$$|\lambda(u - u_{h,\Lambda})| \leq \inf_{\mu \in \mathcal{N}'_{\Psi,N}} \|\lambda - \mu\|_{\mathcal{N}_\Psi(\Omega)'} \cdot \inf_{s \in \mathcal{N}_{\Psi,N}} \|u - s\|_{\mathcal{N}_\Psi(\Omega)} \leq P_{\Psi,\Lambda_N}(\lambda) \|u\|_{\mathcal{N}_\Psi(\Omega)}. \tag{2.90}$$

Proof. Since every $\lambda_j \in \Lambda_N$ satisfies $\lambda_j(u - u_{h,\Lambda}) = 0$, so does every $\mu \in \mathcal{N}'_{\Psi,N}$. Choose any $\lambda \in \mathcal{N}_\Psi(\Omega)'$. Then we have

$$|\lambda(u - u_{h,\Lambda})| = \inf_{\mu \in \mathcal{N}'_{\Psi,N}} |(\lambda - \mu)(u - u_{h,\Lambda})| \leq \inf_{\mu \in \mathcal{N}'_{\Psi,N}} \|\lambda - \mu\|_{\mathcal{N}_\Psi(\Omega)'} \cdot \|u - u_{h,\Lambda}\|_{\mathcal{N}_\Psi(\Omega)}. \tag{2.91}$$

Now for any $s = \sum_{i=1}^N \alpha_i \lambda_i \Psi(\cdot, \mathbf{x}) \in \mathcal{N}_{\Psi,N}$, we get

$$(u - u_{h,\Lambda}, s)_{\mathcal{N}_\Psi(\Omega)} = (u - u_{h,\Lambda}, \sum_{i=1}^N \alpha_i \lambda_i \Psi(\cdot, \mathbf{x}))_{\mathcal{N}_\Psi(\Omega)} = \sum_{i=1}^N \alpha_i \lambda_i (u - u_{h,\Lambda}) = 0. \tag{2.92}$$

Since $s \in \mathcal{N}_{\Psi,N}$ is orthogonal to $u - u_{h,\Lambda}$ in $\mathcal{N}_\Psi(\Omega)$, we have $\|u - u_{h,\Lambda}\|_{\mathcal{N}_\Psi(\Omega)} = \inf_{s \in \mathcal{N}_{\Psi,N}} \|u - s\|_{\mathcal{N}_\Psi(\Omega)} \leq \|u\|_{\mathcal{N}_\Psi(\Omega)}$ which gives our desired result. □

The next Theorem gives a transformation result for the power function.

Theorem 2.3.4. *Suppose that* $\Lambda = \{\lambda_1, \dots, \lambda_N\} \subseteq \mathcal{N}_{\Psi_L}(\Omega)'$ *and* $\lambda \in \mathcal{N}_{\Psi_L}(\Omega)$ *are given. Then*

$$P_{\Psi_L,\Lambda}(\lambda) = P_{\Psi,\Lambda\circ L}(\lambda \circ L).$$

Proof. The proof follows directly by the definition of the power function and Theorem 2.3.2. We have

$$\begin{aligned} P_{\Psi,\Lambda\circ L}(\lambda \circ L) &= \inf_{\mu \in \text{span}\{\Lambda\circ L\}} \|\lambda \circ L - \mu\|_{\mathcal{N}_\Psi(\Omega)'} = \inf_{\mu \in \text{span}\{\Lambda\}} \|\lambda \circ L - \mu \circ L\|_{\mathcal{N}_\Psi(\Omega)'} \\ &= \inf_{\mu \in \text{span}\{\Lambda\}} \|\lambda - \mu\|_{\mathcal{N}'_{\Psi_L}} = P_{\Psi_L,\Lambda}(\lambda). \end{aligned} \tag{2.93}$$

□

Finally, we have a so-called splitting technique for power functions, which allows us to bound a power function on a set of functionals Λ by the power function on a smaller set of functionals $\Lambda_1 \subset \Lambda$.

Theorem 2.3.5. *(Wendland [66], Chapter 16) Suppose that* $\Lambda = \cup_j \Lambda_j$ *where* $\Lambda_j \subset \mathcal{N}'_\Psi$ *is a finite set of functionals for all* j*. Then*

$$P_{\Psi,\Lambda}(\lambda) \leq P_{\Psi,\Lambda_j}(\lambda)$$

for all j *and* $\lambda \in \mathcal{N}_\Psi(\Omega)'$*.*

Proof. Choose any $\lambda \in \mathcal{N}_\Psi(\Omega)'$. We use the definition of the power function and the fact that $\text{span}\{\Lambda_j\} \subset \text{span}\{\Lambda\}$ to get

$$\begin{aligned} P_{\Psi,\Lambda}(\lambda) = \inf_{\mu \in \text{span}\{\Lambda\}} \|\lambda - \mu\|_{\mathcal{N}_\Psi(\Omega)'} \leq \inf_{\mu \in \text{span}\{\Lambda_j\}} \|\lambda - \mu\|_{\mathcal{N}_\Psi(\Omega)'} \\ = P_{\Psi,\Lambda_j}(\lambda). \end{aligned} \tag{2.94}$$

□

The importance of this Theorem is that it allows us to consider functionals separately in the interior of the domain Ω and on the boundary $\partial\Omega$, which will be helpful when deriving the error estimates since we will be able to derive the bounds independently on Ω and $\partial\Omega$.

2.3.1 Deriving an error bound for symmetric meshless collocation

We now show how to bound the error bound for symmetric collocation by using the Theorems from the previous section along with the error bound result from Theorem 2.2.15. The procedure for obtaining the error bound will be to derive the bounds separately on Ω and $\partial\Omega$ and then use a maximum principle to get a global error bound.

We begin by defining sets of point evaluation functionals on Ω and $\partial\Omega$ as $\Delta_1 = \{\delta_{\mathbf{x}_1}, \ldots, \delta_{\mathbf{x}_n}\}$ and $\Delta_2 = \{\delta_{\mathbf{x}_{n+1}}, \ldots, \delta_{\mathbf{x}_N}\}$. We then let $\Lambda_1 = \{\delta_{\mathbf{x}_1} \circ L, \ldots, \delta_{\mathbf{x}_n} \circ L\} = \Delta_1 \circ L$, $\Lambda_2 = \Delta_2$ and set $\Lambda := \Lambda_1 \cup \Lambda_2$.

To get an error bound in Ω, choose any $\mathbf{x} \in \Omega$, and define $\lambda = (\delta_{\mathbf{x}} \circ L) \in \mathcal{N}_\Psi(\Omega)'$. By Theorems 2.3.3, 2.3.5, and 2.3.4, we have the following string of inequalities.

$$\begin{aligned} |\lambda(u - u_{h,\Lambda})| &\leq P_{\Psi,\Lambda}(\delta_{\mathbf{x}} \circ L)\|u\|_{\mathcal{N}_\Psi(\Omega)} 2.3.3) \\ &\leq P_{\Psi,\Lambda_1}(\delta_{\mathbf{x}} \circ L)\|u\|_{\mathcal{N}_\Psi(\Omega)} \text{ (thm. 2.3.5)} \\ &= P_{\Psi,\Delta_1 \circ L}(\delta_{\mathbf{x}} \circ L)\|u\|_{\mathcal{N}_\Psi(\Omega)} \qquad (2.95) \\ &= P_{\Psi_L,\Delta_1}(\delta_{\mathbf{x}})\|u\|_{\mathcal{N}_\Psi(\Omega)} \text{ (thm. 2.3.4)} \\ &= P_{\Psi_L,\mathcal{X}_1}(\mathbf{x})\|u\|_{\mathcal{N}_\Psi(\Omega)}. \end{aligned}$$

where $P_{\Psi_L,\mathcal{X}_1}(\mathbf{x})$ is the power function for the SPD kernel $\Psi_L \in \mathcal{N}_{\Psi_L}(\Omega)$ originally defined in (2.46) with kernel Ψ_L and $\alpha = (0,0)$, namely

$$\Big[P_{\Psi_L,\mathcal{X}_1}(\mathbf{x})\Big]^2 = \|\Psi_L(\cdot, \mathbf{x}) - \sum_{i=1}^{n} \tilde{v}_i(\mathbf{x})\Psi_L(\cdot, \mathbf{x}_i)\|^2_{\mathcal{N}_{\Psi_L}(\Omega)}. \qquad (2.96)$$

This shows that

$$|Lu(\mathbf{x}) - Lu_{h,\Lambda}(\mathbf{x})| \leq P_{\Psi_L,\mathcal{X}_1}(\mathbf{x})\|u\|_{\mathcal{N}_\Psi(\Omega)} \qquad (2.97)$$

for any $\mathbf{x} \in \Omega$.

We can apply a similar string of inequalities to arrive at an error bound on the boundary $\partial\Omega$. Let $\lambda = \delta_{\mathbf{x}}$ for any $\mathbf{x} \in \partial\Omega$. Then we have

$$\begin{aligned} |\lambda(u - u_{h,\Lambda})| &\leq P_{\Psi,\Lambda}(\delta_{\mathbf{x}})\|u\|_{\mathcal{N}_\Psi(\Omega)} \\ &\leq P_{\Psi,\Lambda_2}(\delta_{\mathbf{x}})\|u\|_{\mathcal{N}_\Psi(\Omega)} \\ &= P_{\Psi,\Delta_2}(\delta_{\mathbf{x}})\|u\|_{\mathcal{N}_\Psi(\Omega)} \\ &= P_{\Psi,\mathcal{X}_2}(\mathbf{x})\|u\|_{\mathcal{N}_\Psi(\Omega)}. \end{aligned} \tag{2.98}$$

This shows that

$$|u(\mathbf{x}) - u_{h,\Lambda}(\mathbf{x})| \leq P_{\Psi,\mathcal{X}_2}(\mathbf{x})\|u\|_{\mathcal{N}_\Psi(\Omega)}, \tag{2.99}$$

for any $\mathbf{x} \in \partial\Omega$. At first glance, we see that if we can bound the power function by a constant $C > 0$ times the saturation parameter $h_{\partial\Omega,\mathcal{X}_2}$, we will get our desired result. However, it is not so clear how to bound $P_{\Psi,\mathcal{X}_2}(\mathbf{x})$ since it deals with arguments $\mathbf{x}$ which lie on the boundary $\partial\Omega$, where a cone condition is not satisfied. Thus we need a couple more assumptions on Ω before we can proceed to bound $P_{\Psi,\mathcal{X}_2}(\mathbf{x})$. We first need that Ω is polygonal.

Definition An open bounded set $\Omega \subseteq \mathbb{R}^2$ is said to be a simple polygonal set if it is the intersection of a finite number of half spaces. A half space in $\mathbb{R}^2$ is a set $H_{a,b} = \{\mathbf{x} \in \mathbb{R}^2 \ : \ a^T\mathbf{x} < b\}$ with $a \in \mathbb{R}^2/\{0\}$ and $b \in \mathbb{R}$. A domain $\Omega \subset \mathbb{R}^2$ is said to be a polygonal domain if it is the union of a finite number of simple polygonal set.

The useful property of polygonal domains that we will need is that the boundary of a polygonal domain is the union of a finite number of lines. To bound $P_{\Psi,\mathcal{X}_2}(\mathbf{x})$, we will take advantage of this fact, along with the following Lemma.

Lemma 2.3.1. *Let $\Omega \subseteq \mathbb{R}^2$ be any measurable set and suppose that $\Psi \in C^{2k}(\Omega \times \Omega)$ is an SPD kernel defined by $\Psi(\mathbf{x}, \mathbf{y}) := \Psi_0(\mathbf{x} - \mathbf{y})$ for some radial function Ψ_0. Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega$ is a set of pairwise distinct points. Furthermore, suppose that $T : \mathbb{R}^2 \mapsto \mathbb{R}^2$ is a bijective affine mapping, i.e. $T\mathbf{x} = S\mathbf{x} + c$, $\mathbf{x} \in \mathbb{R}^2$, with an invertible matrix $S \in \mathbb{R}^{2\times 2}$ and a constant $c \in \mathbb{R}^2$. Then the following relation holds for the power function:*

$$P_{\Psi,\mathcal{X}}(\mathbf{x}) = P_{\Psi\circ S^{-1},T(\mathcal{X})}(T\mathbf{x}), \quad \forall \mathbf{x} \in \Omega. \tag{2.100}$$

Here $T(\mathcal{X})$ denotes the set $\{T\mathbf{x}_1, \ldots, T\mathbf{x}_N\}$ and $\Psi \circ S^{-1}$ denotes $\Psi(S^{-1}\mathbf{x}, S^{-1}\mathbf{y}) := \Psi_0(S^{-1}(\mathbf{x} - \mathbf{y}))$ for any $\mathbf{x}, \mathbf{y} \in \Omega$.

Proof. Let $\mathcal{W} = T(\mathcal{X})$ and $\mathbf{w} = T\mathbf{x}$ for any $\mathbf{x} \in \Omega$. Since $\mathbf{w}_i - \mathbf{w}_j = T\mathbf{x}_i - T\mathbf{x}_j = S(\mathbf{x}_i - \mathbf{x}_j)$, we obviously have that

$$(\Psi\circ S^{-1})(\mathbf{w}_i, \mathbf{w}_j) = \Psi(S^{-1}\mathbf{w}_i, S^{-1}\mathbf{w}_j) := \Psi_0(S^{-1}(\mathbf{w}_i - \mathbf{w}_j)) = \Psi_0(S^{-1}S(\mathbf{x}_i - \mathbf{x}_j)) = \Psi(\mathbf{x}_i, \mathbf{x}_j). \tag{2.101}$$

Thus the collocation matrices $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{T(\mathcal{X})}$ for the kernels Ψ and $\Psi \circ S^{-1}$, respectively, are the same (recall that $\mathcal{A}_{\mathcal{X}}[i, j] = \Psi(\mathbf{x}_i, \mathbf{x}_j)$). Furthermore, for any given $\mathbf{x} \in \Omega$, we have that the vector $R_{\Psi,\mathcal{X}}(\mathbf{x}) := (\Psi(\mathbf{x}, \mathbf{x}_1), \ldots, \Psi(\mathbf{x}, \mathbf{x}_N))^T$ is the same as $R_{\Psi\circ S^{-1},T(\mathcal{X})}(\mathbf{w}) = (\Psi(S^{-1}\mathbf{w}, S^{-1}\mathbf{w}_1), \ldots, \Psi(S^{-1}\mathbf{w}, S^{-1}\mathbf{w}_1))^T$. These two results imply that the interpolant functions $\tilde{v}_j(\mathbf{x})$ and $\tilde{v}_j(T\mathbf{x})$, $j = 1, \ldots, N$, defined by $\tilde{\mathbf{v}}(\mathbf{x}) = \mathcal{A}_{\mathcal{X}}^{-1} R_{\Psi,\mathcal{X}}(\mathbf{x})$ and

$$\tilde{\mathbf{v}}(T\mathbf{x}) = \mathcal{A}_{T(\mathcal{X})}^{-1} R_{\Psi\circ S^{-1},T(\mathcal{X})}(T\mathbf{x})$$

(originally defined in (2.35)) are the same. Now by the definition of the power function, we get

$$
\begin{aligned}
&[P_{\Psi,\mathcal{X}}(\mathbf{x})]^2 = \Psi(\mathbf{x},\mathbf{x}) - 2\sum_{i=1}^{N} \tilde{v}_i(\mathbf{x})\Psi(\mathbf{x},\mathbf{x}_i) + \sum_{i,j=1}^{N} \tilde{v}_i(\mathbf{x})\tilde{v}_j\Psi(\mathbf{x}_i,\mathbf{x}_j)\\
&= \Psi(S^{-1}\mathbf{w},S^{-1}\mathbf{w}) - 2\sum_{i=1}^{N} \tilde{v}_i(T\mathbf{x})\Psi(S^{-1}\mathbf{w},S^{-1}\mathbf{w}_i) + \sum_{i,j=1}^{N} \tilde{v}_i(T\mathbf{x})\tilde{v}_j(T\mathbf{x})\Psi(S^{-1}\mathbf{w}_i,S^{-1}\mathbf{w}_j)\\
&= [P_{\Psi\circ S^{-1},T\mathcal{X}}(T\mathbf{x})]^2,
\end{aligned}
\tag{2.102}
$$

which finishes the proof. □

As we will see, this Lemma enables us to compute an error bound on the boundary $\partial\Omega$. We can finally give the desired error estimates for the boundary-valued elliptic problem (2.64).

Theorem 2.3.6. *([66], 15.15) Let $\Omega \subseteq \mathbb{R}^2$ be a polygonal domain and $\Psi \in C^{2k}(\Omega\times\Omega)$ for $k \geq 1$ be an SPD kernel. Suppose that the boundary-valued problem*

$$
\begin{aligned}
Lu &= f \quad \text{in } \Omega\\
u &= g \quad \text{on } \partial\Omega
\end{aligned}
\tag{2.103}
$$

has the unique solution $u \in \mathcal{N}_\Psi(\Omega) \subset C^k(\Omega)$ for a given $f \in C(\Omega)$ and $g \in C(\partial\Omega)$. Let $u_{h,\Lambda}$ be the interpolant $u_{h,\Lambda} = \sum_{i=1}^{n}\alpha_i L_2\Psi(\mathbf{x},\mathbf{x}_i) + \sum_{i=n+1}^{N}\alpha_i\Psi(\mathbf{x},\mathbf{x}_i)$ where the coefficients α_i satisfy the linear system (2.72). Then the following error estimates

$$
\begin{aligned}
|Lu(\mathbf{x}) - Lu_{h,\Lambda}(\mathbf{x})| &\leq Ch^{k-2}_{\Omega,\mathcal{X}_1}\|u\|_{\mathcal{N}_\Psi(\Omega)}, \quad \mathbf{x}\in\Omega\\
|u(\mathbf{x}) - u_{h,\Lambda}(\mathbf{x})| &\leq Ch^{k}_{\partial\Omega,\mathcal{X}_2}\|u\|_{\mathcal{N}_\Psi(\Omega)}, \quad \mathbf{x}\in\partial\Omega
\end{aligned}
\tag{2.104}
$$

are satisfied for a sufficiently dense set of collocation points $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ and constant $C > 0$.

Proof. Firstly, the kernel Ψ_L corresponding to the elliptic operator L is an SPD kernel and in $C^{2k-4}(\Omega \times \Omega)$, as was shown above Theorem 2.3.2. The assumptions on Ψ imply that $C_{\Psi_L}(\mathbf{x})$ from the Theorem 2.2.15 is uniformly bounded on all of $\overline{\Omega}$, and we can thus apply the error estimate from Theorem 2.2.15 with $|\alpha| = 2$ to get a bound on the power function $P_{\Psi_L,\mathcal{X}_1}(\mathbf{x})$ in terms of $C_{\Psi_L}(\mathbf{x})$ and $h_{\Omega,\mathcal{X}_1}$ leading to

$$|Lu(\mathbf{x}) - Lu_{h,\Lambda}(\mathbf{x})| \leq P_{\Psi_L,\mathcal{X}_1}(\mathbf{x})\|u\|_{\mathcal{N}_\Psi(\Omega)} \leq Ch_{\mathcal{X}_1}^{k-2}\|u\|_{\mathcal{N}_\Psi(\Omega)}, \quad \mathbf{x} \in \Omega \tag{2.105}$$

where the constant $C > 0$ satisfies $C_{\Psi_L}(\mathbf{x}) < C$ for any $\mathbf{x} \in \Omega$.

For the boundary, as shown in (2.99) we already have

$$|u(\mathbf{x}) - u_{h,\Lambda}(\mathbf{x})| \leq P_{\Psi,\mathcal{X}_2}(\mathbf{x})\|u\|_{\mathcal{N}_\Psi(\Omega)}, \quad \mathbf{x} \in \partial\Omega. \tag{2.106}$$

We proceed to bound the power function $P_{\Psi,\mathcal{X}_2}(\mathbf{x})$ in terms of $h_{\mathcal{X}_2}$. However, as already mentioned, we cannot directly apply Theorem 2.2.15 since it requires a cone condition in the region where $\mathbf{x}$ comes from, and $\partial\Omega$ does not satisfy a cone condition. Fortunately, we can take a slightly different route by noticing that since $\partial\Omega$ is polygonal, it is a collection of a finite number of lines $H \subset \bar{H} = \{\mathbf{y} \in \mathbb{R}^2 \ : \ a^T\mathbf{y} = b\}$, each of which can be mapped to $\mathbb{R}$ where a cone condition is trivially satisfied. For each line $H \in \partial\Omega$, we construct a bijective affine mapping $T : H \mapsto \mathbb{R}$, to a line segment on the real line, call it $T(H)$. The mapping is defined by $T\mathbf{x} = S\mathbf{x} + c$ for some $c \in \mathbb{R}^2$ and an invertible matrix $S \in \mathbb{R}^{2\times 2}$. We now proceed to bound the error on any of these line segments. To this end, let H be any line segment of $\partial\Omega$. Define the set of collocation points $\mathcal{Y} = \mathcal{X}_2 \cap H = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ (we assume for simplicity that $\mathcal{Y}$ contains at least one point) and let $\mathcal{W} := T(\mathcal{Y}) = \{w_1, \ldots, w_M\}$ be the image of T applied to the points (thus $\mathcal{W}$ are points on the real line). If

$h_{\partial\Omega,\mathcal{X}_2}$ is sufficiently small then we can find for any $\mathbf{x} \in H$ a $\mathbf{y}_j \in \mathcal{Y}$ such that $\|\mathbf{x} - \mathbf{y}_j\|_2 \leq 2h_{\mathcal{X}_2}$. Hence we have, for $w = T\mathbf{x}$,

$$\|w - w_j\|_2 = \|T\mathbf{x} - T\mathbf{y}_j\|_2 = \|S(\mathbf{x} - \mathbf{y}_j)\|_2 \leq \|S\|\|\mathbf{x} - \mathbf{y}_j\|_2 \leq Ch_{\mathcal{X}_2}$$

which means that $h_{T(H),\mathcal{W}} \leq h_{\mathcal{X}_2}$, where $h_{T(H),\mathcal{W}}$ is the point saturation measure on $T(H)$ with respect to $\mathcal{W}$. Now since the set $T(H) \subset \mathbb{R}$ satisfies a cone condition, by Theorem 2.2.15 with $\alpha = 0$, we can bound the power function $P_{\Psi\circ S^{-1},\mathcal{W}}(w) \leq Ch^k_{T(H),\mathcal{W}} \leq Ch^k_{\partial\Omega,\mathcal{X}_2}$ for some constant $C > 0$ dependent on Ψ. Now we can apply Lemma 2.3.1 to get a bound estimate on the boundary. For $\mathbf{x} \in H$, we have using 2.99 and Lemma 2.3.1,

$$\begin{aligned} |u(\mathbf{x}) - u_{h,\Lambda}(\mathbf{x})| &\leq P_{\Psi,\mathcal{X}_2}(\mathbf{x})\|u\|_{\mathcal{N}_\Psi(\Omega)} \leq P_{\Psi,\mathcal{Y}}(\mathbf{x})\|u\|_{\mathcal{N}_\Psi(\Omega)} \\ &\leq P_{\Psi\circ S^{-1},\mathcal{W}}(w)\|u\|_{\mathcal{N}_\Psi(\Omega)} \leq Ch^k_{\partial\Omega,\mathcal{X}_2}\|u\|_{\mathcal{N}_\Psi(\Omega)} \end{aligned} \tag{2.107}$$

Since this can be done for any hyperplane H, and since the number of hyperplanes is finite, this gives us the desired result. □

As we see in Theorem 2.3.6, a good approximation in the interior should require the set $\mathcal{X}_1$ to be finely discretized, such that the error bound in the interior is of the same order as on the boundary. This means the interior should more finely discretized than the boundary, and a good choice is obviously $h^{k-2}_{\Omega,\mathcal{X}_1} \approx h^k_{\partial\Omega,\mathcal{X}_2}$.

Since L is of elliptic type, we can invoke the maximum principle for elliptic operators (cf. Grisvard [30]) which states that

$$\|u - u_{h,\Lambda}\|_{L_\infty(\Omega)} \leq \|u - u_{h,\Lambda}\|_{L_\infty(\partial\Omega)} + C\|Lu - Lu_{h,\Lambda}\|_{L_\infty(\Omega)} \tag{2.108}$$

for a sufficiently large constant $C > 0$ dependent on the coefficient functions $a_{j,k}$ of the operator L. This leads to the following corollary.

Corollary 2.3.1. *If in addition to the assumptions made in Theorem (2.3.6), we have*

$$h = \max\{h_{\Omega,\mathcal{X}_1}, h_{\partial\Omega,\mathcal{X}_2}\}.$$

Then by using the maximum principle for the elliptic operator L in Ω, we have

$$\|u - u_{h,\Lambda}\|_{L_\infty(\Omega)} \leq Ch^{k-2}\|u\|_{\mathcal{N}_\Psi(\Omega)}. \tag{2.109}$$

2.4 Numerical Experiments of Meshless Collocation

2.4.1 Introduction

In this section, we illustrate the implementation and numerical properties related to the symmetric meshless collocation (SMC) method for numerically solving elliptic PDEs. Through a series of experiments designed to investigate the issues of convergence and approximation ability along with the versatility of the method, we attempt to demonstrate that the symmetric meshless collocation method can be used as a robust and easy to implement alternative to the standard finite-element method for numerically solving boundary-valued elliptic PDEs. In particular, we wish to gain insight into the numerical rate of convergence for the collocation method as well as investigate the dependence of the convergence rate on the smoothness of the kernel Ψ and if the numerical convergence rate agrees with the theoretical one given by equation 2.109.

We are also interested in the question of approximation refinement in the collocation method. Namely, how should the number of collocation nodes in the domain be refined to produce a better approximation without the expense of sacrificing stability? We will compare approximation ability of the meshless collocation method with different quasi-uniform collocation node distributions (namely random, uniform, or Gaussian) in Ω and $\partial\Omega$ to determine if there is an optimal distribution that one should use.

Lastly, we would like to compare the symmetric meshless collocation method with the standard finite-element method on a few different test problems to assess the differences in numerical convergence and stability issues. One of the major advantages of meshless collocation over traditional and generalized finite element methods (see [7] for example) is that meshless collocation does not require numerical quadrature for integration since integration is not performed in the collocation method. A grand challenge in the current trend of generalized finite element methods comes in selecting optimal quadrature points and weights for the numerical integration over supports of the underlying basis functions. The introduction of quadrature then of course leads to additional numerical errors which can propogate throughout the global approximation. The fact that no integration is done in collocation greatly simplifies the implementation of the method and allows much freedom in selecting the basis kernels and the placement of collocation nodes. Furthermore, since no triangularization or rectangularization of the domain is required, collocation can be used on much more general smooth manifolds without the problem of domain discretization errors.

Before we investigate the implementation and performance of the meshless collocation method with numerical experiments and compare with the finite-element method, we must first discuss the type of symmetric positive definite kernels that will be used for constructing the native space $\mathcal{N}_\Psi(\Omega)$ and consequently, the finite dimensional space $\mathcal{N}_{\Psi,N}(\Omega)$.

2.4.2 Choice of Reproducing Kernels

A critical component of the symmetric meshless collocation method for the numerical solution of elliptic PDEs is the choice of symmetric positive definite kernel used to construct the native space $\mathcal{N}_\Psi(\Omega)$. In this numerical study, we will restrict ourselves to a class of SPD kernels $\Psi(\mathbf{x},\mathbf{y}) := \Psi_0(\mathbf{x}-\mathbf{y})$ which have compact support and are built from radial functions that can in fact be represented as polynomials on the interval $[0,1]$. This class of kernels was developed by Wendland in [63] and have been shown in the recent literature to have powerful approximation ability along with fast summation techniques (see Wendland [66], Chapter 9).

As mentioned, the kernels are built from radial functions $\Psi_0 \in L^1(\mathbb{R}^2) \cap C(\mathbb{R}^2)$, where $\Psi_0(\cdot) := \psi(\|\cdot\|)$, that have the form

$$\psi(r) = \begin{cases} p(r) & 0 \le r \le 1 \\ 0 & r > 1 \end{cases} . \tag{2.110}$$

The function p is a univariate polynomial of the form $p(r) = \sum_{j=0}^{m} c_j r^j$ with $c_m \neq 0$. The degree of ψ (and consequently Ψ_0) is m.

By construction, it is easy to see that Ψ_0 has compact support on $\|\mathbf{x}\| \le 1$ (the norm $\|\cdot\|$ will be taken to be the standard Euclidean l^2 norm unless otherwise noted) and thus the kernel $\Psi(\mathbf{x},\mathbf{y}) := \Psi_0(\mathbf{x}-\mathbf{y})$ has compact support since only values such that $\|\mathbf{x}-\mathbf{y}\| \le 1$ for $\mathbf{x},\mathbf{y} \in \mathbb{R}^2$ will be nonzero. Furthermore, by introducing a *fixed scaling parameter* $\epsilon > 0$, we can shrink or expand the support of Ψ as desired simply by defining

$$\Psi_\epsilon(\mathbf{x}, \mathbf{y}) := \Psi_0(\frac{\mathbf{x} - \mathbf{y}}{\epsilon}).$$

Changing the support of Ψ_ϵ will ultimately have an effect on the approximation since a larger support for Ψ will imply that more collocation points in $\mathcal{X} \subset \Omega$ will be used in determining the coefficients in the collocation approximation. Unfortunately, despite attempts in the literature, there is currently no analytical method of deriving an optimal shape parameter $\epsilon > 0$ for a given kernel Ψ and domain Ω with collocation points $\mathcal{X} \subset \Omega$ which minimizes the $L^\infty(\Omega)$ error norm. The only way to obtain a near optimal shape parameter is numerically. We will see in the next subsection how the shape parameter $\epsilon > 0$ influences the accuracy of the meshless collocation approximation. Obviously, the optimal shape parameter is also clearly dependent on the kernel used as well as the collocation nodes.

We give a few examples of Wendland's compactly supported functions $\psi_k(r)$ along with their associated smoothness space $C^{2k}(\mathbb{R}^2)$. Here the function $(1-r)_+$ means 0 if $1-r < 0$ and $(1-r)$ otherwise. A brief overview of the construction and

Table 2.1: Examples of ψ_k

$d \leq 3$	$\psi_1(r) = (1-r)_+^4(4r+1)$	C^2
$d \leq 3$	$\psi_2(r) = (1-r)_+^6(35r^2+18r+3)$	C^4
$d \leq 3$	$\psi_3(r) = (1-r)_+^8(32r^3+25r^2+8r+1)$	C^6

the properties of Wendland's compactly supported kernels can be found in Appendix C.10. For a complete detailed analysis of Wendland's compactly supported kernels, [66] Chapter 9 is recommended.

2.4.3 Implementation and numerical experiments for Elliptic Boundary-Valued Problems

In this section we describe the implementation of the SMC method in detail and apply it to a few boundary-valued elliptic problems to assess the robustness of the symmetric meshless collocation method for different distributions of collocation nodes, different boundaries, and different smoothness characteristics of the given data $f \in C(\Omega)$.

In our first experiment, we consider the simple Poisson problem with Dirichlet boundary condition

$$\begin{aligned} \Delta u(x,y) &= -\frac{5}{4}\pi^2 \sin(\pi x)\cos(\frac{\pi y}{2}), \quad (x,y) \in \Omega = [0,1]^2 \\ u(x,y) &= \sin(\pi x), \quad (x,y) \in \Gamma_1 \\ u(x,y) &= 0, \quad (x,y) \in \Gamma_2, \end{aligned} \tag{2.111}$$

where $\Gamma_1 = \{(x,y) : 0 \leq x \leq 1,\ y = 0\}$ and $\Gamma_2 = \partial\Omega/\Gamma_1$. The exact solution is given as $u(x,y) = \sin(\pi x)\cos(\frac{\pi y}{2})$.

Before we begin discussing the computational results, we discuss the implementation of the SMC method introduced in section 2.3 while giving explicit formulas for the collocation matrices based on the basis functions provided in the previous subsection.

Recall that the SMC approach seeks to approximate the solution u by the expansion

$$u_h = \sum_{j=1}^{n} \alpha_j L_2 \Psi(\cdot, \mathbf{x}_j) + \sum_{j=n+1}^{N} \alpha_j \Psi(\cdot, \mathbf{x}_j)$$

where $\mathbf{x}_j$ for $1 \leq j \leq n$ are the interior collocation nodes and $\mathbf{x}_j$ for $n+1 \leq j \leq N$ are the boundary collocation nodes. The approximating SPD kernel is Wendland's compactly supported kernel defined by $\Psi(\cdot, \mathbf{x}_j) := \psi_k(\| \cdot - \mathbf{x}_j)\|_2)$ with ψ_k being defined in Table 2.1. Since $L = \Delta$ in this example, we have

$$u_h = \sum_{j=1}^{n} \alpha_j \Delta \psi(\| \cdot - \mathbf{x}_j\|_2) + \sum_{j=n+1}^{N} \alpha_j \psi(\| \cdot - \mathbf{x}_j\|_2).$$

Substituting this expansion into the boundary-value problem (2.111) and enforcing the approximation to satisfy the PDE at the interior and boundary collocation nodes $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ we get

$$\begin{aligned} Lu_h(\mathbf{x}_j) &= f(\mathbf{x}_j),\ \mathbf{x}_j \in \mathcal{X}_1, \\ u_h(\mathbf{x}_j) &= g(\mathbf{x}_j),\ \mathbf{x}_j \in \mathcal{X}_2, \end{aligned} \tag{2.112}$$

we then arrive to the system

$$\begin{pmatrix} A & C \\ C^T & D \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} f|_{\mathcal{X}_1} \\ g|_{\mathcal{X}_2} \end{pmatrix}, \tag{2.113}$$

where the block matrices are defined by

$$\begin{aligned} A_{i,j} &= \Delta^2 \psi(\|\mathbf{x}_i - \mathbf{x}_j\|_2),\ 1 \leq i, j \leq n, \\ C_{i,j} &= \Delta \psi(\|\mathbf{x}_i - \mathbf{x}_j\|_2),\ n+1 \leq i \leq N, 1 \leq j \leq n, \\ D_{i,j} &= \psi(\|\mathbf{x}_i - \mathbf{x}_j\|_2),\ n+1 \leq i, j \leq N, \end{aligned} \tag{2.114}$$

and the vectors $f|_{\mathcal{X}_1}$ and $g|_{\mathcal{X}_2}$ are the given interior and boundary functions, respectively, evaluated on the collocation nodes. As an example, if we let $\phi := \phi_3 \in C^6(\Omega)$ from Table (2.1), then the functions in the block matrices of (2.114) are given by

$$\begin{aligned} \psi(r) &= (1-r)_+^8(32r^3+25r^2+8r+1) \\ \Delta\psi(r) &= 44(1-r)_+^6(88r^3+3r^2-6r-1) \\ \Delta^2\psi(r) &= 1056(1-r)_+^4(297r^3-212r^2+16r+1). \end{aligned} \tag{2.115}$$

Due to the symmetric positive definite property of the collocation matrix $\mathcal{A}_{\mathcal{X}}$, the coefficients $\boldsymbol{\alpha}$ in (2.113) can readily be obtained by either a direct method or a conjugate gradient method. Of course, the decision to use one method over the other should be influenced by the number of collocation nodes in $\mathcal{X}_N$. For smaller problems $N \sim \mathcal{O}(10^3)$, a direct method is usually very fast and efficient. In larger problems, the condition number of the matrix plays a role, and a conjugate gradient method should be employed. We discuss the stability and conditioning issues in the last section of the chapter.

2.4.3.1 Experiment 1

In our first experiment with the collocation method, our goal is to determine how the numerical approximation u_h depends on the distribution of collocation nodes in the domain Ω and boundary $\partial\Omega$. We will consider three types of collocation node distributions: a random distribution, a uniform distribution, and a Gauss-Lobatto-Legendre (GLL) distribution. Furthermore, to asses the numerical convergence, we let N range from 9 to 1089 for the total number of collocation nodes in $\mathcal{X}$ for all three distributions.

Figure 2.1 depicts the three different types of node distributions in Ω and $\partial\Omega$. The random distribution of nodes was computed simply by using a uniform random number generator in the interval $(0,1)$ and taking two draws for each (x,y) pair.

Table 2.2 shows the L^∞ errors for increasing N. As one can deduce, the uniform node distributions performs slightly better than the random and Gaussian distributions, although the differences between the three are marginal. For such a smooth problem on a square domain, we expect all three distributions to produce similar numerical results.

According to Theorem 2.3.6, the $L^\infty(\Omega)$ error of $|u - u_h|$ should behave like $Ch^{k-|\alpha|}\|u\|_{\mathcal{N}_\Psi(\Omega)}$ where in this case $k = 3$ (since $\Psi_0 \in C^6(\mathbb{R}^2)$) and $|\alpha| = 2$. For the uniform distribution of nodes, we can approximate the saturation parameter h for the set of nodes easily as it is approximately given by $1/\sqrt{N}$. We can approximate the norm of $\|u\|_{\mathcal{N}_\Psi(\Omega)}$ by simply projecting $u \in C(\Omega)$ onto $\mathcal{N}_{\Psi,N}(\Omega)$ giving $I_\mathcal{X} u$ for really large N and then computing $\|I_\mathcal{X} u\|^2_{\mathcal{N}_\Psi(\Omega)} = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \Psi(\mathbf{x}_i, \mathbf{x}_j)$. Here we let $N = 5,000$ giving $\|u\|_{\mathcal{N}_\Psi(\Omega)} \approx 5.2946$. This implies that the constant $C > 0$ in the $L^\infty(\Omega)$ error estimate ranges from $C \approx .0138$ for $N = 9$ to $C \approx 4.6730e-007$ for $N = 1089$.

Figure 2.2 shows the pointwise error (left) and the solution using 25 randomly scattered points in Ω. Local areas in Ω which are not as well covered by nodes gives the largest pointwise error as shown in the left plot.

Figure 2.3 depicts the poinwise errors for the random and uniform node distribtions in the case $N = 1089$. We see that the uniform grid of nodes locally

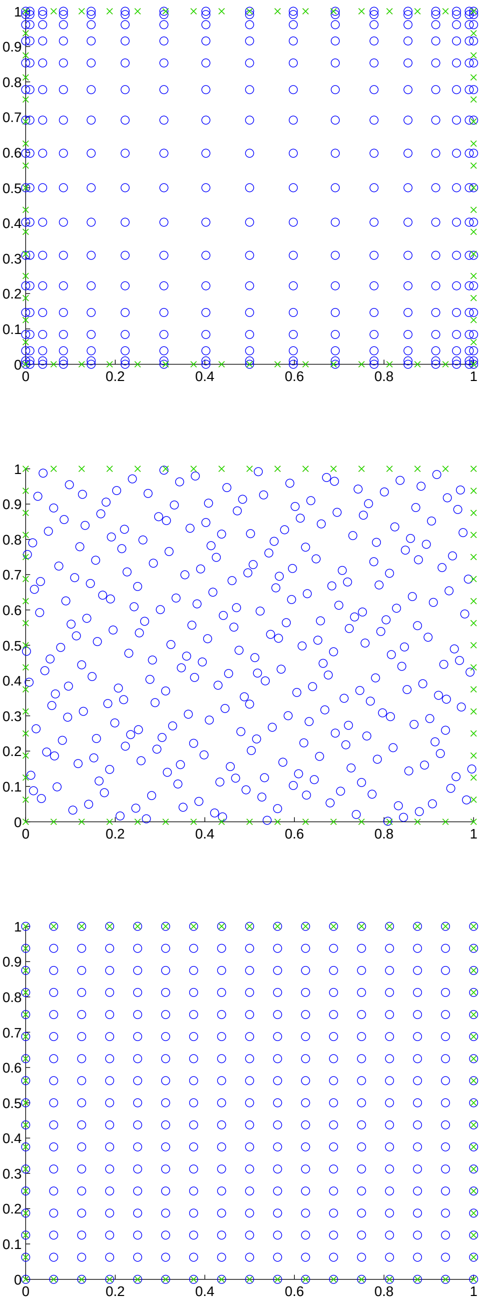


Figure 2.1: Examples of different node distributions (Gaussian, random, uniform).

Table 2.2: Numerical convergence of meshless collocation using $\psi_3 \in C^6$ radial function for kernel Ψ on random, Gaussian, and uniform collocation grids.

N	Random	Gaussian	Uniform
9	5.866837e-003	2.307935e-002	2.307935e-002
25	4.757992e-004	2.571894e-003	7.666688e-004
81	1.828029e-004	2.571894e-003	3.519404e-005
120	3.825086e-005	2.395948e-005	1.784373e-005
160	9.275238e-005	5.348320e-005	3.182937e-005
200	8.138042e-005	2.758392e-005	8.983736e-006
250	2.181343e-005	3.201837e-005	5.984730e-006
289	7.051553e-006	2.395948e-006	1.655152e-006
400	2.099321e-006	1.109739e-006	8.189373e-007
500	1.297749e-006	8.347463e-007	6.393703e-007
1089	8.681275e-008	7.334611e-008	7.080371e-008

approximates the data much better than in the random node distribution.

We now want to compare the approximation ability of the C^6 radial function ψ_3 with the C^4 radial function $\psi_2 \in C^4$ defined in Table 2.1. According to Thereom 2.3.6, the rate of convergence in the interior of the domain should be to the order of Ch^{k-2} where in this case $k = 2$. Table 2.3 shows the L^∞ errors for increasing N on the same collocation nodes as in example from Table 2.2. Clearly, the numerical rate of convergence is much better than the theoretical one offered in the Theorem.

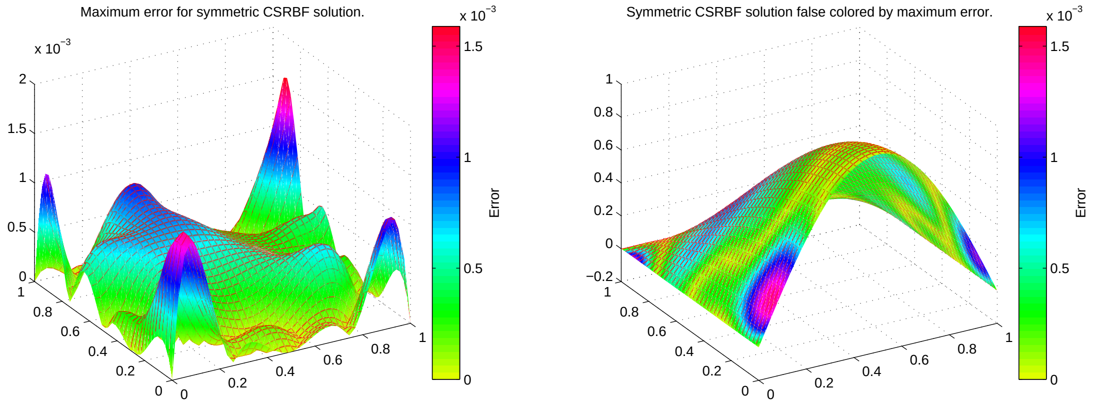


Figure 2.2: Pointwise error and approximation u_h using 25 randomly scattered points in Ω.

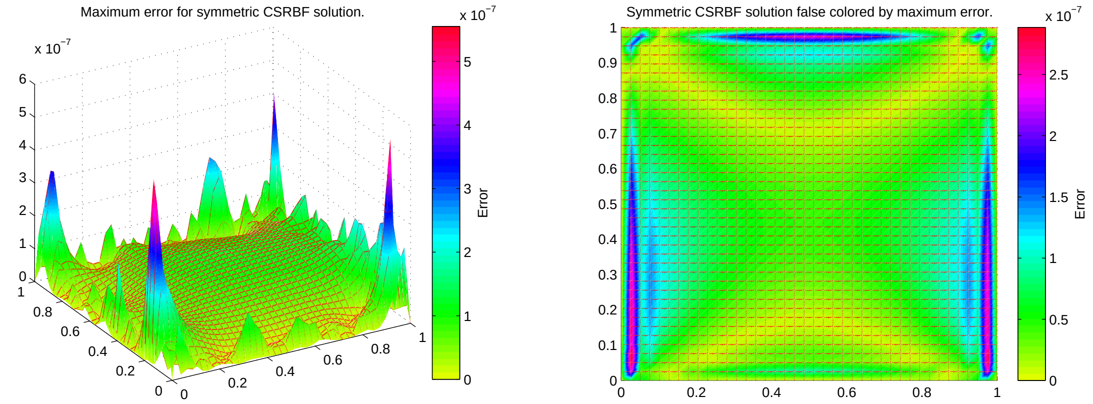


Figure 2.3: Pointwise error of 1089 collocation nodes for random (left) and uniform (right) distribution.

We can expect however that the convergence rate for ψ_2 is not as robust as ψ_3 do to the fact that the approximation $u_{h,\Lambda}$ is only in C^2 as opposed to C^4.

Table 2.3: Numerical convergence of meshless collocation using $\psi_2 \in C^4$ radial function for kernel Ψ on random, Gaussian, and unfiform collocation grids.

N	Random	Gaussian	Uniform
9	9.866837e-002	6.307935e-002	5.537935e-002
25	2.757992e-003	1.214294e-002	7.418338e-003
81	3.128029e-004	2.913423e-003	3.519404e-004
120	1.545086e-004	3.565948e-004	1.784373e-004
160	8.275238e-005	8.348323e-005	3.182937e-005
200	6.138042e-005	6.258391e-005	8.145673e-005
250	3.181343e-005	3.201837e-005	5.814730e-005
289	2.051553e-005	1.455942e-005	1.655152e-005
400	9.099321e-006	7.109739e-006	8.189373e-006
500	8.297749e-006	6.347463e-006	6.393703e-006
1089	7.681275e-006	5.334611e-006	1.080371e-006

We conclude that if a smooth solution to the elliptic problem is expected, then there is hardly an advantage to using ψ_3 over ψ_2 except for the slightly faster numerical convergence rate.

2.4.3.2 Experiment 2

In the next numerical experiment, we investigate the approximation ability of SMC in the case in which Ω is non-convex. We continue to consider the elliptic boundary-value problem (2.111) and take the domain Ω to be L-shaped as in figure 2.4. The problem is now

$$\begin{aligned}
\Delta u(x,y) &= -\frac{5}{4}\pi^2 \sin(\pi x)\cos(\frac{\pi y}{2}), \quad (x,y) \in \Omega, \\
u(x,y) &= \sin(\pi x), \quad 0 \leq x \leq 2, y = 0, \\
u(x,y) &= 0, \quad x = 0, 0 \leq y \leq 2, x = 2, 0 \leq y \leq 1, \\
u(x,y) &= \cos(\frac{\pi}{2})\sin(\pi x), \quad 1 \leq x \leq 2, y = 1, \\
u(x,y) &= 0, \quad 0 \leq x \leq 1, y = 2, \\
u(x,y) &= \cos(\pi y/2), \quad x = 1, 1 \leq y \leq 2,
\end{aligned} \tag{2.116}$$

Furthermore, we continue to utilize the C^6 compactly supported radial function ψ_3. We expect the numerical convergence of u_h to be the same as in the previous example since Theorem 2.3.6 only requires Ω to be polygonal, and not necessarily convex. Figure 2.4 shows two different collocation node configurations for the L-shaped domain.

Table 2.4 shows the L^∞ errors for increasing N. This time, we see a better improvement in convergence for the uniform grid as opposed to the random distribution of nodes. From this experiment, we clearly see that the non-convexity of Ω does not affect the approximation ability of the collocation method. Comparing tables 2.2 and 2.4, we see that the numerical convergence rates for the uniform grids are very similar. The approximation in the square domain is only slightly better.

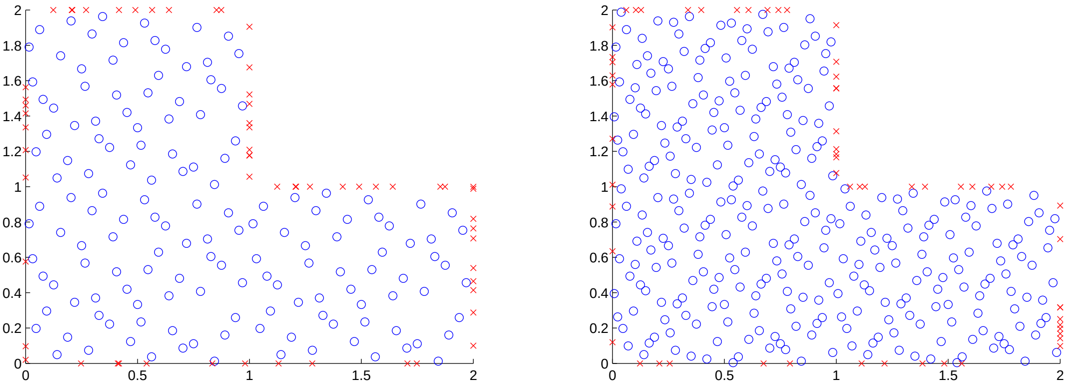


Figure 2.4: L-shaped domain with two different node configurations.

The plot in figure 2.5 show the SMC solution with $N = 100$ random scattered nodes in Ω along with the pointwise error plots for $N = 25$ and $N = 100$ random scattered nodes in figure 2.6. Since the scattered nodes are distributed more densly with $N = 100$ we see the collocation solution is much smoother.

We now want to study the convergence of u_h to the solution u for the same problem (2.111) in the L-shaped domain where we keep the internal nodes $\mathcal{X}_1$ fixed while refining only the boundary nodes $\mathcal{X}_2$. We want to see how the distribution of nodes on $\partial\Omega$ affects the global approximation. To do this, we initialize 30 randomly scattered nodes in Ω and 3 equidistant nodes on each boundary segment of $\partial\Omega$ where giving 18 total boundary nodes (total of 6 boundary segments for L-shaped domain). The refinement on the boundary approximation is done by adding 3 nodes to each boundary segment and then recomputing the approximation. Table 2.5 shows the errors as the total number of nodes on each boundary segment increases. The N_2 in the table represents the total number of nodes on the boundary $\partial\Omega$. The second column is the L^∞ error with 30 random nodes in Ω and the fourth column is the

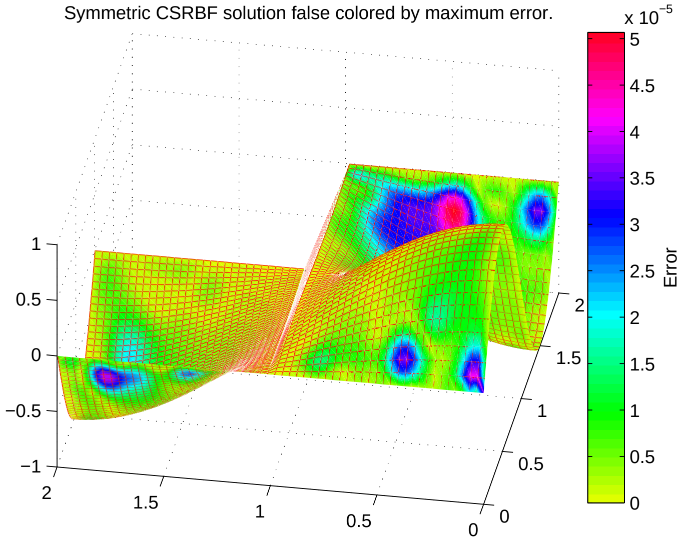


Figure 2.5: L-shaped domain approximation with $N = 100$.

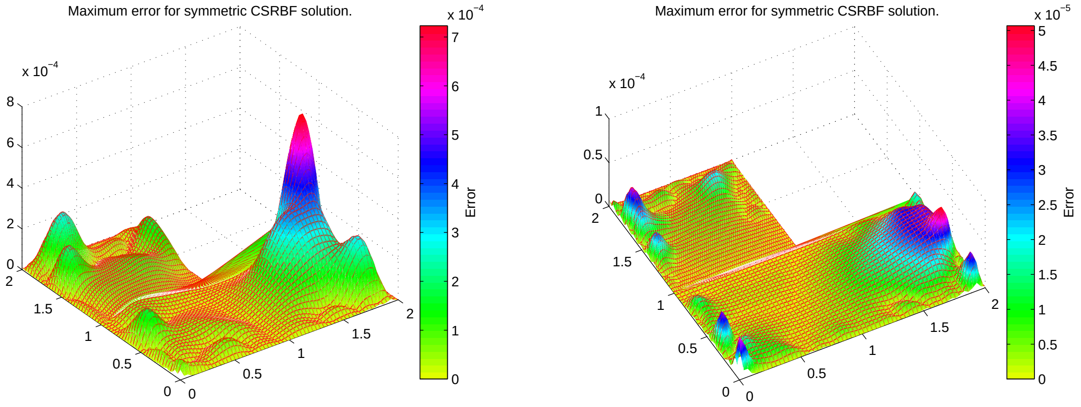


Figure 2.6: L-shaped domain pointwise error plot in two different node configurations with $N = 25$ and $N = 100$.

Table 2.4: Numerical convergence rate for elliptic problem on L-shaped domain.

N	Random	Uniform
50	8.521359e-004	3.7628298e-004
100	9.141483e-005	7.1284928e-005
150	2.990301e-005	4.7840294e-005
200	1.868492e-005	2.0847294e-005
250	1.875683e-005	6.7392034e-006
300	1.812304e-005	3.3871937e-006

L^∞ error with 120 random nodes. Figures 2.7 depict the refinement exclusively on the boundary of Ω.

As we can see, with the internal nodes fixed and the boundary nodes increasing, the improvement in numerical convergence is remarkable, although the rate begins to slow after about 36 total nodes, or 6 per boundary segment. In the 120 total internal node case, a similar convergence rate is seen. This experiment leads us to conclude that the approximation on the boundary influences dramatically the approximation in the interior of the domain. This contradicts the remark made directly below the proof of Theorem 2.3.6 where the we stated that interior node distribution should more finely discretized than the boundary, and a good choice is $h^{k-2}_{\Omega,\mathcal{X}_1} \approx h^k_{\partial\Omega,\mathcal{X}_2}$. Clearly in this example we see that this is indeed not the case and in fact the boundary $\partial\Omega$ should have a smaller saturation parameter $h_{\partial\Omega,\mathcal{X}_2}$ than that of the interior.

Table 2.5: Numerical convergence of L-shaped domain problem for boundary refinement.

N_2	L^∞ error	N_2	L^∞ error
12	5.231837e-002	36	3.7628298e-004
18	1.190819e-002	54	7.1284928e-005
36	9.556934e-004	72	4.7840294e-005
54	7.017452e-004	90	2.0847294e-005
72	6.942045e-004	108	6.7392034e-006

2.4.3.3 Experiment 3

In our third example, we wish to inquire about the approximating robustness of symmetric collocation in the case of a more general elliptic PDE with variable coefficients. We first consider a Helmoltz-type elliptic boundary value problem with smooth variable coefficients and Dirichlet boundary conditions:

$$\frac{\partial}{\partial x}\Big(a(x,y)\frac{\partial}{\partial x}u(x,y)\Big)+\frac{\partial}{\partial y}\Big(a(x,y)\frac{\partial}{\partial y}u(x,y)\Big)+u(x,y) = f(x,y), \quad (x,y) \in \Omega = [0,1]^2 \tag{2.117}$$

$$u(x,y) = 0, \quad (x,y) \in \Gamma = \partial\Omega \tag{2.118}$$

where $f(x,y) = -16x(1-(1-x)(3-2y)e^{x-y}) + 32y(1-y)(3x^2+y^2-x-2)$, and the coefficients are given by $a(x,y) = 2-x^2-y^2$, and $b(x,y) = e^{x-y}$ with exact solution $u(x,y) = 16x(1-x)y(1-y)$.

The operator L in this example is of course more complex than in the previous case and as before, one must first compute both $L_2\Psi(\mathbf{x},\mathbf{y}) = L\psi(\|\mathbf{x}-\mathbf{y}\|_2)$ and

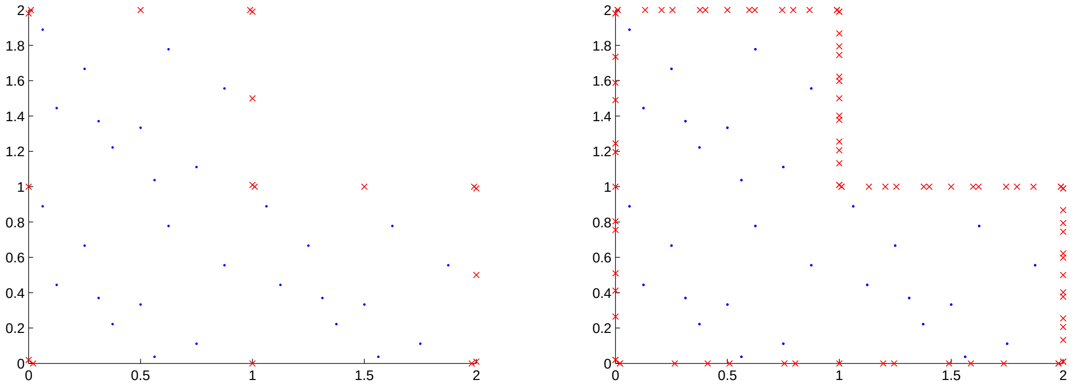


Figure 2.7: L-shaped domain with 30 interior nodes and 18 (left) and 54 (right) boundary nodes.

$L_1L_2\Psi(\mathbf{x},\mathbf{y}) = L^2\psi(\|\mathbf{x}-\mathbf{y}\|_2)$. The approximation to the solution $u(x,y)$ is then given by

$$u_h = \sum_{j=1}^{n} \alpha_j L\psi(\|\cdot - \mathbf{x}_j\|_2) + \sum_{j=n+1}^{N} \alpha_j \psi(\|\cdot - \mathbf{x}_j\|_2) \tag{2.119}$$

where

$$\begin{aligned} L\psi(\|\cdot - \mathbf{x}_j\|_2) = &\Big(a_x(x,y)\psi_x(\|\cdot - \mathbf{x}_j\|_2) + a(x,y)\psi_{xx}(\|\cdot - \mathbf{x}_j\|_2)\Big) \\ &+ \Big(a_y(x,y)\psi_y(\|\cdot - \mathbf{x}_j\|_2) + a(x,y)\psi_{yy}(\|\cdot - \mathbf{x}_j\|_2)\Big) \end{aligned} \tag{2.120}$$

and $\psi_x, \psi_{xx}, \psi_y, \psi_{yy}$ are the first and second order partial derivatives with respect to x and y. Since the basis function ψ is radial, we must use the chain rule to evaluate both partial derivatives. Thus, with $\mathbf{x} = (x,y)$ and $r = \|\mathbf{x}\| = \sqrt{x^2+y^2}$, we have

$$\frac{\partial}{\partial x}\psi(\|\mathbf{x}\|) = \frac{\partial}{\partial r}\psi(r)\frac{\partial}{\partial x}r(\mathbf{x}) = \frac{\partial}{\partial r}\psi(r)\frac{x}{\sqrt{x^2+y^2}} = \frac{x}{r}\frac{\partial}{\partial r}\psi(r)$$

and

$$\frac{\partial^2}{\partial x^2}\psi(\|\mathbf{x}\|) = \frac{x^2}{r^2}\frac{\partial^2}{\partial r^2}\psi(r) + \frac{y^2}{r^3}\frac{\partial}{\partial r}\psi(r).$$

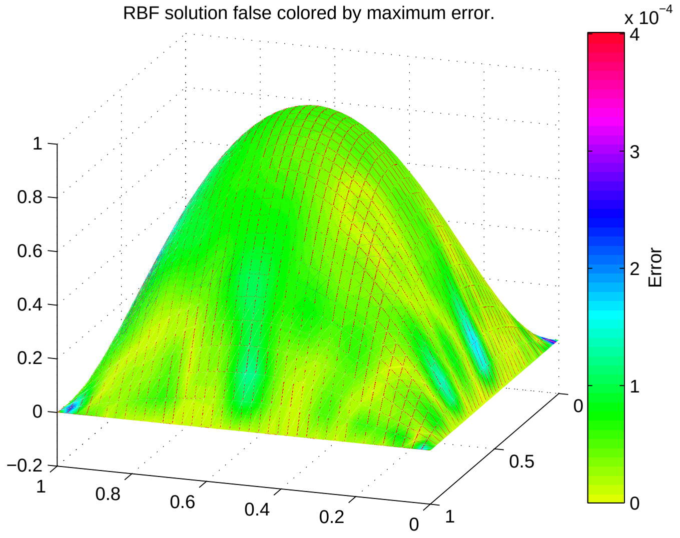


Figure 2.8: Collocation solution with $N = 286$ random nodes.

To test the numerical convergence of the SMC method on the Helmholtz equation with variable coefficients, we approximate the solution on the series of random collocation node distributions with $N = 9, 25, 81, 286, 1086$. Figure 2.8 shows the collocation solution with $N = 286$ randomly scattered nodes in Ω and on the boundary and the figures in 2.9.

Due to the boundedness and continuity of the variable coefficients $a(x, y)$ and $b(x, y)$ in Ω, the operator L is clearly elliptic and thus meets the conditions for the hypothesis of Theorem 2.3.6. The convergence result should therefore hold in this example. Table 2.6 shows the L^∞ error for an increasing number of nodes in the domain and boundary.

The rate of convergence is nearly identical to the rate of the non-variable coefficient elliptic boundary value problem discussed in the previous experiments. The same collocation node configurations were used from 2.2 and alhough we see

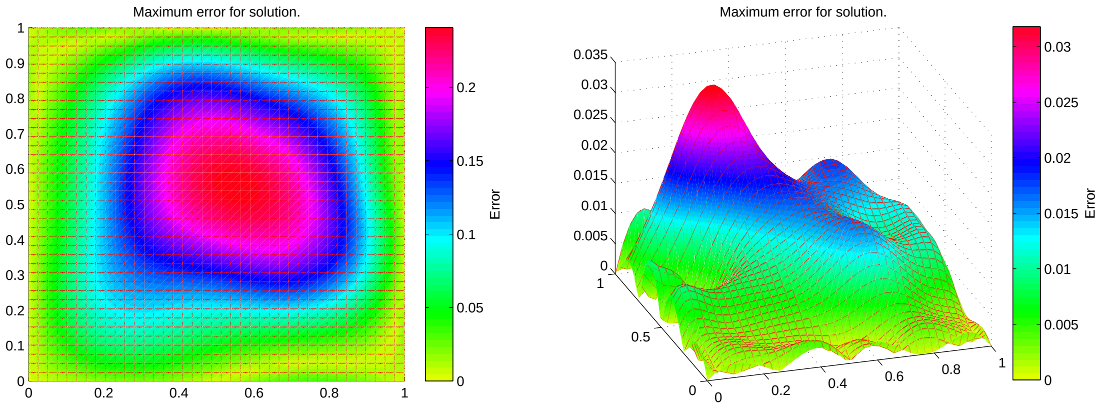


Figure 2.9: Pointwise error for $N = 9$ (left) and $N = 25$ (right) collocation nodes in Ω.

marginally larger L^∞ errors for the same collocation node configuration, the rate of convergence is still highly robust. We can safely conclude that the SMC is successful when applied to elliptic problems with variable coefficients, however at the cost of computing the differential operator L twice. But this can usually be accomplished easily with the help of a symbolic mathematics software package such as Maple.

2.4.4 Comparison with finite-element method

We continue the numerical section on the SMC method by comparing it computionally with the finite-element method (FEM) for three different elliptic problems. One of the themes of this thesis is to demonstrate numerically that meshless collocation methods can be an attractive alternative to classic FEM methods which utilize either a mesh, numerical quadrature, or both. We want to compare the numerical convergence of both methods along with the approximation properties of the solutions on different domains that could potentially be challenging for either approximation method. In all of the experiments, we use the finite element solver

Table 2.6: Numerical convergence for variable coefficient elliptic problem on three different collocation grids.

N	Random	Gaussian	Uniform
9	1.125837e-001	2.307935e-001	8.342139e-001
25	1.157992e-002	2.571894e-003	7.366688e-002
81	1.374529e-003	2.571894e-003	1.519404e-003
120	8.105086e-004	2.395948e-004	1.784373e-004
160	4.275238e-004	5.348320e-004	7.182937e-004
200	2.138042e-004	2.758392e-004	5.983736e-004
250	9.181343e-005	3.201837e-005	8.984730e-005
289	8.105108e-005	9.395948e-005	9.655152e-005
400	4.199321e-006	1.109739e-006	8.189373e-006
500	2.143453e-006	8.347463e-007	6.393703e-007
1089	7.083571e-008	9.334611e-008	9.080371e-008

toolkit in Matlab to construct piecewise linear element solutions.

2.4.4.1 Experiment 1

In our first experiment, we wish to solve the elliptic problem given in polar coordinates on the unit disk $r = \sqrt{x^2 + y^2} \leq 1$

$$\begin{aligned} -\Delta u(r,\theta) &= 4 - \frac{1}{r}, 0 \leq \theta \leq 2\pi \; r < 1 \\ u(1,\theta) &= 0, \; 0 \leq \theta \leq 2\pi \end{aligned} \tag{2.121}$$

where $u(r, \theta) = r(1 - r)$ is the exact solution.

The difficulty in approximating the solution for both methods lies in capturing the sharp point at $r = 0$. For the finite element approximation, we use piecewise linear elements on a uniform triangularization of the disk. The boundary of the disk using the edges of triangles will of course have a big impact on the resulting approximability. We chose 5 different meshes consisting of 50, 200, 500, 800, and 1080 triangles in the units disk. Figures 2.10 and 2.11 show the approximate solution using the finite element approximation. One can clearly see that the approximability at the center point $r = 0$ improves greatly as the number of piecewise linear elements increase.

We suspect that the finite-element method will be able to better handle the approximation at the center of the disk due to the fact that the collocation method seeks a solution in a finite dimensional subspace of $\mathcal{N}_\Psi(\Omega)$ which is a relatively smooth space due to the smoothness of the kernel Ψ. In order to verify this, we compute the collocation approximation where we take the set of collocation nodes to be the nodes of the triangles in the finite-element mesh. The figures in 2.12 show the meshless approximation at different angles for the first two sets of collocation nodes

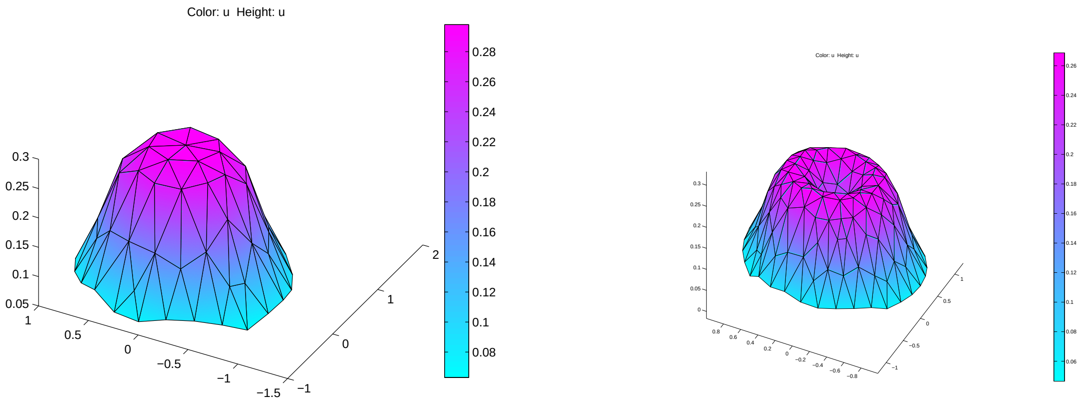


Figure 2.10: Piecewise-linear Finite element solution with 153 (left) and 310 (right) nodes from mesh.

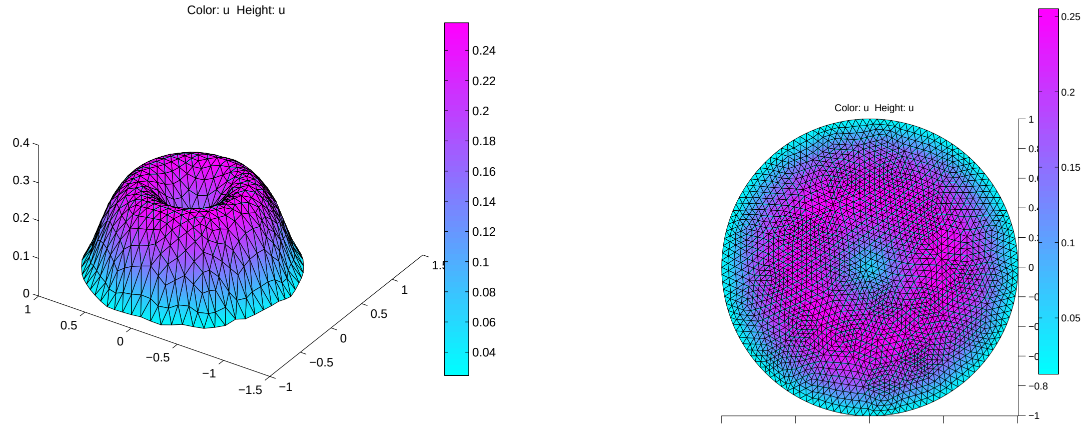


Figure 2.11: Piecewise-linear Finite element solution with 577 (left) and 789 (right) nodes from mesh.

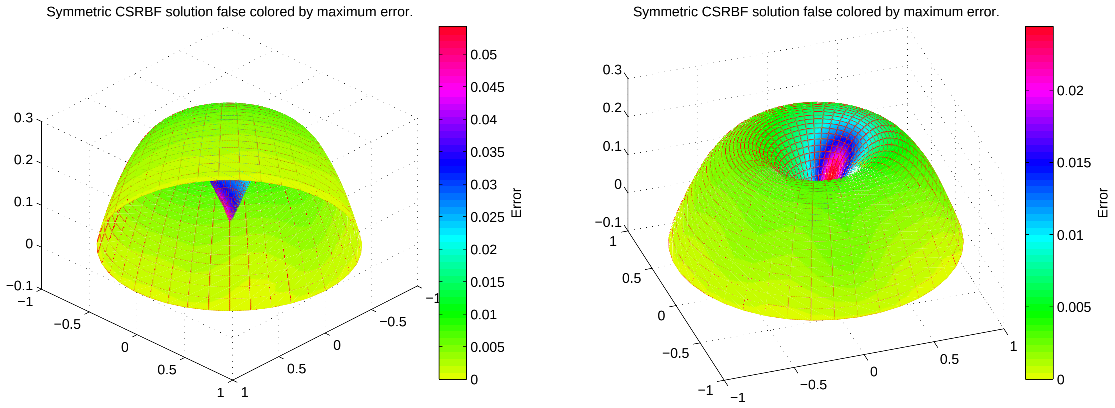


Figure 2.12: Meshless collocation solution with 153 (left) and 310 (right) collocation nodes in Ω.

generated from the first two element meshes. Again, since we are using smooth kernels in C^6 for the native space of the meshless collocation, we cannot expect the approximation at $r = 0$ to improve greatly where the sharp point is located. As one can clearly see, the largest errors in magnitude come near the center point of the disk. This was not the case with the finite element solution.

To compare the performance of the two methods, we evaluate the L^∞ error for each approximation at the corner nodes of each triangle element. Table 2.7 shows the results of the error analysis and we see that the SMC approximation is slightly better by a factor of about $c10^{-1}$ where $c < 1$ is some constant. The meshless method handles the boundary much better of course since no piecewise linear approximation of the unit disk is being done. Most of the larger errors in both approximations come at the center of the disk where the sharp point is centered. Since smooth kernels are being used in the SMC approximation, the method has much difficulty in approximating the sharp point at $r = 0$. This is clearly seen in the convergence rate of the L^∞ error since the rate steadily declines for an increase in nodes. The finite-element solution, however, sees a much more consistent rate of convergence due to its piecewise linear approximation.

Table 2.7: Comparing L^∞ errors of the FEM and SMC method.

Nodes	FEM L^∞ error	SMC L^∞ error
153	1.262e-001	1.902e-002
310	5.831e-001	9.283e-003
577	3.763e-002	3.349e-003
789	6.32de-003	2.012e-003
1180	1.023e-003	9.711e-004

2.4.4.2 Experiment 2

For the second example, we now wish to compare the numerical solution of both approximation methods in a problem where the domain Ω is non-smooth and non-convex. We consider the elliptic problem

$$\begin{gathered} -\Delta u(x,y) = f(x,y), \ (x,y) \in \Omega \\ u(x,y) = \sin(5r^2)\cos(10r^2) \ (x,y) \in \partial\Omega, \end{gathered} \tag{2.122}$$

where $f(x,y) = -\left(\frac{1}{r}\frac{d}{dr} + \frac{d^2}{d^2r}\right)(\sin(5r^2)\cos(10r^2))$, and again $r = \sqrt{x^2+y^2}$. The Ω we consider is a union of 4 different ovals creating the flower shaped domain shown in figure 2.13.

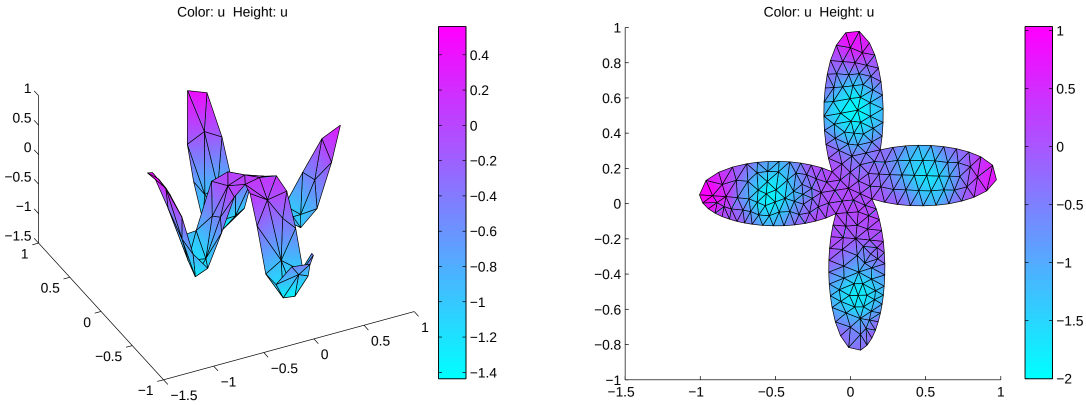


Figure 2.13: Finite element solution with 50 and 150 triangle discretizion Ω.

We again consider a uniform triangularization of the domain using 5 different meshes consisting of 50, 150, 400, 800, and 1500 triangles where the basis functions are piecewise linear across the elements.

To compare the performance of the two methods, we again evaluate the L^∞ error for each approximation at the corner nodes of each triangle element on the finite element mesh. Thus the collocation nodes used in the meshless approximation correspond to the triangle corner nodes of each element and the L^∞ error is computed on the same points. This is shown in Table 2.8.

Since the analytic solution is quite smooth, we see that the SMC performs better than the finite element method. The numerical convergence rate of the the meshless approximation, compared with the previous convergence rates, is only slightly dampened by the nonsmoothness of the domain. From these two examples, we can expect that the smoother the solution, the faster the convergence and more accurate solution for the SCM compared to FEM. We've seen that the smoothness and convexity of the domain, does not have much of an effect on the convergence rate

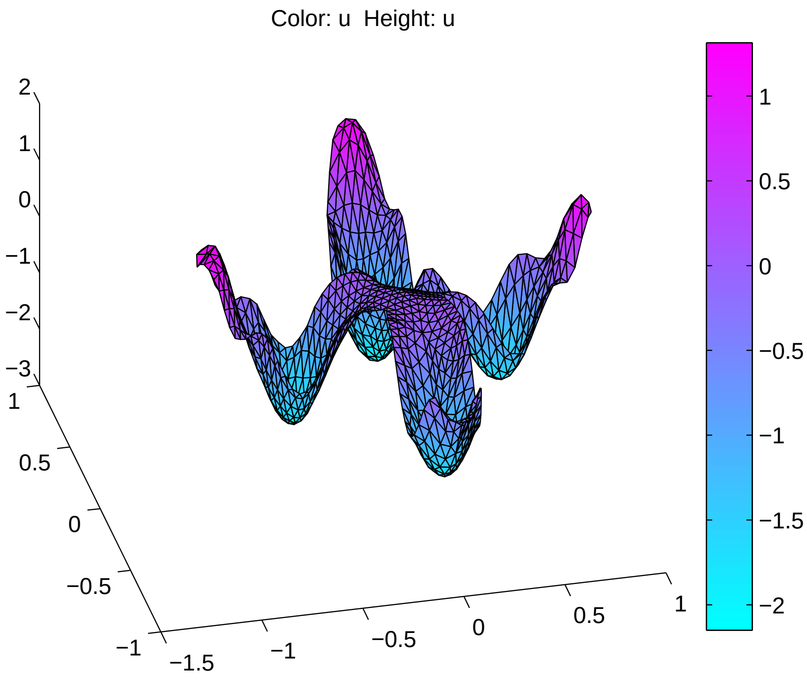


Figure 2.14: Finite element solution with 913 nodes from mesh in Ω.

either for meshless collocation. This is most likely due to the fact that no meshing of the domain is necessary and only dependent on the collocation nodes. In the FEM approximation, if a nonsmooth solution is expected, we saw that the numerical convergence rate is higher and the SMC numerical convergence rate slows significantly.

We have thus seen that the advantage of SMC over FEM, besides the fact that no mesh or numerical quadrature is needed, is in the simplicity of the implementation. Furthermore, for smooth problems we can expect the SMC method to be more accurate and converge faster. One drawback however is if singularities or discontinuities in the solution. We have seen that the SMC method will fail to converge in this case whereas the finite element approximation with piecewise linear elements

Table 2.8: Comparing L^∞ error of flower-shaped domain problem.

Nodes	FEM L^∞ error	SMC L^∞ error
76	1.1983e-001	1.1263e-002
253	7.7472e-002	6.7123e-003
913	2.6393e-003	4.2102e-004
3457	1.1635e-005	7.6162e-006
6890	8.2721e-006	2.1298e-008

has a much easier time handling such problems. There is one issue about the MC method that we still have yet to investigate which is the issue of numerical stability. We discuss this in the next final part of this numerical section.

2.4.5 Accuracy/Numerical Stability Trade-off

A well known problem in meshless collocation methods is the trade-off between the accuracy and the stability of the meshless collocation method. Theorem 2.3.6 in the previous section introduced a pointwise convergence rate for symmetric collocation based on the saturation parameter h of the collocation set $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$. The numerical experiments demonstrated that in fact, a much better rate can be achieved, even on nonsmooth and nonpolygonal domains. However, as we will now show, there is a tradeoff between these robust numerical convergence rates and the stability of the approximation.

Generally, the *stability* of the meshless collocation method is measured in terms of the condition number of the collocation matrix $\mathcal{A}_{\mathcal{X}}$ where the condition number is given as $\text{cond}(\mathcal{A}_{\mathcal{X}}) = \|\mathcal{A}_{\mathcal{X}}\|_2 \|\mathcal{A}_{\mathcal{X}}^{-1}\|_2$ which in most numerical applications is estimated by the ratio of the largest to smallest eigenvalue. We define $\lambda_{\min}(\mathcal{A}_{\mathcal{X}}) = \inf_{\boldsymbol{\alpha} \in \mathbb{R}^N} \frac{\boldsymbol{\alpha}^T \mathcal{A}_{\mathcal{X}} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{\alpha}} > 0$ as the minimum eigenvalue of the collocation matrix $\mathcal{A}_{\mathcal{X}}$ and is positive due to the positive definite property of $\mathcal{A}_{\mathcal{X}}$. To clearly see why $\lambda_{\min}(\mathcal{A}_{\mathcal{X}})$ is important in the stability of the collocation process, we know that if $\boldsymbol{\alpha}$ satisfies the collocation problem $\mathcal{A}_{\mathcal{X}} \boldsymbol{\alpha} = \mathbf{f}$ where $\mathbf{f} = (f|_{\mathcal{X}_1}, g|_{\mathcal{X}_2})$ is the interior and boundary data, then $\boldsymbol{\alpha}^T \mathcal{A}_{\mathcal{X}} \boldsymbol{\alpha} = \boldsymbol{\alpha}^T$ and hence

$$(\boldsymbol{\alpha}^T \boldsymbol{\alpha})^{1/2} = \|\boldsymbol{\alpha}\|_2 \leq \frac{1}{\lambda_{\min}(\mathcal{A}_{\mathcal{X}})} \|\mathbf{f}\|_2.$$

Thus the closer $\lambda_{\min}(\mathcal{A}_{\mathcal{X}})$ is to zero, then the less we know about the solution vector $\boldsymbol{\alpha}$ and the larger the condition number of $\mathcal{A}_{\mathcal{X}}$. In the following example, we will demonstrate that $\lambda_{\min}(\mathcal{A}_{\mathcal{X}})$ does in fact get closer to zero as h decreases.

We will attempt to demonstrate numerically that the greater the accuracy desired in the approximation, the closer the collocation matrix $\mathcal{A}_{\mathcal{X}}$ is to becoming ill-conditioned and eventually non-inversible using standard matrix inversion routines. To investigate this inverse relationship between the stability and convergence, we compare the evaluation of the L^∞ norm of the power function with respect to the differential operator L and the smallest eigenvalue of the collocation matrix $\mathcal{A}_{\mathcal{X}}$. Recall from the previous section that the power function for the modified kernel Ψ_L evaluated anywhere on Ω is defined by $P^2_{\Psi_L, \mathcal{X}_1}(\mathbf{x}) = \|\Psi_L(\cdot, \mathbf{x}) - \sum_{j=1}^{n} \tilde{v}_j(\mathbf{x}) \Psi_L(\cdot, \mathbf{x}_j)\|^2_{\mathcal{N}_{\Psi_L}(\Omega)}$ where $\mathcal{X}_1$ is the set of interior collocation nodes. The L^∞ norm of the power function $\|P_{\Psi_L, \mathcal{X}_1}\|_{L^\infty}$ was proven to be bounded by Ch^{k-2} for some constant C dependent on the kernel Ψ and where $k \geq 2$ is the smoothness of the kernel. The power function measures how well the finite sum $\sum_{j=1}^{n} \tilde{v}_j \Psi_L(\cdot, \mathbf{x}_j)$ approximates the kernel $\Psi_L(\cdot, \mathbf{x})$ in the native space. Of course, as the number of points in $\mathcal{X}$ increases, the power function should go pointwise to 0. We show that this is indeed the case, but at the cost of decreasing the smallest eigenvalue $\lambda_{\min}(\mathcal{A}_{\mathcal{X}})$ and thus rendering the matrix $\mathcal{A}_{\mathcal{X}}$ ill-conditioned.

To compute the power function in these numerical experiments, we use the fact that

$$\begin{aligned} &\|\Psi_L(\cdot, \mathbf{x}) - \sum_{j=1}^{n} \tilde{v}_j(\mathbf{x})\Psi_L(\cdot, \mathbf{x}_j)\|^2_{\mathcal{N}_{\Psi_L}(\Omega)} = \\ &\qquad \Psi_L(\mathbf{x}, \mathbf{x}) - 2\sum_{j=1}^{n} \tilde{v}_j(\mathbf{x})\Psi(\mathbf{x}, \mathbf{x}_j) + \sum_{j,i=1}^{n} \tilde{v}_i(\mathbf{x})\tilde{v}_j(\mathbf{x})\Psi_L(\mathbf{x}_i, \mathbf{x}_j). \end{aligned} \tag{2.123}$$

This is written in matrix-vector form as

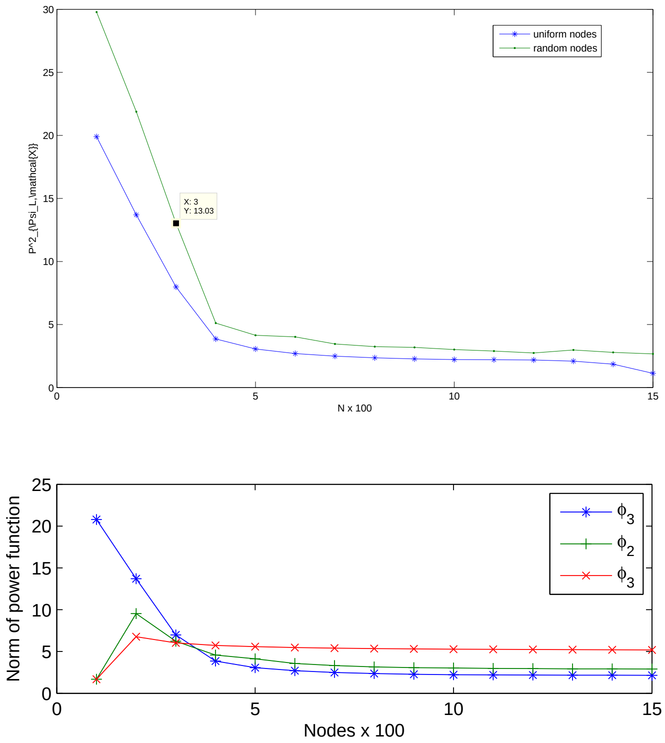
$$\tilde{\mathbf{v}}(\mathbf{x})^T \mathcal{A} \tilde{\mathbf{v}}(\mathbf{x}) - 2\tilde{\mathbf{v}}(\mathbf{x})^T \mathbf{R}(\mathbf{x}) + \Psi_L(\mathbf{x}, \mathbf{x}), \tag{2.124}$$

with matrix $\mathcal{A}[i, j] = \Psi_L(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{R}(\mathbf{x}) = (\Psi_L(\mathbf{x}, \mathbf{x}_1), \ldots, \Psi(\mathbf{x}_n))^T$. But since $\mathcal{A}\tilde{\mathbf{v}}(\mathbf{x}) = \mathbf{R}(\mathbf{x})$ by definition of the vector $\tilde{\mathbf{v}}(\mathbf{x})$, we can reduce the power function to

$$P_{\Psi_L, \mathcal{X}_1}(\mathbf{x}) = (\Psi_L(\mathbf{x}, \mathbf{x}) - \tilde{\mathbf{v}}(\mathbf{x})^T \mathbf{R}(\mathbf{x}))^{1/2}. \tag{2.125}$$

In these numerical examples we continue to use the positive definite kernel Ψ defined by the radial function $\psi_3 \in C^6(\mathbb{R}^2)$. We compute the power function on 15 different sets of collocation nodes containing $N = 100, 200, \ldots, 1500$ evenly distributed nodes in the square domain $\Omega = [0, 1]^2$. We compare this with $N = 100, 200, \ldots, 1500$ distributed nodes as well. Figure 2.15 shows the plot of the L^∞ norm of the power

Figure 2.15: (Left) L^∞ norm of the power function as N increases for uniform random grids. (Right) Condition number of $\mathcal{A}_\mathcal{X}$ on uniform and random grids.



function for increasing N. The plot on the right plots the condition number of the collocation matrix $\mathcal{A}_\mathcal{X}$ as N increases. Notice that the condition number quickly grows and then tapers off as N gets large past $N = 1000$. Since the nodes are getting closer and closer, with h decreasing, this implies the the rows are becoming more linearly dependent, thus rendering $\mathcal{A}_\mathcal{X}$ ill-conditioned.

Figure 2.16 shows two different angles of the plots of the power function $P^2_{\Psi_L,\mathcal{X}}$ for a uniform grid of collocation nodes $\mathcal{X}$ where $N = 100, 200, 300, 400$. The plots are layed on top of each other to see the effect of $P^2_{\Psi_L,\mathcal{X}}$ converging to 0 everywhere in Ω pointwise as N increases, as it should.

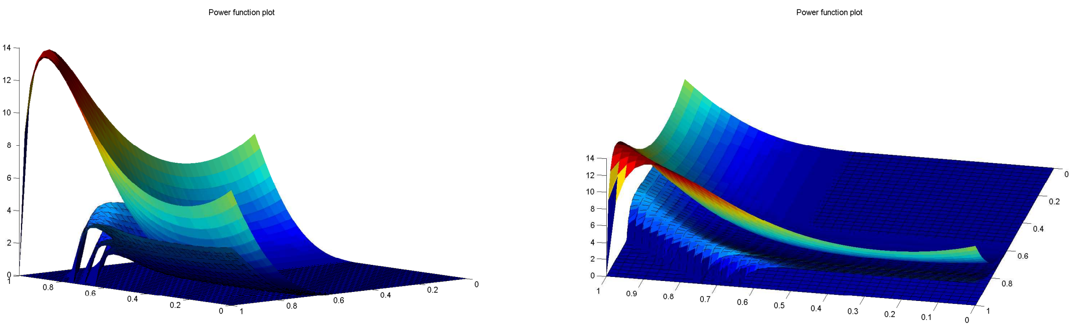


Figure 2.16: Power function on Ω for 4 different sets $\mathcal{X}$ with $N = 100, 200, 300, 400$. Two different angles shown.

The fact that as the points in $\mathcal{X}$ get closer together, the rows of $\mathcal{A}_{\mathcal{X}}$ become more linear dependent suggest that the smallest eigenvalue $\lambda_{\min}(\mathcal{A}_{\mathcal{X}})$ would be dependent on the node separation distance given by $q_{\mathcal{X}_N} := \frac{1}{2}\min_{j\neq k}\|\mathbf{x}_j^N - \mathbf{x}_k^N\|_2$. Wendland shows in [69] that in the case of compactly supported kernels constructed from Wendland's compactly supported radial functions ψ_k from Table 2.1 that the bound

$$\lambda_{\min}(\mathcal{A}_{\mathcal{X}}) \geq Cq^{2k-2|\alpha|-2}, \tag{2.126}$$

holds for some constant $C > 0$ where $|\alpha|$ is the order of the differential operator L. Thus according to Wendland's estimate, less smooth kernels should produce higher lower bounds for the smallest eighenvalue, but at the cost of having slower converging collocation approximations.

In this final experiment of the numerical section, we wish to investigate Wendland's lower bound for the minimal eigenvalue numerically. In this experiment however, to allow the use of the additional kernel ψ_1, we simply compute smallest eigenvalue for the case in which all the functionals are point evaluation functionals

$\lambda_j = \delta_{\mathbf{x}_j}$ in the interior instead of $\lambda_j = (\delta_{\mathbf{x}_j} \circ L)$. In this case, the lower bound for the minimal eigenvalue should simply be $Cq_{\mathcal{X}_N}^{2k-2}$. We compute the power function $P^2_{\Psi,\mathcal{X}}(\mathbf{x}) = \|\Psi(\cdot, \mathbf{x}) - \sum_{j=1}^{N} \tilde{v}_j(\mathbf{x})\Psi(\cdot, \mathbf{x}_j)\|^2_{\mathcal{N}_{\Psi_L}(\Omega)}$ for each kernel $\Psi_0(\cdot) = \psi_k(\|\cdot\|)$ and compare the smallest eigenvalue with the L^∞ norm of the power function and the separation distance $q_{\mathcal{X}_N}$. The interpolation matrix $\mathcal{A}_{\mathcal{X},k}$ then has entries $\mathcal{A}_{\mathcal{X},k}[i,j] = \Psi(\mathbf{x}_i, \mathbf{x}_j) = \psi_k(\|\mathbf{x}_i - \mathbf{x}_j\|)$. We compute this on 15 different uniform node grids and compare the performance in figure 2.17 for the ψ_1, ψ_2, and ψ_3, respectively. The plots show how the norm of the power function decreases as the condition number increases. In each plot, the rate of each kernel ψ_1, ψ_2, and ψ_3 is compared. Ideally, we would like to see the norm of the power function decrease at the same rate the condition number increases. However, this is not exactly the case.

We do see that in fact that the less smooth the kernel, the better conditioned the resulting interpolation matrix will be, but at the cost of a slightly less accurate solution. This phenomenon should be carefully considered when deciding what kernel should be used in computing numerical solutions to PDEs using meshless collocation. When one wants an accurate solution with a very limited number of collocation nodes in the domain, these numerical experiments suggest that a smoother kernel should be used to construct the native space and consequently the solution. However, if the solution is expected to be nonsmooth, the kernel built from ψ_1 should be considered as it will allow for higher amount of collocation nodes with a lower condition number.

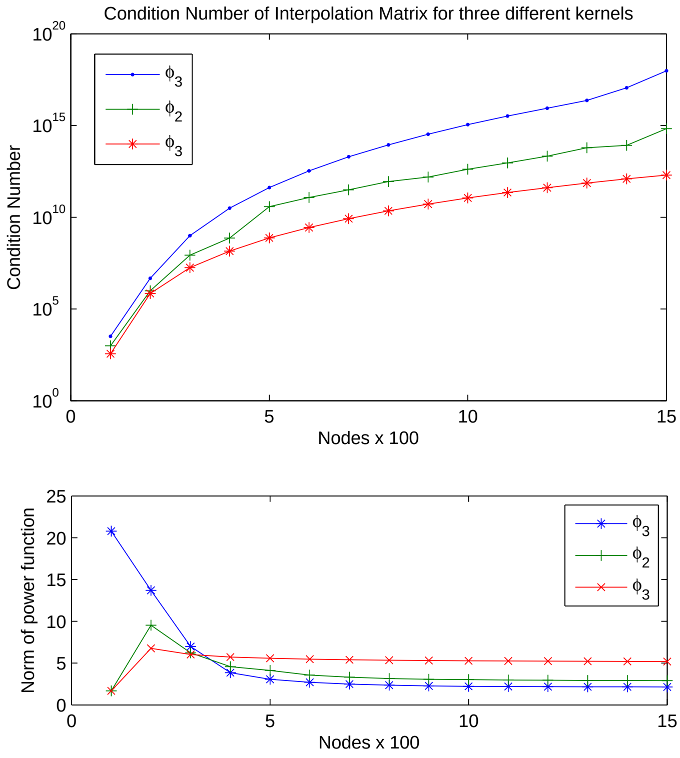


Figure 2.17: Condition numbers (top) and power function norm (bottom) for the three different kernels defined by ψ_1, ψ_2, ψ_3.

Chapter 3

A Meshless Collocation Method for the Rotational Shallow Water Equations on the Sphere

3.1 Introduction

In this second part of the thesis we construct a meshless collocation approximation method for the rotational shallow water equations on the sphere and apply it to a small suite of tests designed to characterize its strengths and weaknesses. The first part of the chapter deals with some theoretical issues of regional approximation that have been considered in past literature such as Staniforth [52] when developing a new regional modeling technique. We highlight some results of the theoretical issues which we later apply when discussing our numerical experiments. We then discuss the shallow-water model and its discretization on the sphere. Given that there are many ways of discretizing the nonlinear system of equations on the sphere, we aim at developing a so-called cubed-sphere model and discuss its advantages over other types of discretizations. The third part of the chapter then discusses in detail the implementation of meshless collocation on the shallow-water equations where we adopt a semi-implicit time stepping scheme which results in a mathematical structure well suited for high-performance parallelization. In the next chapter, we give an efficient implementation of the parallelization of the model where we also discuss hardware and software considerations that we use to run numerical simulations. The simulations are based on the standardized test cases proposed by Williamson et al. [71] primarily for experimenting with new approximation methods for the shallow-water equations on the sphere. The computational results are given which summarize the potential of the meshless collocation method for high-performance geophysical fluid dynamics applications. We then conclude with a final discussion on some future research endeavors stemming from the meshless collocation method proposed in this chapter.

3.2 Theoretical issues of regional approximation in global models

One of the primary goals of numerical climate and weather prediction is to construct high-resolution approximations over the largest possible areas in the domain of interest for a long period of time. Our main motivation in this chapter is to show computationally that the construction of a high-resolution approximation over large and smaller regional areas for a long period of time is possible with meshless collocation methods.

It is generally considered impractical in most meteorological models to use high resolution uniformly over the globe due to storage and computational time costs. Because of this, the study of regional approximation methods in global models for weather and climate has gained massive attention over the past two decades. High resolution is required over regions of interest namely to sufficiently represent small scales and processes which can affect the atmosphere's global evolution. Staniforth, Cote and others (see e.g. [18], [52]) claim that constructing regional approximations of the regions of interest can be done based on the fact that meteorological systems move with finite phase speeds. However, the size of the regional area largely depends on the simulated or forecasted time period. As suggested in [52] and shown computationally in [56], this would primarily be due to insufficient resolutions at the boundaries of the regional approximation causing poorly resolved features to propagate inwards at phase speeds dependent on the model and time stepping discretization. If mesh structures change too abruptly however (i.e. from coarse to fine and then to coarse), propagating waves governed by the hyperbolic system of equations can begin to see the addition of spurious modes stemming from the inconsistency between the mesh sizes. This will globally cause instability in the approximation over long periods of integration time if the mesh is not properly constructed. It is thus a grand importance when testing regional approximation schemes to not only vary the size and resolution of the localized region, but also to enforce a uniform transition from fine to coarse resolution and to test the time integration interval and its effect on the regional approximation accuracy.

Most of the current methods for regional approximation utilized in global climate and similar geophysical models stem from the area of mesh (or grid) refinement. In this type of methodology, the mesh in the regional domain of interest is refined to a desired mesh size while the mesh size increases by small factors moving away from the fixed region. Some numerical experiments using global GCMs in the past (e.g. [2]) have demonstrated using this type of mesh refinement that a decrease in horizontal grid size improved the predictability of large, already well-resolved waves such as Rossby waves.

In this thesis however, we will propose a different approach to regional approximation through the use of meshless collocation. As we hope to demonstrate, the advantage to using meshless collocation for both global and regional approximation of the shallow-water dynamics comes from the fact that no numerical quadrature and remeshing will be needed due to the nature of collocation. This greatly simplifies the approximation and refinement at any time step. In the next sections, we will discuss the underlying shallow-water model along with its cubed-sphere domain discretization and the semi-implicit time stepping method which will be used for the time evolution of the shallow-water model. Both provide an integral part of the proposed geophysical model in this thesis.

3.3 The Shallow Water model

Being the simplest form of motion equations that can approximate the horizontal structure of the atmosphere or the dynamics of oceans, the shallow-water equations have been used as a robust testing model in atmospheric and oceanic sciences. The solutions can represent certain types of motion including Rossby waves and inertia-gravity waves while describing an incompressible fluid subject to gravitational and rotating acceleration. The governing equations for the inviscid flow of a thin layer of fluid in 2-D are the horizontal momentum and continuity equations for the velocity field the geopotential height. We will not discuss the derivation of the shallow-water model in this section. For a good derivation of the shallow water equations on the sphere, the interested reader is referred to Kalnay [34].

While there are many different ways of defining the shallow-water equations, we focus in this model on cubed-sphere geometry originally proposed by Sadourny in [43] and used in other global models in recent years such as [58] and [57]. We begin by a brief review of the cubed-sphere while adopting notational conventions from [58].

3.3.1 Implementation of the Cubed-Sphere

As a visual aid, the cubed-sphere is constructed as follows. Consider a cube inscribed inside a sphere where each corner of the cube is a point in the sphere and where each face of the cube is subdivided into N_E subregions. The goal is to project each face of the cube onto the sphere and in effect, obtain a quasi-uniform spherical grid of $6 \times N_E$ subregions which can be further subdivided into many subregions of interest. In the mapping of the cube to sphere, each face of the cube is constructed with a local coordinate system and employs metric terms for transforming between the cube and the sphere which we define next. This will allow computations to be done on a unit square for each face of the cube, which will prove to be much easier than working in spherical coordinates for the meshless collocation.

Let (α, β) be equal angular coordinates such that $-\pi/4 \leq \alpha, \beta \leq \pi/4$. Then any x_1 and x_2 on a face P_i of the cube is related through $x_1 = \tan\alpha, x_2 = \tan\beta$. We denote $\mathbf{r}$ the corresponding position vector on the sphere with longitude λ and latitude θ. For such an equiangular projection, we define basis vectors $\mathbf{a}_1 = \mathbf{r}_\alpha$ and

$\mathbf{a}_2 = \mathbf{r}_\beta$ which may be written as

$$\mathbf{r}_\alpha = \frac{1}{\cos^2\alpha}\mathbf{r}_{x_1}, \qquad \mathbf{r}_\beta = \frac{1}{\cos^2\beta}\mathbf{r}_{x_2} \tag{3.1}$$

where $\mathbf{r}_{x_1}$ and $\mathbf{r}_{x_2}$ are defined as

$$\begin{aligned} \mathbf{r}_{x_1} &= \Big(\cos\theta\,\lambda_{x_1},\ \theta_{x_1}\Big) \\ \mathbf{r}_{x_2} &= \Big(\cos\theta\,\lambda_{x_2},\ \theta_{x_2}\Big). \end{aligned}$$

The metric tensor $g_{ij},\ i,j \in [1,2]$ can be derived as

$$\begin{aligned} g_{ij} = \mathbf{a_i}\cdot\mathbf{a_j} &= \begin{bmatrix} \mathbf{r}_{x_1}\cdot\mathbf{r}_{x_1} & \mathbf{r}_{x_1}\cdot\mathbf{r}_{x_2} \\ \\ \mathbf{r}_{x_2}\cdot\mathbf{r}_{x_1} & \mathbf{r}_{x_2}\cdot\mathbf{r}_{x_2} \end{bmatrix}, \\ &= \frac{1}{r^4\cos^2\alpha\cos^2\beta}\begin{bmatrix} 1+\tan^2\alpha & -\tan\alpha\tan\beta \\ -\tan\alpha\tan\beta & 1+\tan^2\beta \end{bmatrix}, \\ &= \tilde{\mathbf{A}}^T\tilde{\mathbf{A}}, \end{aligned}$$

where $r^2 = 1+\tan^2\alpha+\tan^2\beta$ and the Jacobian of the transformation and the matrix $\tilde{\mathbf{A}}$ are, respectively,

$$\sqrt{g} = [\det(g_{ij})]^{1/2} = \frac{1}{r^3\cos^2\alpha\cos^2\beta},$$

and

$$\tilde{\mathbf{A}} = \begin{bmatrix} \cos\theta\ \lambda_\alpha & \cos\theta\ \lambda_\alpha \\ \\ \theta_\alpha & \theta_\beta \end{bmatrix}.$$

In order to transform between the cube and the sphere, an explicit form of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{A}}^{-1}$ are needed for each face of the cube which are derived by the local coordinate system on each face. First, the matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{A}}^{-1}$ are given as

$$\tilde{\mathbf{A}} = \mathbf{A} \begin{bmatrix} 1/\cos^2\alpha & 0 \\ 0 & 1/\cos^2\beta \end{bmatrix}, \tag{3.2}$$

and

$$\tilde{\mathbf{A}}^{-1} = \begin{bmatrix} \cos^2\alpha & 0 \\ 0 & \cos^2\beta \end{bmatrix} \mathbf{A}^{-1}. \tag{3.3}$$

In order to complete the transformation from the cube to the sphere, the explicit form of the matrix $\mathbf{A}$ is needed which is dependent on the cube face. For the lateral faces P_1 through P_4, $\mathbf{A}$ is the same. Using the global spherical coordinates λ and θ, the values are

$$\lambda_{x_1} = \cos^2\lambda, \qquad \lambda_{x_2} = 0$$

$$\theta_{x_1} = \sin\theta \, \sin\lambda \, \cos\theta \, \cos\lambda, \qquad \theta_{x_2} = \cos^2\theta \, \cos\lambda,$$

which gives

$$\mathbf{A} = \cos\theta \, \cos\lambda \begin{bmatrix} \cos\lambda & 0 \\ -\sin\lambda \, \sin\theta & \cos\theta \end{bmatrix}, \tag{3.4}$$

and

$$\mathbf{A}^{-1} = \sec\theta\,\sec\lambda \begin{bmatrix} \sec\lambda & 0 \\ \\ \tan\lambda\,\tan\theta & \sec\theta \end{bmatrix}. \tag{3.5}$$

The top panel P_5 of the cubed

$$\lambda_{x_1} = \cos\lambda\tan\theta, \qquad \lambda_{x_2} = \sin\lambda\tan\theta,$$

$$\theta_{x_1} = -\sin\lambda\,\sin^2\theta, \qquad \theta_{x_2} = \cos\lambda\,\sin^2\theta,$$

which gives

$$\mathbf{A} = \sin\theta \begin{bmatrix} \cos\lambda & \sin\lambda \\ \\ -\sin\lambda\,\sin\theta & \sin\theta\,\cos\lambda \end{bmatrix}, \tag{3.6}$$

and

$$\mathbf{A}^{-1} = \frac{1}{\sin^2\theta} \begin{bmatrix} \sin\theta\cos\lambda & -\sin\lambda \\ \\ \sin\lambda\,\sin\theta & \cos\theta \end{bmatrix}. \tag{3.7}$$

And finally, similar to the top face, the bottom face P_6 metric is defined as

$$\lambda_{x_1} = -\cos\lambda\tan\theta, \quad \lambda_{x_2} = \sin\lambda\tan\theta,$$

$$\theta_{x_1} = \sin\lambda\,\sin^2\theta, \quad \theta_{x_2} = \cos\lambda\,\sin^2\theta$$

which gives

$$\mathbf{A} = \sin\theta \begin{bmatrix} -\cos\lambda & \sin\lambda \\ \\ \sin\lambda\,\sin\theta & \sin\theta\,\cos\lambda \end{bmatrix}, \tag{3.8}$$

and

$$\mathbf{A}^{-1} = \frac{1}{\sin^2\theta}\begin{bmatrix} -\sin\theta\cos\lambda & -\sin\lambda \\ \\ \sin\lambda\,\sin\theta & \cos\theta \end{bmatrix}. \tag{3.9}$$

Finally, while using the definition of g_{ij} given in (3.2), we can write transformations between covariant and contravariant components of a vector $\mathbf{u} = (u_1, u_2)$ as

$$\begin{bmatrix} u_1 \\ \\ u_2 \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ \\ g_{21} & g_{22} \end{bmatrix}\begin{bmatrix} u^1 \\ \\ u^2 \end{bmatrix}, \quad \begin{bmatrix} u^1 \\ \\ u^2 \end{bmatrix} = \begin{bmatrix} g^{11} & g^{12} \\ \\ g^{21} & g^{22} \end{bmatrix}\begin{bmatrix} u_1 \\ \\ u_2 \end{bmatrix}. \tag{3.10}$$

With the metric terms defined, we can now write the shallow water equations in the curvilinear coordinates system to be integrated on the cubed-sphere. In such a coordinate system, the shallow-water equations can be written as follows

$$\boxed{\begin{aligned}
\frac{\partial u_1}{\partial t} &= -\Big[u^2 g(f+\zeta) + \frac{\partial}{\partial x^1}\Big(\frac{1}{2}u_1u^1 + u_2u^2\Big) + \frac{\partial \eta}{\partial x^1}\Big],\\
\frac{\partial u_2}{\partial t} &= -\Big[u^1 g(f+\zeta) + \frac{\partial}{\partial x^2}\Big(\frac{1}{2}u_2u^2 + u_2u^2\Big) + \frac{\partial \eta}{\partial x^2}\Big],\\
\frac{\partial \eta'}{\partial t} &= -\Big(\sum_{j=1}^{2}(u^j\frac{\partial \eta}{\partial x^j} + \frac{\eta}{g}\frac{\partial}{\partial x^j}(g\ u^j))\Big)
\end{aligned}}$$

with initial conditions given by an initial velocity field and geopotential surface height η' $\mathbf{u}(\lambda,\theta) = \mathbf{f}(\lambda,\theta)$ and $\eta'(\lambda,\theta) = g(\lambda,\theta)$, respectively. We will assume both fields are continuous across the sphere in latitude and longitude.

Here, we have defined $\eta = \eta' + \eta_0$ where η_0 is the fixed bottom topography and η' is the displaced geopotential height, $f = 2\omega \sin\theta$ is the Coriolis force (ω is the rotation of earth) and

$$g\zeta = \frac{\partial u_2}{\partial x^1} - \frac{\partial u_1}{\partial x^2}$$

is the relative vorticity. Covariant and contravariant vectors are defined through the short-hand metric tensor notation $u^i = g^{ij}u_j$, $g^{ij} = (g_{ij})^{-1}$. We note that the metric terms g^{ij} can be computed and stored prior to the time stepping of the equations once the discretization of the cube has been resolved. An example of a discretized cubed-sphere is shown in figures (3.1) at two different angles. On each face of the cube, we used a 8×8 Gauss-Lobatto-Legendre distribution of points.

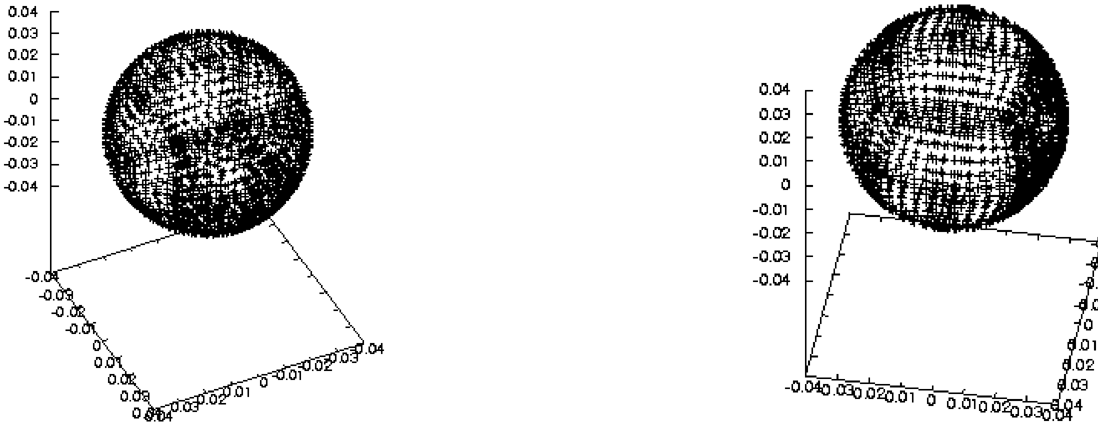


Figure 3.1: Discretized Cubed-Sphere

Only knowledge of the collocation points are needed to compute the metric terms for evaluating the velocity and geopotential fields on the sphere. We discuss the efficient implementation in Fortran 90 of the metric terms and the covariant and contravariant vectors for mapping between the cube and the sphere in the next chapter.

3.3.2 Semi-Implicit Time Discretization

As an integral part of our geophysical model, the semi-implicit time stepping scheme which we discuss in this section has many computational advantages. Semi-implicit time stepping schemes were first used in atmospheric models in order to alleviate the problem of stability constraints ultimately due to the fast moving gravity waves in the discrete shallow water equations [58]. They have been successfully applied for allowing an increase in the time step without affecting the atmospherically important Rossby waves. Furthermore, since the fast modes in the spherical shallow water system are ultimately associated with the small scale gravity waves, semi-implicit integration is highly desirable for climate models where the small scale gravity waves are not significant to the global large scale dynamics. Such a semi-implicit method is described in this section and was originally proposed in the spectral element model developed in [58].

As we will show later in this section, the resulting time discrete formulation of the shallow water equations using the semi-implicit method leads to solving an elliptic Helmholtz problem for the geopotential at each time step. A preconditioned conjugate gradient method is then used to solve the system since the resulting discrete Helmholtz operator is symmetric positive definite.

The semi-implicit time stepping is composed of an explicit leapfrog scheme for the advection terms combined with a Crank-Nicholson scheme for the gradient and divergence terms. Adopting the difference notation $\delta u_j = u_j^{n+1} - u_j^{n-1}$ and $\delta\eta = \eta^{n+1} - \eta^{n-1}$, the time discretized shallow water equations in curvilinear coordinates can be written as

$$\delta u_j + \Delta t \frac{\partial}{\partial x^j}(\delta\eta) \quad = \quad 2\Delta t\Big[\frac{\partial}{\partial x^j}(\eta^{n-1}) + f_u^{j,n}\Big],$$

$$\delta\eta + \Delta t \frac{\eta_0}{g}\sum_{j=1}^{2}\frac{\partial}{\partial x^j}(g\delta u^j) \quad = \quad 2\Delta t\Big[\frac{\eta_0}{g}\sum_{j=1}^{2}\frac{\partial}{\partial x^j}(gu^{j,n-1}) + f_\eta^n\Big],$$

where the tendencies $f_u^{i,n}$ and f_η^n contain nonlinear terms along with the Coriolis term and are evaluated explicity as

$$f_u^{j,n} = u^{k,n}g(f+\zeta^n) + \frac{\partial}{\partial x^j}\Big(\frac{1}{2}u_1u^1 + u_2u^2\Big)^n,$$

and

$$f_\eta^n = -(\sum_{j=1}^{2} u^{j,n}\frac{\partial\eta^n}{\partial x^j} + \frac{\eta'^n}{g}\sum_{j=1}^{2}\frac{\partial}{\partial x^j}(gu^{j,n-1})).$$

Now to put the equations in the form we are interested in, we bring the implicit terms to the left hand side of the equation and the explicit terms to the right, we end up with the time discrete evolution form of the shallow water equations

$$u_j^{n+1} + \Delta t\frac{\partial}{\partial x^j}(\eta)^{n+1} = u_j^{n-1} - \Delta t\frac{\partial}{\partial x^j}(\eta)^{n-1} + 2\Delta t f_u^{j,n} \tag{3.11}$$

$$\eta^{n+1} + \Delta t\frac{\eta_0}{g}\sum_{j=1}^{2}\frac{\partial}{\partial x^j}(gu^{j,n+1}) = \eta^{n-1} - \Delta t\frac{\eta_0}{g}\sum_{j=1}^{2}\frac{\partial}{\partial x^j}(gu^{j,n-1}) + 2\Delta t f_\eta^n \tag{3.12}$$

which is now in the form needed for spatial discretization using meshless collocation. Due to the Crank-Nicholson terms, the storage of two or more previous time steps is needed. In order to compute the first two time steps, $n = 1$ and $n = 2$, two forward explicit steps with very small time step Δt can be applied to the shallow-water equations where the nonlinear terms are on the right-hand side of the equation. For example, with initial states of the velocity and geopotential fields given as (u_1^0, u_2^0) and η^0, a small initial step Δt to get (u_1^1, u_2^1) and η^1 can be computed as

$$u_j^1 \quad = \quad u_j^0 - \Delta t \frac{\partial}{\partial x^j}(\eta^0) + \Delta t f_u^{j,0},$$

$$\eta^1 \quad = \quad \eta^0 - \Delta t \frac{\eta_0}{g} \sum_{j=1}^{2} \frac{\partial}{\partial x^j}(g u^{j,0}) + \Delta t f_\eta^0.$$

We note however that the Δt used in this step should be magnitudes lower than the Δt used in the semi-implicit stepping. We discuss this issue further in the next section on the initialization and implementation of the model.

The overall performance of this semi-implicit method is reduced to the performance of a robust approach for solving equations (3.11) in an efficient manner for obtaining the solution at the $n+1$ time step. Of course, this is highly dependent on the spatial discretization of the variables $\mathbf{u} = (u_1, u_2)$ and η. In the next section, we propose a meshless collocation method for approximating these variables at each time step, where the collocation method will rely heavily on the theory from chapter 2.

3.4 Meshless Collocation for the shallow-water equations

With the shallow water system of equations cast into discrete time-stepping form, we can now approximate spatially across the cubed-sphere. A direct approach to solving the equations in (3.11) using the meshless collocation method from the previous chapter is not so clear. In fact, if one does a direct discretization of each variable using SPD kernels as shown in chapter 2, the resulting linear system would be nontrivial to solve since it would be nonsymmetric and not necessarily positive definite. It might even be singular. In order to guarantee a unique solution at every time step, a different approach must be taken to cast (3.11) into a form in which a unique solution is obtainable and feasibly computable.

In this section, we propose a new meshless collocation approach for discretizing and solving (3.11) in a robust and computationally fast manner. The method will take advantage of compactly supported symmetric positive definite (SPD) kernels and the theory of SPD kernels presented in the previous chapter. We first give a general approach to the method and then discuss specific implementational issues related to the compactly supported SPD kernels. In the next chapter, we then demonstrate an efficient parallelization of the method in Fortran 90.

In this new meshless collocation approach for approximating the shallow water equations on the sphere, we will use a distribution of collocation nodes for both the velocity and geopotential fields. Let $\Omega = \cup_{k=1}^{6} P_k$ where P_k, $k = 1, \ldots, 6$ are the 6 faces of the inscribed cube in the sphere and let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a distribution of N collocation nodes on Ω for the velocity and the geopotential fields.

Following the theory of chapter 2, in this meshless collocation construction we consider a symmetric positive definite kernel $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$ such that $\Psi \in C^{2k}(\Omega \times \Omega)$ and construct the following discrete Native space on Ω

$$\mathcal{N}_{\Psi,N} = \text{span}\{\Psi(\cdot, \mathbf{x}_1), \ldots, \Psi(\cdot, \mathbf{x}_N)\}. \tag{3.13}$$

We approximate each component of the velocity field $\mathbf{u}^n = (u_1^n, u_2^n)$ and the geopotential η^n at any time step $n > 0$ by a reproducing kernel expansion so that $u^n_{1,N}, u^n_{2,N}, \eta^n_N \in \mathcal{N}_{\Psi,N}$. Thus we seek vectors $\boldsymbol{\alpha}^{n,u_1}$, $\boldsymbol{\alpha}^{n,u_2}$, $\boldsymbol{\alpha}^{n,\eta} \in \mathbb{R}^N$ at each time step $n > 0$ such that

$$\begin{aligned} u^n_{1,N}(\cdot) &= \sum_{j=1}^{N} \alpha_j^{n,u_1} \Psi(\cdot, \mathbf{x}_j) \\ u^n_{2,N}(\cdot) &= \sum_{j=1}^{N} \alpha_j^{n,u_2} \Psi(\cdot, \mathbf{x}_j) \\ \eta^n_N(\cdot) &= \sum_{j=1}^{N} \alpha_j^{n,\eta} \Psi(\cdot, \mathbf{x}_j). \end{aligned} \tag{3.14}$$

As required by collocation, we enforce all three of these expansions to satisfy the time discrete equations at the collocation nodes $\mathcal{X}$. This amounts to enforcing

$$\begin{aligned} u^{n+1}_{i,N}(\mathbf{x}_k) + \Delta t \frac{\partial}{\partial x^i}(\eta_N)^{n+1}(\mathbf{x}_k) &= \\ u^{n-1}_{i,N}(\mathbf{x}_k) - \Delta t \frac{\partial}{\partial x^i}(\eta_N)^{n-1}(\mathbf{x}_k) &+ 2\Delta t f_u^{i,n}(\mathbf{x}_k),\ \forall \mathbf{x}_k \in \mathcal{X} \\ \eta^{n+1}_N(\mathbf{x}_k) + \Delta t \frac{\eta_0}{g} \sum_i^2 \frac{\partial}{\partial x^i}(g u^i_N)^{n+1}(\mathbf{x}_k) &= \\ \eta^{n-1}_N(\mathbf{x}_k) - \Delta t \frac{\eta_0}{g} \sum_i^2 \frac{\partial}{\partial x^i}(g u^i_N)^{n-1}(\mathbf{x}_k) &+ 2\Delta t f_\eta^n(\mathbf{x}_k),\ \forall \mathbf{x}_k \in \mathcal{X} \end{aligned} \tag{3.15}$$

for each time step $n > 0$.

Although equations (3.15) are in the desired form, we do not directly attempt to solve them since the resulting system would be nonsymmetric, very large, and possibly not even solvable for a unique solution at each time step. We instead take a different approach where we eventually decouple the velocity and geopotential approximations, consequently replacing the system (3.15) by a smaller one which is in fact symmetric, positive definite, and easier to solve via conjugate gradient methods.

In order to solve for all three components more efficiently, we begin by casting (3.15) in a slightly different form using a nodal approximation for each component. This is done as follows. For simplicity of exposition, we demonstrate the method on a simple continuous scalar function on Ω and then apply the method to the components in (3.15). To this end, let $u \in C(\Omega)$ be any continuous function and consider approximating u by some $u_N \in \mathcal{N}_{\Psi,N}$ which satisfies $u_N(\mathbf{x}_k) = u(\mathbf{x}_k)$ for all $\mathbf{x}_k \in \mathcal{X}$. It was shown in Chapter 2 that this can be accomplished by letting $u_N(\cdot) = \sum_{j=1}^N \alpha_j \Psi(\cdot, \mathbf{x}_j)$ and then requiring $\sum_{j=1}^N \alpha_j \Psi(\mathbf{x}_k, \mathbf{x}_j) = u(\mathbf{x}_k)$ for all $\mathbf{x}_k$. Thus we arrive at the system $\mathcal{A}\boldsymbol{\alpha} = \mathbf{u}$ where $\mathbf{u} = (u(\mathbf{x}_1), \ldots, u(\mathbf{x}_N))$ and so $\boldsymbol{\alpha} = \mathcal{A}^{-1}\mathbf{u}$.

Now suppose we wish to approximate $\frac{\partial u}{\partial x^1}$ and $\frac{\partial u}{\partial x^2}$ given the data $u(\mathbf{x}_1), \ldots, u(\mathbf{x}_N)$. Since $u(\mathbf{x}_k) = u_N(\mathbf{x}_k)$ at any $\mathbf{x}_k \in \mathcal{X}$, we have

$$\frac{\partial u}{\partial x^1}(\mathbf{x}_k) = \sum_{j=1}^N \alpha_j \left(\frac{\partial}{\partial x^1}\right)_1 \Psi(\mathbf{x}_k, \mathbf{x}_j), \ \forall \mathbf{x}_k \in \mathcal{X}$$

where $\left(\frac{\partial}{\partial x^1}\right)_1$ denotes the derivative acting on the first argument of the kernel Ψ. In matrix form, we write

$$\mathbf{u}_{x^1} = \mathcal{A}_{x^1}\boldsymbol{\alpha} \tag{3.16}$$

where $\mathbf{u}_{x^1} = (\frac{\partial u}{\partial x^1}(\mathbf{x}_1), \ldots, \frac{\partial u}{\partial x^1}(\mathbf{x}_N))$ and

$$\mathcal{A}_{x_1} = \begin{pmatrix} \left(\frac{\partial}{\partial x^1}\right)_1 \Psi(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \left(\frac{\partial}{\partial x^1}\right)_1 \Psi(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \left(\frac{\partial}{\partial x^1}\right)_1 \Psi(\mathbf{x}_N, \mathbf{x}_1) & \cdots & \left(\frac{\partial}{\partial x^1}\right)_1 \Psi(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix}. \tag{3.17}$$

Now since $\boldsymbol{\alpha} = \mathcal{A}^{-1}\mathbf{u}$, substituting into (3.16) we have

$$\mathbf{u}_{x^1} = \mathcal{A}_{x^1}\mathcal{A}^{-1}\mathbf{u} = D_1\mathbf{u}. \tag{3.18}$$

Notice that the $N \times N$ matrix $D_1 = \mathcal{A}_{x^1}\mathcal{A}^{-1}$ is actually a differentiation matrix. Thus to approximate $\frac{\partial u}{\partial x^1}$ pointwise on $\mathcal{X}$ we simply construct the matrix D_1 using the inverse of the collocation matrix $\mathcal{A}$ and the matrix $\mathcal{A}_{x^1}$. A similar construction for the differentiation matrix $D_2 = \mathcal{A}_{x^2}\mathcal{A}^{-1}$ for derivatives in the x^2 variable can also readily be accomplished. Using these differentiation matrices, the pointwise values of the gradient field of a scalar function η and the divergence of a vector velocity field $\mathbf{u}$ can now easily be approximated as $(D_1\bar{\eta}, D_2\bar{\eta})$ and $D_1\bar{u}_1 + D_2\bar{u}_2$, where $\bar{\eta}$, $\bar{u}_1$, and $\bar{u}_2$ is the vector of values at the collocation nodes.

Using the differentiation matrices just defined, we can now cast the equations (3.15) into nodal form. We let D_1 and D_2 be the differentiation matrices with respect to the collocation nodes $\mathcal{X}$ and define the $2N \times 2N$ matrix $\mathbf{D} = [D_1 D_2]$ then (3.15) can be written as

$$\begin{aligned} \mathbf{u}^{n+1} + \mathbf{D}^T\boldsymbol{\eta}^{n+1} &= \mathbf{u}^{n-1} - \mathbf{D}^T\boldsymbol{\eta}^{n-1} + \mathbf{R}^n_{\mathbf{u}} \\ \boldsymbol{\eta}^{n+1} + \mathbf{D}\mathbf{G}_t\mathbf{u}^{n+1} &= \boldsymbol{\eta}^{n-1} + \mathbf{D}\mathbf{G}_t\mathbf{u}^{n-1} + \mathbf{R}^n_{\eta} \end{aligned} \tag{3.19}$$

for each time step $n > 0$, where $\mathbf{u}^n = (\tilde{u}^n_1, \tilde{u}^n_2)$ and $\boldsymbol{\eta}^n$ are vectors of length $2N$ and N, respectively of the nodal values at the collocation nodes in $\mathcal{X}$. The vectors $\mathbf{R}^n_{\mathbf{u}} \in \mathbb{R}^{2N}$ and $\mathbf{R}^n_{\eta} \in \mathbb{R}^N$ are the vectors of the values at the collocation nodes of the nonlinear and Coriolis forcing terms at time step n. The matrix $\mathbf{G}_t$ is a diagonal matrix containing the terms $\Delta t \frac{\eta_0}{g}$.

The velocity-geopotential decoupling can now be accomplished by first writing the the system in matrix-vector form

$$\begin{bmatrix} \mathbf{I}^t & -\mathbf{D}^T \\ & \\ & \\ \mathbf{D} & \mathbf{I}^t \end{bmatrix} \begin{bmatrix} \mathbf{u}^{n+1} \\ \\ \eta^{n+1} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_u^n \\ \\ R_\eta^n \end{bmatrix}, \tag{3.20}$$

where we have defined the matrices and vectors

$$\begin{aligned} \mathbf{I}^t &= \mathbf{I}/\Delta t, \\ \\ \tilde{\mathbf{I}}^t &= \mathbf{I}/(\Delta t \eta_0/g) \end{aligned} \tag{3.21}$$

with $\mathbf{I}$ being the identity matrix, and

$$\begin{aligned} \mathbf{R}_{\mathbf{u}}^n &= (\mathbf{u}^{n-1} - \mathbf{D}^T\boldsymbol{\eta}^{n-1} + \mathbf{R}_{\mathbf{u}}^n)/\Delta t \\ \mathbf{R}_\eta^n &= (\boldsymbol{\eta}^{n-1} + \mathbf{D}\mathbf{G}_t\mathbf{u}^{n-1} + \mathbf{R}_\eta^n)/\Delta t. \end{aligned} \tag{3.22}$$

A Helmholtz problem for the geopotential perturbation at time step $n+1$ is obtained by first writing the expression for the velocity solution at time step $n+1$ yielding

$$\mathbf{u}^{n+1} = \Delta t(\mathbf{R}_{\mathbf{u}}^n + \Delta t\mathbf{D}^T\boldsymbol{\eta}^{n+1}), \tag{3.23}$$

and then applying back-substitution to obtain an equation for the geopotential

$$\boldsymbol{\eta}^{n+1} + \Delta t^2\eta_0\mathbf{D}\mathbf{D}^T\boldsymbol{\eta}^{n+1} = \mathbf{R}'_\eta, \tag{3.24}$$

where

$$\mathbf{R}'_\eta \equiv \mathbf{R}_\eta^n - \Delta t\eta_0\mathbf{D}\mathbf{D}^T\mathbf{R}_{\mathbf{u}}^n. \tag{3.25}$$

Once the geopotential is computed from (3.25), the velocity field is updated at time step $n+1$ using equation (3.23).

Notice that the system in (3.24) is of Helmholtz-type due to the fact that the $N \times N$ matrix defined by $\mathbf{H} = (\mathbf{I} + \Delta t^2 \eta_0 \mathbf{D}\mathbf{D}^T)$ is symmetric and positive definite. Furthermore, since $\mathbf{D}^T \boldsymbol{\eta}^{n+1}$ approximates the gradient field of the η variable evaluated at the collocation points, it is easy to see that the differentiation matrix $\mathbf{D}$ then approximates the divergence of $\mathbf{D}^T \boldsymbol{\eta}^{n+1}$. We can thus view $\mathbf{D}\mathbf{D}^T$ as the Laplacian operator on η.

Using this semi-implicit time stepping algorithm to decouple the velocity and geopotential fields, we have effectively reduced the $3N \times 3N$ nonsymmetric system of equations in (3.19) at every time step to an $N \times N$ system. Moreover, since the matrix $\mathbf{H}$ is symmetric and positive definite, we can solve (3.24) using an efficient preconditioned conjugate gradient method to approximate the geopotential $\boldsymbol{\eta}^{n+1}$ and consequently the velocity field through equation (3.23). In the numerical section of this chapter, we discuss an efficient preconditioned conjugate gradient method for solving (3.24) in parallel using a message passing interface in Fortran 90.

3.4.1 Implementation of the shallow water model

In this section we discuss an efficient implementation of the approximation scheme for the shallow water equations on cubed-sphere described in the previous section. Since the core of each time step obviously comes in solving an $N \times N$ system for the geopotential, the robustness and computational speed of the meshless collocation is heavily dependent on a robust preconditioner for the conjugate gradient method. Before we discuss solving the Helmholtz problem for the geopotential, the initialization issues of the model need to be addressed.

As with any collocation method, the choice of the SPD kernel for constructing the discrete Native space is crucial for this meshless collocation scheme applied to the shallow water equations. In this thesis, due to there powerful approximation ability and fast summation techniques, we have chosen to use the compactly supported radial functions developed by Wendland in [64] and reviewed in Appendix C. In particular, we are interested the C^4 Wendland function defined by $\phi_2(r) = (1-r)_+^6(35r^2+18r+3)$ since its collocation matrix is better conditioned than the ϕ_3 radial function and has better approximation ability than the ϕ_1 function as demonstrated in section 2.4.3. An additional parameter which we will make use of is the inclusion of the so-called shape parameter $\epsilon > 0$. Using the ϕ_2 radial function, we can scale the support of the associated SPD kernel by simply defining $\Psi(\mathbf{x}, \mathbf{y}; \epsilon) := \phi_2(\|\mathbf{x}-\mathbf{y}\|/\epsilon)$. The support of the kernel can thus be expanded or contracted depending on the value of epsilon. We will use this feature to appropriately construct the interpolation matrix $\mathcal{A}$ for Ψ in such a way so that the inverse of $\mathcal{A}_{\mathcal{X}}$ can be computed quickly. We will show how this is done in the following.

The meshless collocation spaces are initialized on the entire cube by distributing N collocation nodes on $\Omega = \cup_{i=1}^{6} P_i$ and then constructing the discrete native space $\mathcal{N}_{\Psi,N}(\Omega) = \text{span}\{\Psi(\cdot, \mathbf{x}_j), \ldots, \Psi(\cdot, \mathbf{x}_N)$ that will approximate both components of the velocity and the geopotential. The interpolation matrix $\mathcal{A}_{\mathcal{X}}$ based on the collocation nodes $\mathcal{X}$ is computed by constructing each row as follows. For each $j = 1, \ldots, N$, choose the shape parameter $\epsilon > 0$ such that at most $m > 1$ nearest neighbors to $\mathbf{x}_j$ are included in the support $\Psi(\cdot, \mathbf{x}_j; \epsilon)$, for some fixed m. This can be achieved, for example, by using the nearest neighbor search algorithm presented in Wendland Chapter 14. We reproduce it in the Appendix C. Once an $\epsilon > 0$ has been chosen such that at most m nearest neighbors are in the support of $\Psi(\cdot, \mathbf{x}_j; \epsilon)$, the interpolation matrix $\mathcal{A}_{\mathcal{X}}$ is then built row by row where the nonzero entries corresponding to row i are the values of the kernel $\Psi(\mathbf{x}_i, \mathbf{x}_j; \epsilon)$, for all $\mathbf{x}_j \in \mathcal{X}$ such that $\|\mathbf{x}_i - \mathbf{x}_j\| \leq \epsilon$. If the collocation nodes are clustered correctly, The resulting matrix should then have a concentration of values along the diagonal, with a bandwidth of m, and zeros in all other entries. For example, with $m = 4$, the matrix could resemble something like

$$\mathcal{A}_{\mathcal{X}} = \begin{pmatrix} \Psi(\mathbf{x}_1, \mathbf{x}_1; \epsilon) & \Psi(\mathbf{x}_1, \mathbf{x}_2; \epsilon) & \Psi(\mathbf{x}_1, \mathbf{x}_3; \epsilon) & \cdots & 0 \\ \Psi(\mathbf{x}_2, \mathbf{x}_1; \epsilon) & \Psi(\mathbf{x}_2, \mathbf{x}_2; \epsilon) & \Psi(\mathbf{x}_2, \mathbf{x}_3; \epsilon) & \cdots & 0 \\ \Psi(\mathbf{x}_3, \mathbf{x}_1; \epsilon) & \Psi(\mathbf{x}_3, \mathbf{x}_2; \epsilon) & \Psi(\mathbf{x}_3, \mathbf{x}_3; \epsilon) & \Psi(\mathbf{x}_3, \mathbf{x}_4; \epsilon) & \cdots \\ 0 & 0 & \Psi(\mathbf{x}_4, \mathbf{x}_3; \epsilon) & \Psi(\mathbf{x}_4, \mathbf{x}_4) & \cdots \\ \vdots & \ddots & \Psi(\mathbf{x}_i, \mathbf{x}_{i-1}; \epsilon) & \Psi(\mathbf{x}_i, \mathbf{x}_i; \epsilon) & \vdots \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & \cdots & \Psi(\mathbf{x}_N, \mathbf{x}_{N-3}; \epsilon) & \Psi(\mathbf{x}_N, \mathbf{x}_{N-2}; \epsilon) & \cdots \end{pmatrix}.$$

The next step in the initialization is to compute the inverse of the interpolation matrix $\mathcal{A}_{\mathcal{X}}$. Since the matrix might be quite large rendering direct Gaussian elimination methods infeasible, we propose an efficient parallel algorithm in the next chapter based on a so-called Connected Schur's Component algorithm originally developed in Mahmood et al [39].

Once $\mathcal{A}^{-1}$ has been computed, the differentiation matrices D_1 and D_2 with respect to the collocation nodes are constructed by first building the matrices

$$\mathcal{A}_{x^1} = \begin{pmatrix} \Psi_{x^1,1}(\mathbf{x}_1, \mathbf{x}_1; \epsilon) & \Psi_{x^1,1}(\mathbf{x}_1, \mathbf{x}_2; \epsilon) & \Psi_{x^1,1}(\mathbf{x}_1, \mathbf{x}_3; \epsilon) & \cdots \\ \Psi_{x^1,1}(\mathbf{x}_2, \mathbf{x}_1; \epsilon) & \Psi_{x^1,1}(\mathbf{x}_2, \mathbf{x}_2; \epsilon) & \Psi_{x^1,1}(\mathbf{x}_2, \mathbf{x}_3; \epsilon) & \cdots \\ \vdots & \ddots & \ddots & \vdots \end{pmatrix},$$

and

$$\mathcal{A}_{x^2} = \begin{pmatrix} \Psi_{x^1 21}(\mathbf{x}_1, \mathbf{x}_1; \epsilon) & \Psi_{x^2,1}(\mathbf{x}_1, \mathbf{x}_2; \epsilon) & \Psi_{x^1,1}(\mathbf{x}_1, \mathbf{x}_3; \epsilon) & \cdots \\ \Psi_{x^2,1}(\mathbf{x}_2, \mathbf{x}_1; \epsilon) & \Psi_{x^2,1}(\mathbf{x}_2, \mathbf{x}_2; \epsilon) & \Psi_{x^2,1}(\mathbf{x}_2, \mathbf{x}_3; \epsilon) & \cdots \\ \vdots & \ddots & \ddots & \vdots \end{pmatrix},$$

The differentiation matrices D_1 and D_2 are then computed as shown before by $D_1 = \mathcal{A}_{x^1}\mathcal{A}^{-1}$ and $D_2 = \mathcal{A}_{x^2}\mathcal{A}^{-1}$. It is worthwhile to note that due to the banded diagonal structure of the matrices $\mathcal{A}_{x^1}$ and $\mathcal{A}_{x^2}$, the differentiation matrices can be quickly computed by only summing on the elements which are nonzero. This significantly reduces the cost of the matrix-matrix multiplication.

The final step in initialization of the method before starting the time stepping process of the semi-implicit discretization is to compute the metric tensor terms evaluated at the collocation nodes for mapping the velocity and geopotential fields to the sphere. In the next chapter, we show how this can efficiently be done using data structures in Fortran 90.

Once the metrics and Coriolis forcing evaluated at each collocation node has been resolved, the semi-implicit time stepping procedure can be initiated. Recall from the previous section that in order to compute th first semi-implicit steps, the initial and first step need to be stored. Due to its simplicity, we propose a simple forward explicit time step with very small Δt. With initial states of the velocity and geopotential fields evaluated on the collocation nodes $\mathcal{X}$ of the cube, and represented by the vectors $\mathbf{u}^0 \in \mathbb{R}^{2N}$ and $\boldsymbol{\eta}^0 \in \mathbb{R}^N$, the first step can be computed as

$$\mathbf{u}^1 \;=\; \mathbf{u}^0 - \Delta t \mathbf{D}^T \boldsymbol{\eta}^0 + \Delta t \mathbf{R}_{\mathbf{u}}^0,$$

$$\boldsymbol{\eta}^1 \;=\; \boldsymbol{\eta}^0 - \Delta t \mathbf{G}\mathbf{D}\mathbf{u}^0 + \Delta t \mathbf{R}_{\boldsymbol{\eta}}^0.$$

As already mentioned, due to the nature of explicit time stepping, the Δt used in this step should be magnitudes lower than the Δt used in the semi-implicit stepping.

Once the vectors $\mathbf{u}^1$ and $\boldsymbol{\eta}^1$ have been computed, the semi-implicit method can be utilized. For each time step $n > 1$, the computation of the nonlinear and Coriolis forcing terms are done by using the nodal values of each field from the previous time step along with the differentiation matrices and block identity matrices. For example, to compute the vorticity of the velocity field at time n,

$$\zeta^n = \frac{\partial u_2^n}{\partial x^1} - \frac{\partial u_1^n}{\partial x^2}$$

we have using the differentiation matrices

$$\boldsymbol{\zeta}^n = \mathbf{D}\mathbf{E}\mathbf{u}^n$$

where $\mathbf{E}$ is the $2N \times 2N$ block permutation matrix defined by

$$\mathbf{E} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix},$$

and $\mathbf{I}$ is the $N \times N$ identity matrix. Similarly, the nonlinear term $\frac{\partial}{\partial x^j}\left(\frac{1}{2}u_1 u^1 + u_2 u^2\right)^n$ can be approximated at the collocation nodes by first computing the vector

$$\mathbf{v}^n = (\mathbf{I}\;\mathbf{I})(\mathbf{u}^n \otimes \bar{\mathbf{u}}^n) \in \mathbb{R}^N$$

where $\bar{\mathbf{u}}^n$ is the vector of contravariant values of $\mathbf{u}^n$, $\otimes$ represents the element by element multiplication operator, and $(\mathbf{I}\ \mathbf{I})$ is the $N \times 2N$ block matrix formed by two consecutive $N \times N$ identity matrices. Applying the differentiation matrix $D_j \mathbf{v}^n$, $j = 1, 2$, then gives the values of $\frac{\partial}{\partial x^j}\left(\frac{1}{2}u_1 u^1 + u_2 u^2\right)^n$ at the collocation nodes.

3.4.2 Complete Algorithm

We conclude this chapter with the complete algorithm for the discrete evolution of the rotational shallow water equations on the sphere using the meshless collocation method proposed in this chapter. The algorithm presented below summarizes all the steps discuss in the chapter.

Algorithm 5.1 Input: N, $\mathcal{X} \subset \Omega$, compactly supported SPD Ψ, time steps $K > 1$

Initialization:

- Compute covariant and contravariant metric terms and Coriolis forcing at collocation nodes $\mathcal{X}$
- Assemble interpolation matrix $\mathcal{A}$ using compactly supported SPD kernel $\Psi : \Omega \times \Omega \mapsto \mathbb{R}$. Choose shape parameter ϵ so that $\mathcal{A}$ is a band matrix with bandwidth $m > 1$
- Compute $\mathcal{A}^{-1}$ and differentiation matrices $D_1 = \mathcal{A}_{x^1}\mathcal{A}^{-1}$ and $D_2 = \mathcal{A}_{x^2}\mathcal{A}^{-1}$.
- Evaluate ICs $\mathbf{u}^0(\lambda, \theta)$ and $\eta^0(\lambda, \theta)$ at collocation nodes on cube to get vectors $\mathbf{u}^0$ and $\boldsymbol{\eta}^0$. Use forward explicit step to get $\mathbf{u}^1$ and $\boldsymbol{\eta}^1$.

For $n = 1 \ldots K$, do

- Compute vorticity $\boldsymbol{\zeta}^n = \mathbf{DEu}^n$ and advection terms

$$D_j\mathbf{v}^n, \quad \mathbf{v}^n = (\mathbf{I}\,\mathbf{I})(\mathbf{u}^n \otimes \bar{\mathbf{u}}^n),\ j = 1, 2$$

- Assemble right-hand side vectors, $\mathbf{R}_{\mathbf{u}}^n$ and $\mathbf{R}_{\eta}^n$, of Helmholtz system.

- Solve for Helmholtz equation for geopotential using conjugate gradient method

$$\boldsymbol{\eta}^{n+1} + \Delta t^2 \eta_0 \mathbf{D}\mathbf{D}^T \boldsymbol{\eta}^{n+1} = \mathbf{R}'_{\eta},$$

where

$$\mathbf{R}'_{\eta} \equiv \mathbf{R}_{\eta}^n - \Delta t \eta_0 \mathbf{D}\mathbf{D}^T \mathbf{R}_{\mathbf{u}}^n.$$

- Update velocity field with

$$\mathbf{u}^{n+1} = \Delta t(\mathbf{R}_{\mathbf{u}}^n + \Delta t \mathbf{D}^T \boldsymbol{\eta}^{n+1}),$$

The computational core of this algorithm of course comes in computing the inverse of the band matrix $\mathcal{A}$ with bandwidth $m > 1$ and the conjugate gradient subroutine for approximating the geopotential at each time step. In the next chapter, we discuss the parallel implementation of the above algorithms using the mpi libraries and give an analysis of their total number of operations and their parallel time complexity. Effective and robust data structures using the modular paradigm in Fortran 90 for the velocity and geopotential fields will also be discussed. In the final part of the chapter we will then present a suite of numerical experiments aimed at demonstrating the high-performance capabilities of semi-implicit, meshless collocation discretization of the rotational shallow water equations.

Chapter 4

The High-Performance Parallel Computation of the Meshless Collocation Method for the Shallow-Water Equations on the Sphere

4.1 Introduction

Since the goal of this thesis is focused on the demonstrating the mathematical and numerical properties of meshless collocation methods, our aim in this chapter is to provide numerical verification and validation of the meshless collocation scheme applied to the rotational shallow-water equations on the sphere that was introduced in the previous chapter. We would like to show computationally that this proposed model can compete with existing high performance methods for approximating the shallow-water equations such as the SEAM (spectral-element atmospheric model) developed at NCAR all while highlighting the advantages and disadvantages of the method.

The chapter is organized as follows. The first part of the chapter introduces two parallel algorithms used in the implementation of the semi-implicit meshless collocation method proposed in the previous chapter. The algorithms are utilized in the two computationally intensive steps of the model; 1) finding the inverse of a band matrix and 2); the parallel conjugate gradient method for the solution of the Helmholtz problem. The algorithms will be discussed in detail using pseudocode along with specific data structure examples. We follow up with examples in Fortran 90 of data structures used for the velocity and geopotential fields along with the covariant/contravariant metric structures used for mapping the cube to sphere. Hardware issues in the parallelization of the model aimed at giving insight into optimal computation time for the model will also be discussed.

Finally, to conclude the chapter, we give a suite of computational experiments to validate the effectiveness of the proposed shallow water model. For this, we will use some standard tests provided by Williamson et al. [71] that have been used for the past two decades for assessing computational accuracy and time benchmarks.

4.2 Parallel Algorithms

In this section we propose some parallel algorithms for the high-performance computation of the semi-implicit meshless collocation approximation of the shallow water model. Specifically, we introduce a parallel algorithm for computing the inverse of a band matrix based on an algorithm originally developed in Mahmood et al. [39]. We then give an analysis of the algorithm's operation complexity as a function of the matrix size.

The second parallel algorithm that we discuss is a parallel conjugate gradient method. Since the system $\mathbf{H}\boldsymbol{\eta} = \mathbf{R}$ must be solved at every time step, where $\mathbf{H}$ is a symmetric and positive definite matrix, a fast conjugate gradient method must be used to approximate $\boldsymbol{\eta}$. As we demonstrated in the previous chapter, the efficiency and robustness of the meshless collocation model relies on the semi-implicit time-stepping scheme which in turn relies on a fast method for approximation $\boldsymbol{\eta}$. The parallel conjugate gradient method has the potential to be a fast and robust solver, but at the cost of finding an effective preconditioner and knowing how to distribute the positive definite matrix and vectors among the processors in an optimal manner. We will limit our discussion to the parallel conjugate gradient method and defer preconditioner issues to the numerical experiment section of the chapter.

Both algorithms are based on the assumption that we have a cluster of processor nodes where each node has one or multiple processor and capable of storing, reading, and writing data with all other nodes using an array of data broadcasting and gather routines. This is a standard assumption in high performance parallel computing and enables the use of the *message passing interface* library (mpi) of Fortran 90.

4.2.1 Fast parallel computation of inverting a band matrix

Recall from section 3.4.1 that in order to compute the differentiation matrices D_1 and D_2, the inverse of the interpolation matrix $\mathcal{A}$ must be computed. Since the $N \times N$ interpolation matrix might be infeasible to directly invert using global SPD kernels, we chose Wendland's compactly supported radial functions and selected the support parameterized by the shape parameter $\epsilon > 0$ such that at most $m > 1$ nodes are in the support. The resulting matrix,

$$\mathcal{A}_{\mathcal{X}} = \begin{pmatrix} \Psi(\mathbf{x}_1, \mathbf{x}_1; \epsilon) & \Psi(\mathbf{x}_1, \mathbf{x}_2; \epsilon) & \Psi(\mathbf{x}_1, \mathbf{x}_3; \epsilon) & \cdots & 0 \\ \Psi(\mathbf{x}_2, \mathbf{x}_1; \epsilon) & \Psi(\mathbf{x}_2, \mathbf{x}_2; \epsilon) & \Psi(\mathbf{x}_2, \mathbf{x}_3; \epsilon) & \cdots & 0 \\ \Psi(\mathbf{x}_3, \mathbf{x}_1; \epsilon) & \Psi(\mathbf{x}_3, \mathbf{x}_2; \epsilon) & \Psi(\mathbf{x}_3, \mathbf{x}_3; \epsilon) & \Psi(\mathbf{x}_3, \mathbf{x}_4; \epsilon) & \cdots \\ 0 & 0 & \Psi(\mathbf{x}_4, \mathbf{x}_3; \epsilon) & \Psi(\mathbf{x}_4, \mathbf{x}_4) & \cdots \\ \vdots & \ddots & \Psi(\mathbf{x}_i, \mathbf{x}_{i-1}; \epsilon) & \Psi(\mathbf{x}_i, \mathbf{x}_i; \epsilon) & \vdots \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & \cdots & \Psi(\mathbf{x}_N, \mathbf{x}_{N-3}; \epsilon) & \Psi(\mathbf{x}_N, \mathbf{x}_{N-2}; \epsilon) & \cdots \end{pmatrix},$$

was shown to be banded, with a bandwidth $m > 1$. It turns out that this band matrix is easier to invert using a parallel version of the so-called connectivity of Schur's complement (CSC) algorithm. The idea behind this fast banded matrix inversion is to properly use the connectivity of Schur's complements (CSC) resulting in an algorithm with time complexity $\mathcal{O}(n\log(n))$ where $n < N$ is a fraction of N.

An outline of the algorithm can be stated as follows. We initiate the algorithm by first transforming the $N \times N$ banded matrix into a tridiagonal block matrix of $n \times n$ block matrices of size $b \times b$ where $b = (m-1)/2$. To do this we take blocks of size $b \times b$ of the original matrix and extend down through the diagonal. This new block matrix is tridiagonal and we can now proceed into computing the inverse of each individual smaller $b \times b$ matrix that can be done in $\mathcal{O}(b^3)$ time on one processor. An example of a banded matrix with bandwidth $m = 5$ and its block tridiagonal representation is given in figure 4.1.

The successive steps in the CSC algorithm then considers a string of calculations where inverses of the block matrices are computed and then the computation

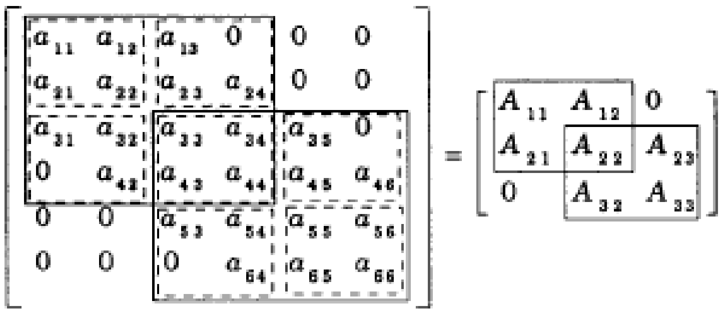

a_{11} a_{12} a_{13} 0 0 0
a_{21} a_{22} a_{23} a_{24} 0 0
a_{31} a_{32} a_{33} a_{34} a_{35} 0
0 a_{42} a_{43} a_{44} a_{45} a_{46}
0 0 a_{53} a_{54} a_{55} a_{56}
0 0 0 a_{64} a_{65} a_{66}
=
A_{11} A_{12} 0
A_{21} A_{22} A_{23}
0 A_{32} A_{33}

Figure 4.1: Example of 6 × 6 matrix with bandwidth 5 and its 3 × 3 block tridiagonal matrix representation.

of the Schur's complements are followed. We outline the CSC steps as follows.

- Compute the forward inverse of the band
- Compute the back inverse of the band
- Compute the inverse outside the band

In the first two steps above, the neighboring blocks are connected through the Schur's complement. We illustrate the three steps of the algorithm with the 3 × 3 block tridiagonal matrix shown in figure 4.1. We will denote the inverse of **A** in this 3 × 3 block matrix as

In the first step, the inverse of each diagonal block matrix is computed and the successive Schur's compliment to compute the remaining diagonals are computed. This process is shown in figure 4.2.

In the second step of the algorithm, the outer block matrices of the tridiagonal matrices are computed by beginning with the last diagonal block matrix and sweeping up towards through the diagonal to the first diagonal block matrix commu-

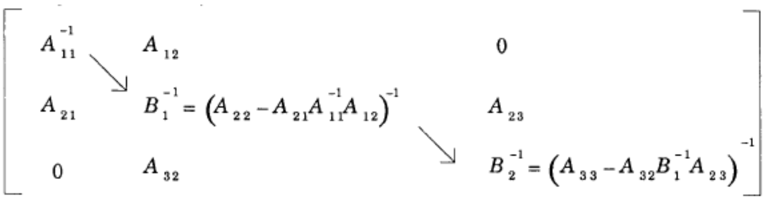

A_{11}^{-1}
A_{12}
0
A_{21}
$B_1^{-1} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$
A_{23}
0
A_{32}
$B_2^{-1} = (A_{33} - A_{32}B_1^{-1}A_{23})^{-1}$

Figure 4.2: Diagram of of the forward inverse process on a simple 3×3 block matrix.

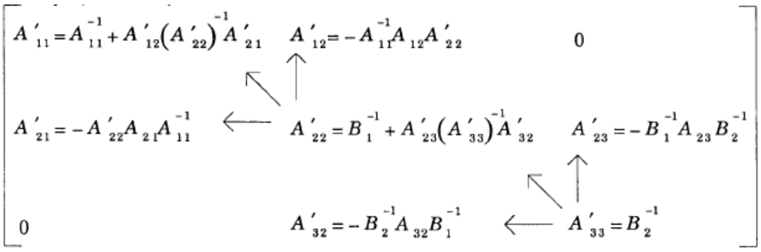

$A'_{11} = A_{11}^{-1} + A'_{12}(A'_{22})^{-1}A'_{21}$
$A'_{12} = -A_{11}^{-1}A_{12}A'_{22}$
0
$A'_{21} = -A'_{22}A_{21}A_{11}^{-1}$
$A'_{22} = B_1^{-1} + A'_{23}(A'_{33})^{-1}A'_{32}$
$A'_{23} = -B_1^{-1}A_{23}B_2^{-1}$
0
$A'_{32} = -B_2^{-1}A_{32}B_1^{-1}$
$A'_{33} = B_2^{-1}$

Figure 4.3: Example of the back inverse process on a 3×3 block matrix.

nicating each diagonal matrix to its neighbors to be used in successive computations of the newly updated diagonal block matrices. The steps involved in back inverses are as follows. Assume there are $n > 1$ block diagonals.

- Define $B_n^{-1} = A'_{nn}$, the final diagonal block matrix.

- For $1 < i < n$, compute the above-diagonal block matrix A'_{ii} at the i-th block diagonal matrix as $A'_{ii} = B_{i-1}^{-1} - A'_{(i)(i+1)}(A_{(i+1)(i+1)})^{-1}A'_{(i+1)(i)}$.

- Compute the left block matrix at the i-th block diagonal as $A'_{(i)(i-1)} = -A'_{ii}A_{(i)(i-1)}A^{-1}_{(i-1)(i-}$ and the above block matrix as $A'_{(i)(i-1)} = -A'_{(i-1)(i-1)}A_{(i-1)(i)}A_{ii}^{-1}$

To illustrate these steps with the 3×3 example, figure shows each block matrix computation for the diagonal and off-diagonal block matrices.

Notice that due to the dependence on the the neighboring block matrices at each diagonal block matrix, this step must also be performed in serial on a mas-

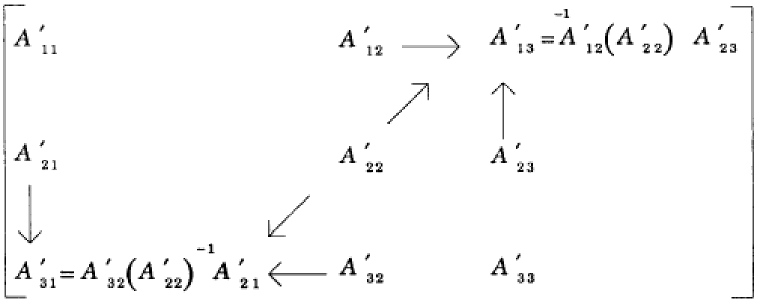

A'_{11}
A'_{12}
$A'_{13} = A'_{12}(A'_{22})^{-1} A'_{23}$
A'_{21}
A'_{22}
A'_{23}
$A'_{31} = A'_{32}(A'_{22})^{-1} A'_{21}$
A'_{32}
A'_{33}

Figure 4.4: Example of the outside inverse computations on a 3 × 3 block matrix.

ter processor that communicates with the other processor nodes using a broadcast routine. Thus the operation count for this step will be $3n * (b^3)$.

The third and final step of the algorithm is the most complex bu can be done on multiple processors; $(n-2)/2$ processors allocated for the lower part and $(n-2)/2$ for the upper part of the tridiagonal matrix. Each processor computes a matrix of the form $XY^{-1}Z$ where Y is the diagonal neighbor, X is the column neighbor block matrix, and Z is the row neighbor block matrix (left or right depending on if in upper or lower part of the matrix). To illustrate, we give an example for the 3 × 3 example once again.

The result is a cascading effect outward toward the upper-right and lower-left corners of the block matrix where in each step outward, only $n - 2/2$ block matrices are remaining and thus only $n - 2/2$ processors are needed, while data communicating with the previously computed matrices are needed.

To find the total operation count, we first consider the serial version of the algorithm. If we let $b = (m-1)/2$ be the size of each block matrix and mult(b), add(b), inv(b) be the operation count for multiplying, adding, and computing inverse of $b \times b$ matrices, then

$$\begin{aligned}&\big\{(n-1)[\text{inv}(b)+6\text{mult}(b)+2\text{add}(m)]\big\} \text{ inverse inside band}\\&+2\big\{(n-1)\text{inv}(b)-\text{inv}(b)+\frac{\text{mult}(m)}{2}\big[\log(n-1)\big](n-1)\\&+\frac{\text{mult}(m)}{2}\big[n-2\big](n-1)\big\}, \text{outside of band}\end{aligned}$$

Assuming that the multiplication and and the inverse of the $b \times b$ matrix can be done in $\mathcal{O}(b^3)$ operation counts, then the entire serial algorithm has between n^2b and $2n^2b$ operations as the block size m increases. If $n >> b$ then the total time complexity of the algorithm is $\mathcal{O}(n^2)$, where again, n is the number of the block matrices in the diagonal.

Now to consider the parallel implementation of this algorithm, we assume $n >> b$ and that we have $(n-2)/2$ processors. In the initialization of the algorithm, the block matrices are partitioned such that every processor node has access to reading and writing the surrounding neighbor block matrices. The i-th diagonal block matrix B_i^{-1} is computed by first computing the block matrix $B_i = A_{ii} - A_{(i-1)(i)}B_{i-1}^{-1}A_{(i)(i-1)}$ and then computing the inverse on one processor. Since the block matrices will be very small (with $b = 4, 5,$ or 6 in practice) this can be done directly. A similar strategy for the second step is also required since this step must be done in serial as well.

The third part of the algorithm is where we utilize the parallel message passing interface. We begin by allocating to the $n-2/2$ processors two $b \times b$ block matrices from the first off-band diagonal (one block matrix from upper and one from lower off-band matrix) to compute all the necessary inverses and matrix products in $(O)(b^3)$ time. Once these inverses are computed in parallel in $\mathcal{O}(b^3)$ time. In the next step, the inverse matrices in the next off-band diagonal are computed, needing only $n-2/4$ processors, and again done in $\mathcal{O}(b^3)$ time. Since the number of block matrices to compute divides in half as the algorithm sweeps out through the off-band block matrices, the time complexity for computing all off-band block matrices in step three is $\mathcal{O}(\log(n) \cdot b^3)$. Assuming $(n-2)/2$ processors, the total number of operations for the entire algorithm is $nb^3 + 3nb^3 + 2n\log(n)b^3$, or $\mathcal{O}(n \log n)$ if $n >> b$.

4.2.2 Parallel Conjugate Gradient Method

The first parallel algorithm that we discuss is an approach to conjugate gradient iterations for approximating solution to $Ax = b$, where A is symmetric and positive definite using message passing on a cluster of connected processor nodes. To motivate the use of a parallel conjugate gradient method, we briefly review the computational iterations involved.

Initializing x_1 as an initial guess, and setting $r_1 = b - Ax_1$, $p_1 = \frac{A^T r_1}{\|A^T r_1\|_2^2}$, we perform for $k = 2, 3, \ldots$

$$\begin{aligned} &\alpha_k = \frac{1}{(Ap_k, Ap_k)}, \quad x_k = x_{k-1} + \alpha_k p_{k-1} \\ &r_k = r_{k-1} - \alpha_k A p_{k-1}, \quad \beta_k = \frac{1}{\|A^T r_k\|_2^2}, \quad p_k = p_{k-1} + \beta_k A^T r_k. \end{aligned}$$

We iterate until $\|r_k\|/\|b\|_2 \leq \epsilon$ where ϵ is usually chosen to be close to machine precision. The search vectors p_k must be computed sequentially since one direction in the conjugate gradient depends on previous directions. The matrix-vector multiplication operations can clearly be parallelized using an efficient partitioning and distribution of the matrices and vectors.

There are a number of methods to determine the data decomposition and distribution for the basic matrix/vector operations. One thing to keep in mind of course is that the distribution of the matrix A across the processors should be consistent with the distribution of the vectors x and b.

Throughout the assembly of the algorithm. We assume that the message passing will be done using the *master-slave* parallel paradigm. In this way, we have a master processor capable of holding the matrix A, and the input and solution vector x. The global reading and writing operations are done uniquely through the master processor. The slave processors carry out the multiplication and addition arithmetic operations. Matrix A on the master processor will stay unchanged throughout the entire number of iterations. Each slave node has then at its disposal a unique part of matrix A and when the need of computing one of the two operations arise, it receives an appropriate vector from the master process.

We consider a standard row-wise partition of the matrix onto p processors such that $p < N$. The matrix is partitioned row-wise onto the processors by taking for $1 \leq i \leq p$ the i-th consecutive set of N/p of matrix A and distributing it to processor i. Each process only stores N/p complete rows of A and a N/p portion of the resulting vector b. Since the vector x needs to be multiplied by every row of the matrix A, every process needs the entire vector. Using the interface mpi, this requires an all-to-all broadcast. This broadcast takes place among p processes and involves messages of size N/p. Each process then multiplies all N/p rows by the vector x and stores the N/p results in a temporary vector.

Using this row-wise matrix-vector paradigm, we can now adjust the computations in the original conjugate gradient method to produce the parallel algorithm. Of course, we need to aim at minimizing the communication time between the processors. Notice that since the Helmholtz matrix $\mathbf{H}$ is symmetric positive definite, in this algorithm, we have $A = A^T$ which greatly simplifies the parallel cg method. We thus propose the following strategy:

- Initialization: Broadcast initial guess x_0, and the i-th set of N/p rows of A and entries of b to processor i. Compute the N/p entries of $r_1 = b - Ax_0$ on each process and send to master to gather entries.

- Compute $p_1 = \frac{Ar_1}{\|Ar_1\|_2^2}$

- For $k = 2, \ldots$ do

- Compute vector $t = Ap_k$, broadcast to master, then set on master processor $\alpha_k = \frac{1}{\|t\|_2^2}, \quad x_k = x_{k-1} + \alpha_k p_k$

- Compute $r_k = r_{k-1} - \alpha_k A p_{k-1}$ using row-wise partition. Broadcast the i-th N/p portion of r_k to master processor, gather the entries of r_k on master processor; broadcast to all processes.

- Compute vector $t = Ar_k$, broadcast to master, and then set $p_k = p_{k-1} + \frac{t}{\|t\|_2^2}$

- Check on master process $\|r_k\|/\|b\|_2$.

The time complexity of each iteration of this algorithm is a function of the row matrix-vector products on each process along with the broadcast and gather routines amongst the processors. The communications used for the all-to-all broadcast and gather routines are clearly hardware dependent. For the matrix-vector routines after communication, each process spends time N^2/p multiplying the N/p partitions of the matrix A.

In the numerical experiment section of this chapter, we give results for the timings and scalability of this algorithm. In particular, we will be interested in investigating the communication time between process for the mpi routines in some experimental integrations of the model.

4.3 Implementation of velocity and geopotential field

In this section, we discuss some implementation issues concerning the velocity and geopotential fields. In particular, we discuss some of the data structures which were implemented in our model using Fortran 90 and the computation of the metrics necessary for mapping both fields to and from the sphere.

We begin with the structures for computing the geopotential and velocity. Below is Fortran 90 code which implements data structure in a Fortran 90 module. We first define vectors of size $nX \times nY \times 2 \times 4$ for the velocity and $nX \times nY \times 4$ for the geopotential. The parameters nX and nY denote the number of collocation nodes in the x and y direction on each face of the cube. The last dimension is used to store previous time steps for the semi-implicit method. The second group of data structures features the exact coordinate values of the nodes on the sphere and cubed (in Cartesian coordinate system). These are given in code below.

3

```
integer,public, parameter :: nX = 40

integer,public, parameter :: nY = 40


!==========Data structure for random node element
type, public :: meshless_elem_state_t
     sequence
     real (kind=real_kind) :: v(nX,nY,2,4)  !velocity
     real (kind=real_kind) :: p(nX,nY,4)    !geopotential
end type

!=========Identification tag of element
type, public :: meshless_elem_t
     sequence
     integer(kind=int_kind) :: LocalId
     integer(kind=int_kind) :: GlobalId

! ==========Coordinate values of element points on sphere/cube

type (spherical_polar_t) :: spherev(nX,nY)

type (spherical_polar_t) :: spherep(nX,nY)
type (cartesian2D_t)     :: cartv(nX,nY)

type (cartesian2D_t)     :: cartp(nX,nY)
```

The next group of data structures are the 2×2 metric terms evaluated at the collocation nodes. Furthermore, the interpolation matrix of the SPD kernel and its inverse are stored in separate vectors for each the geopotential and velocity.

3

```
!====================================
! Metric terms
!====================================

real (kind=real_kind)    :: met(2,2,nX,nY)
real (kind=real_kind)    :: metinv(2,2,nX,nY)
real (kind=real_kind)    :: metdet(nX,nY)
real (kind=real_kind)    :: metdetp(nX,nY)
real (kind=real_kind)    :: rmetdetp(nX,nY)
real (kind=real_kind)    :: D(2,2,nX,nY)
real (kind=real_kind)    :: Dinv(2,2,nX,nY)
```

```
!======================================
! Interpolation matrix terms
!======================================

real (kind=real_kind)    :: mp(nX,nY)
real (kind=real_kind)    :: mv(nX,nY)
real (kind=real_kind)    :: rmv(nX,nY)

!======================================
! Coriolis term
!======================================

real (kind=real_kind)    :: fcor(nX,nY)
```

To compute the covariant vector values of the velocity, we first extract the state of the velocity on the sphere at all the nodes on each face and multiply each component by the metric g_{ij}. Similarly, Similarly, the values of $g\mathbf{v}$ are calculated by multiplying the value of the determinant at each node by the velocity components. An example is shown below.

3

```
  do j=1,nX
```

```
     do i=1,nY

     !Extract value of velocity at i,j collocation node
     v1 = face(ie)%state%v(i,j,1,1,n0)
     v2 = face(ie)%state%v(i,j,2,1,n0)

     !Multiply by the metric evaluated at i,j-the node
     vco(i,j,1) = face(ie)%met(1,1,i,j)*v1 + face(ie)%met(1,2,i,j)*v2
     vco(i,j,2) = face(ie)%met(2,1,i,j)*v1 + face(ie)%met(2,2,i,j)*v2

     gv(i,j,1) = face(ie)%metdet(i,j)*v1
     gv(i,j,2) = face(ie)%metdet(i,j)*v2

     end do
  end do
```

With some of the data structures discussed in the implementation of the model in Fortran 90, we finally discuss the hardware used in the forthcoming simulations in final section of the thesis. Optimizing the parallelization of the model is a pertinant task and shall be discussed as well.

4.4 Numerical Experiments

4.4.1 Verification and Validation

The grand challenge in scientific computation when proposing a new computational model to provide quantitatively accurate predictions is proving to the respective community the verification and validation of the computational model. With any new computational model, verification deals with the confirmation of the mathematical accuracy of the calculations along with the error-free operation of the underlying software. As this is ultimately a mathematical and computer science challenge, providing tests to secure such verification can take a large percentage of the total work put into the model. The second problem faced by a new model is validation which asks for confirmation that the physical model used for the computation is a correct representation of the real physical phenomena, which is ultimately an experimental challenge.

As it is well established that the spherical rotational shallow-water equations represent a simplified model of the dynamics of the atmosphere, Williamson et al. [71] have proposed a series of eight test cases for the equations in spherical geometry. It is proposed by the authors that in order to have any type of success with a new numerical scheme for an a climate model, successful integrations of the numerical scheme with these test cases are imperative. The purpose of the tests are to examine the sensitivities of a numerical scheme with many computational challenges faced in atmospheric modeling such as stabilization of the scheme for large time steps over a long period of time, the pole problem, simulating flows which have discontinuous first-derivatives in the potential vorticity, and simulating flows over mountain topographies.

In the final numerical section of the thesis, we visit four different test cases arranged to challenge the meshless collocation method for the shallow water equations introduced in the previous chapter. The test cases are in increasing complexity and realism and achieve more subtle aspects of atmospheric flows. The challenges for the numerical schemes range from testing the long-term steady state global accuracy to the approximation ability of regional data with large gradients. A comparison with a spectral-element approximation of the rotational shallow water model offered by the NCAR *homme* project [32] and a high-resolution spherical harmonic solution provided by Williamson will provide insight into the ability, advantages, and disadvantages of the meshless collocation model. In particular we are interested in verifying and validating the numerical model with the following characteristics in mind:

- Approximability: How robust and accurate is the approximability of the numerical method compared to that of spectral-element and spherical harmonic methods?

- Validity: How well do the actual numerical solutions computed satisfy the underlying conservation principles of the continuous model over long integration periods?

- Scalability: How well does the meshless collocation method provide scalability in the computational time complexity of the overall simulation time? Namely, what is the overall computational speedup when additional processor nodes are utilized?

- Local Refinement: Does meshless collocation provide an efficient means for refining the approximation locally during the time evolution of the model?

For a new numerical model to successfully become integrated as a dynamical core for larger general circulation models, these four model characteristics must certainly be well studied and compared to other numerical models. The first two items in the list are clearly the more important properties that a new numerical method must endow; if the numerical model isn't at least as accurate as pre-existing models nor satisfies the underlying conservation principles of the model, then the numerical method will very likely become useless. The second two properties are the two most important aspects in the high-performance computational environment of geophysical models. As already discussed, local refinement to achieve smaller scale approximations in certain regions of interest without the addition of noise across the boundaries of the refinement is a desirable goal in global baroclinic general circulation models. In test case 5 of the numerical suite, we demonstrate local refinement of our meshless collocation method and show that it achieves approximation refinement without adding noise.

All these issues will be discussed further in the various test cases where numerical tests will be provided to validate the claims.

4.4.2 Test Case 2: Global Steady State Nonlinear Zonal Geostrophic Flow

As the second test case of the seven provided by Williamson et. al, a steady state flow to the full non-linear shallow water equations is prescribed and the challenge for a numerical scheme is simply to test the numerical stability with respect to l_1, l_∞ errors over time. Since the flow is steady, the numerical scheme should be able to integrate the model for many steps without instabilities. Case 2 also provides a benchmark for timing various machines and different number of processors. It exercises the complete set of equations while benefiting from a steady state analytical solution and thus no extra computations are required during integration. For computational timing purposes a11 extra output processes of the data should be removed to achieve optimal computation time.

The constant non-divergent velocity field is given by

$$u = u_0(\cos\theta\,\cos\alpha + \cos\lambda\,\sin\theta\,\sin\alpha),$$

$$v = -u_0 \sin\lambda\,\sin\alpha,$$

where $u_0 = 2\pi a/(12days)$. The geopotential field is given by

$$\eta = gh_0 - \Big(a\Omega u_0 + \frac{u_0^2}{2}\Big)(-\cos\lambda\cos\theta\,\sin\alpha + \sin\theta\,\cos\alpha)^2,$$

with constant $gh_0 = 2.94 \times 10^4 m^2/s^2$ (m/s = meters/second).

4.4.2.1 Numerical Convergence

For this numerical experiment, we take full advantage of the steady state solution property of this test case by constructing convergence tests of different collocation node arrangements. We apply some of the ideas given in Staniforth [52] to examine the properties of the meshless collocation over long time integrations.

In these experiments, our integration takes place over a large time interval of 100 days where we will vary the location and size of the regional approximation while performing L_∞ error analysis. The reason we choose the L_∞ norm to compute the errors is motivated by the desire to seek the largest localized pointwise error due to the pointwise approximation nature of meshless collocation. The collocation nodes that we utilize for each integration are the Gauss-Lobatto-Legendre distributions (see numerical section of chapter 2) on 5 of the faces and a random collocation node distribution will be used on local regions of the sixth face. The reason we choose two different distributions for the collocation nodes on the cube-sphere will be to emphasize the independence of the meshless collocation method with the given node distributions.

To initialize our first experiment, we divide the cubed-sphere into 24 total rectangular regions (4 regions on each face) and compute the metric tensors and Coriolis forcing at 64 Gauss-Lobatto-Legendre points in each rectangle using the data structures from section 4.3 giving a total of 64×4 nodes on each face of the cubed-sphere. We then allocate on one face of the cubed-sphere a distribution of 64 randomly placed nodes. Figure (4.5) shows the cubed-sphere tiled onto the rectangle

$[0, \pi] \times [0, 2\pi]$.

To assess the numerical convergence of the model as the approximation is refined, we increase the number of collocation nodes in each rectangle and re-integrate the model in time for 100 days. After every rotation of the sphere, we record the L_∞ error using the exact solution and plot the error over the 100 days for each mesh. In figure (4.6) we can clearly see the convergence of the approximation for an increase in the total number of collocation nodes. This meshless collocation convergence is known as the h-convergence, where we recall that h is node saturation parameter of the global node distribution $\mathcal{X}$. Furthermore, we can deduce that the location of the meshless regional approximation has no effect on the global convergence of the method, and that the errors grow fast initially, but level off after about 40 days of integration.

We deduce from the 3 different time stepping trials that the meshless collocation method is seemingly stable with a satisfactory convergence results for the refinement in approximation using a time step $\Delta t = 1/288$.

4.4.2.2 Scalability of Model

The final quantitative aspect of the meshless collocation shallow water model that we would like to investigate using this test case is the computational scalability of our numerical method. An important role in the construction of a high-performance computational model for geophysical fluid dynamics on the sphere is the scalability of the numerical model in regards to the parallelization of the model

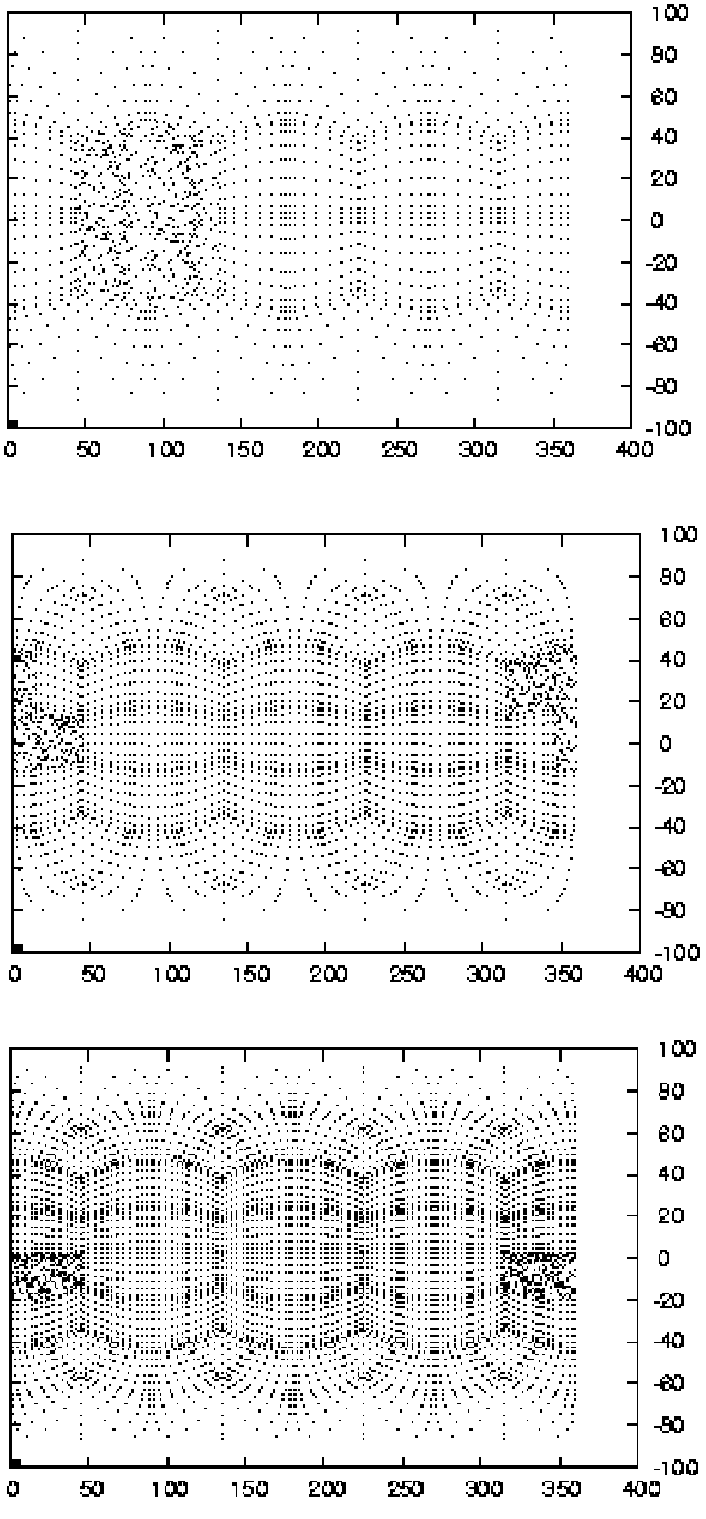


Figure 4.5: Mesh arrangements for the meshless collocation approximation on the sphere. The cube is tiled using 24 regions and Gauss Lobatto Legendre collocation nodes are distributed on each region (top) along with two additional refinements (middle) and (bottom).

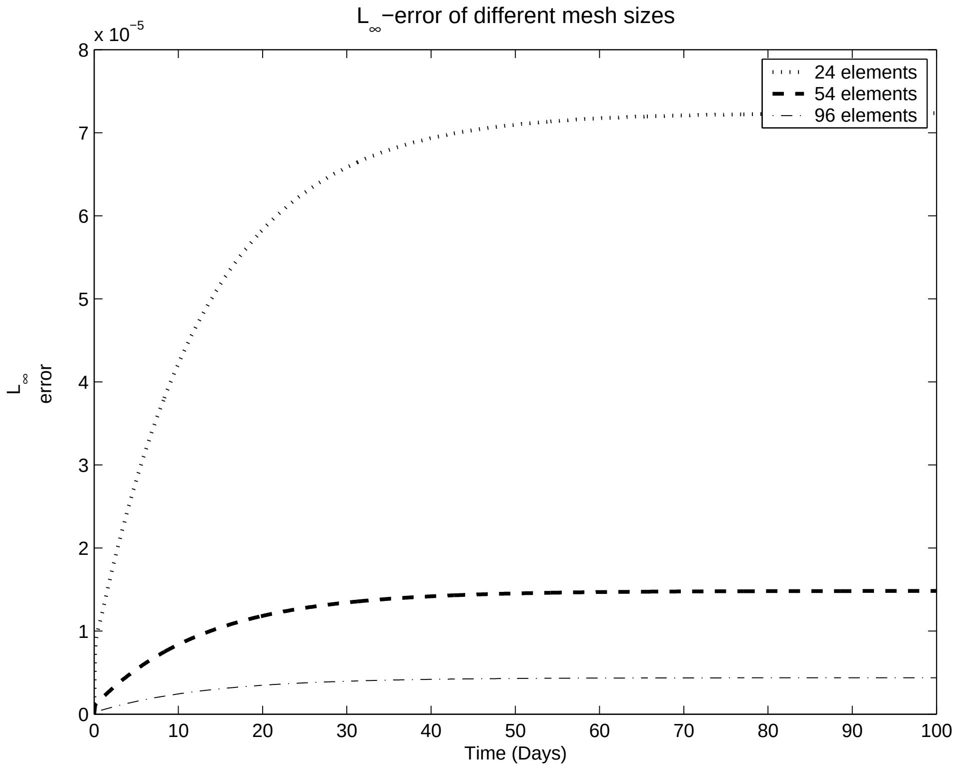


Figure 4.6: L_∞ error of the three mesh configurations. We use a time step Δt of 1/288, namely 288 time steps per rotation of the sphere.

throughout the course of the time integration of the model. How does the overall time complexity improve when additional processes are added to the computational core. This is mainly a question of achieving optimal parallelization of the numerical methods algorithms. Achieving optimal performance on any hardware platform requires a properly designed layout of data structures and algorithms to take advantage of the full capacity of the hardware.

As discussed previously, the main computational core of our model is in constructing the inverse of the interpolation matrix $\mathcal{A}_{\mathcal{X}}$ and in performing the conjugate gradient steps during the time evolution. We now assess how well the parallel algorithms scale with respect to an increase in the number of processors. After the necessary parts of the model were parallelized using mpi for the parallel algorithms from section 4.2, model test runs took place on three different high-performance machines, *Seaborg* and *Jacquard* at NERSC and *Palm* at NASA Goddard. *Jacquared* is a 640-CPU Opteron cluster (320-dual processor nodes) running the Linux operating system **SUSE**. With a processor clock speed of 2.2GHz and a theoretical peak performance of 4.4 Gflops/s, *Jacquard* was the fastest of the three computers for a long integration of the model.

Throughout the runs, 10,000 collocation nodes were initialized on 25 uniform regions on each face of the cube with an equal amount of nodes allocated to each face of the cube. Table (4.1) shows the statistics of a run through test case 2 of the simulation for 200 days (28,300) time steps using the semi-implicit time stepping method. The the total time in μ-seconds of the initialization of the cubed-sphere with meshless collocation including the computation of the metric terms and the inverse of the SPD kernel interpolation matrix $\mathcal{A}_{\mathcal{X}}$ using the parallel CSC algorithm is shown in the second column. Boundary communication between processor nodes is shown in the third column and the total time in seconds of the model run is shown in the last column. It is interesting to note that about 95 percent of the total communication of the model was done during the Helmholtz solver which uses `MPI_allreduce` and `MPI_alltoall`, so it is not surprising to see that the total run time achieved a minimum when there were 2-3 regions of the cubed-sphere allocated to a processor.

From the clocked performance of the parallel model, we see that the optimal amount of processor nodes is obtained at 8. A dramatic improvement from one node to 8 is obtained, however due to increased communicated between the processors as the number of nodes increases, the total run time of the integration is slowed marginally. In order to see if we can increase the performance of the model run, we increase the total number of collocation nodes of the model, while keeping the number of time steps the same. Table 4.2 shows the model integration using 15,000 collocation nodes on the cube and and 28,300 time steps.

With a higher number of collocation nodes, the results of initialization, communication, and total run time shows an improvement in the parallelization efficiency. With a 50 percent increase in collocation nodes, the minimal total run time over all the different processor node layouts has increased by only 43 percent, giving much better computational efficiency. This is most likely due to better scaling properties of the matrix and vector partition layout in the parallel conjugate gradient steps.

We now compare the scalability of the meshless collocation model with the

Table 4.1: Initialization and total boundary communication time in μ-seconds and total runtime of model in seconds for 10,000 collocation nodes and 28,300 time steps.

NP	Init.	Communication	Total time (sec.)
1	9085	49.68	396.11
2	8134	963.08	232.96
4	7727	1673.52	164.27
6	43950	2481.28	162.85
8	122247	2687.13	**160.65**
10	11680	2863.83	168.76
12	72927	3265.94	**161.96**
14	14378	3473.93	173.38
16	707678	3698.40	185.17.11
18	23657	4014.71	189.41
20	18611	4056.69	191.04

homme spectral-element model developed at NCAR. The scaling properties of the *homme* spectral element model were taken from Tribbia et al [32].

Figure 4.7 shows the overall computational time speedup scale for the meshless collocation method against the spectral-element method for a time integration of 100 days. In the left plot, the spectral element solution has 25 elements per face while using Legendre polynomials of order 8, giving a total of 9600 integration nodes. For the meshless collocation, to prevent biases using different grids, we compute the

Table 4.2: Initialization and total boundary communication time in μ-seconds and total runtime of model in seconds for 15,000 collocation nodes and 28,300 time steps.

NP	Init.	Communication	Total time (sec.)
1	17082	110.68	692.16
2	14121	1362.01	381.10
4	12219	2621.11	294.72
6	67951	4019.38	272.42
8	202241	5112.67	250.12
10	22682	6078.82	249.76
12	92921	6015.94	237.11
14	32258	6389.93	**229.21**
16	907272	6698.40	241.11
18	40258	7134.12	256.12
20	38634	7366.54	261.32

collocation solution at each time step using the integration nodes as the collocation nodes. The right plot features 36 elements per face with Legendre polynomials of order 8 for a total of 18816 nodes.

The slightly larger speedup factors for higher amount of processor nodes in the collocation approximation is due to the fact that there is less communication between processors in the gather and scatter *mpi* operations than in the spectral element case. For example, the edges of elements and at the cube corners must be

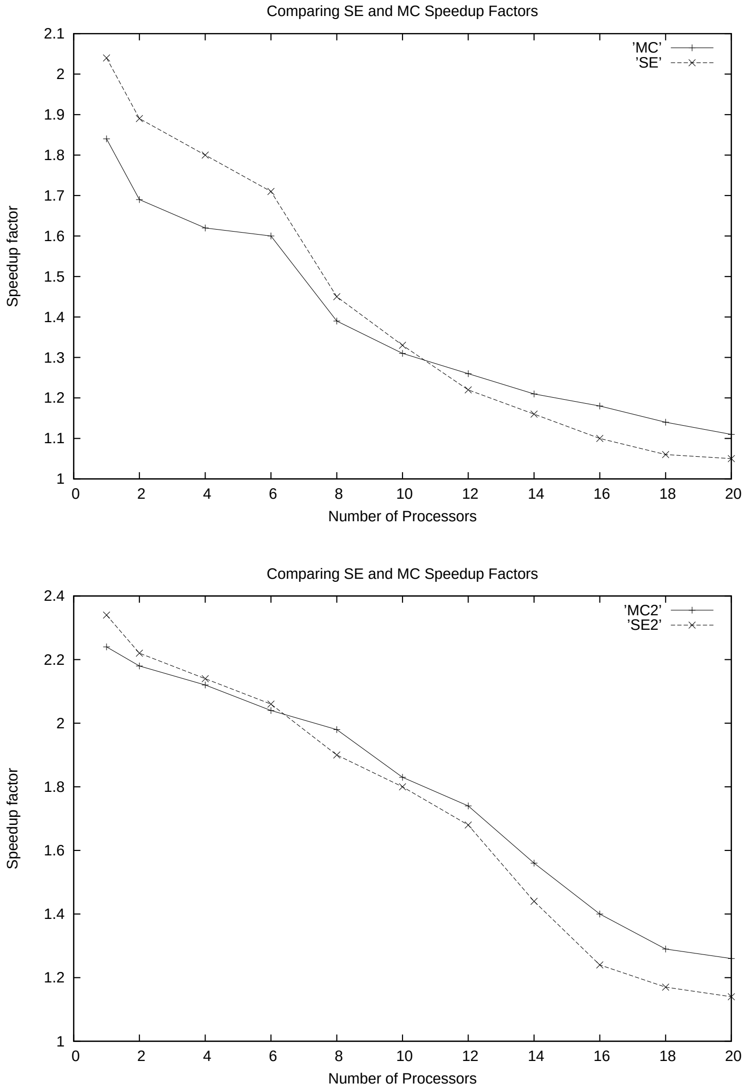


Figure 4.7: Scalability factors for the spectral element (SE) and meshless collocation (MC) solutions. Left 9600 nodes; right 18816 nodes.

averaged in order to obtain a solution at each time step in the SE approximation. This might require additional gather and scatter routines. The MC method does not require such averaging and thus spends less time communicating. We clearly see that our meshless collocation method competes with the spectral element model on scalability issues. However, it would be difficult to assess a comparison with finite element methods in general since the parallelization in other implementations of finite element type models might be better optimized.

4.4.3 Test Case 5: Flow over Mountain

This case was designed to study the different effects of local refinement methods and is the first case of the more difficult last three cases where the exact solutions are not known. It features a sharp gradient change in the geopotential, a feature to which we will pay particularly close attention. It features the zonal flow from test case 2 presented above with $\alpha = 0$, but with the obstruction of a local mountain. The geopotential is also given as in test case 2 but with mean height $h_0 = 5400m$. The mountain height and radius in meters is given by

$$h_s = 2000(1 - 9r/\pi),$$

where $r^2 = \min\{R^2, (\lambda - \lambda_c)^2 + (\theta - \theta_c)^2\}$ and centered at $\lambda_c = -\pi/2$ and $\theta_c = \pi/6$.

Because the flow impinges on a mountain, the resulting flow after short time is no longer steady. The mountain introduces waves into the flow causing the interaction of waves from the nonlinearity of the equations. Figures 4.8 and 4.9 shows isolines of the geopotential height at various times. The numerical solution was computed using 1560 uniform nodes on the cubed sphere, 260 nodes per face, with a time step of $\Delta t = 1/288$.

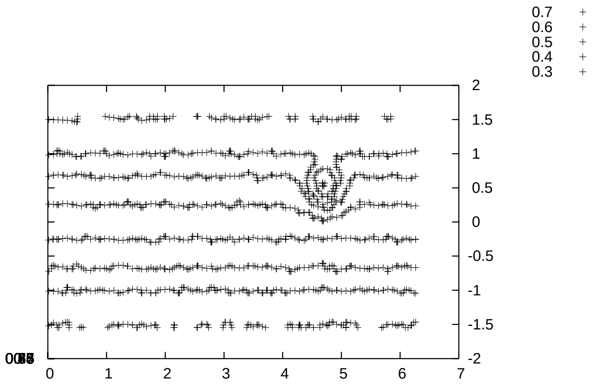
4.4.3.1 Validity of Meshless Collocation Solution: Computing Invariants

In our first experiment with this test case, we wish to assess approximation accuracy and validate the meshless collocation numerical solution as a reliable solution to the rotational shallow water equations. To do this, we will use a series of integral invariants to measure certain conservation properties over long periods of time integration. The invariant is computed up to a given time $t > 0$ as follows. For an integrable function $h : [0, 2\pi] \times [-\pi/2, \pi/2]$, we define the operator I by $I(h) = \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} h(\lambda, \theta) \cos\theta d\theta d\lambda$. Now for a given time $t > 0$, we will compute the normalized global invariants given by

$$I_i(\xi(t)) = \big(I[\xi(\lambda, \theta, t)] - I[\xi(\lambda, \theta, 0)]\big)/I[\xi(\lambda, \theta, 0)] \tag{4.1}$$

- $i = 1$: Total energy $\xi = 0.5\eta'\mathbf{v} \cdot \mathbf{v} + \frac{1}{2}(\eta^2 - \eta_0^2)$,
- $i = 2$: Potential enstrophy $\xi = 0.5(\zeta + f)^2/\eta'$,
- $i = 3$: Mass $\xi = \eta'$,
- $i = 4$: Vorticity $\xi = \zeta$,

We compute the invariants over a 200 day integration period with the effects of a locally refined grid. In the first 200 day simulation, we used the uniform grid from



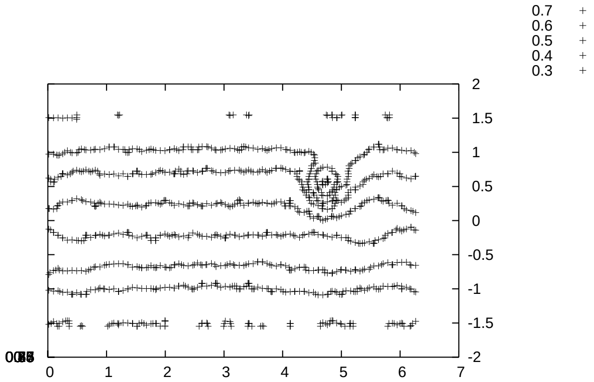


Figure 4.8: Geopotential height isolines of days 1 and 3 of model integration. The model used 1560 uniform collocation nodes over entire cubed-sphere. As clearly seen, steady state flow from test case 2 becomes nonlinear due to the obstructive mountain.

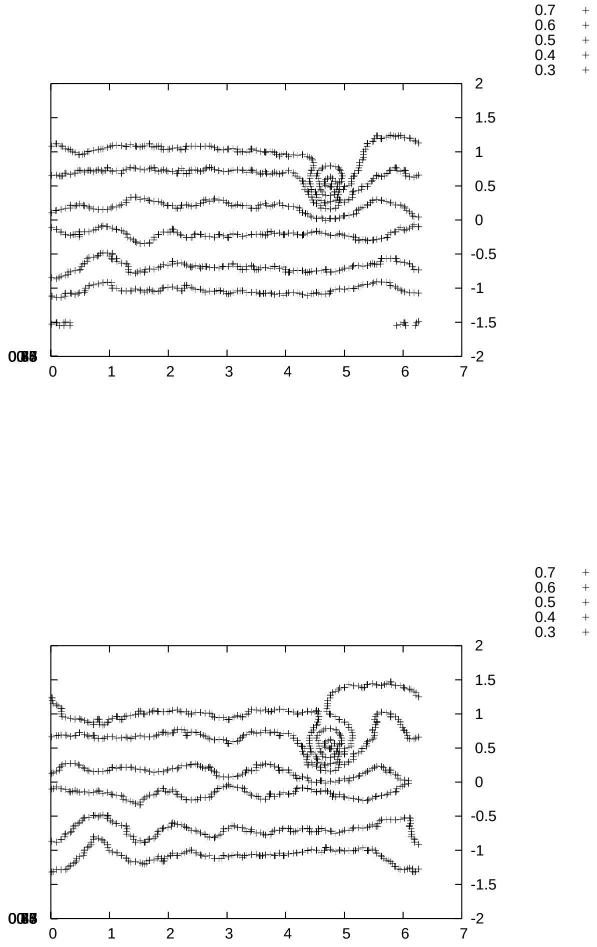


Figure 4.9: Geopotential height isolines of days 5 and 8 of model integration. Flow is clearly nonlinear at this point and large gradients form near the mountain.

figure 4.10 (top) and computed the integral invariants using a standard Gaussian quadrature routine over each face of the cube. Next, to see the effects of local refinement in the domain, we add additional collocation nodes to faces 4, 5, and 6 of the cube and compute the same integral invariants over the 500 day period. Face 4 and 6 of the cubed-sphere feature the local dynamics surrounding the mountain, thus it is very important to observe if there is any additional noise. Each simulation uses a small time step of $\Delta t = 1/846$, or 846 time steps per rotation of sphere, to perform the time stepping. Plots of the invariants over the second grid (middle) are in figures 4.11. In the final run, we add an additional 400 collocation nodes to face 4 of the cube to get the grid featured in figure 4.10 (bottom).

Because the invariants are normalized using the initial conditions of the problems, the numerical results of the invariant plots over time ideally should be close to zero as time goes on. However, due to the small numerical errors building in the solution over time, this cannot be the case. We can only hope that the invariants improve when we locally refine the grid.

We see that an improvement in all the computed invariants is accomplished due to the local refinement which capture the smaller scales around the mountain. This is an important characteristic of the model since we can now deduce that through the use of local refinement, higher accuracy of the smaller subscales that might feature sharp gradients can be achieved such as in this impinging mountain case .

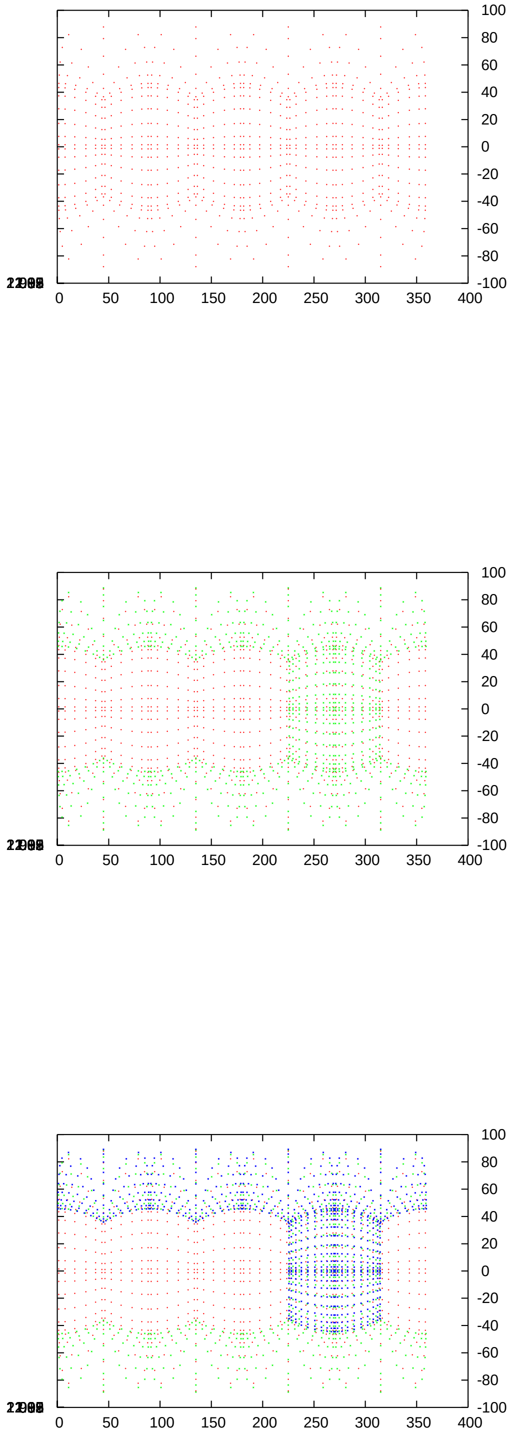


Figure 4.10: Collocation node arrangements I, II, and III. The grid of nodes is refined locally on II and then refined again on III.

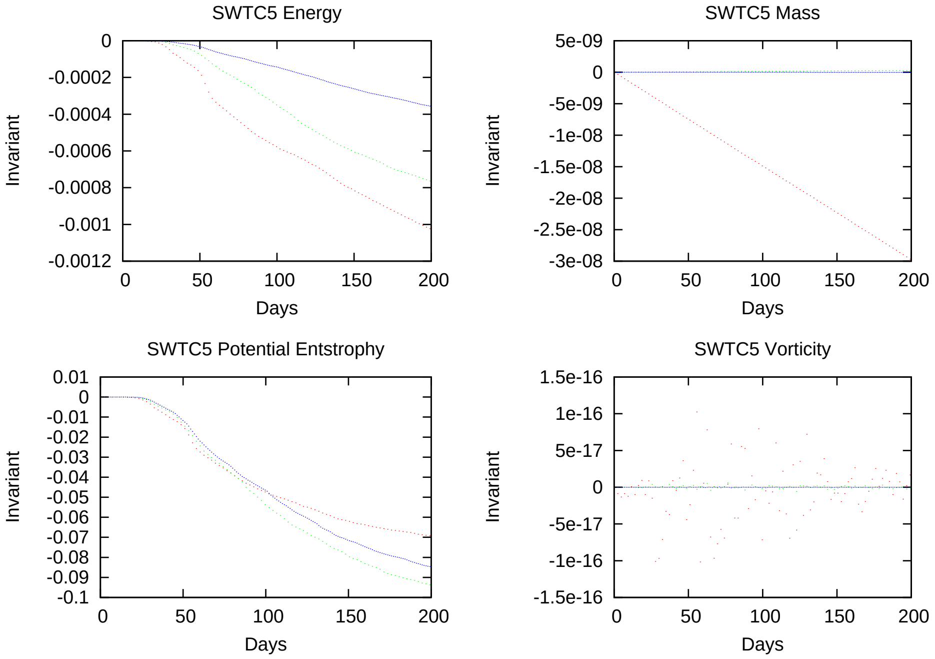


Figure 4.11: Invariants of the MC solution over 200 days of integration for different node sets. (Red) Collocation node set I, (Green) Collocation node set II, (Blue) Collocation node set III.

4.4.3.2 Local Refinement in Meshless Collocation

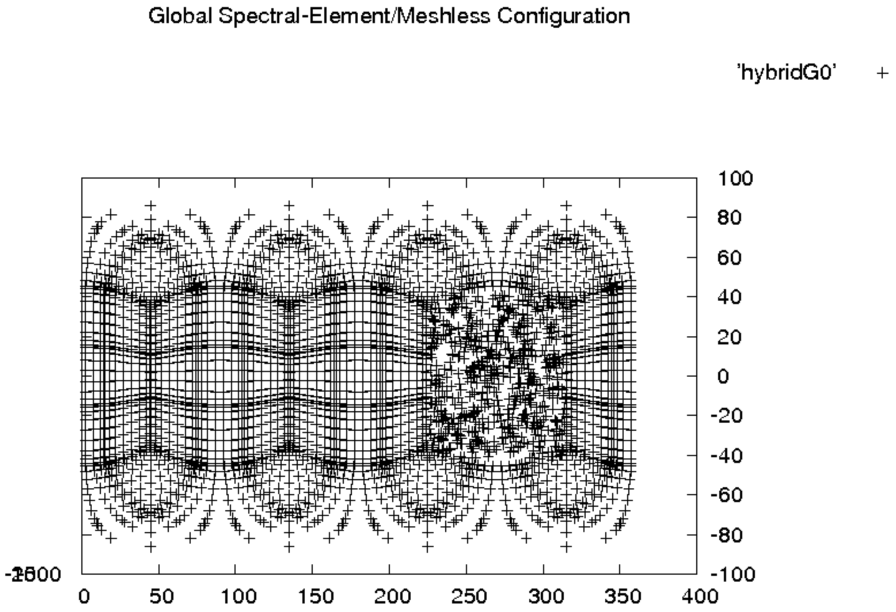
In the next experiment, we investigate the numerical solution behavior of the regional meshless approximation given on a local region of the cubed-sphere. To do this, we begin by initializing the cubed sphere with 9 rectangular regions on each face and distribute 64 GLL collocation nodes on each region. Furthermore, we allocate on the second face of the cubed-sphere a collection of N random points and compare the time stepping evolution with three different N values $400, 900$ and then 1200. The figures in 4.12 show the initial condition of the cubed-sphere layout at two different angles with the second face allocated to 900 nodes.

As seen in the figures, the local region we are focused on with a large number of nodes on the second face of the cubed-sphere was chosen to include the region directly surrounding the mountain located at $(\lambda_c, \theta_c) = (-\pi/2, \pi/6)$ which is the more difficult region in this particular test case to approximate.

Plots of the approximate solution to the geopotential field at 44 and 100 days are shown in figures 4.13 with the contours of the approximated geopotential at 44 and 100 days given in figures 4.14.

As clearly seen, the regional approximation of face 2 handles the fast moving waves without the presence of instabilities at the interface between the neighboring faces of the cube. To see more clearly the regional meshless approximation at 44 and 100 days, a plot of the approximation strictly on a localized region of face 2 is given in figures 4.15.

We now examine the convergence of the meshless approximation inside the lo-



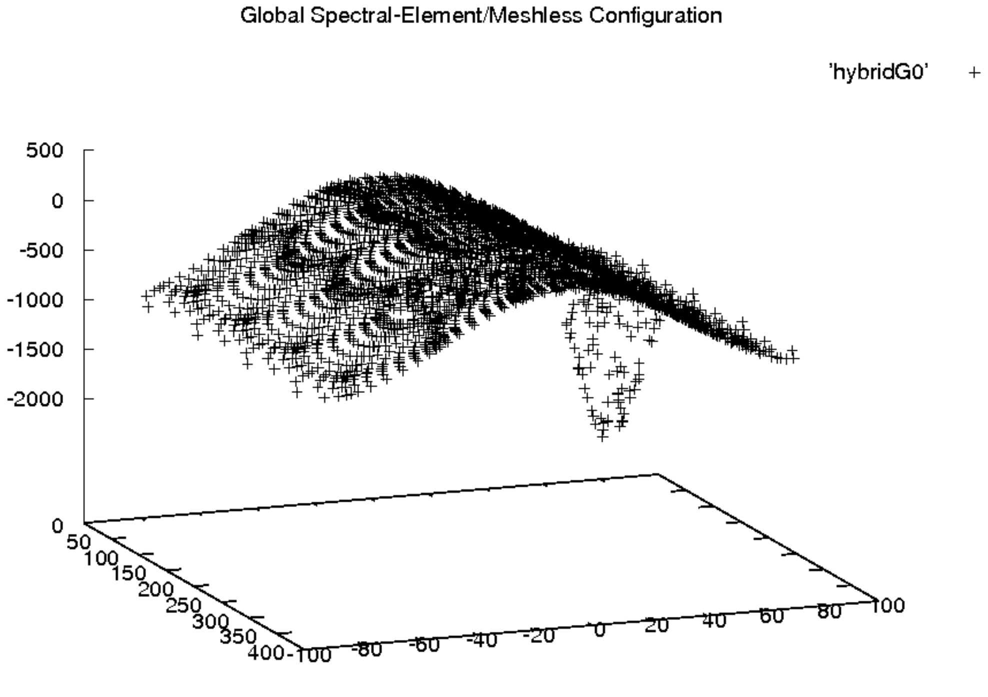


Figure 4.12: Initial condition of cubed-sphere layout at two different angles with 900 nodes allocated to face 2.

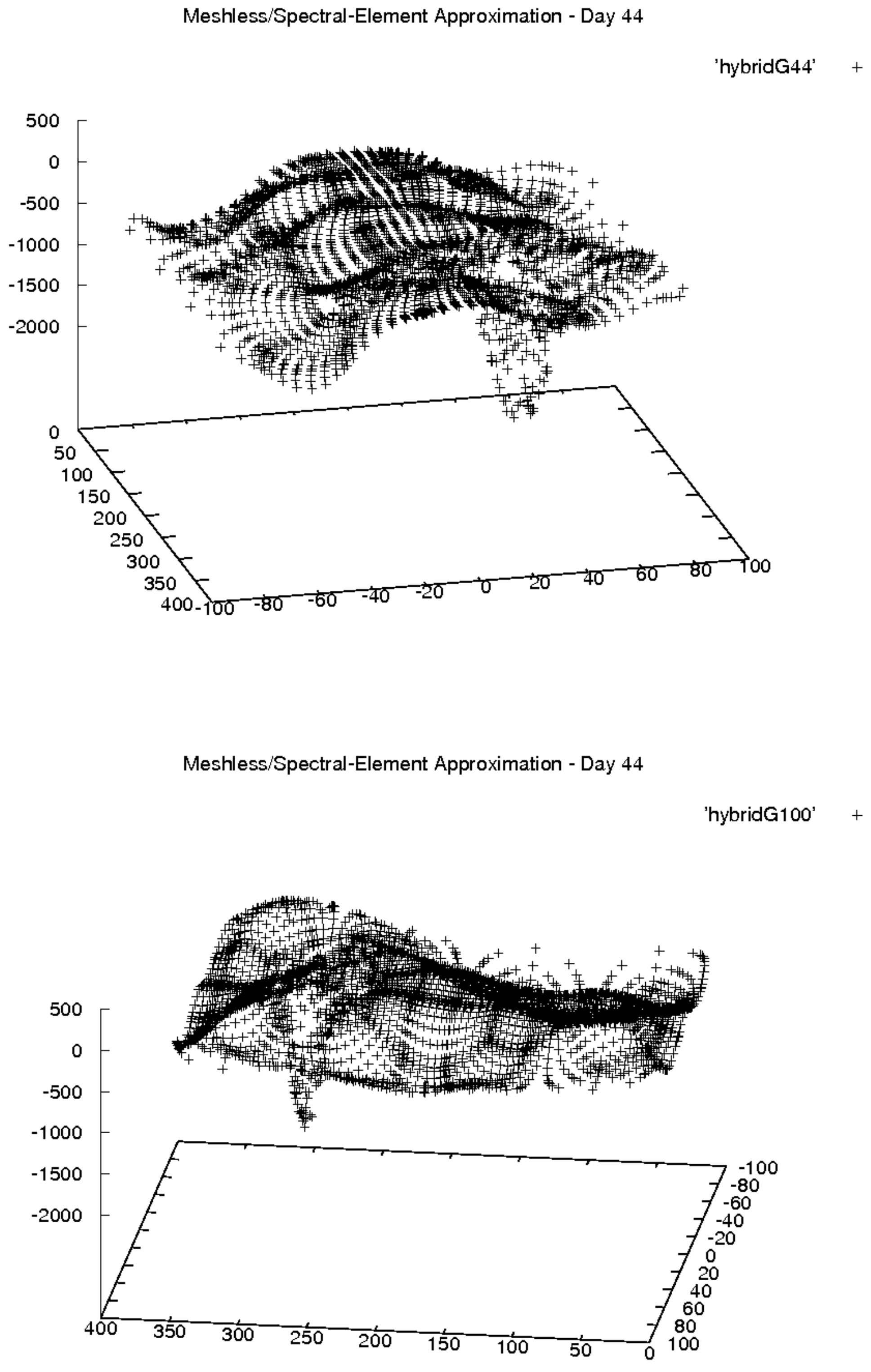


Figure 4.13: Plots of the geopotential meshless collocation approximation after 44 and 100 days.

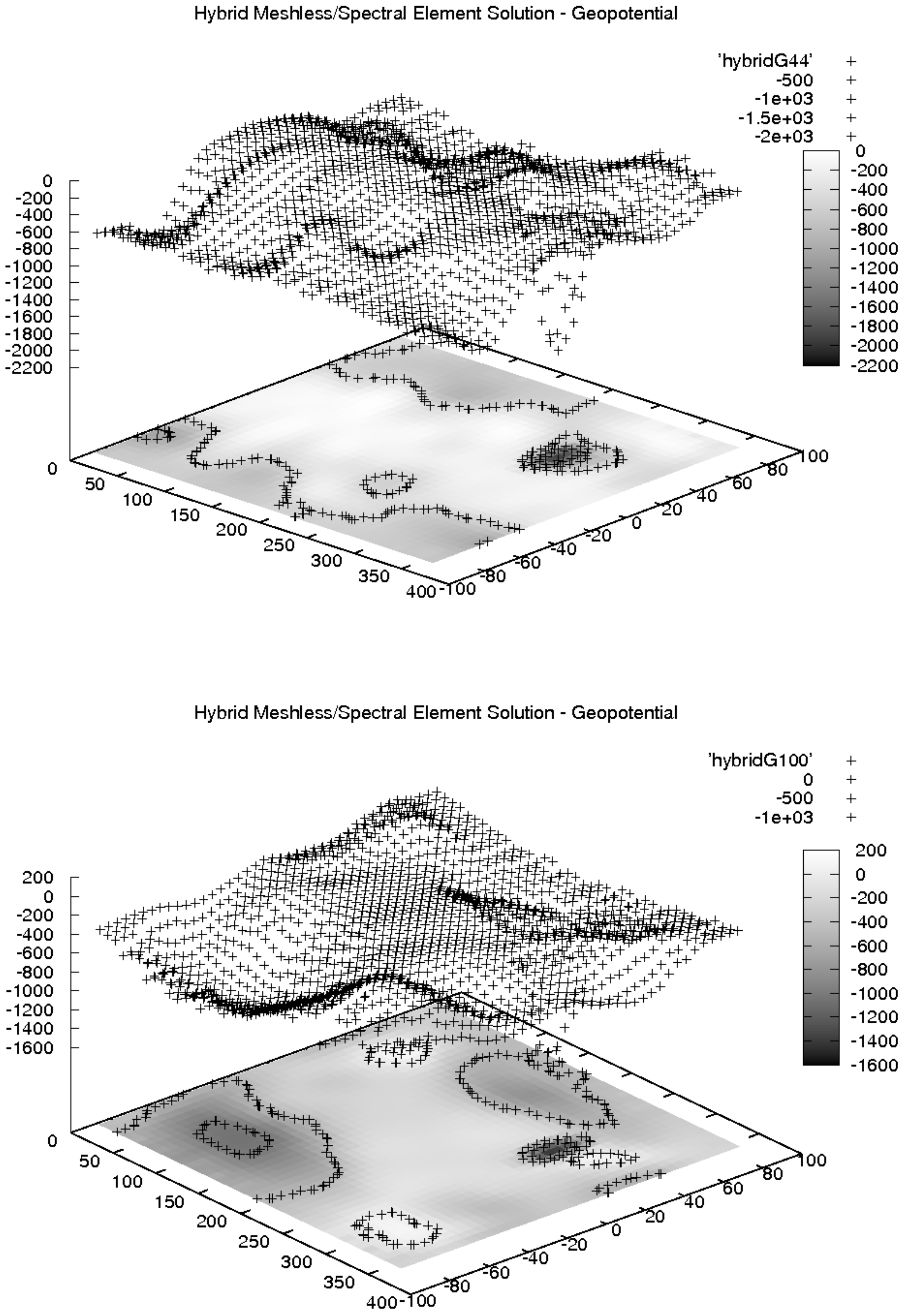


Figure 4.14: Plots and contours of the geopotential approximation after 44 and 100 days

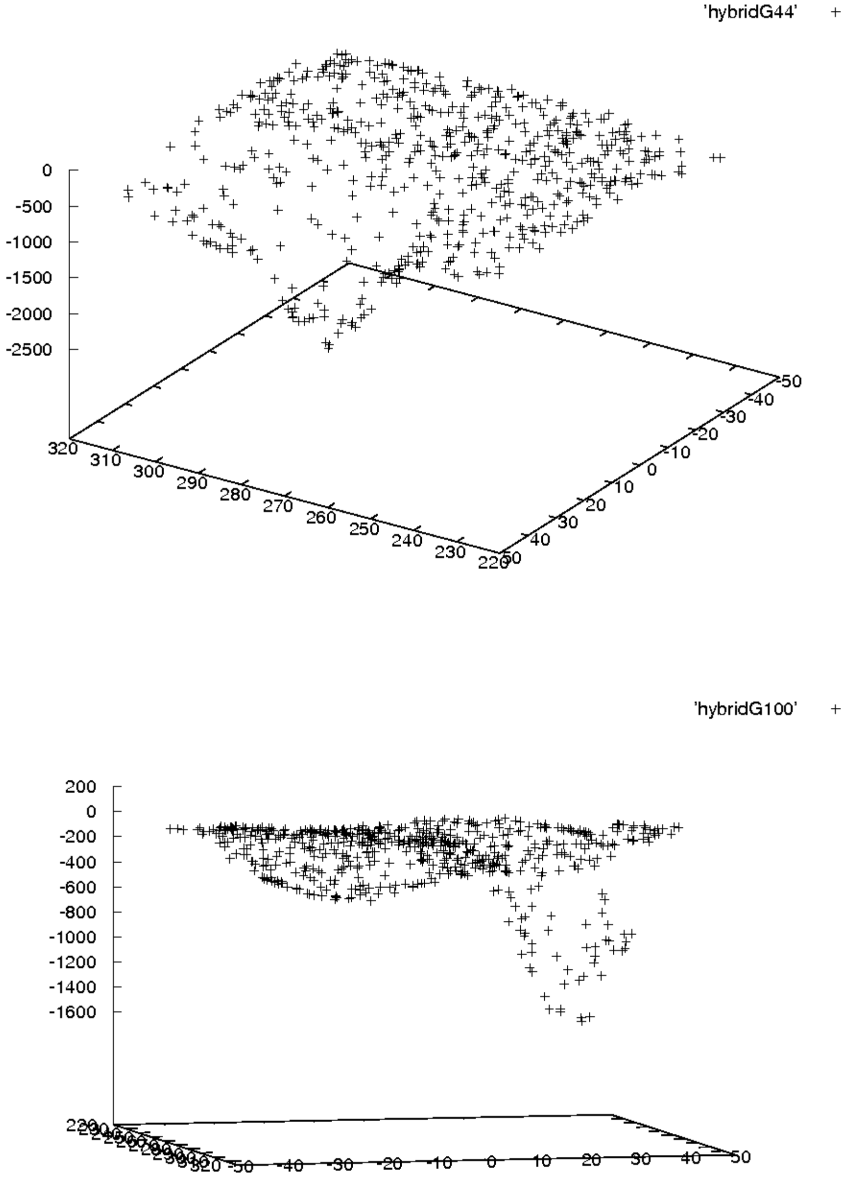


Figure 4.15: A zoom-in on the regional meshless approximation at 44 and 100 days

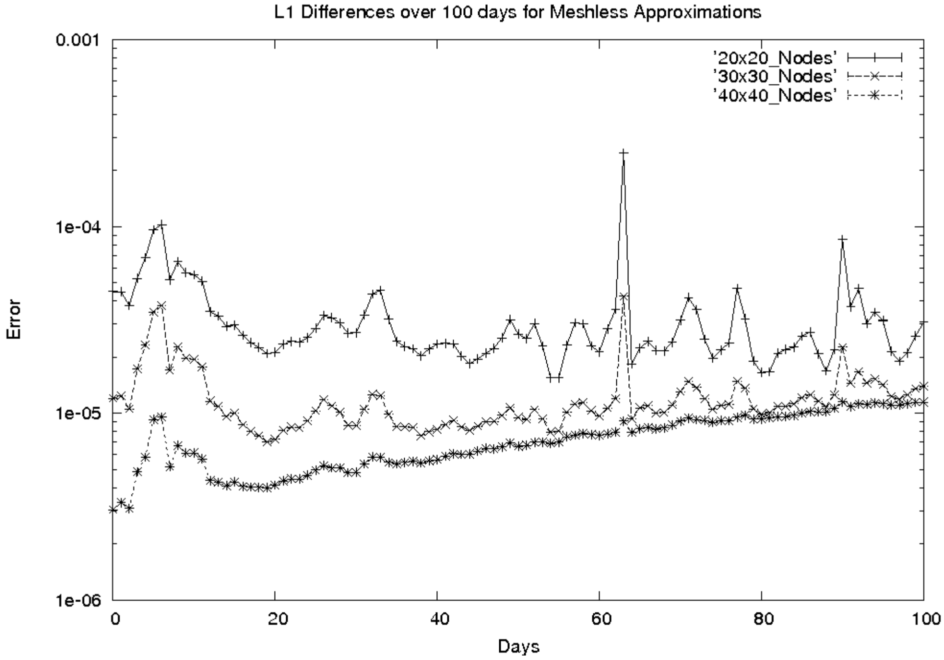
cal region on face 2 by comparing the three different approximations using $N = 400$, 900, and 1200 nodes in the local region to a high-order global spectral approximation using 64 elements on each face with Legendre polynomials of degree 10 for the geopotential field. We compare the convergence of the meshless approximation on the local region to the standard spectral element approximation on the same region using $4, 9$ and 16 elements and plotting the L_1 differences on face 2 with the high-order 64 element approximation. This way, we can compare the regional meshless approximation with the standard spectral element approximation in the same region. The L_1 relative differences between the meshless approximation of the geopotential and the high-order spectral element approximation in the local region is given by

$$\|\eta_N - \eta_M\|_{l_1(P_2)} = \frac{1}{N_M} \sum_{j=1}^{N} \frac{|\eta_N(\mathbf{x}_j) - \eta_M(\mathbf{x}_j)|}{|\eta_N(\mathbf{x}_j)|} \tag{4.2}$$

where, again, $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is the set of collocation nodes on the local region of face 2 and $\eta_N(\mathbf{x}_i), \eta_M(\mathbf{x}_i)$ are the meshless and spectral element approximations, respectively, evaluated at node $\mathbf{x}_i$.

The L_1 differences over time for 100 days are shown in figures 4.16 for the meshless approximation (top) and spectral element approximation (bottom).

Comparing the two different approximation methods, we can see that the meshless collocation approximation performs rather similar to the spectral element approximation in the same region despite being less consistent in the approximation ability at each time step. The 4 element approximation, which amounts to 256 Gaussian-Legendre-Lobatto (GLL) nodes, performs slightly better than the 400 node



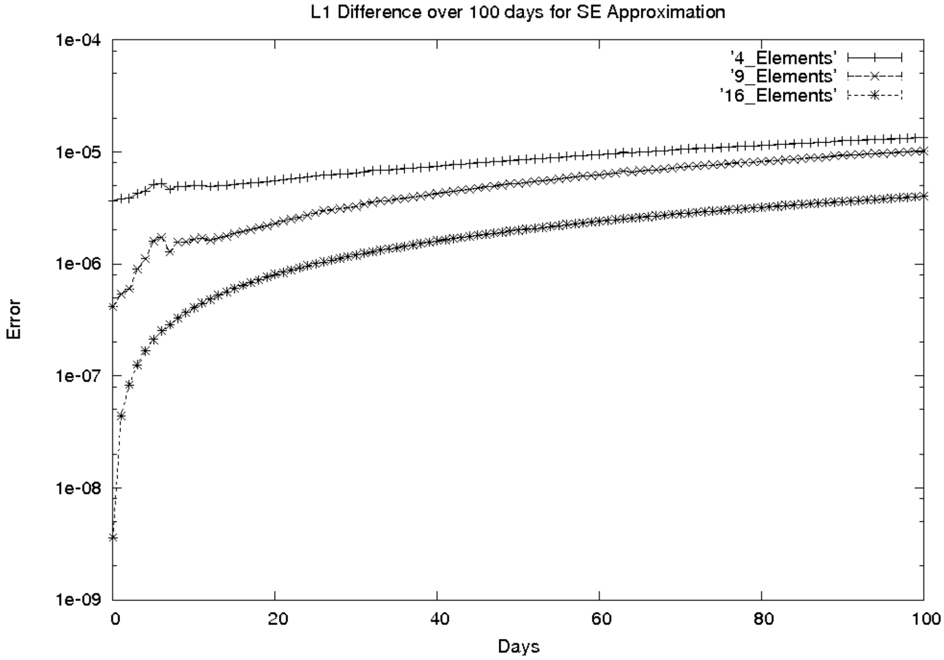


Figure 4.16: L1 differences of regional meshless approximation with 400, 900, 1200 nodes to high-order 64 element solution (top). L1 differences of SE approximation with 4,9, and 16 elements to high-order 64 element solution (bottom)

meshless approximation, while the 900 node meshless approximation performs better than the spectral element approximation of 9 elements (576 GLL nodes). The point of this experiment was to show that local refinement in collocation achieved by simply adding additional nodes in a local region improves the approximability of the solution in that region. Comparing to the high-resolution spectral element model, we see that this is indeed the case. Furthermore, we see that not only do we accomplish an improvement in approximability, but we also see that the meshless collocation performs at least as well as the spectral element approximation of roughly the same resolution.

4.4.4 Test Case 6: Rossby-Haurwitz Waves

Another interesting test case for the numerical solution to the shallow water equations on the sphere features an initial condition for the velocity which is actually the analytical solution to the non divergent nonlinear Barotropic equation on the sphere, given as a vorticity equation. The velocity components are given as

$$u = a\omega \cos\theta + aK \cos^{R-1}\theta(R\sin^2\theta - \cos^2\theta)\cos R\lambda,$$

$$v = -aKR\cos^{R-1}\theta \sin\theta \sin R\lambda,$$

and an initial geopotential field of the form

$$\eta = gh_0 + a^2 A(\theta) + a^2 B(\theta)\cos R\lambda + a^2 C(\theta)\cos(2R\lambda),$$

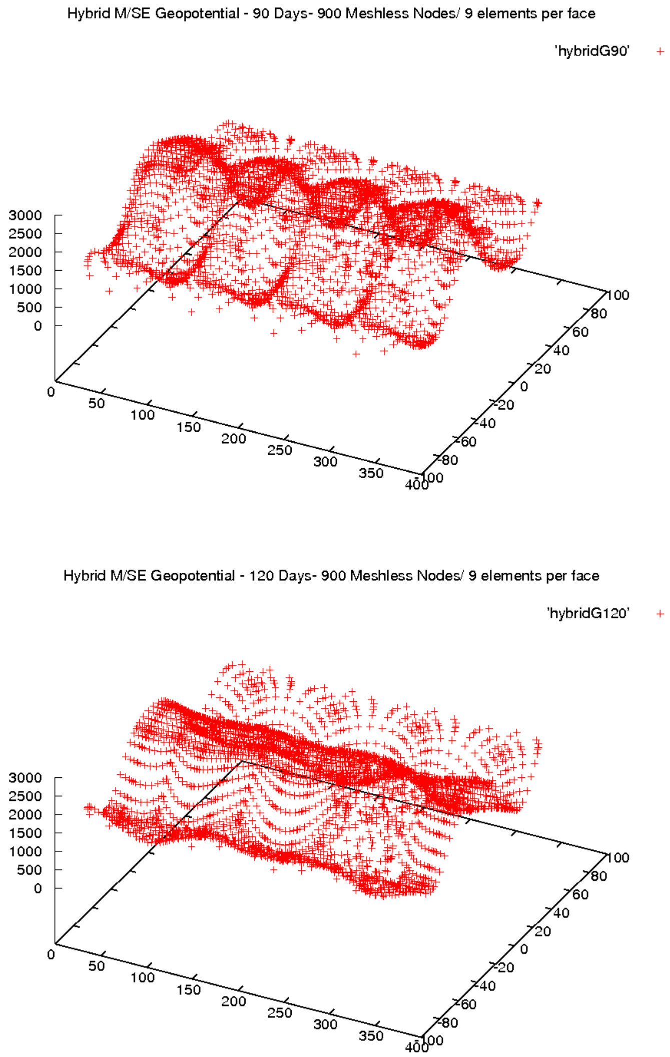
where A, B and C are given in [71]. Constants used in this test are $\omega = K = 7.848 \times 10^{-6}/sec$ and $h_0 = 8 \times 10^3 m$.

As originally proposed, these waves were expected to evolve nearly steadily with wavenumber 4. However, Thuburn and Li showed in [59] that this case is actually weakly unstable in that it will eventually break down once perturbed. The rapid breakdown of the basis wave state usually occurs after about 100 days depending on the model parameters and approximation used. In this experiment, since no local phenomena will be taking place, we instead focus on the global properties of meshless collocation. To this end, we initiate our study by using an arrangement of 4 to 9 regional areas per face allocated with different arrangements of collocation nodes. Our experiment will be to determine how the total amount of collocation nodes on the cubed-sphere effect the steady wave advection state of the geopotential before a transition takes place to an unstable wave pattern.

Figures in 4.17 show the geopotential in its initial steady state (top) at 90 days right before its transition to an unstable state (bottom), which is at about 120 days. The collocation solution was computed using 9 regions per face with GLL nodes on all faces except face 4, where a random distribution was given. The time step for all the simulations was kept small at $\Delta t = 1/864$ to minimize the effects of time discretization errors in the wave advection.

After about the 120th day, the waves in the solution over time significantly change and finally show the effects of the nonlinearity in the equations. The waves have a much more nonlinear structure as different wave numbers are interacting in the opposite zonal direction.

Figure 4.18 shows a plot of the geopotential after 200 days using the same collocation node arrangement as before.



Hybrid M/SE Geopotential - 90 Days- 900 Meshless Nodes/ 9 elements per face

Hybrid M/SE Geopotential - 120 Days- 900 Meshless Nodes/ 9 elements per face

Figure 4.17: Plot of the geopotential at 90 days and then at 120 days. The collocation solution was computed using 9 regions per face with GLL nodes on all faces except face 4, where a random distribution was given.

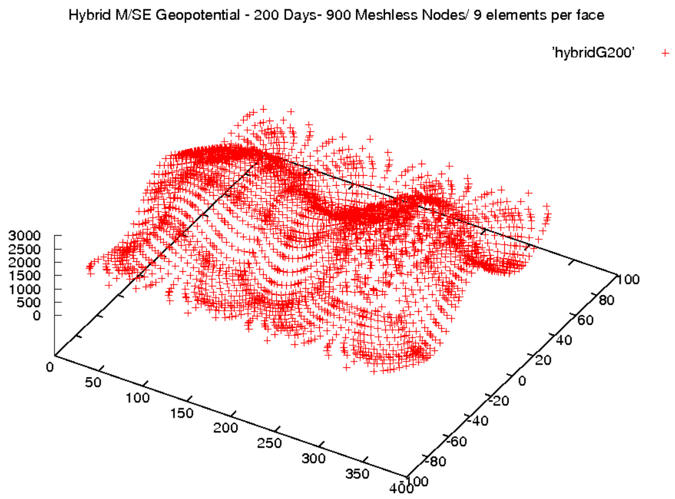


Figure 4.18: Plot of the geopotential at 200 days.

The figures in 4.19 show the contours of the geopotential after 90, 120, and 200 days. Notice the wave structure at 200 days is now completely different than at 90 days after undergoing the nonlinear wave breakdown at around 120 days.

We notice that an increase in the number of nodes renders has very little effect on the day in which the breakup of stable wave flow begins. The lowest order of the approximations sees a breakdown at around the 90th day whereas the highest order of approximation sees a breakdown around the 95th day. This is most likely due to the sensitivity of the meshless solution as the amount of nodes in the domain increases. Additional noise in the lower order approximation adds noise to the global solution, causing a changing of wave structure earlier. However, once the wave structure has changed, we see a long-term pattern emerging which suggests that second steady-state in the traveling waves has been developed. The fine-scale spectral model using spherical harmonics which was examined in [59] demonstrates

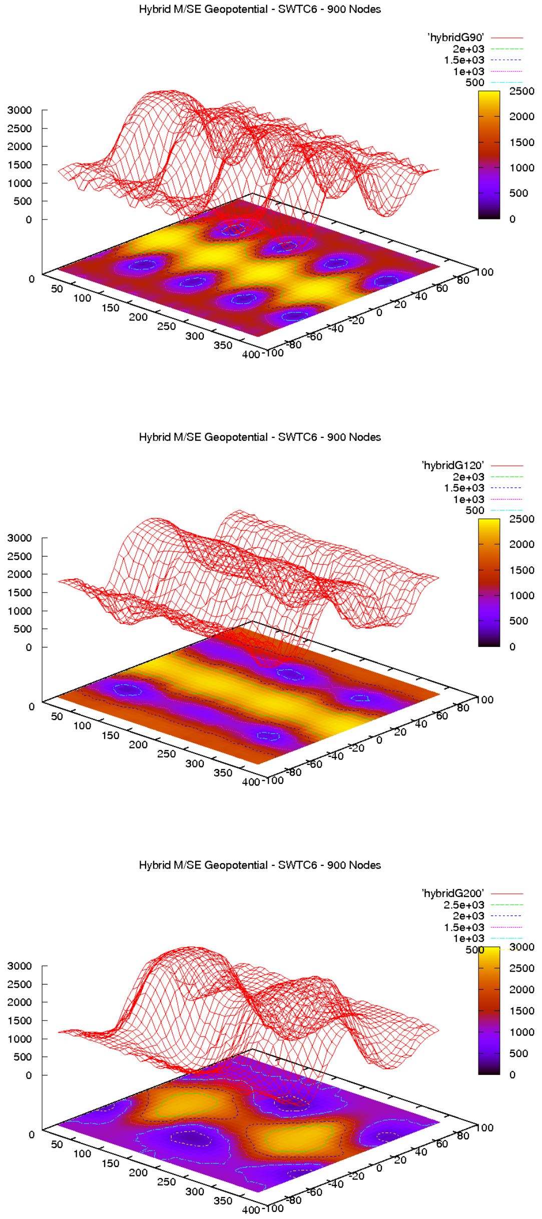


Figure 4.19: Plot of the contours of the geopotential at 90, 120, and 200 days.

a breakdown at around 80 days.

Clearly visible in the plot over 220 days is period in which the approximate solution over time incorporates the nonlinear effects of the equations. After the 20 day period of wave breakdown, the solution becomes stable again and a wave solution propagates around the sphere in the opposite direction, a phenomenon which was discovered in the Thiburn and Li paper [59].

4.4.4.1 Comparing with spectral-element solution

As we did in the previous Test Case, we compare the different meshless approximation using 400, 900, and 1200 nodes on each face of the cubed-sphere to a high-resolution spectral element approximation offered by NCAR homme [32]. Figure 4.20 shows the L_1 differences given by formula 4.2 for three different integrations over a period of 220 days.

We see that the collocation solution improves with the refinement in number of nodes. However, with the higher resolution, we see that the Rossby-Haurwitz wave structure is better captured resulting in an oscillatory L_1 difference between the high-resolution spectral element solution. This is to be expected since the higher resolution in the meshless collocation solution admits finer scales of the the wave which are not present in the lower resolution approximation.

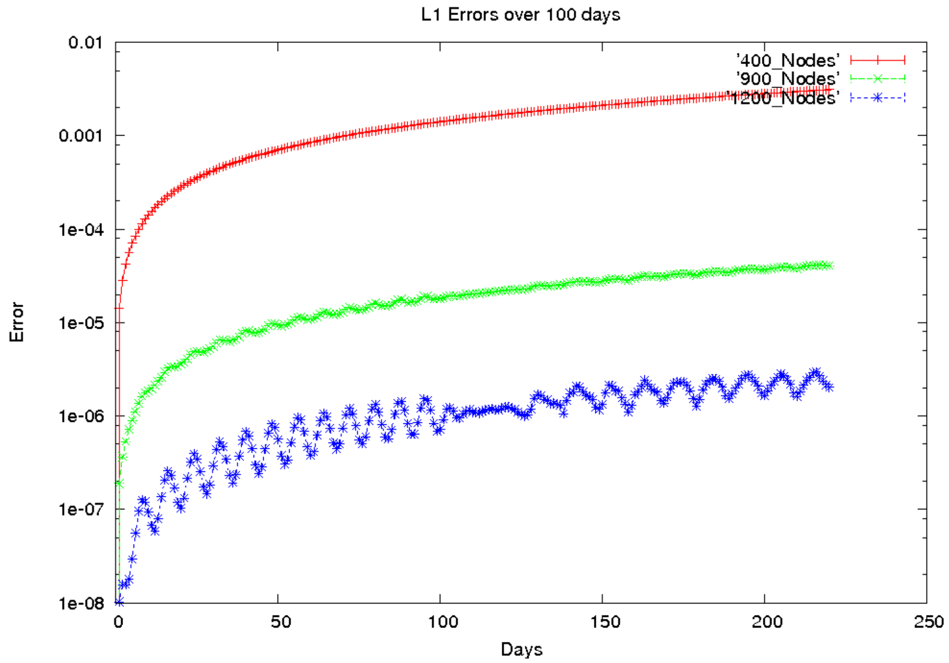


Figure 4.20: Plot of L1 differences between the meshless approximation on the face 2 and the high-order 64-element spectral element solution in the same region.

4.4.5 Test Case 8: Barotropic Instability

Our final test case comes from Galewsky et al. [26]. As a supplement to the standard test suite W92, Galewsky et al. proposed a new test case which is intended to provide a more realistic atmospheric flow in regards to the global and local dynamics. Specifically, the test case was motivated by the need for tight vorticity gradients at subgrid scales. It also does not require any code modification during time stepping other than the setup of the initial conditions.

The initial condition consists of a simple zonal flow representing a zonal tropospherical jet perturbed by a zonal 'bump' to influence the development of barotropic instability. Following Galewsky, the basic flow is a zonal jet **u** given by a function of latitude λ

$$\mathbf{u}(\lambda) = \mathbf{u}_0 \cos^2\Big(\frac{\pi(\lambda - \lambda_0)}{2\gamma}\Big), \tag{4.3}$$

where $\mathbf{u}_0 = 80ms^{-1}$ is the maximum zonal velocity, $\lambda_0 = \pi/4$ is the meridional offset of the jet in radians and $\gamma = \pi/18$ is a nondimensional parameter that controls the width of the jet. The geopotential hight η is obtained from the zonal flow by numerically integrating the balance equation:

$$\eta(\lambda) = g\eta_0 - \int_0^{\pi} a\mathbf{u}(\lambda')\Big(f + \frac{\tan(\lambda')}{a}\mathbf{u}(\lambda')\Big)d\lambda', \tag{4.4}$$

where a is the radius of the earth. Finally, to initiate the baratropic instability, a perturbation is added zonal flow by adding a localized bump to the balanced geopotential height field given by

$$\Phi'(\lambda, \theta) = \hat{\Phi}\text{sech}^2(\alpha(\theta - \theta_0))\,\text{sech}^2(\beta(\lambda - \lambda_0)). \tag{4.5}$$

After a few time steps, due to the geopotential field not being in complete balance with the zonal velocity, growing barotropic instability causes tight gradients in the vorticity field that evolve on the timescale of days. If the time step is small enough, gravity waves radiate away from the perturbed zone during the first several hours of the integration. The challenge of this test case is thus to capture two different types of timescale dynamics with meshless collocation: (1) fast gravity waves that develop on timescales of minutes and hours; (2) slower vorticity dynamics that evolve on timescales of days.

Figures in (4.21) show the contour map of the initial conditions of the problem. Top figure shows the zonal profile of the initial zonal wind. Bottom figure shows the initial balanced height field with a superimposed perturbation. Contour lines are set at 100m.

4.4.5.1 Capturing Fast Gravity Waves

Our first experiment explores the ability of meshless collocation to capture the small scale dynamics that are present in the fast gravity waves. To do this we first initialize the meshless collocation grid by tiling the cubed-sphere with 150 regions, each endowed with 100 GLL collocation nodes. In order to capture the small timescale dynamics of the model, we chose a small time step of 15 seconds. To see the fast moving gravity waves, we plot the divergence field during the first 10 hours of integration, illustrating gravity wave propagation from the initial perturbation. Figures 4.22 and 4.23 show the divergence field at 4, 6, 8 and 10 hours.

The effects of the small scale gravity waves are also seen in the geopotential height field anomaly. After the balanced initial height is perturbed during the first few time steps, the effects of the gravity waves propagate from the initial perturbation. This is seen in figures 4.24, showing the first 8 and 10 hours of integration.

The vorticity field also presents very interesting dynamics. Fine scale vorticities are seen after about 60 days with zonal interactions taking place just after a few days. As clearly seen in the figures, the collocation method with the resolution of nearly 15,000 collocation nodes with a time step of 15 seconds captures in great detail the small scale dynamics of the vorticity. Figures in 4.25 and 4.26 show the

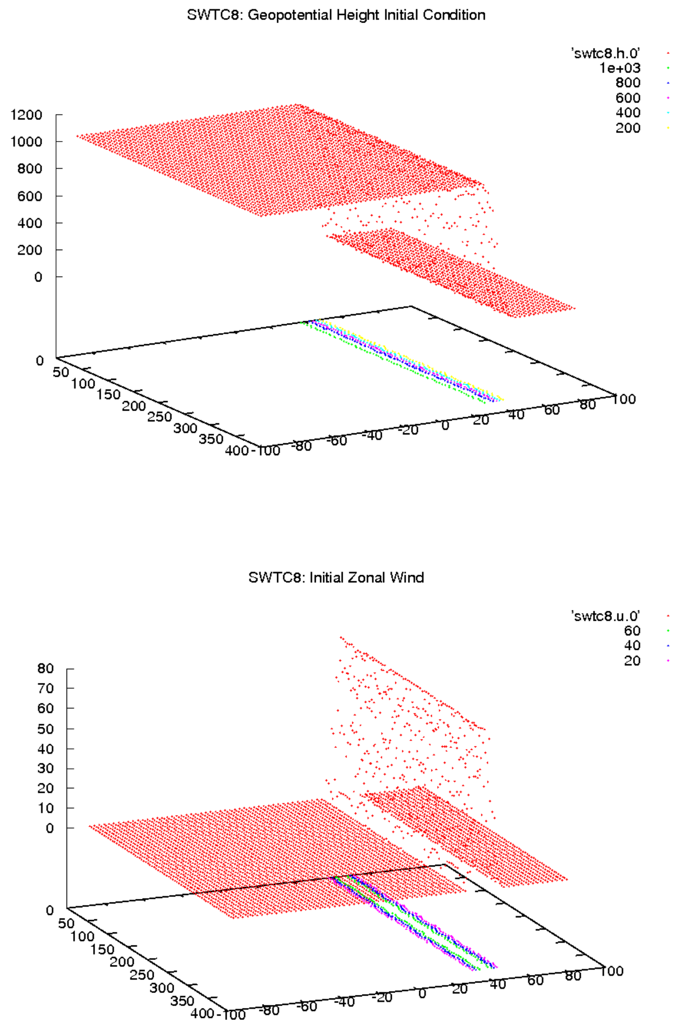
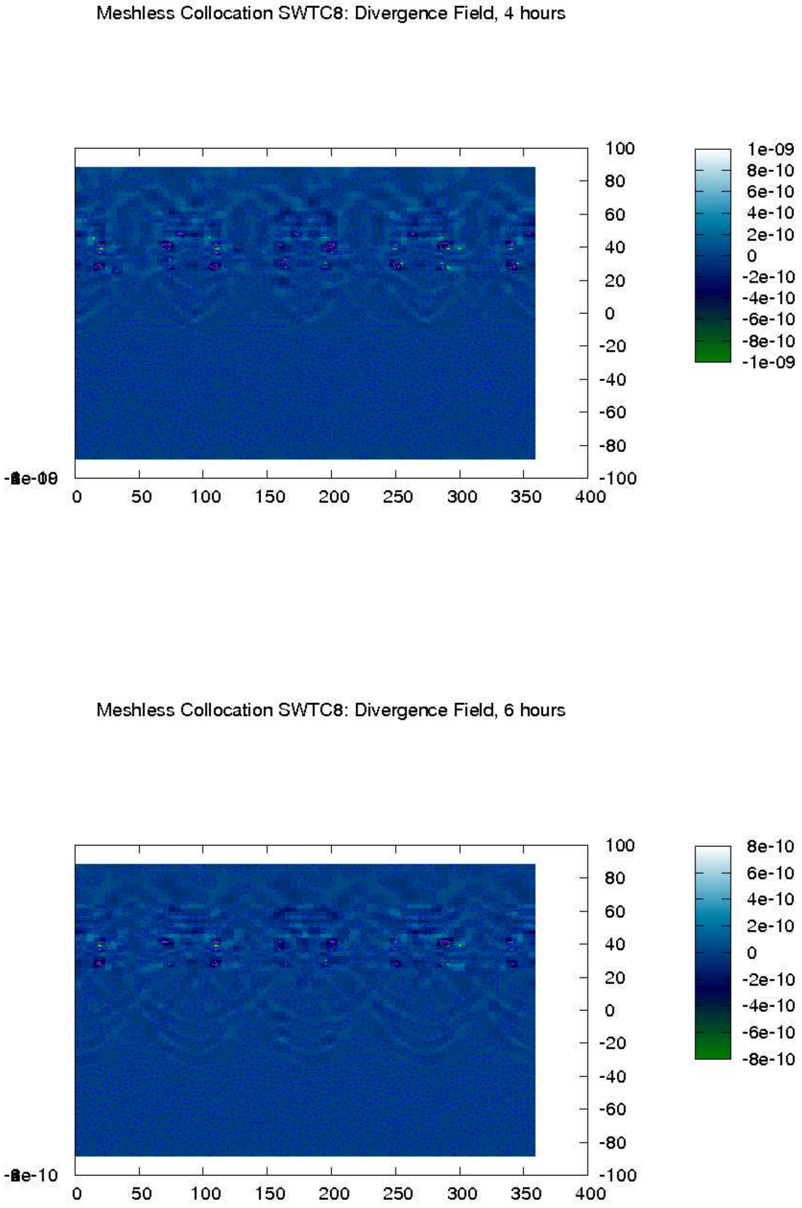
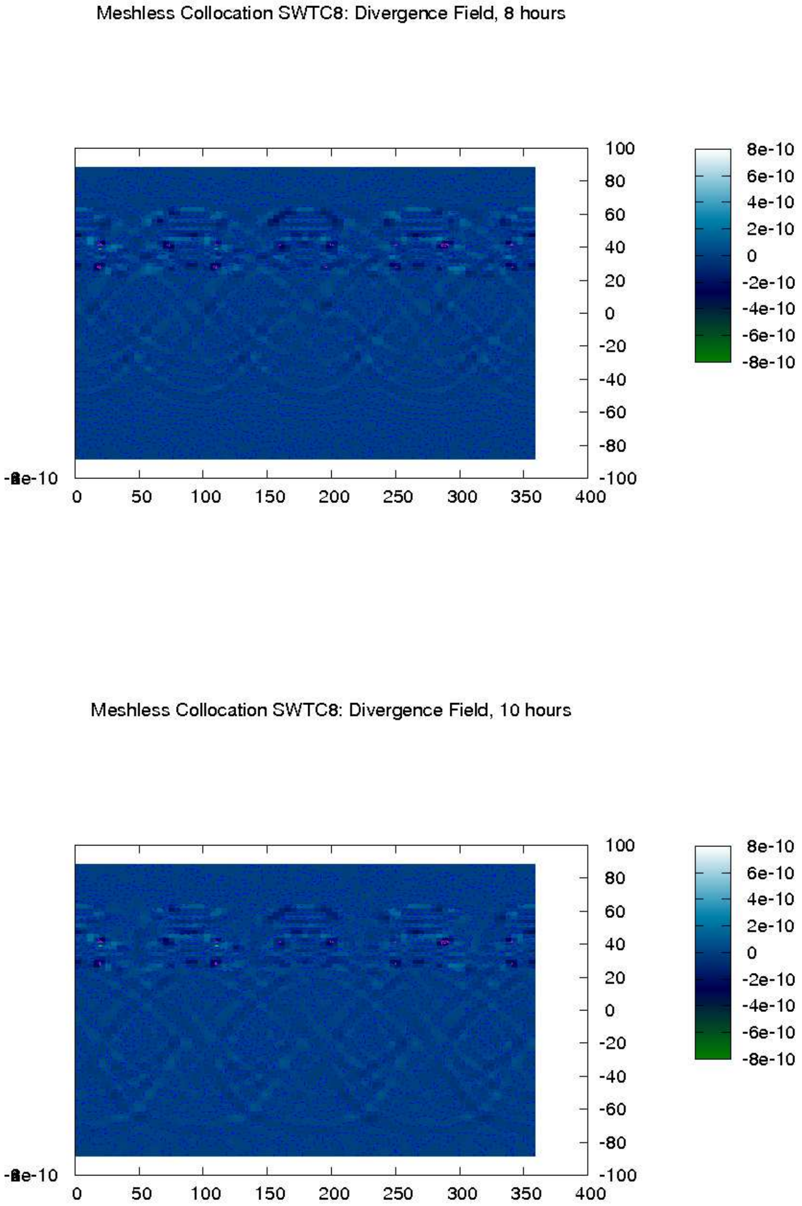


Figure 4.21: Initial conditions of the problem. Zonal profile of the initial zonal wind and initial balanced height field with a superimposed perturbation.


Meshless Collocation SWTC8: Divergence Field, 4 hours

Meshless Collocation SWTC8: Divergence Field, 6 hours

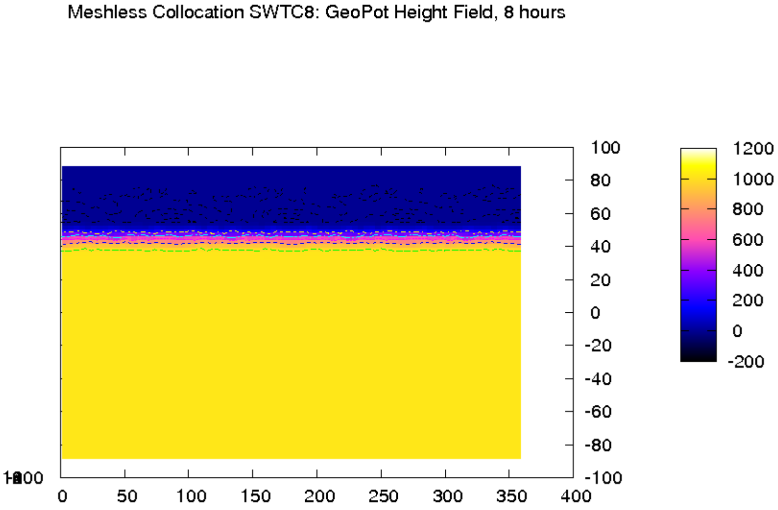
Figure 4.22: Divergence field illustrating gravity wave propagation from the initial perturbation at 4 and 6 hours. Results from meshless collocation approximation with time step 15 seconds.



Meshless Collocation SWTC8: Divergence Field, 8 hours

Meshless Collocation SWTC8: Divergence Field, 10 hours

Figure 4.23: Divergence field illustrating gravity wave propagation from the initial perturbation at 8 and 10 hours. Results from meshless collocation approximation with time step 15 seconds.



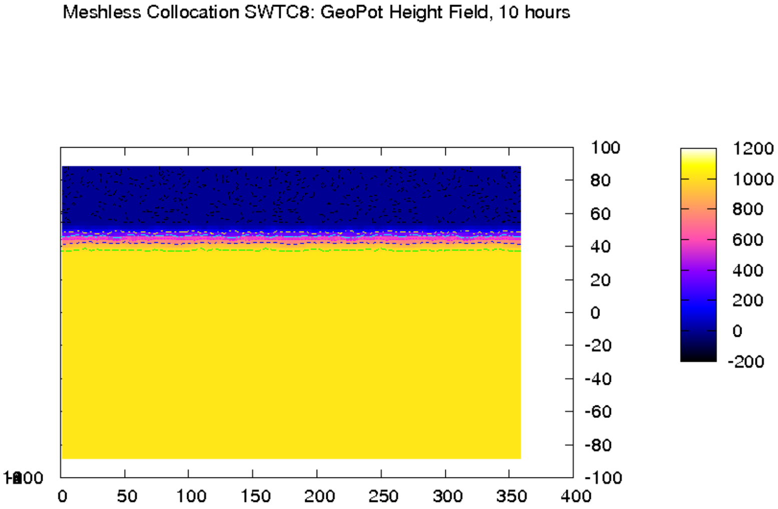


Figure 4.24: Height field illustrating gravity wave propagation from the initial perturbation at 8 and 10 hours. Results from meshless collocation approximation with time step 15 seconds.

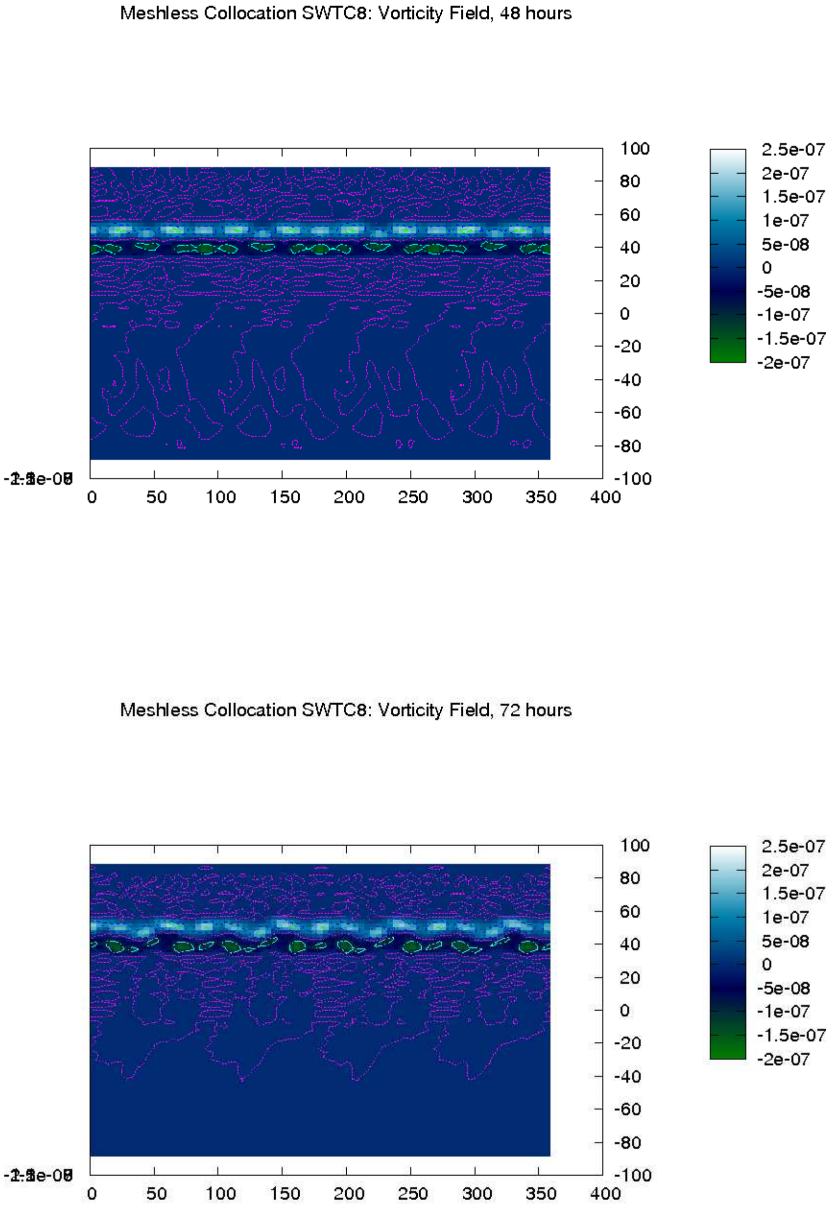
vorticity ζ after 48, 72, 80, and 90 hours. The results are very similar to the ones found in the original paper by Galewsky et al.

In comparing the results of these simulations with the ones of Galewsky et al., we can conclude that meshless collocation does a remarkable job approximating small scale dynamics such as fast-moving gravity waves, even when employing the semi-implicit time stepping scheme. Larger timescale dynamics are also clearly captured by the method as seen in the plots of the vorticity field. All parameters of the model with the exception for grid resolution ([26] uses a high-resolution finite-difference scheme) were recreated to compute the solutions in near identical conditions.

4.5 Conclusion

In this chapter, we proposed and developed a framework for the parallelization and high-performance computation of the meshless collocation model for the shallow water equations introduced in chapter 3. Based on the merging of several numerical tools including a new meshless collocation scheme, semi-implicit time stepping, and the efficient parallelization of inverting a band matrix, we have attempted to demonstrate that the model could be an attractive alternative to tradition spectral/finite-element methods for solving a large-scale geophysical models.

The introduction of the semi-implicit time stepping method combined with the approximation using symmetric positive definite kernels having compact support has lead to a model approximating the nonlinear geophysical equations that has desirable


Meshless Collocation SWTC8: Vorticity Field, 48 hours

Meshless Collocation SWTC8: Vorticity Field, 72 hours

Figure 4.25: Meshless collocation solution of vorticity field with nearly 15,000 collocation nodes and a time step of 15 seconds. Plot of field after 48 and 72 hours.

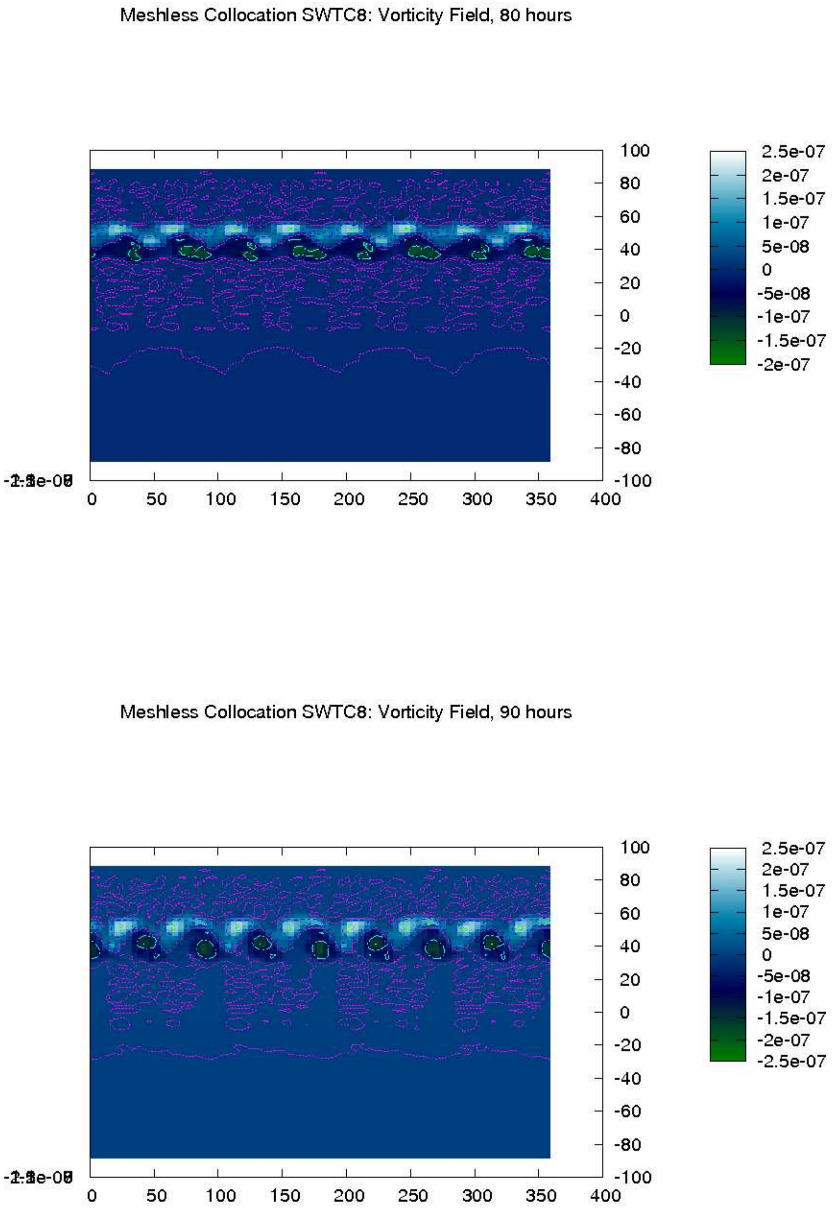


Figure 4.26: Meshless collocation solution of vorticity field with nearly 15,000 collocation nodes and a time step of 15 seconds. Plot of field after 80 and 90 hours.

mathematical properties suitable for parallelization. One attractive feature was that with the semi-implicit method, we were able to transform the problem of solving for the velocity and geopotential concurrently to only solving for the geopotential, which greatly eliminated much of the computational burden. Furthermore, due to the structure of the meshless collocation approach that we have proposed, the resulting system for computing the geopotential at each time step, with a symmetric and positive definite matrix $\mathbf{H} = \mathbf{I} + c\mathbf{D}^T\mathbf{D}$, was shown to be efficiently handled by a parallelized conjugate gradient method.

One advantage of the method that clearly comes into play is obviously the fact that since no numerical quadrature is needed due to the nature of collocation, no mesh is needed. As was shown in the numerical experiments, the collocation points can be fairly arbitrary, as long as the separation distance $q_{\mathcal{X}}$ and the saturation parameter $h_{\mathcal{X},\Omega}$ are relatively equal. In fact, in many of the simulations such as the ones in test case 6, a hybrid grid was used that combined collocation nodes from a spectral-element mesh with randomly scattered nodes. This is an attractive feature for the proposed collocation method since some application in geophysical sciences might have data sites available that are randomly scattered and not uniform, thus no interpolation of the data to the randomized sites would be necessary resulting in no loss of approximation precision.

A high-performance Fortran 90 software suite was developed for the model for use on distributed memory parallel processors with the message passing interface libraries (mpi). With a focus on transparent data structures and optimized matrix-vector multiplication, the parallel algorithms demonstrated remarkable scalability in function with number of collocation nodes on the cubed sphere. This was shown in the beginning of the section where the number of internodal communications and overall clock times were recored as the number of total processors increased. Compared with the spectral element method, the speedup factors were slight better due to less *scatter* and *gather* mpi calls in the conjugate gradient subroutine. This is ultimately due to the fact that no averaging along element boundaries is needed since the basis functions in the collocation solution have compact support. This is a great advantage that the proposed meshless collocation method has over finite/spectral element methods for large scale problems.

Some of the disadvantages and difficulties that we encountered with the proposed meshless collocation scheme is in defining appropriate parameters for the compactly supported basis functions. Since such freedom is given in choosing the parameters such as the location of the collocation nodes and shape parameter, additional work before actually solving a given problems might be necessary. Choosing the appropriate number of collocation nodes N, its distribution (uniform, Gaussian, random) and the ϵ shape parameter that controls the bandwidth of the interpolation matrix is usually simply done by trial-and-error. If the shape parameter is chosen too large or too small, we found that the iterations of the parallel conjugate gradient method for solving the Helmholtz problem converge very slowly. In practice, this is clearly one of the disadvantages of the meshless collocation method. As we had mentioned in the second the chapter, an easy but tedious approach to finding a near optimal shape parameter ϵ for a given collocation node distribution is by a trial-and-error on a smaller problem where an exact solution is known. In our approach, we chose one face of the cube and simply cycled through an array of shape parameters and selected the one that minimizes the L^∞ error of the interpolation problem $I_{\mathcal{X}} f(\mathbf{x}_j) = f(\mathbf{x}_j)$ for all collocation nodes $\mathbf{x}_j \in \mathcal{X}$, where f is a given continuous function.

Once the near optimal parameters for the meshless collocation method have been chosen however, the method is fairly easy to implement. Given that no mesh data structures or quadrature rules are necessary and that the computation of the interpolation matrix inversion to construct the differentiation matrices at initialization can be done rapidly in parallel using the CSC algorithm from the beginning of the chapter.

Future research in the this meshless collocation approach to numerical geophysical dynamics includes the implementation of a threading parallel paradigm in addition to the message passing interface. Recent research direction in large scale computational models have reported better scalability in the parallelization of matrix-vector operations with the integration of a multiple threading interface such as OpenMP, or Pthread. These libraries enable multiple threading on shared memory. Combining the multiple threading with message passing produces a *hybrid* technique for parallel computing. One obvious way the threading could be used is to speed up summations inside matrix-vector products on each processor node. Of course, this would require each node to have multiple processor with access to the same memory. However, this is very common on many high-performance clusters/supercomputers.

Due to the success of the performance by the meshless collocation method in all four test cases, a strong case has been made for future work to implement the approximation method in an atmospheric dynamical core model. Due to the mathematical structure of the collocation method, coupling the numerical scheme with physics forcing packages should not be too much of a challenge. Thereafter, employing the meshless collocation method to solve the primitive equations of the atmosphere, where vertical dynamics will be accomplished via finite-differencing, is of high interest. With such success in numerical convergence, robustness of local refinement, and scalability of model, we hope the meshless collocation method can become an attractive computational tool for upcoming research trends in high-performance atmospheric modeling.

Chapter A

Results from Functional Analysis

A.1 Banach and Hilbert Spaces

We introduce the following notation and definitions. Let X and Y be Banach spaces with respective dual spaces denoted by X', Y', the space of linear continuous functionals defined on X and Y, respectively. The dual pairing of a functional $x' \in X'$ and an element $x \in X$ will be denoted by using a bracket notation of either $\langle x', x\rangle_{X'\times X}$ or $\langle x, x'\rangle_{X\times X'}$. Both will be taken to mean the value $x'(x)$ of the functional $x' \in X'$ acting on an element $x \in X$. We will omit the reference to $X \times X'$ in the notation when there is a clear understanding of the spaces being used.

Let $B \in \mathcal{L}(X, Y')$ denote a bounded linear operator from X into Y' with its adjoint $B^* \in \mathcal{L}(Y, X')$ from Y into X' defined by

$$\langle B^*y, x\rangle_{X'\times X} := \langle y, Bx\rangle_{Y\times Y'} \quad \forall y \in Y, x \in X. \tag{A.1}$$

In addition, we denote the range of B by $R(B) := B(X)$ and the *kernel* of B by $N(B) := \{x \in X \ : \ Bx = 0\}$.

An important Theorem for Banach spaces X endowed with an inner product (Hilbert space) is the classical Riesz Theorem.

Theorem A.1.1. *Let X be a real Hilbert space with inner product $(\cdot,\cdot)_X$. Then for* *any $f \in X'$, there exists a unique $u \in X$ such that*

$$\langle f, v\rangle = (u, v)_X, \quad \forall v \in X. \tag{A.2}$$

Furthermore, this defines an operator $R : X' \mapsto X$ given by $R : f \mapsto u$ that is an isometric isomorphism. R is referred to as the Riesz operator.

Proof. K. Yosida [73], Ch. III/6 □

A second classical result is the so-called Lax-Milgram Theorem.

Theorem A.1.2. *Let X be a real Hilbert space and let $A \in \mathcal{L}(X, X')$ be a linear coercive (or X-elliptic) operator (i.e. there exists $\gamma > 0$ such that $\langle Au, u\rangle \geq \gamma\|u\|_X$ for all $u \in X$). Then for any $f \in X'$, there exists a unique $u \in X$ such that*

$$\langle Au, v\rangle = \langle f, v\rangle, \quad \forall v \in X; \tag{A.3}$$

u satisfies $\|u\|_X \leq \frac{1}{\gamma}\|f\|_{X'}$.

Proof. K. Yosida [73], Ch. III/7. □

Now let $X_1, \ldots, X_N$ be N Hilbert spaces.

Lemma A.1.1. $\mathbf{X} := \Pi_i^N X_i$ *is a Hilbert space with inner product* $(\mathbf{x}, \mathbf{y})_{\mathbf{X}} := \sum_{i=1}^N (x_i, y_i)_{X_i}$ *for* $\mathbf{x} = (x_1, \ldots, x_N), \mathbf{y} = (y_1, \ldots, y_N) \in \mathbf{X}$.

The proof is obtained by a simple verification of the axioms for Hilbert spaces. Now we note that the dual of a Hilbert space is itself a Hilbert space.

Lemma A.1.2. *Let X be a real Hilbert space. Then X' is a Hilbert space with inner product*

$$(f, g)_{X'} = (R^{-1}f, R^{-1}g)_X, \quad f, g \in X, \tag{A.4}$$

where R is the Riesz operator defined in (A.1.1).

For a closed subspace $V \subset X$ we define the *orthogonal complement* to V.

Definition For a Hilbert space X, the *orthogonal complement* of a closed subspace $V \subset X$ is defined by

$$V^{\perp} := \{u \in X \mid (u, v)_X = 0,\ \forall v \in V\}.$$

The following Lemma states that any Hilbert space can be decomposed into two disjoint Hilbert spaces, which is important the analysis of saddle point problems as we will see.

Lemma A.1.3. *Let X be a Hilbert space and $V \subset X$ be a closed subspace. Then the Hilbert space $V^{\perp}$ is a closed subspace of X and the decomposition $X = V \oplus V^{\perp}$ holds.*

Proof. K. Yosida [73], ch. III/1. □

We can easily show using Theorem A.1.1 that in the case X is a Hilbert with a closed subspace $V \subset X$, the polar set $V^0 \subset X'$ and the orthogonal complement $V^{\perp} \subset X$ are linked by the Riesz operator R. To see this, we first have by definition

$$V^0 := \{x' \in X' \mid \langle x', v \rangle = 0,\ \forall v \in V\}.$$

Now using Theorem A.1.1, for any $x' \in V^0 \subset X'$, there exists a unique $u \in X$ such that

$$\langle x', v \rangle = (u, v)_X = 0,\ \forall v \in V.$$

Thus $Rx' = u \in V^{\perp}$.

Let X be a Banach space with dual X' and second dual $(X')'$ (the space of linear functionals on X'). For convenience we simply denote $(X')'$ by X''. We first define the mapping $J : X \mapsto X''$ by $\langle x', J(x)\rangle_{X'\times X''} = \langle x', x\rangle_{X'\times X}$ for every $x \in X$ and $x' \in X'$. That is, J maps x to the functional on X' given by evaluation at x.

Definition A Banach space X is said to be reflexive if $R(J) = X''$, i.e. for every $x'' \in X''$, there is an element $x \in X$ such that $J(x) = x''$.

Remark. A Hilbert space is reflexive ([73], Chapter III, 6).

Theorem A.1.3. *Hahn-Banach Theorem Suppose M is a proper subspace of a Banach space X. If $m' \in M'$, then there exists a functional $x' \in X'$ such that $\|x'\| = \|m'\|$ and $\langle x', x\rangle = \langle m', x\rangle$ for $x \in M$.*

Proof. Taylor [55], Theorem 4.3-A. □

We now give an important consequence of the well known Hahn-Banach Theorem.

Theorem A.1.4. *Let X be a Banach space and let X_0 be a closed proper subspace of X. Suppose there exists an $x_1 \in X$ such that $x_1 \notin X_0$ and the distance from x_1 to X_0 is $h > 0$. Then there exists a functional $x' \in X'$ such that $\langle x', x_1\rangle = h$, $\|x'\| = 1$, and $\langle x', x\rangle = 0$ if $x \in X_0$.*

Proof. (Taylor, Theorem 4.3-D) Let $x_1 \in X$ such that $x_1 \notin X_0$ and $dist(x_1, X_0) = h$. We first define the closed subspace M generated by X_0 and x_1, namely $M = span\{X_0, x_1\} \subset X$. Any element of M can be represented in the form $m = \alpha x_1 + x$, $x \in X_0$, where α and x are uniquely determined by m.

Consider the functional m' which acts on an element $m \in M$ by $\langle m', m\rangle = \alpha h$. We first show that indeed $m' \in M'$ and $\|m'\| = 1$ by showing that both $\|m'\| \leq 1$ and $\|m'\| \geq 1$ hold. Firstly, if $\alpha \neq 0$, then $\|m\| = \|\alpha x_1 + x\| = \|-\alpha(-\alpha^{-1}x - x_1)\| \geq |\alpha| h$ since $\alpha^{-1}x \in X_0$ and $dist(x_1, X_0) = h$. Thus $|\langle m', m\rangle| = |\alpha| h \leq \|m\|$. Consequently, $m' \in M'$ and $\|m'\| \leq 1$. Now to show that $\|m'\| \geq 1$, we note that for any $\epsilon > 0$, there exists an $x \in X_0$ such that $\|x - x_1\| < h + \epsilon$. Let $y = \frac{x - x_1}{\|x - x_1\|}$. Then $y \in M$, $\|y\| = 1$ and $|\langle m', y\rangle| = \frac{h}{\|x - x_1\|} > \frac{h}{h+\epsilon}$. This means that $\|m'\| \geq \frac{h}{h+\epsilon}$ for any $\epsilon > 0$ since $h > 0$ and so it follows that $\|m'\| \geq 1$.

Lastly, we have from the representation of elements from M that $\langle m', x_1\rangle = h$ (since $\alpha = 1$), and $\langle m', x\rangle = 0$ for any $x \in X_0$ (since $\alpha = 0$). We apply the Hahn-Banach Theorem to conclude the existence of an element $x' \in X'$ such that $\|x'\| = \|m'\|$ and $\langle x', x_1\rangle = \langle m', x_1\rangle = h$ and $\langle x', x\rangle = \langle m', x\rangle = 0$ for $x \in X_0$. This completes the proof. □

A.2 Sobolev Spaces

We begin with the following vector spaces of continuous functions. To reduce notation, we use the standard multi-index notation, i.e. for each vector $\alpha := (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ we define $|\alpha| := \sum_{i=1}^d \alpha_i$ and

$$D^\alpha \phi := \frac{\partial^{|\alpha|}\phi}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

for $|\alpha| \geq 0$ and sufficiently smooth functions $\phi : \mathbb{R}^d \mapsto \mathbb{R}$.

Definition Let $\Omega \subset \mathbb{R}^d$ be an open subdomain and $m \in \mathbb{N}_0$.

- The set of continuous functions on Ω is given by

$$C(\Omega) := \{\phi : \Omega \mapsto \mathbb{R} \,|\, \phi \, is \, continuous\};$$

- The set of m-times continuously differentiable functions on Ω is given by

$$C^m(\Omega) := \{\phi : \Omega \mapsto \mathbb{R} \,|\, D^\alpha \phi \in C(\Omega), \; \forall |\alpha| \leq m\};$$

- Let $C^m(\overline{\Omega})$ be the set of functions $C^m(\Omega)$ which with its derivatives of order $\leq m$ can be extended continuously to the boundary $\partial\Omega$ of Ω.

- Let $D(\Omega) := C_0^\infty(\Omega)$ be the set of functions in $C^\infty(\Omega)$ which have compact support in Ω.

$C^m(\overline{\Omega})$ is a Banach space with norm

$$\|\phi\|_{C^m(\overline{\Omega})} := \max_{|\alpha| \leq m} \sup_{\mathbf{x} \in \Omega} |D^\alpha \phi(\mathbf{x})|$$

for a bounded open set $\Omega \subset \mathbb{R}^d$ (cf. R. Adams [1], ch. 1.26).

Next we define the *Holder Space* of functions which we need to describe the smoothness of boundaries.

Definition For $0 \leq \lambda \leq 1$ and $m \in \mathbb{N}_0$, the Holder space $C^{m,\lambda}(\overline{\Omega})$ consists of functions in $C^m(\overline{\Omega})$ which satisfy

$$\|\phi\|_{C^{m,\lambda}(\overline{\Omega})} := \|\phi\|_{C^m(\overline{\Omega})} + \sum_{|\alpha|=m} \sup_{\mathbf{x},\mathbf{y} \in \Omega, \mathbf{x} \neq \mathbf{y}} \frac{|D^\alpha \phi(\mathbf{x}) - D^\alpha \phi(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|^\lambda} < \infty. \qquad \text{(A.5)}$$

Now smoothness classes of boundaries $\partial\Omega$ can be given.

Definition Let $\Omega \subset \mathbb{R}^d$ be an open subdomain and $m \in \mathbb{N}_0$ with $0 \leq \lambda \leq 1$. We say that its boundary $\partial\Omega$ is of class $C^{m,\lambda}$ if the following conditions are satisfied: For every $\mathbf{x} \in \partial\Omega$, there exists a neighborhood V around $\mathbf{x}$ in $\mathbb{R}^d$ and new orthogonal coordinates $\{y_1, \ldots, y_d\}$ such that V is a hypercube in the new coordinates:

$$V = \{(y_1, \ldots, y_d) \,|\, -a_i < y_i < a_i, 1 \leq i \leq d\} :$$

and there exists a function $\varphi \in C^{m,\lambda}(V')$ with

$$V' = \{(y_1, \ldots, y_{d-1}) \,|\, -a_i < y_i < a_i, 1 \leq i \leq d-1\} :$$

and such that

$$\begin{aligned} |\varphi(y')| &\leq \frac{1}{2}a_d, \quad \forall y' = (y_1, \ldots, y_{d-1}) \in V'. \\ \Omega \cap V &= \{y = (y', y_d) \in V \,|\, y_d < \varphi(y')\}. \qquad (A.6) \\ \partial\Omega \cap V &= \{y = (y', y_d) \in V \,|\, y_d = \varphi(y')\}. \end{aligned}$$

A boundary of class $C^{0,1}$ is called *Lipschitz* boundary.

We now give a definition of Sobolev Spaces of integer orders on bounded domains.

Definition For an open bounded domain Ω, the Sobolev Space $H^k(\Omega)$, $k \in \mathbb{N}_0$, is defined by

$$H^k(\Omega) := \{u \in L^2(\Omega) \,|\, D^\alpha u \in L^2(\Omega), \; \forall |\alpha| \leq k\} \qquad (A.7)$$

and is endowed with the scaler inner product

$$(u, v)_{H^k(\Omega)} = (u, v)_{k,\Omega} := \sum_{|\alpha| \leq k} \left(D^\alpha u, D^\alpha v\right)_\Omega, \quad \forall u, v \in H^k(\Omega) \qquad (A.8)$$

and corresponding norm $\|u\|_{H^k(\Omega)} = \|u\|_{k,\Omega} := \sqrt{(u, u)_{k,\Omega}}$.

Theorem A.2.1. *The space $H^k(\Omega)$ is a Hilbert space.*

Proof. See R. A. Adams [1], ch. 3.2. □

We also need the *seminorm*

$$|u|_{k,\Omega} := \Big(\sum_{|\alpha|=k}\int_\Omega (D^\alpha u)^2 dx\Big)^{1/2}, \quad \forall u \in H^k(\Omega) \tag{A.9}$$

Definition The space $H_0^k(\Omega)$, $k \in \mathbb{N}_0$ is given by the completion of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{k,\Omega}$.

Clearly $H_0^k(\Omega)$ is a Hilbert space with respect to the inner product $(\cdot,\cdot)_{k,\Omega}$ which includes functions u which vanish on the boundary of Ω. The following embedding theorem from R.A. Adams ([1], ch. 5.4) is frequently used.

Theorem A.2.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary and suppose that $k \in \mathbb{N}_0$. Then the following continuous embedding holds true. For $2k > d$, we have $H^k(\Omega) \subseteq C(\overline{\Omega})$.*

We can also give another definition of the Sobolev space for noninteger s on the entire domain $\mathbb{R}^d$. This is given as follows.

Definition Let $s \in \mathbb{R}$. $H^s(\mathbb{R}^d)$ is a Hilbert space of elements $u \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ such that the Fourier transform of u, $\hat{u}$, is a measurable function and $(1+\|\xi\|_2^2)^{s/2}\hat{u} \in L^2(\mathbb{R}^d)$. The scalar product is given by

$$(u,v)_s = \int_{\mathbb{R}^d} (1+\|\xi\|_2^2)^s \hat{u}(\xi)\hat{v}(\xi)d\xi, \quad u,v \in H^s(\mathbb{R}^d)$$

with corresponding norm

$$\|u\|_s^2 = (u,u)_s.$$

We have the following Sobolev embedding theorem.

Theorem A.2.3. *Let $k \in \mathbb{N}$ and $s \in \mathbb{R}$ such that $s > k + d/2$. Then the space $H^s(\mathbb{R}^d)$ is continuously embedded in the space $C^k_{\to 0}(\mathbb{R}^d)$ which is defined as the space functions $u \in C^k(\mathbb{R}^d)$ such that $\lim_{|\mathbf{x}|\to\infty} D^\alpha u(\mathbf{x}) = 0$ for any $|\alpha| \leq k$, and equipped with the norm $|u|_k = \sum_{|\alpha|\leq k} \sup_{\mathbb{R}^d} |D^\alpha u|$.*

Proof. cf. R. Adams [1] □

B.3 Local Polynomial Reproduction

We discuss the construction of local polynomial reproduction on a compact set $\Omega \subseteq \mathbb{R}^2$. We consider polynomial spaces in two dimensions and total degree $m > 0$ restricted to Ω which will be denoted as $P_m^2 := P_m^2|_\Omega$ with dimension Q. As usual, we will denote by $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega$ a set of N pairwise distinct scattered points.

The basic definition of a local polynomial reproduction can be stated as follows.

Definition (LPR) A process that defines for every $\mathcal{X} \subseteq \Omega$ with point saturation parameter $h := h_{\Omega,\mathcal{X}}$ a family of functions $U := U_{\mathcal{X}} = \{u_1, \ldots, u_N\}$, $u_j : \Omega \mapsto \mathbb{R}$, is called a local polynomial reproduction of degree m on Ω if there exists constants $h_0, c_1, c_2 > 0$ such that for any $\mathbf{x} \in \Omega$

1. $\sum_{j=1}^N p(\mathbf{x}_j)u_j(\mathbf{x}) = p(\mathbf{x}), \quad \forall p \in P_m^2|_\Omega,$
2. $\sum_{j=1}^N |u_j(\mathbf{x})| \leq c_1$
3. $u_j(\mathbf{x}) = 0, \quad \text{if } \|\mathbf{x} - \mathbf{x}_j\|_2 > c_2 h$

are satisfied for any $\mathcal{X} \subseteq \Omega$ with $h \leq h_0$.

The first and third conditions justify the local polynomial reproduction while the second condition ensures that the so-called Lebesgue functions defined as $\sum_j |u_j(\mathbf{x})|$ are bounded uniformly over Ω.

Note that in the case $m = 0$ for the polynomial space P_m^2, namely the space of constants, local polynomial reproduction is easily satisfied. For any given $\mathbf{x} \in \Omega$, if one simply chooses an $\mathbf{x}_j$ which minimizes $\|\mathbf{x} - \mathbf{x}_j\|_2$ and sets $u_j(\mathbf{x}) = 1$, $u_k(\mathbf{x}) = 0$

for any $k \neq j$, then the reproduction conditions are satisfied. For the rest of the appendix, we will consider any $m \geq 1$.

It turns out that the existence for local polynomial reproduction can be shown for sets Ω which satisfy an interior cone condition. The recipe to construct local polynomial reproduction on such sets will be discussed in this appendix. We will closely follow the framework for local polynomial reproduction found in Wendland [66].

B.4 Cone condition and properties

To begin, we first recall the definition of the interior cone condition for the set Ω, which will play an important role. We then discuss some properties of cones and sets Ω which satisfy cone conditons.

Definition The set Ω satisfies an interior cone condition if there exists an angle $\theta \in (0, \pi/2)$ and a radius $r > 0$ such that for every $\mathbf{x} \in \Omega$ a unit vector $\xi(\mathbf{x})$ exists such that the cone

$$C(\mathbf{x}, \xi(\mathbf{x}), \theta, r) := \{\mathbf{x} + \lambda \mathbf{y} : \mathbf{y} \in \mathbb{R}^2, \|\mathbf{y}\|_2 = 1, \mathbf{y}^T \xi(\mathbf{x}) \geq \cos(\theta), \lambda \in [0, r]\} \quad \text{(B.10)}$$

is contained in Ω.

Note: We will usually omit the dependence of the unit vector ξ on $\mathbf{x}$ and simply write ξ.

The importance of the cone condition comes from the fact that a cone around the point $\mathbf{x}$ which we denote by $C(\mathbf{x}) := C(\mathbf{x}, \xi, \theta, r)$ is a convex set. This will enable us to create line segments inside the cone $C(\mathbf{x})$ where the line segments are not only in $C(\mathbf{x})$, but also in Ω due to the property that $C(\mathbf{x}) \subset \Omega$.

Lemma B.4.1. *A cone $C(\mathbf{x}) := C(\mathbf{x}, \xi, \theta, r)$ is a convex set.*

Proof. Without loss of generality, assume $\mathbf{x} = 0$, so the vertex of the cone is at zero. Let $\mathbf{x}_0$ and $\mathbf{x}_1$ be any two points in the cone and define the line segment $l(t) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ for $0 \leq t \leq 1$. We must show that every point on the line segment generated by l is in $C(\mathbf{x})$. Firstly, for two vectors $\xi \in \mathbb{R}^2$ and $\mathbf{z} \in \mathbb{R}^2$ with $\|\xi\|_2 = 1$, denote the angle between ξ and $\mathbf{z}$ by $\angle(\xi, \mathbf{z}) = \cos^{-1}(\mathbf{z}^T\xi/\|\mathbf{z}\|_2) = \theta$.

Now since $\mathbf{x}_0$ and $\mathbf{x}_1$ are in the cone, we have that $\angle(\mathbf{x}_0, \xi) \leq \theta$ and $\angle(\mathbf{x}_0, \xi) \leq \theta$. Since $\mathbf{x}_0$ and $\mathbf{x}_1$ are in the cone, we can find $\lambda_0 \leq r$ and $\lambda_1 \leq r$ such that $\mathbf{x}_0 = \lambda_0\mathbf{y}_0$ and $\mathbf{x}_1 = \lambda_1\mathbf{y}_1$ where $\mathbf{y}_0$ and $\mathbf{y}_1$ are unit vectors in C. Thus for any fixed $t \in [0,1]$ we have $\mathbf{z} := l(t) = (1-t)\lambda_0\mathbf{y}_0 + t\lambda_1\mathbf{y}_1$. Now $\angle(\mathbf{z}, \xi) = \cos^{-1}(\mathbf{z}^T\xi/\|\mathbf{z}\|_2)$ which implies that

$$((1-t)\lambda_0\mathbf{y}_0 + t\lambda_1\mathbf{y}_1)^T\xi \geq \|(1-t)\lambda_0\mathbf{y}_0 + t\lambda_1\mathbf{y}_1\|_2 \cos(\theta) \tag{B.11}$$

giving

$$((1-t)\lambda_0\mathbf{y}_0)^T\xi + (t\lambda_1\mathbf{y}_1)^T\xi = \|((1-t)\lambda_0\mathbf{y}_0)\|_2 \cos(\theta) + \|t\lambda_1\mathbf{y}_1\|_2 \geq \|(1-t)\lambda_0\mathbf{y}_0 + t\lambda_1\mathbf{y}_1\|_2 \cos(\theta) \tag{B.12}$$

and thus the $\angle(\mathbf{z}, \xi) \leq \theta$. To show that the length of the vector $\mathbf{z}$ is less than r, we have by the Triangle inequality $\|\mathbf{z}\|_2 = \|(1-t)\lambda_0\mathbf{y}_0 + t\lambda_1\mathbf{y}_1\|_2 \leq (1-t)\lambda_0 + t\lambda_1 \leq ((1-t)+t)r = r$. □

Another nice property of the cone condition is the following geometric property which will be used later.

Lemma B.4.2. *Suppose $C(\mathbf{x}) := C(\mathbf{x}, \xi, \theta, r)$ is a cone. For any $h \leq r/(1 + \sin\theta)$ the closed ball $B(\mathbf{y}, h\sin\theta)$ with center $\mathbf{y} = \mathbf{x} + h\xi$ and radius $h\sin\theta$ is contained in the cone $C(\mathbf{x})$.*

Proof. Without restriction, assume $\mathbf{x} = 0$. If $\mathbf{z} \in B$, then $\|\mathbf{x} - \mathbf{z}\|_2 = \|\mathbf{z}\|_2 \leq \|\mathbf{z} - \mathbf{y}\|_2 + \|\mathbf{y}\|_2 \leq h\sin\theta \leq r$. Thus the ball B is contained in the larger ball centered at zero with radius r. We now need to show now that $B(\mathbf{y}, h\sin\theta)$ is contained in the correct segment of the bigger ball centered at zero with radius r. Suppose that this is not the case. Then we can find a $\mathbf{z} \in B$ with $\mathbf{z} \notin C$, meaning that $\mathbf{z}^T\xi \leq \|\mathbf{z}\|_2 \cos\theta$. But this implies since $h\sin\theta$ is the radius of B that

$$h^2\sin^2\theta \geq \|\mathbf{z} - \mathbf{y}\|^2 = \|\mathbf{z} - h\xi\|^2 = \|\mathbf{z}\|^2 + h^2 - 2h\mathbf{z}^T\xi > \|\mathbf{z}\|^2 + h^2 - 2h\|\mathbf{z}\|\cos\theta. \tag{B.13}$$

Now subtracting $h^2\sin^2\theta$ from both sides of the inequality, we get

$$\begin{aligned} 0 &> \|\mathbf{z}\|^2 + h^2(1 - \sin^2\theta) - 2h\|\mathbf{z}\|\cos\theta \\ &= \|\mathbf{z}\|^2 + h^2\cos^2\theta - 2h\|\mathbf{z}\|\cos\theta \\ &= (\|\mathbf{z}\| - h\cos\theta)^2 \leq 0, \end{aligned} \tag{B.14}$$

and thus a contradiction. So $\mathbf{z} \in C$. □

Consequently as a corollary of this Lemma, we know that if $\mathbf{z}$ is a point in the ball, then the whole line segment $\mathbf{x} + t(\mathbf{z} - \mathbf{x})/\|\mathbf{z} - \mathbf{x}\|_2$, $t \in [0, r]$ is contained in the cone due to the convexity of the cone. This will be important later.

We now give an example of a set that satisfies an interior cone condition.

Lemma B.4.3. *Every ball with radius $\delta > 0$ satisfies an interior cone condition with radius $\delta > 0$ and angle $\theta = \pi/3$.*

Proof. Without restriction we assume the ball is centered at 0. For any $\mathbf{x} \neq 0$ in the ball, we choose the direction $\xi = -\mathbf{x}/\|\mathbf{x}\|$. A point in the cone is given by $\mathbf{x} + \lambda\mathbf{y}$ with some $\|\mathbf{y}\| = 1$ and $\mathbf{y}^T\xi \geq \cos\pi/3 = 1/2$ and $0 \leq \lambda \leq \delta$. Now we show that this point is still in the ball. We have

$$\|\mathbf{x} + \lambda\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \lambda^2 - 2\lambda\|\mathbf{x}\|\xi^T\mathbf{y} \leq \|\mathbf{x}\|^2 + \lambda^2 - \lambda\|\mathbf{x}\|.$$

The last expression equals $\|\mathbf{x}\|(\|\mathbf{x}\| - \lambda) + \lambda^2$ which can be bounded by $\lambda^2 \leq \delta^2$ in the case $\|\mathbf{x}\| \leq \lambda$. In the case $\lambda \leq \|\mathbf{x}\|$ then we can transform the last expression to $\lambda(\lambda - \|\mathbf{x}\|) + \|\mathbf{x}\|^2$ by swapping roles of λ and $\|\mathbf{x}\|$. Now since $\lambda - \|\mathbf{x}\| \leq 0$, we get $\lambda(\lambda - \|\mathbf{x}\|) + \|\mathbf{x}\|^2 \leq \|\mathbf{x}\|^2 \leq \delta^2$. Thus $\mathbf{x} + \lambda\mathbf{y}$ is in the ball. □

Now we have a second calculation that we will need.

Lemma B.4.4. *Let $C = C(\mathbf{x}_0, \xi, \theta, r)$ be a cone with angle $\theta \in (0, \pi/5]$ and radius $r > 0$. Define the point*

$$\mathbf{z} = \mathbf{x}_0 + \frac{r}{1 + \sin\theta}\xi.$$

Then we have for any $\mathbf{x} \in C$ the bound $\|\mathbf{x} - \mathbf{z}\| \leq \frac{r}{1+\sin\theta}$.

Proof. Without restriction we can assume $\mathbf{x}_0 = 0$. With $0 < \theta \leq \pi/5$, we have

$2\cos\theta/(1+\sin\theta) \geq 1$. Calculating $\|\mathbf{x}-\mathbf{z}\|^2$ for any $\mathbf{x} \in C$, we have

$$\begin{aligned}
\|\mathbf{x}-\mathbf{z}\|^2 &= \|\mathbf{x}\|^2 + \|\mathbf{z}\|^2 - 2\mathbf{z}^T\mathbf{x} \\
&= \|\mathbf{x}\|^2 + \frac{r^2}{(1+\sin\theta)^2} - 2\frac{r}{(1+\sin\theta)}\mathbf{x}^T\xi \\
&\leq \|\mathbf{x}\|^2 + \frac{r^2}{(1+\sin\theta)^2} - 2\frac{r\cos\theta}{(1+\sin\theta)}\|\mathbf{x}\|^2 \\
&\leq \|\mathbf{x}\|^2 - r\|\mathbf{x}\| + \frac{r^2}{(1+\sin\theta)^2} \\
&= \|\mathbf{x}\|(\|\mathbf{x}\| - r) + \frac{r^2}{(1+\sin\theta)^2} \\
&\leq \frac{r^2}{(1+\sin\theta)^2}.
\end{aligned} \tag{B.15}$$

□

Are final result of the cone condition property is to show that a cone satisfies and interior cone condition.

Lemma B.4.5. *Suppose $C = C(\mathbf{x}_0, \xi, \theta, r)$ is a cone with radius $r > 0$ and angle $\theta \in (0, \pi/5]$. Then C satisfies a cone condition with angle $\hat{\theta} = \theta$ and radius*

$$\tilde{r} = \frac{3\sin\theta}{4(1+\sin\theta)}r.$$

Proof. Without loss of generality we assume $\mathbf{x}_0 = 0$. Define $r_0 := \frac{\sin\theta}{(1+\sin\theta)}r$ and set the point $\mathbf{z} := (r - r_0)\xi = \frac{r}{(1+\sin\theta)}\xi$. By Lemma B.4.2, the ball $B(\mathbf{z}, r_0)$ is contained in the cone C. From Lemma B.4.3, we know that we can find for any $\mathbf{x} \in B(\mathbf{z}, r_0)$ a cone with angle $\pi/3 > 0$ and radius $r_0 > (3/4)r_0 = \tilde{r}$. This means we can find a cone for any point inside this ball. Now we need to show the cone condition for points outside of this ball but still in the cone.

We fix a point $\mathbf{x} \in C$ with $\mathbf{x} \notin B(\mathbf{z}, r_0)$, giving $\|\mathbf{x}-\mathbf{z}\| \geq r_0$. We define the direction of the proiuposed interior cone to be $\zeta = (\mathbf{z}-\mathbf{x})/\|\mathbf{x}-\mathbf{z}\|$. We have to show that any point $\mathbf{y} = \mathbf{x} + \lambda_0\eta$ with $\lambda_0 \in [0, \tilde{r}]$, and direction $\|\eta\| = 1$ with $\eta^T\zeta \geq \cos\theta$ lies in C. Define $\lambda := \|\mathbf{z} - \mathbf{x}\| \cos\theta + [r_0^2 + \|\mathbf{z} - \mathbf{x}\|^2(\cos^2\theta - 1)]^{1/2}$ which is in $\mathbb{R}$ since $\|\mathbf{x} - \mathbf{z}\| \leq r_0/(\sin\theta)$ by Lemma B.4.4 and this means that

$$\begin{aligned} r_0^2 + \|\mathbf{x} - \mathbf{z}\|^2(\cos^2\theta - 1) &= r_0^2 - \|\mathbf{x} - \mathbf{z}\|^2 \sin^2\theta \\ &\geq r_0^2 - \sin^2\theta(r_0^2/(\sin^2\theta)) = 0. \end{aligned} \tag{B.16}$$

Now restricting the angle θ to $(0, \pi/5)$ gives $\lambda \geq \|\mathbf{x}-\mathbf{z}\| \cos\theta \geq r_0 \cos\theta \geq 3r_0/4 = \tilde{r}$. This means that if the point $\mathbf{x} + \lambda\eta$ is contained in C, then so is $\mathbf{x} + \lambda_0\eta$ by the convexity of C, which gives the cone property. We show that $\mathbf{x} + \lambda\eta$ is in the ball $B(\mathbf{z}, r_0) \subset C$. This is done by the following calculation. Let $\eta \in \mathbb{R}^2$ with $\|\eta\|_2 = 1$ and $\eta^T\zeta \geq \cos\theta$. Then we have

$$\begin{aligned} \|\mathbf{z} - (\mathbf{x} + \lambda\eta)\|^2 &= \|\mathbf{z} - \mathbf{x}\|^2 + \lambda^2 - 2\lambda\eta^T(\mathbf{z} - \mathbf{x}) \\ &= \|\mathbf{z} - \mathbf{x}\|^2 + \lambda^2 - 2\lambda\eta^T\zeta\|\mathbf{z} - \mathbf{x}\| \\ &\leq \|\mathbf{z} - \mathbf{x}\|^2 + \lambda^2 - 2\lambda\cos\theta\|\mathbf{z} - \mathbf{x}\| \\ &= (\lambda - \cos\theta\|\mathbf{z} - \mathbf{x}\|)^2 - \|\mathbf{z} - \mathbf{x}\|^2\cos^2\theta + \|\mathbf{z} - \mathbf{x}\|^2 \\ &= (r_0^2 + \|\mathbf{z} - \mathbf{x}\|^2(\cos^2\theta - 1) - \|\mathbf{z} - \mathbf{x}\|^2\cos^2\theta + \|\mathbf{z} - \mathbf{x}\|^2 \\ &= (r_0^2 + \|\mathbf{z} - \mathbf{x}\|^2(\cos^2\theta - 1) - \|\mathbf{z} - \mathbf{x}\|^2(\cos^2\theta - 1) \\ &= r_0^2 \end{aligned} \tag{B.17}$$

Hence, $\mathbf{x} + \lambda\eta \in B(\mathbf{z}, r_0) \subset C$ which finishes the proof. □

B.5 Norming sets

We will also need the concepts of *unisolvent sets* and *norming sets* which will be used for local polynomial reproduction. The two concepts are directly related to each other as we will see.

Definition Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega \subseteq \mathbb{R}^2$ be a set of N distinct points and let $p, q \in P_m^2$ be any polynomials of degree m or less on Ω. Then $\mathcal{X}$ is P_m^2-unisolvent if $p(\mathbf{x}_j) = q(\mathbf{x}_j)$ for all $\mathbf{x}_j \in \mathcal{X}$ implies that $p = q$, i.e. the two polynomials are identical. We could also equivalently say that $\mathcal{X}$ is P_m^2-unisolvent if 0 is the only polynomial from P_m^2 that vanishes on $\mathcal{X}$.

Definition Let X be a Banach space and $V \subset X$ be a finite dimensional subspace with norm $\|\cdot\|_V$ (we will simply denote the norm for V by $\|\cdot\|$ if the context is clear). Define $W = \text{span}\{\lambda_1, \ldots, \lambda_N\} \subseteq V'$ to be the span of N linear bounded functionals on V (recall that V' is the dual space to V). The set W is called a norming set of V if there exists a $c > 0$ such that

$$\sup_{\lambda \in W,\, \|\lambda\|_{V'}=1} |\lambda(v)| \geq c\|v\|, \quad \forall v \in V \tag{B.18}$$

It is also useful to view the concept of *norming sets* in the sense of a sampling operator. For $W = \text{span}\{\lambda_1, \ldots, \lambda_N\} \subseteq V'$, if we consider the mapping $T : V \mapsto T(V) \subseteq \mathbb{R}^N$ defined by $T(v) = (\lambda_1(v), \ldots, \lambda_N(v)) \subseteq \mathbb{R}^N$ (we will call T the sampling operator), we could also say that the set W is a norming set of V iff the mapping T is injective. To see this, equip $\mathbb{R}^N$ with the l_∞-norm, with the dual space being l_1. Now if T is injective, we can define the norm of the inverse T^{-1} as

$$\|T^{-1}\| = \sup_{0 \neq x \in T(V)} \frac{\|T^{-1}x\|}{\|x\|_\infty} = \sup_{0 \neq v \in V} \frac{\|v\|}{\|T(v)\|_\infty} = \bar{c}. \tag{B.19}$$

The constant $\bar{c}$ will be termed the *norming constant* of the norming set W. The following Lemmas proves that this definition of a norming set is equivalent to definition

B.18.

Lemma B.5.1. *$W = span\{\lambda_1, \ldots, \lambda_N\}$ is a norming set for $V \subset X$ iff the sampling operator $T = (\lambda_1(v), \ldots, \lambda_N(v)) \in \mathbb{R}^N$ is injective.*

Proof. Suppose W is a norming set, but T is not injective. Then there exists a $0 \neq v \in V$ such that $T(v) = (\lambda_1(v), \ldots, \lambda_N(v)) = (0, \ldots, 0)$. On the other hand, since W is a norming set, we have for some $C = (c_1, \ldots, c_N) \in \mathbb{R}^N$ not all 0 and $c > 0$ such that,

$$\sup_{\lambda \in W,\, \|\lambda\|_{V'}=1} |\lambda(v)| = |c_1\lambda_1(v) + \ldots + c_N\lambda_N(v)| = |C \cdot T(v)| \geq c\|v\|.$$

But this is a contradiction since we assumed $T(v) = (\lambda_1(v), \ldots, \lambda_N(v)) = (0, \ldots, 0)$. Thus T must be injective.

Now suppose that T is injective. Since T is injective with norm $\|T^{-1}\| = 1/c_1$ for some constant $c_1 > 0$ and we have for any $0 \neq v \in V$ the bound $\|v\| = \|T^{-1}(T(v))\| \leq \|T^{-1}\|\|T(v)\|_\infty$. Thus

$$\begin{aligned} c_1\|v\| \leq \|T(v)\|_\infty &= \max_{1\leq j\leq N} |\lambda_j(v)| = \max_{1\leq j\leq N} \frac{|\lambda_j(v)|}{\|\lambda_j\|}\|\lambda_j\| \\ &\leq \big(\max_{1\leq j\leq N} \|\lambda_j\|\big) \max_{1\leq j\leq N} \frac{|\lambda_j(v)|}{\|\lambda_j\|} \\ &\leq c_3 \sup_{\lambda\in W} \frac{|\lambda(v)|}{\|\lambda\|}, \end{aligned} \tag{B.20}$$

where $c_3 := \max_j \|\lambda_j\| > 0$. Hence, W is a norming set for V. □

It is clear that we need at least $N \geq \dim V$ functionals to make the operator T injective. We can get along in fact with exactly $N = \dim V$ functionals. But in practice, usually the functionals W and N are given, for example point evaluation functionals which we use later. The natural question to ask is then how many of these functionals are necessary to make not only T injective, but also to control the norm of T and its inverse. This will be addressed in the following.

We now use the concept of norming sets in the following theorem.

Theorem B.5.1. *Suppose $V \subset X$ is a finite dimensional linear space with norm $\|\cdot\|_V$ and that $W = span\{\lambda_1, \dots, \lambda_N\} \subseteq V'$ is a norming set for V. For any $\lambda \in V'$, there exists a vector of real numbers $u = (u_1, \dots, u_N) \in \mathbb{R}^N$ depending on λ such that*

$$\lambda(v) = \sum_{j=1}^{N} u_j \lambda_j(v), \quad \forall v \in V \tag{B.21}$$

and

$$\sum_{j=1}^{N} |u_j| \leq \|\lambda\|_{V'} \|T^{-1}\|, \tag{B.22}$$

where T is the sampling operator determined by the set W.

Proof. We define the linear functional $\bar{\lambda}$ on $T(V)$ by $\bar{\lambda}(z) = \lambda(T^{-1}z)$ for $z \in T(V) \subset \mathbb{R}^N$. It has a norm that is bounded by $\|\bar{\lambda}\| \leq \|\lambda\|_{V'} \|T^{-1}\|$. By the Hahn-Banach theorem (Theorem A.1.3), $\bar{\lambda}$ has a norm-preserving extension $E\bar{\lambda}$ to all of $\mathbb{R}^N$. On $\mathbb{R}^N$ all linear functionals can be represented by the inner product with a fixed vector. Hence there exists $u \in \mathbb{R}^N$ such that

$$(E\bar{\lambda})(z) = (u, z) := \sum_{i=1}^{N} u_j z_j, \quad \forall z \in \mathbb{R}^N.$$

Using this, we observe that

$$\|E\bar{\lambda}\| = \sup_{0 \neq z \in \mathbb{R}^N} \frac{|(u, z)|}{\|x\|_\infty} = \|u\|_1$$

by duality. Since the extension E is norm preserving and $\|\bar{\lambda}\| \leq \|\lambda\|_{V'}\|T^{-1}\|$, we have $\|u\|_1 \leq \|\lambda\|_{V'}\|T^{-1}\|$. Finally, we find for an arbitrary $v \in V$, by setting $z = T(v) \in T(V)$,

$$\lambda(v) = \lambda(T^{-1}z) = \bar{\lambda}(z) = (E\bar{\lambda})(z) = (u, z) = \sum_{i=1}^{N} u_j z_j = \sum_{i=1}^{N} u_j \lambda_j(v), \qquad \text{(B.23)}$$

which is (B.21). This finishes the proof. □

We can now apply this theorem in the context of polynomial reproduction. If we choose $V = P_m^2|_\Omega \subset C^m(\Omega) = X$ endowed with the norm $\|\cdot\|_{\infty,\Omega} := \|\cdot\|_{L^\infty(\Omega)}$ and let $W = \text{span}\{\lambda_1, \dots, \lambda_N\} := \text{span}\{\delta_{\mathbf{x}_1}, \dots, \delta_{\mathbf{x}_N}\} \subset V'$ for the set of point centers $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, where $\delta_{\mathbf{x}_i}$ is the point evaluation functional at the center $\mathbf{x}_i \in \mathcal{X}$, and take $\lambda = \delta_{\mathbf{x}} \in V'$ for $\mathbf{x} \in \Omega$, then we have the following Lemma.

Lemma B.5.2. *The set $W = span\{\delta_{\mathbf{x}_1}, \dots, \delta_{\mathbf{x}_N}\}$ is a norming set for $P_m^2|_\Omega$ if and only if the set $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subseteq \Omega$ is P_m^2-unisolvent.*

Proof. Suppose $W = \text{span}\{\lambda_1, \dots, \lambda_N\}$ where $\lambda_j = \delta_{\mathbf{x}_j}$ is a norming set for $V = P_m^2|_\Omega$ but that $\mathcal{X}$ is not P_m^2-unisolvent, so there exists a $p \in P_m^2$ nonvanishing on Ω such that $p(\mathbf{x}_j) = 0$ for all $\mathbf{x}_j \in \mathcal{X}$. Now since W is a norming set, we can choose $\mathbf{x} \in \Omega$ with $\lambda = \delta_{\mathbf{x}}$ and can find numbers $(u_1(\mathbf{x}), \dots, u_N(\mathbf{x})) \in \mathbb{R}^N$, which are not all zero, such that

$$\lambda(p) = \sum_{j=1}^{N} u_j(\mathbf{x})\lambda_j(p) = \sum_{j=1}^{N} u_j(\mathbf{x})p(\mathbf{x}_j) = p(\mathbf{x}) \neq 0.$$

But since $u_j(\mathbf{x})$ are not all zero, $p(\mathbf{x}_j)$ cannot be all zero. Thus $\mathcal{X}$ is P_m^2-unisolvent.

Now suppose $\mathcal{X}$ is P_m^2-unisolvent and choose any $p \in P_m^2$ which is nonvanishing on Ω. We define the sampling operator $T(p) = (\delta_{\mathbf{x}_1}(p), \dots, \delta_{\mathbf{x}_N}(p))$. Since $\mathcal{X}$ is P_m^2-

unisolvent, the elements of the mapping $T(p) = (p(\mathbf{x}_1), \ldots, p(\mathbf{x}_N))$ are not all zero, thus T is injective. By Lemma B.5.1, since T is injective, W is a norming set for $P_m^2|_\Omega$. This completes the proof. □

It should be clear by now after the previous Lemma that we already have the first two properties in the definition of local polynomial reproduction. To see this, again let $W = \text{span}\{\lambda_j, \ldots, \lambda_N\} := \text{span}\{\delta_{\mathbf{x}_1}, \ldots, \delta_{\mathbf{x}_N}\}$, $V = P_m^2|_\Omega$, and assume $\mathcal{X} \subseteq \Omega$ is P_m^2-unisolvent on Ω. If we choose $\lambda = \delta_\mathbf{x} \in V'$ for $\mathbf{x} \in \Omega$, then for any polynomial $p \in P_m^2 = V$, applying theorem B.5.1, we can find a vector $u(\mathbf{x}) = \{u_1(\mathbf{x}), \ldots, u_N(\mathbf{x})\} \in \mathbb{R}^N$ dependent only on $\mathbf{x}$ such that

$$\lambda(p) = p(\mathbf{x}) = \sum_{j=1}^{N} u_j(\mathbf{x}) p(\mathbf{x}_j)$$

which gives the first property of local polynomial reproduction. Secondly, we have taking the norm on $\mathbb{R}^N$ to be l_∞ and the norm on $(\mathbb{R}^N)'$ to be l_1, we have the second property of local polynomial reproduction

$$\sum_{j=1}^{N} |u_j(\mathbf{x})| = \|u(\mathbf{x})\|_1 \leq \|\delta_\mathbf{x}\| \|T^{-1}\| = \|T^{-1}\| = C_2$$

by using the fact that $\|\delta_\mathbf{x}\| = \sup_{p \in P_m^2, \|p\|_\infty = 1} |p(\mathbf{x})| = 1$.

Now we only need to show the existence of such a vector $u = (u_1(\mathbf{x}), \ldots, u_N(\mathbf{x})) \in \mathbb{R}^N$ that also satisfies the third property, which is the local property of the polynomial reproduction. But first we need to give conditions on the point centers $\mathcal{X} \subseteq \Omega$ such that W defined above is indeed a norming set for the space of polynomials P_m^2 on Ω. This is hopeless in the case of general domains. We need a condition on the domain which will lead to this property. This will be the cone condition.

In this next Theorem, we give conditions on the set $\mathcal{X}$ that are sufficient to make it unisolvent. Furthermore, we give a value for the norming constant in the case $W = \text{span}\{\delta_{\mathbf{x}_1}, \ldots, \delta_{\mathbf{x}_N}\}$.

To prove the norming set property, we will use the fact that any multivariate polynomial in $\Omega \subseteq \mathbb{R}^2$ can be reduced to a univariate polynomial by restricting it to a line segment in Ω. We want to then relate the norm on the univariate polynomial on the line segment to the norm of the multvariate polynomial on Ω. To do this, we have to ensure the line segment will be completely contained in Ω, hence the use of the cone condition.

We will make use of a simple Markov inequality which states that for any $p \in P_m^1$, we have

$$|p'(t)| \leq m^2 \|p\|_{\infty,[-1,1]}, \quad t \in [-1,1].$$

Furthermore, a simple scaling argument shows that for $r > 0$ and all $p \in \mathcal{P}_m^1$, $t \in [0,r]$,

$$|p'(t)| \leq \frac{2}{r} m^2 \|p\|_{\infty,[0,r]}, \quad t \in [0,r].$$

Theorem B.5.2. *Suppose that $\Omega \subseteq \mathbb{R}^2$ is compact and satisfies an interior cone condition with radius $r > 0$ and angle $\theta \in (0, \pi/2)$. Suppose $m \in \mathbb{N}$, $m \geq 1$, is fixed and that the set of centers $\mathcal{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_N\} \subset \Omega$ satisfies*

1. *$h := h_{\mathcal{X},\Omega} \leq \frac{r \sin\theta}{4(1+\sin\theta)m^2}$*

2. *for every ball $B(\mathbf{x}, h) \subseteq \Omega$, there is a center $\mathbf{x}_j \in B(\mathbf{x}, h)$.*

Then $W = span\{\delta_{\mathbf{x}_j} : \mathbf{x}_j \in \mathcal{X}\}$ is a norming set for $\mathcal{P}_m^2|_\Omega$ with norming constant bounded by 2 for the sampling operator $T : P_m^2 \mapsto \mathbb{R}^N$ defined by $T(p) =$

$(\delta_{\mathbf{x}_i}(p))_{\delta_{\mathbf{x}_i} \in W}$.

Proof. Choose an arbitrary $p \in P_m^2$ with $\|p\|_{\infty,\Omega} = 1$. Since Ω is compact, we can find an $\mathbf{x} \in \Omega$ such that $|p(\mathbf{x})| = 1$. Furthermore, since Ω satisfies an interior cone condition we can find a $\xi \in \mathbb{R}^2$ with $\|\xi\|_2 = 1$ such that the cone $C(\mathbf{x}) := C(\mathbf{x}, \xi, \theta, r)$ is completely contained in Ω. Since $h/\sin\theta \le r/(1+\sin\theta)$ we can use Lemma B.4.2 with h replaced by $h/\sin\theta$ to find a ball $B(\mathbf{y}, h) \subset C(\mathbf{x})$ with center $\mathbf{y} = \mathbf{x} + (h/\sin\theta)\xi$. This implies that for an $\mathbf{x}_j \in \mathcal{X}$ with $\|\mathbf{y} - \mathbf{x}_j\|_2 \le h$, $\mathbf{x}_j$ is in the ball $B(\mathbf{y}, h)$, and since the cone is convex, the line segment $\mathbf{x} + t\frac{\mathbf{x}_j - \mathbf{x}}{\|\mathbf{x}_j - \mathbf{x}\|_2}$, $t \in [0, r]$, lies in the cone $C(\mathbf{x}) \in \Omega$. Now we can apply Markov's inequality with $r > \|\mathbf{x}_j - \mathbf{x}\|_2$ to the chosen polynomial p restricted to this line segment in $C(\mathbf{x})$ as

$$\hat{p}(t) := p(\mathbf{x} + t\frac{\mathbf{x}_j - \mathbf{x}}{\|\mathbf{x}_j - \mathbf{x}\|_2}),\ t \in [0, r],$$

We can see that

$$\begin{aligned}|p(\mathbf{x}) - p(\mathbf{x}_j)| = |\hat{p}(\|\mathbf{x} - \mathbf{x}_j\|_2) - \hat{p}(0)| &\le \int_0^{\|\mathbf{x}-\mathbf{x}_j\|_2} |\hat{p}'(t)|dt \le \|\mathbf{x} - \mathbf{x}_j\| \max_{0 \le t \le \|\mathbf{x}-\mathbf{x}_j\|} |\hat{p}(t)'| \\ &\le \|\mathbf{x} - \mathbf{x}_j\|_2 \frac{2}{r} m^2 \|\hat{p}\|_{\infty,[0,r]} \\ &\le h\frac{2(1+\sin\theta)}{r\sin\theta}(m-1)^2\|p\|_{\infty,\Omega} \\ &\le 1/2\end{aligned} \tag{B.24}$$

by using $\|\mathbf{x} - \mathbf{x}_j\|_2 \le \|\mathbf{x} - \mathbf{y}\|_2 + \|\mathbf{y} - \mathbf{x}_j\|_2 \le h + h/\sin\theta = h(1 + 1/\sin\theta)$ with the definition of the saturation measure h and the fact that $\|\mathbf{y} - \mathbf{x}\|_2 \le h/\sin\theta$. This shows that $|p(\mathbf{x}_j)| \ge 1/2$ since we know $|p(\mathbf{x})| = 1$. Now we have by applying

(B.19) with $V = P_m^2$,

$$\|T^{-1}\| = \sup_{p \in P_m^2,\, \|p\|_{\infty,\Omega}=1} \frac{\|p\|_{\infty,\Omega}}{\|T(p)\|_\infty} \leq \frac{1}{\max_{\mathbf{x}_j \in \mathcal{X}} |p(\mathbf{x}_j)|} \leq \frac{1}{1/2} = 2.$$

This proves the theorem. □

Now the next step in our recipe is utilizing Theorems B.5.2 and B.5.1 together. This gives the following Corollary.

Corollary B.5.1. *If $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega$ is P_m^2-unisolvent, then there exists for any $\mathbf{x} \in \Omega$ some real numbers $u_j(\mathbf{x})$ such that $\sum_j |u_j(\mathbf{x})| \leq 2$ and $\sum_j^N u_j(\mathbf{x})p(\mathbf{x}_j) = p(\mathbf{x})$ for all $p \in P_m^2$.*

Now we can finally assemble all these Theorems and Lemmas to give us the final product which is the local version of polynomial reproduction.

Theorem B.5.3. *Suppose that $\Omega \subseteq \mathbb{R}^2$ is a compact set which satisfies the cone condition for some angle $\theta \in (0, \pi/2)$ and radius $r > 0$. For fixed $m \in \mathbb{N}$, there exists constants $h_0, C_1, C_2 > 0$ depending only on $m, \theta,$ and r such that for any set of distinct point centers $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega$ with $h \leq h_0$ and any $\mathbf{x} \in \Omega$, we can find real numbers $u_j(\mathbf{x})$ for $1 \leq j \leq N$, such that*

1. $\sum_{j=1}^N p(\mathbf{x}_j)u_j(\mathbf{x}) = p(\mathbf{x}), \quad \forall p \in P_m^2|_\Omega,$

2. $\sum_{j=1}^N |u_j(\mathbf{x})| \leq C_1$

3. $u_j(\mathbf{x}) = 0,$ *if* $\|\mathbf{x} - \mathbf{x}_j\|_2 > C_2 h$

Proof. Without restriction, we assume that $\theta \leq \pi/5$. We define the constants as

$$C_1 = 2,\ C_2 = \frac{16(1+\sin\theta)^2 m^2}{3\sin^2\theta},\ h_0 = \frac{r}{C_2}.$$

Choose any $\mathbf{x} \in \Omega$. Let $h := h_{\mathcal{X},\Omega}$ be the point saturation measure for the set $\mathcal{X}$. Since $C_2 h \leq r$, the set Ω also satisfies a cone condition with angle θ and radius $C_2 h$. Denote this cone by $C(\mathbf{x}) := C(\mathbf{x}, \xi, \theta, C_2 h) \subset \Omega$. By Lemma B.4.4, the cone $C(\mathbf{x})$ itself satisfies a cone condition with angle θ and radius

$$\tilde{r} = \frac{3 \sin\theta}{4(1+\sin\theta)} C_2 h.$$

Substituting the definition of C_2 in the above equation and solving for h, we get that

$$h = \tilde{r} \frac{\sin\theta}{4(1+\sin\theta)m^2}.$$

Now consider the set of points $\mathcal{Y} = \mathcal{X} \cap C(\mathbf{x})$. To get our desired result, we need to show that $\mathcal{Y}$ is P_m^2-unisolvent on $C(\mathbf{x})$. This amounts to showing that $C(\mathbf{x})$ and $\mathcal{Y}$ satisfy the conditions in Theorem B.5.2, with $\Omega := C(\mathbf{x})$, $\mathcal{X} := \mathcal{Y}$, and h. Firstly, the point saturation measure h of the set $\mathcal{Y}$ obviously satisfies the first condition in Theorem B.5.2. For the second condition, we show that any ball $B(\mathbf{y}, h) \subset C(\mathbf{x})$ contains a point $\mathbf{x}_j \in \mathcal{Y}$. Indeed, by Lemma B.4.2, since $\tilde{r}/4(1+\sin\theta)m^2 \leq \tilde{r}/(1+\sin\theta)$, we know that the ball $B(\mathbf{y}, \tilde{r}\frac{\sin\theta}{4(1+\sin\theta)m^2}) = B(\mathbf{y}, h)$ with center $\mathbf{y} = \mathbf{x} + h\xi$ is in $C(\mathbf{x})$. By definition of the saturation parameter h, we know that there exists at least one $\mathbf{x}_j \in \mathcal{Y}$ which is in $B(\mathbf{y}, h)$. Thus $W = \operatorname{span}\{\delta_{\mathbf{x}_j},\ \mathbf{x}_j \in \mathcal{Y}\}$ is a norming set for $P_m^2|_{C(\mathbf{x})}$. Hence, by Corollary B.5.1 we can find numbers $u_j(\mathbf{x})$ for every j such that $\mathbf{x}_j \in \mathcal{Y}$ satisfying

$$\sum_{\mathbf{x}_j \in \mathcal{Y}} u_j(\mathbf{x}) p(\mathbf{x}_j) = p(\mathbf{x})$$

for all $p \in P_m^2|_{C(\mathbf{x})}$ and

$$\sum_{\mathbf{x}_j \in \mathcal{Y}} |u_j(\mathbf{x})| \leq 2.$$

Finally, to get the third property at the point $\mathbf{x} \in \Omega$, we simply set $u_j(\mathbf{x}) = 0$ for any $\mathbf{x}_j \notin \mathcal{Y}$. This gives properties 1), 2) and 3), and hence a local polynomial reproduction. □

We remark that in the construction for this proof, one might notice that the numbers $u_j(\mathbf{x})$ for $j = 1, \ldots, N$ depend not only on $\mathbf{x}$, but also on the chosen cone $C(\mathbf{x}) \subset \Omega$. For example, if one changes the direction of the cone $C(\mathbf{x})$, the numbers $u_j(\mathbf{x})$ might change since the points in the set $\mathcal{Y} \cap C(\mathbf{x})$ might change as well. For notational convenience however, we have omitted the dependency on the cone $C(\mathbf{x})$ from the numbers $u_j(\mathbf{x})$.

B.6 Local reproduction of derivatives of polynomials

We can also extend this local polynomial reproduction to include the local reproduction of derivatives of polynomials. Norming sets will again be a key ingredient. However, in order to cover the case of local reproduction of derivatives of polynomials, we first need an additional Bernstein inequality for mulitvariate polynomials. In the following Ω is assumed to be open, bounded and connected, which is necessary for estimates on the derivatives. In this case, $\overline{\Omega}$ is compact.

Lemma B.6.1. *Suppose $\Omega \subseteq \mathbb{R}^2$ is bounded and satisfies an interior cone condition with radius $r > 0$ and angle θ. If $p \in P_m^2$ and $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}_0^2$, $|\alpha| := \alpha_1 + \alpha_2 \leq m$,* *is a multi-index then*

$$\|D^\alpha p\|_{\infty,\Omega} \leq \left(\frac{2m^2}{r\sin\theta}\right)^{|\alpha|}\|p\|_{\infty,\Omega}. \tag{B.25}$$

Proof. Assume that ∇p is not identically zero. Let $\mathbf{x}_M$ be the point in $\overline{\Omega}$ that maximizes $\|\nabla p(\mathbf{x})\|_2$ over $\overline{\Omega}$ ($\|\cdot\|_2$ is the standard Euclidean norm). Because Ω satisfies a cone condition, so does $\overline{\Omega}$ and thus $\mathbf{x}_M$ is the vertex of a cone $C(\mathbf{x}_M) \subseteq \overline{\Omega}$ having radius r, an axis along a direction ξ, and an angle θ. Define the unit vector $\eta = \nabla p(\mathbf{x}_M)/\|\nabla p(\mathbf{x}_M)\|_2$ (adjusting the sign of p so that $\eta^T\xi \geq 0$). There is a unit vector ζ pointing into the cone and satisfying $\eta^T\zeta \geq \cos(\pi/2-\theta) = \sin\theta$. Using this we have $\|\nabla p(\mathbf{x}_M)\|\sin(\theta) \leq \nabla p(\mathbf{x}_M)\cdot\zeta = \partial p(\mathbf{x}_M)/\partial\zeta$ and so

$$\|\nabla p(\mathbf{x}_M)\|_2 = \frac{\partial p}{\partial\eta}(\mathbf{x}_M) \leq \csc(\theta)\frac{\partial p}{\partial\zeta}(\mathbf{x}_M). \tag{B.26}$$

Now for $t \in \mathbb{R}$, the polynomial $\tilde{p}(t) := p(\mathbf{x}_M + t\zeta)$ is in $P_M^1(\mathbb{R})$. In particular, it obeys the usual Bernstein inequality mentioned earlier on $0 \leq t \leq r$ given by

$$|\tilde{p}'(t)| \leq \frac{2m^2}{r}\max_{t\in[0,r]}|\tilde{p}(t)| \leq \frac{2m^2}{r}\|p\|_{\infty,\Omega}.$$

Now since $\tilde{p}(0) = (\partial p)/(\partial\zeta)(\mathbf{x}_M)$, we have for all $\mathbf{x} \in \overline{\Omega}$

$$\|\nabla p(\mathbf{x})\|_2 \leq \|\nabla p(\mathbf{x}_M)\|_2 \leq \csc(\theta)\frac{\partial p}{\partial\zeta}(\mathbf{x}_M) \leq \frac{2m^2}{r\sin\theta}\|p\|_{\infty,\Omega}.$$

Finally, noting that $|(\partial p/\partial\mathbf{x}_j)(\mathbf{x})| \leq \|\nabla p(\mathbf{x}_M)\|_2$, and differentiating p α times, we get the desired result. □

Now that we have shown the inequality $\|D^\alpha p\|_{\infty,\Omega} \leq \left(\frac{2m^2}{r\sin\theta}\right)^{|\alpha|}\|p\|_{\infty,\Omega}$, we can use this in Theorem B.5.2 along with Theorem B.5.1 to get the following global version of polynomial reproduction adapted to the derivatives of polynomials. First, let $W = \text{span}\{\lambda_j, \dots, \lambda_N\} := \text{span}\{\delta_{\mathbf{x}_1}, \dots, \delta_{\mathbf{x}_N}\}$ be a norming set for $V = P_m^2|_\Omega$. If we choose $\lambda = (\delta_{\mathbf{x}} \circ D^\alpha) \in V'$ for $\mathbf{x} \in \Omega$, then for any polynomial $p \in P_m^2$, applying Theorem B.5.1, we can find a vector $u^\alpha(\mathbf{x}) = (u_1^\alpha(\mathbf{x}), \dots, u_N^\alpha(\mathbf{x})) \in \mathbb{R}^N$ depending on $\mathbf{x}$ such that

$$D^\alpha p(\mathbf{x}) = \sum_{j=1}^N u_j^\alpha(\mathbf{x}) p(\mathbf{x}_j)$$

which gives the first property of local polynomial reproduction. Now using the bound shown in Lemma B.6.1, we can use this Theorem B.5.1 to get a bound on the Lebesgue function

$$\sum_{j=1}^N |u_j^\alpha(\mathbf{x})| = \|u^\alpha\|_1 \le 2\Big(\frac{2m^2}{r\sin\theta}\Big)^{|\alpha|}.$$

Thus is summarized in the following proposition.

Proposition Let $p \in P_m^2$ and let Ω be a bounded domain satisfying an interior cone condition with radius $r > 0$ and angle θ. Suppose that $h > 0$ and the set $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subseteq \Omega$ satisfy

1. $h \le \frac{r\sin\theta}{4(1+\sin\theta)m^2}$
2. for every $B(\mathbf{x}, h) \subseteq \Omega$ there is a point $\mathbf{x}_j \in \mathcal{X} \cap B(\mathbf{x}, h)$,

then for any multi-index $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}_0^2$ with $|\alpha| := \alpha_1 + \alpha_2 \le m$ there exists real numbers $u_j^\alpha(\mathbf{x})$ such that

$$D^\alpha p(\mathbf{x}) = \sum_{j=1}^N u_j^\alpha(\mathbf{x}) p(\mathbf{x}_j), \quad \forall p \in P_m^2 \tag{B.27}$$

and

$$\sum_{j=1}^N |u_j^\alpha(\mathbf{x})| \le 2\Big(\frac{2m^2}{r\sin\theta}\Big)^{|\alpha|}. \tag{B.28}$$

for all .

Proof. We already demonstrated that equation (B.27) holds using Theorem B.5.1 with $\lambda = (\delta_{\mathbf{x}} \circ D^\alpha) \in V'$ and $W = \{\lambda_1, \ldots, \lambda_N\}$ with $\lambda_j = \delta_{\mathbf{x}_j}$.

To show that (B.28) holds we use Theorem B.5.1 and the fact that W is a norming set for $P_m^2|_\Omega$ to get that $\sum_{j=1}^N |u_j^\alpha(\mathbf{x})| \leq \|\lambda\|_{V'}\|T^{-1}\|$ where T is the sampling operator associated with W. Thus we need to bound $\|\lambda\|_{V'}\|T^{-1}\|$. Firstly, by Theorem B.5.2, we have

$$\|T^{-1}\| = \sup_{p \in P_m^2,\, \|p\|_{\infty,\Omega}=1} \frac{\|p\|_{\infty,\Omega}}{\|T(p)\|_\infty} \leq \frac{1}{\max_{\mathbf{x}_j \in \mathcal{X}} |p(\mathbf{x}_j)|} \leq \frac{1}{1/2} = 2.$$

Now to bound $\|\lambda\|_{V'} = \|(\delta_{\mathbf{x}} \circ D^\alpha)\|_{V'}$, we use Lemma B.6.1, to get

$$\begin{aligned} \|\lambda\|_{V'} &= \max_{p \in P_m^2(\Omega), \|p\|_{\infty,\Omega}=1} |\lambda(p)| = \max_{p \in P_m^2(\Omega), \|p\|_{\infty,\Omega}=1} |D^\alpha p(\mathbf{x})| \\ &\leq \|D^\alpha p\|_{\infty,\Omega} \leq \left(\frac{2m^2}{r \sin\theta}\right)^{|\alpha|} \|p\|_{\infty,\Omega} = \left(\frac{2m^2}{r\sin\theta}\right)^{|\alpha|}. \end{aligned} \tag{B.29}$$

Thus we get our desired result. □

Now we can use this to get the local version. Using the fact that Ω (and consequently $\overline{\Omega}$) satisfies an interior cone condition, we can use the above proposition and proceed exactly as in Theorem B.5.3 to get the following local version of local polynomial reproduction for derivatives.

Theorem B.6.1. *Suppose that $\Omega \subseteq \mathbb{R}^2$ is a bounded and satisfies the cone condition for some angle $\theta \in (0, \pi/2)$ and radius $r > 0$. For fixed $m \in \mathbb{N}$, there exists constants $h_0, C_2^\alpha > 0$ depending only on m, θ, and r such that for any set of distinct point centers $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subseteq \Omega$ with $h \leq h_0$ and any $\mathbf{x} \in \Omega$, we can find real numbers $u_j^\alpha(\mathbf{x})$, for $1 \leq j \leq N$ such that*

1. $\sum_{j=1}^N p(\mathbf{x}_j) u_j^\alpha(\mathbf{x}) = D^\alpha p(\mathbf{x}), \quad \forall p \in P_m^2|_\Omega,$

2. $\sum_{j=1}^{N} |u_j^\alpha(\mathbf{x})| \leq C_1^\alpha h^{-|\alpha|}$

3. $u_j^\alpha(\mathbf{x}) = 0,$ *if* $\|\mathbf{x} - \mathbf{x}_j\|_2 > C_2^\alpha h$

Proof. Without restriction, we assume that $\theta \leq \pi/5$. We define the constants as

$$C_1^\alpha = 2\Big(\frac{1}{2(1+\sin\theta)}\Big)^{|\alpha|}, \; C_2^\alpha = \frac{16(1+\sin\theta)^2 m^2}{3\sin^2\theta}, \; h_0 = \frac{r}{C_2^\alpha}.$$

Choose any $\mathbf{x} \in \Omega$. Let $h := h_{\mathcal{X},\Omega}$ be the point saturation measure for the set $\mathcal{X}$. Since $C_2^\alpha h \leq r$, the set Ω also satisfies a cone condition with angle θ and radius $C_2^\alpha h$. Denote this cone by $C(\mathbf{x}) := C(\mathbf{x}, \xi, \theta, C_2^\alpha h) \subset \Omega$. By Lemma B.4.4, the cone $C(\mathbf{x})$ itself satisfies a cone condition with angle θ and radius

$$\tilde{r} = \frac{3\sin\theta}{4(1+\sin\theta)} C_2^\alpha h.$$

Substituting the definition of C_2^α in the above equation and solving for h, we get that

$$h = \tilde{r}\frac{\sin\theta}{4(1+\sin\theta)m^2}.$$

Now consider the set of points $\mathcal{Y} = \mathcal{X} \cap C(\mathbf{x})$. We already demonstrated in the proof of Theorem B.5.3 that the conditions

1. $h \leq \frac{\tilde{r}\sin\theta}{4(1+\sin\theta)m^2}$

2. for every $B(\mathbf{y}, h) \subseteq C(\mathbf{x})$ there is a point $\mathbf{x}_j \in \mathcal{Y}$,

are satisfied. Thus, by applying Proposition B.6 with $\Omega := C(\mathbf{x})$ and $\mathcal{X} := \mathcal{Y}$, we can find numbers $u_j^\alpha(\mathbf{x})$ for every j such that $\mathbf{x}_j \in \mathcal{Y}$ satisfying

$$\sum_{\mathbf{x}_j \in \mathcal{Y}} u_j^\alpha(\mathbf{x}) p(\mathbf{x}_j) = D^\alpha p(\mathbf{x})$$

for all $p \in P_m^2|_{C(\mathbf{x})}$ and

$$\sum_{\mathbf{x}_j \in \mathcal{Y}} |u_j^\alpha(\mathbf{x})| \le 2\Big(\frac{2m^2}{\tilde{r}\sin\theta}\Big)^{|\alpha|}.$$

Now since $h = \tilde{r}\frac{\sin\theta}{4(1+\sin\theta)m^2}$, and $C_1^\alpha := 2\Big(\frac{1}{2(1+\sin\theta)}\Big)^{|\alpha|}$, then we see that

$$\sum_{\mathbf{x}_j \in \mathcal{Y}} |u_j^\alpha(\mathbf{x})| \le 2\Big(\frac{2m^2}{\tilde{r}\sin\theta}\Big)^{|\alpha|} = C_1^\alpha h^{-|\alpha|},$$

which is the second property. Finally, to get the third property at the point $\mathbf{x} \in \Omega$, we simply set $u_j^\alpha(\mathbf{x}) = 0$ for any $\mathbf{x}_j \notin \mathcal{Y}$. This gives properties 1), 2) and 3), and hence a local polynomial reproduction of derivatives. □

C.7 Legendre Polynomials and Wendlend's Kernels

C.8 Legendre Polynomial Expansions

(Note on notation: In this appendix, we will frequently employ the notation $(\cdot,\cdot)_{L^2(I)}$ to mean the $L^2(I)$ inner product on the interval $I = (-1, 1)$ and $(\cdot,\cdot)_{r,I}$ to mean the inner product on the Sobolev space $H^r(I)$. Thus, using this notation $(u, v)_{L^2(I)}$ and $(u, v)_{0,I}$ are equivalent expressions for any $u, v \in L^2(I)$.)

Define P_N on I to be the space of polynomials of degree less than or equal to N restricted to the interval I. The Legendre polynomial of degree n on the unit interval $I = (-1, 1)$ is given by

$$L_n(x) = \frac{(-1)^n}{2^n n!}((1-x^2)^n)^{(n)} \tag{C.30}$$

and is the n-th eigenfunction of the Legendre differential equation

$$((1-x^2)L_n'(x))' + \lambda_n L_n(x) = 0, \quad x \in I \tag{C.31}$$

with eigenvalue $\lambda_n = n(n+1)$. Some useful properties are

$$\begin{aligned} &|L_n(x)| \leq 1, \quad x \in [-1,1] \\ &|L_n'(x)| \leq \frac{1}{2}n(n+1), \quad x \in [-1,1] \\ &L_n(\pm 1) = (\pm 1)^n \\ &L_n'(\pm 1) = \frac{1}{2}(\pm 1)^{n+1} n(n+1) \end{aligned} \tag{C.32}$$

The set of Legendre polynomials is an L^2-orthogonal system on I satisfying the orthogonality condition

$$\int_I L_l(x) L_m(x) dx = (l + \frac{1}{2})^{-1} \delta_{l,m}.$$

where $\delta_{l,m} = 0$ if $l \neq m$ and 1 otherwise, so $\|L_n\|_{L^2(I)} = (l + \frac{1}{2})^{-1/2}$. Thus, an orthonormal $L^2(I)$ system can be obtained by multiplying each L_n by its normalizing factor $1/\|L_n\|_{L^2(I)} = \sqrt{(n + 1/2)}$. (See Szego [54] for proofs of these properties).

We can also show that $\|L_n'\|^2_{L^2(I)} = n(n+1)$. Using integration by parts, we have

$$\begin{aligned} \int_I (L_n'(x))^2 dx &= -\int_I L_n(x)(L_n')'(x)dx + (L_nL_n')|_{-1}^{1} \\ &= -\int_I L_n(x)(L_n')'(x)dx + (L_n'(1) - (-1)^n L_n'(-1)). \end{aligned} \tag{C.33}$$

We first notice that $\int_I L_n(x)(L_n')'(x)dx = 0$ since $(L_n')'$ is an $n-2$ degree polynomial and L_n is orthogonal to any polynomial in $P_{n-2}(I)$ (if $n < 2$, $(L_n')' = 0$). Now using the property $L_n'(\pm 1) = (\pm 1)^{n+1} n(n+1)/2$, we get $(L_n'(1) - (-1)^n L_n'(-1)) = n(n+1)/2 - (-1)^{2n+1} n(n+1)/2 = n(n+1)$ and so it follows

$$\|L_n'\|^2_{0,I} = n(n+1). \tag{C.34}$$

The Legendre expansion of a function $v \in L^2(I)$ is given as $v(x) = \sum_{n=0}^{\infty} a_n L_n(x)$ with $a_n = (n + 1/2)\int_I v(x)L_n(x)dx$ where "=" is understood in the sense that

$$\lim_{p\to\infty} \|v - \sum_{n=0}^{p} a_n L_n\|_{L^2(I)} = 0. \tag{C.35}$$

For a given $N > 0$, the L^2-orthogonal projection $\Pi_N : L^2(I) \mapsto P_N(I)$ is a mapping such that for any $v \in L^2(I)$ we have

$$(v - \Pi_N v, \phi)_{L^2(I)} = 0, \quad \forall \phi \in P_N(I). \tag{C.36}$$

Since the set of Legendre polynomials L_n for $0 \leq n \leq N$ provides a basis for P_N on I, (C.36) is equivalent to finding coefficients $a_0, \ldots, a_N \in \mathbb{R}$ in the expansion $\Pi_N v = \sum_{n=0}^{N} a_n L_n$ such that (C.36) holds for all $\phi \in \text{span}\{L_0, \ldots, L_N\}$.

In order to prove an error estimate for Legendre polynomial approxmation in $L^2(I)$, we first will need a density argument and a Sobolev interpolation Theorem (proof can be found in Adams [1].

Lemma C.8.1. *Let $(a, b) \subset \mathbb{R}$ and $r \geq 0$ be any integer. The space $C^\infty([a, b])$ is dense in $H^r(a, b)$.*

In other words, functions in $H^r(a, b)$ can be approximated arbitrarily well by infinitely differentiable functions on $[a, b]$ in the distance induced by the norm of $H^r(a, b)$.

Theorem C.8.1. *Let $0 < s < 1$ and $k \geq 1$ be an integer. Then*

$$H^{k+s}(a, b) = [H^k(a, b), H^{k+1}(a, b)]_{s,2} \tag{C.37}$$

and the norms are equivalent.

Proof. Brenner and Scott [15], Theorem 14.2.3. □

This Theorem states that if $v \in H^{k+1}(a, b)$, then $v \in H^{k+s}(a, b)$ with $0 < s < 1$ where the space H^{k+s} is given by the interpolation between the spaces H^k and H^{k+1}.

The following Lemma is an error estimate for the difference between v and $\Pi_N v$ in $L^2(I)$.

Lemma C.8.2. *For any real $r \geq 0$, there is a constant $c > 0$ such that*

$$\|v - \Pi_N v\|_{0,I} \leq cN^{-r}\|v\|_{r,I}, \quad \forall v \in H^r(I) \tag{C.38}$$

Proof. In the case $r = 0$, we have $\|v - \Pi_N v\|_{0,I} \leq \|v\|_{0,I}$ which is trivially satisfied. For the moment, lets assume that r is an even integer, namely $r = 2m$ where $m \geq 1$, and that $v \in C^\infty([-1,1])$. Let A be the 2nd order differential operator defined by $Av = -((1-x^2)v')'$. Using equation (C.31), it follows that $AL_n = n(n+1)L_n$. For the left side of the bound (C.38), we have $\|v - \Pi_N v\|_{0,I}^2 = \sum_{n=N+1}^\infty a_n^2 \|L_n\|_{0,I}^2$. Using the definition of the Legendre coefficients, by an application of integration by parts we get for any $n > N$

$$\begin{aligned} a_n &= \frac{1}{\|L_n\|_{0,I}^2} \int_I v(x)L_n(x)dx = \frac{1}{n(n+1)\|L_n\|_{0,1}^2} \int_I v(x)AL_n(x)dx \\ &= -\frac{1}{n(n+1)\|L_n\|_{0,1}^2} \int_I (1-x^2)v'(x)L_n'(x)dx + (1-x^2)v'(x)L_n(x)|_{-1}^1. \end{aligned} \tag{C.39}$$

Since $v \in C^\infty([-1,1])$, all derivatives are bounded and continuous on $[-1,1]$ and thus $\lim_{x\to\pm1}(1-x^2)v'(x)L_n(x) = 0$. Applying integration by parts once more, we have

$$a_n = \frac{1}{n(n+1)\|L_n\|_{0,I}^2} \int_I Av(x)L_n(x)dx$$

. Iterating this procedure m times yields

$$a_n = \frac{1}{n^m(n+1)^m\|L_n\|^2} \int_I A^m v(x)L_n(x)dx \tag{C.40}$$

and thus

$$\begin{aligned} \|v - \Pi_N v\|_{0,I}^2 &= \sum_{n=N+1}^\infty \Big(\frac{1}{n^m(n+1)^m\|L_n\|_{0,1}^2} \int_I A^m v(x)L_n(x)dx\Big)^2 \|L_n\|_{0,1}^2 \\ &\le cN^{-4m} \sum_{n=0}^\infty \|L_n\|_{0,1}^2 (\frac{\int_I A^m v(x)L_n(x)dx}{\|L_n\|_{0,1}^2})^2 \\ &\le cN^{-4m}\|A^m v\|_{0,I}^2. \end{aligned} \tag{C.41}$$

Since A^m is a differential operator of order $2m$, taking the square root of both sides in (C.41) and applying the inequality $\|A^m v\|_{0,I} \le c\|v\|_{2m,I}$ for some constant $c > 0$, the desired bound is achieved for even $r > 0$ and $v \in C^\infty([-1,1])$. To obtain the result for $v \in H^{2m}(I)$, we recall from Lemma C.8.1 that $C^\infty([-1,1])$ is dense in $H^{2m}(I)$, thus the result holds for $v \in H^{2m}(I)$. For $r > 0$ between even integers, namely $2m-2 < r < 2m$, we apply the Sobolev space interpolation $[H^{2m-2}(I), H^{2m}(I)]_{s,2} = H^{2ms+(1-s)(2m-2)}$ for $0 < s < 1$ to get the desired result. □

Using Legendre polynomials, we can show an inverse inequality on the space of polynomials P_N.

Theorem C.8.2. *For every polynomial $v \in P_N(I)$, there exists a constant $C > 0$ such that*

$$\|v'\|_{0,I} \leq CN^2 \|v\|_{0,I}. \tag{C.42}$$

Proof. Any $v \in P_N$ can be expanded into a Legendre series given by $v(x) = \sum_{n=0}^N a_n L_n(x)$ where

$$\|v\|_{0,I}^2 = \sum_{n=0}^N \frac{1}{n+1/2}|a_n|^2.$$

Now using (C.34), we have

$$\begin{aligned}\|v'\|_{0,I}^2 = \|\sum_{n=0}^N a_n L_n'\|_{0,1}^2 &\leq N \sum_{n=0}^N |a_n|^2 \|L_N'\|_{0,I}^2 \leq N \sum_{n=0}^N |a_n|^2 (n(n+1)) \\ &\leq N \sum_{n=0}^N \big(\frac{1}{n+1/2}|a_n|^2 n(n+1)(n+1/2)\big) \\ &\leq N^2(N+1)(N+1/2) \sum_{n=0}^N \frac{1}{n+1/2}|a_n|^2 \\ &\leq N^2(N+1)(N+1/2)\|v\|_{0,I}^2\end{aligned} \tag{C.43}$$

so it follows that

$$\|v'\|_{0,I}^2 \leq 3N^4 \|v\|_{0,I}^2$$

which gives $\|v'\|_{0,I} \leq CN^2 \|v\|_{0,I}$. □

We now define a quadrature rule based on Legendre polynomials for integrating functions in $P_{2N-1}(I)$ and $L^2(I)$ (see Canuto [17] for derivations).

Theorem C.8.3. Gauss-Lobatto Integration *Let $\{\xi_i\}_{i=0}^N$ denote the zeros of $q(x) = L_{N+1}(x) + aL_N(x) + bL_{N-1}(x)$ with $-1 = \xi_0 < \cdots < \xi_N = 1$, where the parameters a, b are chosen such that $q(-1) = q(1) = 0$. Then one can find $N+1$ constants $\rho_0, \ldots, \rho_N$ such that*

$$\int_I p(x)dx = \sum_{i=0}^{N} \rho_i p(\xi_i), \quad \forall p \in P_{2N-1}. \tag{C.44}$$

Thus the numerical quadrature scheme on I is exact when integrating any polynomial $p \in P_{2N-1}$.

Now, given a function $v \in L^2(I)$, an approximate quadrature scheme using the Gauss-Lobatto-Legendre (GLL) rule is

$$\int_I u(x)dx \approx \sum_{i=0}^{N} \rho_i u(\xi_i) \tag{C.45}$$

where $\rho_0, \ldots, \rho_N$ and the zeros $\xi_0, \ldots, \xi_N$ are defined as follows:

$$\begin{aligned} &\xi_i : \xi_0 = -1, \xi_N = 1, \text{zeros of } L_N'(x),\ 1 \leq i \leq N-1 \\ &\rho_i : \frac{2}{2(N+1)(L_N(\xi))^2},\ 0 \leq i \leq N \end{aligned} \tag{C.46}$$

C.9 Two-dimensional Legendre Expansions

Let $Q = (-1,1)^2$. We denote the space P_N^2 on the square Q as

$$P_N^2(Q) = \{v = \sum_{0 \leq i,j \leq N} a_{ij} x_1^i x_2^j, \quad a_{ij} \in \mathbb{R},\ (x_1, x_2) \in Q\}. \tag{C.47}$$

This is a **tensor product space** of one-dimensional polynomials of degree N with $dim(P_N^2(Q)) = (N+1)^2$. (i.e. The space of functions which are polynomials of degree $\leq N$ in each variable. For simplicity in notation, we define $K := (N+1)^2$)

For $\mathbf{x} := (x_1, x_2) \in Q$, we define the k-th order Legendre polynomial as $L_k(\mathbf{x}) := L_i(x_1) \cdot L_j(x_2)$ where $k := i * N + j$ for $0 \leq i, j \leq N$ and L_i (L_j) are the ith (jth) degree Legendre polynomials on $I = (-1, 1)$. Then span$\{L_k(\cdot),\ 0 \leq k \leq K\}$ provides a basis for $P_N^2(Q)$. Furthermore, it can easily be shown using the orthogonal property of L_n in $L^2(I)$ that the two dimensional Legendre basis provides an orthogonal basis for $L^2(Q)$. We can normalize the k-order Legendre polynomial $L_k(\mathbf{x}) = L_i(x_1) \cdot L_j(x_2)$ by multiplying by the factor $(i + 1/2)^{1/2}(j + 1/2)^{1/2}$. Thus $\{L_k^*(\mathbf{x})\}_{k=0}^{\infty}$ where $L_k^*(\mathbf{x}) = (i + 1/2)^{1/2}(j + 1/2)^{1/2} L_i(x_1) L_j(x_2)$ is an orthonormal family for $L^2(Q)$.

The Legendre expansion of a function $v \in L^2(Q)$ is given as $v(\mathbf{x}) = \sum_{k=0}^{\infty} a_k L_k^*(\mathbf{x})$ with $a_k = \int_Q v(\mathbf{x}) L_k^*(\mathbf{x}) d\mathbf{x}$. The projection $\mathbf{\Pi}_N : L^2(Q) \mapsto P_N^2(Q)$ is defined by the unique element $\mathbf{\Pi}_N v \in P_N^2(Q)$ such that

$$(v - \mathbf{\Pi}_N v, \phi)_{L^2(I)} = 0, \quad \forall \phi \in P_N^2(Q). \tag{C.48}$$

This is equivalent to finding coefficients $a_0, \ldots, a_K \in \mathbb{R}$ in the expansion $\mathbf{\Pi}_N v(\mathbf{x}) = \sum_{k=0}^{K} a_k L_k^*(\mathbf{x})$ such that (C.36) holds for all $\phi \in \text{span}\{L_0^*, \ldots, L_N^*\}$.

The following Lemma is the two-dimensional version of Lemma C.8.2.

Lemma C.9.1. *For any real $r \geq 0$, there exists a constant $C > 0$ such that*

$$\|v - \mathbf{\Pi}_N v\|_{0,Q} \leq C N^{-r} \|v\|_{r,Q}, \quad \forall v \in H^r(Q). \tag{C.49}$$

Proof. The proof is nearly indentical to the proof of Lemma C.8.2. □

C.10 Wendland's Compactly Supported Kernels

In this section we review the construction and some properties of Wendland's compactly supported radial kernels which are used in the hybrid spectral-element/meshless approximation method. The many advantages of using Wendland's compactly supported functions for partial differential equations include their approximation ability, their fast summation ability on a set of collocation nodes $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \Omega \subset \mathbb{R}^d$, and as we will see, the fact that they can be represented as a univariant polynomial on a local domain in any dimension $d \in \mathbb{N}$. They are also symmetric and positive definite functions and whose Fourier transform decays like $(1 + \|\xi\|_2^2)^{-s}$ for some $s > d/2$.

We consider the functions $\Psi(\mathbf{x}) = \phi(r)$ with $r = \|\mathbf{x}\|$ which have the form

$$\phi(r) = \begin{cases} p(r) & 0 \leq r \leq 1 \\ 0 & r > 1 \end{cases} \tag{C.50}$$

where p is a univariate polynomial of the form $p(r) = \sum_j^m c_j r^j$ with $c_m \neq 0$. We will denote m the degree of Ψ or ϕ.

The construction of the compactly supported functions begins with the trucated power function $\Phi_l(r) = (1-r)_+^l$ where $(1-r)_+$ denotes $(1-r)$ if $r < 1$ and 0 otherwise. It was shown in [65] that $\Phi^l(r)$ is a positive definite function in $\mathbb{R}^d$ when $l \geq d/2 + 1$. Next we introduce the operator I and its inverse D by acting on any radial function $\phi(r)$ as

$$(I\phi)(r) = \int_r^\infty t\phi(t)dt \tag{C.51}$$

and

$$(D\phi)(r) = -\frac{1}{r}\phi'(r). \tag{C.52}$$

Using the truncated power function $\Phi_l(r)$, a class of positive definite compactly supported functions can be defined as

$$\phi_{d,k} = I^k \Phi_{\text{floor}(d/2+k+1)} \tag{C.53}$$

where floor(x) denotes the largest integer less than or equal to x. For simplicity in notation and without loss of generality, we will assume $d = 2m$ for some integer m, so we can supress the notation floor. In Wendland [65], the following Theorem is proved about the functions $\phi_{d,k}$.

Theorem C.10.1. *The functions $\phi_{d,k}$ induce positive definite functions in $\mathbb{R}^d$ with the form*

$$\phi_{d,k}(r) = \begin{cases} p_{d,k}(r) & 0 \le r \le 1 \\ 0 & r > 1 \end{cases} \tag{C.54}$$

where $p_{d,k}$ is a univariate polynomial of degree $m := floord/2+3k+1$. Furthermore, they belong to the class of functions $C^{2k}(\mathbb{R}^d)$ and possess a Fourier transform $\hat{\phi}_{d,k}(\xi)$ which decays like $(1+\|\xi\|^2)^{-d/2-k-1/2}$ and thus the native space of $\Psi(\mathbf{x},\mathbf{y}) := \Psi(\mathbf{x}-\mathbf{y}) := \phi_{d,k}(r)$ is equivalent in norm to the Sobolev space $H^s(\mathbb{R}^d)$ for $s := d/2+k+1/2$.

Proof. cf. [65], Theorems 1.2 and 2.1 □

We now give an important Thoorem about the polynomial representation of the positive definite function $\phi_{d,k}$.

Theorem C.10.2. *Within its support* $[0,1]$*, the functions* $\phi_{d,k}$ *have the representation*

$$\phi_{d,k}(r) = \sum_{j=0}^{l+2k} c_{j,k}^{(l)} r^j \tag{C.55}$$

with $l := d/2 + k + 1$*. The coefficients* $c_{j,k}^{(l)}$ *of the polynomial can be defined by the recursive relation in* k *for* $0 \leq i \leq k-1$ *by*

$$\begin{aligned} c_{j,0}^{(l)} &= (-1)^j \begin{pmatrix} l \\ j \end{pmatrix}, \quad 0 \leq j \leq l \\ c_{0,i+1}^{(l)} &= \sum_{j=0}^{l+2i} \frac{c_{j,i}^{(l)}}{j+2}, \; c_{1,i+1}^{(l)} = 0, \quad i \geq 0 \\ c_{j,i+1}^{(l)} &= -\frac{c_{j-2,i}^{(l)}}{j}, \quad i \geq 0, \; 2 \leq j \leq l+2i+2. \end{aligned} \tag{C.56}$$

Furthermore, precisely the first k *odd coefficients* $c_{j,k}^{(l)}$ *vanish.*

Proof. cf. [65], Theorem 1.3., [66], Theorem 9.12 □

We give examples of different positive definite compactly supported functions $\phi_{d,k}$ up to a positive constant factor for $d = 1,2,3$ and $k = 1,2,3$ in the following table originally produced in [65].

To conclude this review of compactly supported Wendland functions, we give a result from [64] which states an error estimate in L^∞ for a given function $f \in H^s(\mathbb{R}^d)$ approximation by Wendland functions $\Psi(\cdot, \mathbf{x}) := \phi_{d,k}(\| \cdot - \mathbf{x} \|) = \phi_{d,k}(r)$ given as $I_{\mathcal{X}} f = \sum_{i=1}^{N} \alpha_i \Psi(\cdot, \mathbf{x}_i)$.

Theorem C.10.3. *For* $\phi_{d,k}$*, let* $s = d/2 + k + 1/2$ *with* $k \geq 1$ *and* $d = 1,2$*. For every* $H^s(\mathbb{R}^d)$ *and compact domain* $\Omega \subseteq \mathbb{R}^d$ *which satisfies an interior cone condition, the*

Table C.1: Examples of $\phi_{d,k}$

$d = 1$	$\phi_{1,0}(r) = (1-r)_+$	C^0
$d = 1$	$\phi_{1,1}(r) = (1-r)_+^3(3r+1)$	C^2
$d = 1$	$\phi_{1,2}(r) = (1-r)_+^5(8r^2+5r+1)$	C^4
$d \leq 3$	$\phi_{3,1}(r) = (1-r)_+^4(4r+1)$	C^2
$d \leq 3$	$\phi_{3,2}(r) = (1-r)_+^6(35r^2+18r+3)$	C^4
$d \leq 3$	$\phi_{3,3}(r) = (1-r)_+^8(32r^3+25r^2+8r+1)$	C^6

interpolant $I_{\mathcal{X}} f = \sum_{j=1}^{N} \alpha_j \phi_{d,k}(\| \cdot - \mathbf{x}_j \|)$ *on the points* $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \Omega$ *with point saturation measure* h *satisfies the estimate*

$$\|f - I_{\mathbf{x}} f\|_{L^\infty(\Omega)} \leq C h^{s+1/2} \|f\|_{H^s(\mathbb{R}^d)} \tag{C.57}$$

for a sufficiently small h*. Thus interpolation with* $\phi_{d,k}$ *provides an approximation order of* $s + 1/2$*.*

Proof. cf. [65], Theorem 2.2 □

Bibliography

[1] R.A. Adams. *Sobolev Spaces.* Academic Press Inc., New York, 1975.

[2] Arakawa, A., Lamb, V.R., Computational Design and the basic dynamical processes of the UCLA general circulation model. *Methods in Computational Physics 17*, 1977, p173.

[3] Arakawa, A. Computational design for long-term numerical integration of the equations of fluid motion: two-dimensional incompressible flow: Part 1. *J. Comp. Phys. 1* 1977, p119.

[4] Aronszajin, N. (1950). Theory of Reproducing kernels. *Transaction of the American Mathematical Society* 68, 337-404.

[5] Atluri SN, Shen S, *The Basis of Meshless Domain Discretization: The Meshless Local Petrov Galerkin (MLPG) Method* (2003) Advance in Computational Mathematics

[6] F. Baer, H. Wang, J.J. Tribbia, A. Fournier. Climate Modeling with Spectral Elements, *Monthly Weather Review*, **134** 3610-3624

[7] Babuska, I., Banergee, U., Osborn, J.E. *Survey of meshless and generalized finite element methods: a unified approach*, Acta Numerica, 1-125.

[8] R.K. Beatson, J.B. Cherrie and C.T. Mouat, Fast fitting of radial basis functions: Methods based on preconditioned GMRES iteration, *Adv. Comput. Math. 11* (1999) 253-270.

[9] T. Belytschko, Y. Krongauz, D. Organ, M. Fleming and P. Krysl, *Meshless Methods: an overview and recent developments*, Comp. Meth. Appl. Mech. Eng. (139) 1996, 3-47.

[10] Y. M. Berezansky, *Functional Analysis I.* Birkhauser, Berlin

[11] Bergh, Lofstrom *Interpolation Spaces.* Grundlehren der mathematischen Wissenschaften, Springer- Verlag, Berlin, Heidelberg, New York, 1976

[12] C. Blakely *A New Backus-Gilbert Meshless Method for Initial-Boundary Value Problems.* International Journal on Computational Methods, Vol 3, No. 3, 2006

[13] J.P. Boyd *Chebyshev and Fourier Spectral Methods.* Dover Publications, Second Edition, 1999

[14] D. Braess. *Finite Elemente.* Springer-Verlag, Berlin-Heidelberg-New York, 1992.

[15] S.C. Brenner, L.R. Scott. *The Mathematical Theory of Finite Element Methods.* Springer, New York, 1994.

[16] M.D. Buhmann, *Radial Basis Functions: Theory and Implementations.* Cambridge University Press, Cambridge, 2003.

[17] C. Canuto, M. Hussaini A. Quarteroni and T. Zang *Spectral Methods in Fluid Dynamics* Springer-Verlag, New York, 1987. xvi + 557pp

[18] J. Cote, A. Staniforth. Semi-Lagrangian integration schems for atmospheric models: A review. *Monthly Weather Review* 1991

[19] R. A. DeVore, G. G. Lorentz, *Constructive approximation, A Series of Comprehensive Studies in Mathematics*, Springer-Verlag, Berline Heidelberg, 1993.

[20] G. Fasshauer, *Meshfree Methods*, to appear in Handbook of Theoretical and Computational Nanotechnology, M. Rieth and W. Schommers (eds.), American Scientific Publishers, 2006, 33-97

[21] G.E. Fasshauer, *Matrix-free multilevel moving least-squares methods.* in Approximation Theory, X St. Louis, MO, 2002 271-278.

[22] G.E. Fasshauer, *Approximate moving least-squares approximation: a fast and accurate multivariate approximation method*, in *Curve and Surface Fitting: Saint-Malo 2002* Nashboro Press, 2003 138 - 148.

[23] G.E. Fasshauer, *Appproximate moving least-squares approximation for Time-Dependant PDEs*, in *WCCM V* July 7-12, 2002 Vienna, Austria

[24] G. E. Fasshauer, *On smoothing for multilevel approximation with radial basis functions*, in Approximation Theory IX, Vol.II: Computational Aspects, Charles K. Chui, and L. L. Schumaker (eds.), Vanderbilt University Press, 1999, 55-62.

[25] A. Fournier, M.A. Taylor, J.J. Tribbia, *The Spectral-Element Atmosphere Model: High-Resolution Parallel Computation and Localized Resolution of Regional Dynamics.* Monthly Weather Review 123 (2004)726-748

[26] Galewsky, J., L. M. Polvani, R. K. Scott, 2003: *An initial-value problem to test numerical models of the shallow water equations.* Monthly Weather Review, in review

[27] A. Gelb, E. Tadmor, Adaptive Edge Detectors for Piecewise Smooth Data Based on the Minmod Limiter, CSCAMM Report CS-05-06, 2005

[28] V. Girault, P.A. Raviart. *Finite Element Methods for Navier-Stokes Equations.* Springer-Verlag, Berlin-Heidelberg-New York, 1986.

[29] M. A. Golberg, C. S. Chen, and S. R. Karur. Improved multiquadric approximation for partial differential equations. *Eng. Anal. Bound. Elem.*, 18(1):9-17, July 1996.

[30] P. Grisvard. *Elliptic problems in nonsmooth domains.* Pitman, Marsheld, 1985.

[31] Y. C. Hon and R. Schaback. On unsymmetric collocation by radial basis functions. *Appl. Math. Comput.*, 119(2-3): 177-186, 2001.

[32] `http://www.homme.ucar.edu`

[33] A. Iske, Reconstruction of functions from generalized Hermite-Birkhov data, in Approximation Theory VIII, Vol. 1: *Approximation and Interpolation*, C. Chui, and L. Schumaker (eds.), World Scientic Publishing, Singapore, 1995, 257-264.

[34] E. Kalnay *Atmopheric Modeling, Data Assimilation and Predictibility.* Cambridge University Press, 2003.

[35] E. J. Kansa. Multiquadrics—a scattered data approximation scheme with applications to computational Fluid-dynamics. I. Surface approximations and partial derivative estimates. *Comput. Math. Appl.,* 19(8-9): 127-145, 1990.

[36] G.E. Karniadakis, S.J. Sherwin, *Spectral/hp Element Methods for Computational Fluid Dynamics.* Oxford University Press, 1999, 404 pp.

[37] P. Lancaster and K. Salkauskas, *Surfaces generated by moving least squares methods*, Math. Comp. (37) 1981, 141-158.

[38] L. Ling and E. J. Kansa. A least-squares preconditioner for radial basis functions collocation methods. *Adv. Comput. Math.,* PIPS No: 5271809, 2004.

[39] A. Mahmood, D.J. Lynch, L.D. Phillip. A fast banded matrix inversion using connectivity of Schur's compliments. *IEEE Systems Engineering*, 1991

[40] http://citeseer.ist.psu.edu/karypis95metis.html

[41] F. J. Narcowich, R. Schaback and J.D. Ward. *Multilevel interpolation and approximation*, Appl. Comput. Harmon. Anal. 7 (1999), 243-261.

[42] F.J. Narcowich and J.D. Ward, *Generalized Hermite Interpolation via Matrix-Valued Conditonally Positive Definite Functions*, Math. Comp., 63 (1994), 661-688

[43] Sadourny, R., 1972: *Conservative finite-difference approximations of the primitive equations on quasi-uniform spherical grids.* Mon. Wea. Rev., 100, 136 144.

[44] R. Schaback, *Improved Error Bounds for Scattered Data Interpolation by Radial Basis Functions*, Math. Comp. 68 (1999), 201-216

[45] R. Schaback, *Optimal recovery in translation-invariant spaces of functions*, Annals of Numerical Mathematics, 4 (1997), 547-556.

[46] R. Schaback, *Native Hilbert spaces for radial basis functions, I*, in New Developments in Approximation Theory (Dortmund, 1998) Ser. Numer. Math., 132. Birkhser, Basel, 1999, 255-282.

[47] R. Schaback. *Convergence of unsymmetric kernel based meshless collocation methods.* Preprint from `http://www.num.math.uni-goettingen.de/schaback/research/group.html`, 2005.

[48] R. Schaback, *On the efficiency of interpolation by radial basis functions*, in Surface Fitting and Multiresolution Methods, A. LeMehaute, C. Rabut, and L.L. Schumaker, Venderbilt University Press, Nashville TN, 1997, 309-318.

[49] R. Schaback, *Error estimates and condition numbers for radial basis function interpolation*, Adv. in Comput. Math., 3 (1995), 251-264.

[50] R. Schaback, Z, Wu *Local Error Estimates for Radial Basis Function Interpolation of Scattered Data.* IMA J. Numer. Anal. 13 (1993) pp. 13-27

[51] N. Sivakumar and J. D. Ward, *On the least squares fit by radial functions to multidimensional scattered data*, Numer. Math. 65 (1993), 219-243.

[52] A. Staniforth, J. Cote, An accurate and efficient finite-element global model of the shallow water equations. *Monthly Weather Review* (1990).

[53] M.J. Suarex, and Held, I. H, 1994: *A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. Bull. Amer. Met. Soc.*, 75, 1825 1830.

[54] Szeg, G, *Orthogonal Polynomials* AMS Bookstore, 1975

[55] A. E. Taylor. *Introduction to functional analysis.* New York, John Wiley and Sons, 1958

[56] M. Taylor, J. Tribbia, M. Iskandarani *Performance of a spectral-element atmospheric model on the HP Exemplar SPP2000* NCAR Tech. Rep TN-439 + EDD, 16pp.

[57] Taylor, M., J. Tribbia, and M. Iskandarani, 1997a: *The spectral element method for the shallow water equations on the sphere.* J. Comp. Phys., 130, 92 108.

[58] Thomas, S. J., and R. D. Loft, 2002: Semi-implicit spectral element atmospheric model. *J. Sci. Comput.*, 17, 339-350.

[59] J. Thurburn and Y. Li, Numerical simulation of Rossby-Haurwitz waves, *Tellus A* **52** (2000)(2) pp. 181-189.

[60] H. Triebel. *Interpolation Theory, Function Spaces, and Differential Operators.* 2nd Edition, J. Barth. Publ. Leipzig, Germany, (1995)

[61] `http://nccs.gsfc.nasa.gov/systems.html`

[62] H. Wendland, *Scattered Data Modelling by Radial and Related Functions*, Habilitation thesis, Univ. Gottingen, 2002.

[63] H. Wendland, *Error estimates for interpolation by compactly supported radial basis functions of minimal degree*, J. Approx. Theory (93) 1998, 258-272

[64] H. Wendland, F.J. Narcowich and J.D. Ward, *Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting*, Mathematics of Computation 74 (2005), 743-763

[65] H.Wendland, F.J. Narcowich and J.D. Ward, *Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions*, Constructive Approximation 24 (2006), 175–186.

[66] H. Wendland, *Scattered Data Approximation*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge 2004

[67] H. Wendland, *Piecewise Polynomial, Positive Definite and Compactly Supported Radial Functions of Minimal Degree*, AICM 4 (1995), pp 389 - 396.

[68] H. Wendland, *Sobolev-type error estimates for interpolation by radial basis functions*, in: A. LeMhaut, C. Rabut, and L.L. Schumaker (eds.), Surface Fitting and Multiresolution Methods , 337-344, Vanderbilt University Press, Nashville, TN, 1997

[69] H. Wendland, *On the stability of meshless symmetric collocation for boundary value problems*, BIT Numerical Mathematics 47 (2007), 455-468

[70] A. S. M. Wong, Y. C. Hon, T. S. Li, S. L. Chung, and E. J. Kansa. *Multizone decomposition for simulation of time-dependent problems using the multiquadric scheme.* Comput. Math. Appl., 37(8):23-43, 1999.

[71] Williamson, D. L., Drake, J. B, Hack, J. J., Jakob, R., and Swarztrauber, P. N., *A standard test set for numerical approximations to the shallow water equations in spherical geometry. J. Comput. Phys.*, 102, 211-224.

[72] Z. Wu, *Hermite-Birkhoff interpolation of scattered data by radial functions*, Approximation Theory and its Applications, 8 (1992), 1-10.

[73] K. Yosida, *Functional Analysis*, Springer-Verlag, 1980

[74] X. Zhou, Y. C. Hon, and Jichun Li. *Overlapping domain decomposition method by radial basis functions.* Appl. Numer. Math., 44(1-2):241-255, 2003.

99