# Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

## ABSTRACT

## NUMERICAL METHODS FOR OPTIMAL TRANSPORT AND OPTIMAL INFORMATION TRANSPORT ON THE SPHERE

**by**
**Axel G. R. Turnquist**

The primary contribution of this dissertation is in developing and analyzing efficient, provably convergent numerical schemes for solving fully nonlinear elliptic partial differential equation arising from Optimal Transport on the sphere, and then applying and adapting the methods to two specific engineering applications: the reflector antenna problem and the moving mesh methods problem. For these types of nonlinear partial differential equations, many numerical studies have been done in recent years, the vast majority in subsets of Euclidean space. In this dissertation, the first major goal is to develop convergent schemes for the sphere. However, another goal of this dissertation is application-centered, that is evaluating whether the partial differential equation techniques using Optimal Transport are actually the best methods for solving such problems.

The reflector antenna is an optics inverse problem where one finds the shape of a reflector surface in order to refocus light into a prescribed far-field output intensity. This problem can be solved using Optimal Transport. The moving mesh methods problem is an adaptive mesh technique where one redistributes the density of the vertices of a mesh without tangling the edges connecting the vertices. Both Optimal Transport and Optimal Information Transport approaches can be used in solving this problem.

The Monge Problem of Optimal Transport is concerned with computing the "optimal" mapping between two probability distributions. This actually can define a Riemannian distance between probability measures in a probability space. Another choice of Riemannian metric on this space, the infinite-dimensional Fisher-Rao metric, gives an "information geometric" structure to the space of probability measures. It turns out that a simple partial differential equation can be solved for a mapping that relates to the underlying information geometry given by the

Fisher-Rao metric. Solving for such an "information geometric" mapping is known as Optimal Information Transport.

In this dissertation, a convergence framework is first established for computing the solution to the partial differential equation formulation of Optimal Transport on the sphere. This convergence framework uses geodesic normal coordinates to perform computations in local tangent planes. The numerical scheme also has a control on the Lipschitz constant of the discrete solution, which allows a convergence theorem for consistent and monotone discretizations to be proved in the absence of a comparison principle for the partial differential equation. Then, a finite-difference scheme for the partial differential equation formulation of Optimal Transport on the sphere is constructed which satisfies the hypotheses of the convergence theorem. An explicit formula for the mixed Hessian term is derived for two different cost functions. In order to construct a monotone discretization, discrete Laplacian terms are carefully added into the scheme. Current work has established convergence rates for solutions of monotone discretizations of linear elliptic partial differential equations on compact 2D manifolds without boundary. The goal is to then generalize these linearized arguments for the Optimal Transport case on the sphere.

Computations are performed for the reflector antenna problem. Other *ad hoc* schemes exist for computing the reflector antenna problem, but the proposed scheme is the most efficient provably convergent scheme. Further adaptations are made that allow for the scheme to deal with non-smooth cases more explicitly.

For the moving mesh methods problem, a comparison of computations via Optimal Transport and Optimal Information Transport is performed for the sphere using provably convergent monotone schemes for both computations. These comparisons show the merits of using Optimal Information Transport for some challenging computations. Optimal Information Transport also seems like a natural generalization to other compact 2D surfaces beyond the sphere.

# NUMERICAL METHODS FOR OPTIMAL TRANSPORT AND OPTIMAL INFORMATION TRANSPORT ON THE SPHERE

by
**Axel G. R. Turnquist**

**A Dissertation**
**Submitted to the Faculty of**
**New Jersey Institute of Technology and**
**Rutgers, The State University of New Jersey – Newark**
**in Partial Fulfillment of the Requirements for the Degree of**
**Doctor of Philosophy in Mathematical Sciences**

**Department of Mathematical Sciences**
**Department of Mathematics and Computer Science, Rutgers-Newark**

**May 2022**

**APPROVAL PAGE**

**NUMERICAL METHODS FOR OPTIMAL TRANSPORT AND OPTIMAL INFORMATION TRANSPORT ON THE SPHERE**

**Axel G. R. Turnquist**

---

Dr. Brittany D. Hamfeldt, Dissertation Advisor — Date
Associate Professor of Mathematical Sciences, NJIT

---

Dr. Yassine Boubendir, Committee Member — Date
Professor of Mathematical Sciences, NJIT

---

Dr. Cyrill B. Muratov, Committee Member — Date
Professor of Mathematical Sciences, NJIT

---

Dr. David G. Shirokoff, Committee Member — Date
Associate Professor of Mathematical Sciences, NJIT

---

Dr. Rongjie Lai, Committee Member — Date
Associate Professor of Mathematics, Rensselaer Polytechnic Institute, Troy, NY

## BIOGRAPHICAL SKETCH

**Author:** Axel G. R. Turnquist

**Degree:** Doctor of Philosophy

**Date:** May 2022

**Undergraduate and Graduate Education:**

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2022
- Bachelor of Science in Physics
  University of Washington, Seattle, WA, 2012

**Major:** Mathematical Sciences

**Publications:**

Hamfeldt, B. D. and Turnquist, A. G. R. (2022). On the reduction in accuracy of finite difference schemes on manifolds without boundary. *arxiv.org*. https://arxiv.org/abs/2204.01892

Turnquist, A. G. R. (2021). Adaptive mesh methods on compact manifolds via Optimal Transport and Optimal Information Transport. *arxiv.org*. https://arxiv.org/abs/2111.14276

Hamfeldt, B. D. and Turnquist, A. G. R. (2021). Convergent numerical method for the reflector antenna problem via Optimal Transport on the sphere. *Journal of the Optical Society of America A*, 38: 1704-1713.

Hamfeldt, B. D. and Turnquist, A. G. R. (2021). A convergent finite-difference method for Optimal Transport on the sphere. *Journal of Computational Physics*, 445.

Hamfeldt, B. D. and Turnquist, A. G. R. (2021). A convergence framework for Optimal Transport on the sphere. *arxiv.org*. https://arxiv.org/abs/2103.05739

Turnquist, A. G. R. and Rotstein, H. G. (2018). Quadratization: from conductance-based models to caricature models with parabolic nonlinearities *Encyclopedia of Computational Neuroscience*.

**Presentations:**

Turnquist, A. G. R. (2022, March 31) *Smooth Mesh Redistribution on Manifolds Using PDE Techniques* [Poster Session] Graduate Student Association 3-Minute Research Presentation, New Jersey Institute of Technology, Newark, New Jersey, United States.

Turnquist, A. G. R. (2021, November 15-19) *Convergent numerical schemes for optimal transport with applications on the sphere and beyond* [Conference Presentation] Schrödinger Problem and Mean-field PDE Systems: Computational and Theoretical Advances, Faculty of Sciences, Aix Marseille University, Marseille, France.

Turnquist, A. G. R. (2021, June 20-25) *Optical inverse problems and optimal transport* [Conference Presentation] Entropic Regularization of Optimal Transport and Applications, Banff International Research Station, Banff, Alberta, Canada.

Turnquist, A. G. R. (2021, June 7-11) *Optimal transport on the sphere* [Conference Presentation] Canadian Mathematical Society Summer Meeting StudC Research Session - Optimal Transport and Applications, University of Ottawa, Ottawa, Ontario, Canada.

Turnquist, A. G. R. (2021, June 7-11) *Optimal transport on the sphere* [Poster Session] Canadian Mathematical Society Summer Meeting StudC Research Session - Optimal Transport and Applications, University of Ottawa, Ottawa, Ontario, Canada.

Turnquist, A. G. R. (2021, April 21) *Optimal transport on the sphere* [Poster Session] Dana Knox Student Research Showcase, New Jersey Institute of Technology, Newark, New Jersey, United States.

Turnquist, A. G. R. (2021, April 13) *Solution guarantees for the reflector antenna problem* [Poster Session] Graduate Student Association 3-Minute Research Presentation, New Jersey Institute of Technology, Newark, New Jersey, United States.

Turnquist, A. G. R. (2020, November 9-13) *A convergence framework for the Monge-Ampère PDE on the sphere* [Poster Session] Workshop on Optimal Control, Optimal Transport, and Data Science, University of Minnesota, Minneapolis, Minnesota, United States.

Turnquist, A. G. R. (2020, January 15-18) *Towards convergent finite-difference schemes for the Monge-Ampère PDE on the sphere* [Conference Presentation] Recent Developments in Numerical Methods for PDEs: Joint Mathematics Meeting, Denver Convention Center, Denver, Colorado, United States.

Turnquist, A. G. R. and Leiser, R. (2017, April 19) *Effects of input amplitude and global coupling on network synchrony and entrainment* [Poster Session] Dana Knows Student Research Showcase, New Jersey Institute of Technology, Newark, New Jersey, United States.

感谢我老婆的爱

*I am thankful for my wife loving me*

## ACKNOWLEDGMENTS

I would like to especially thank my dissertation advisor Brittany Hamfeldt for all of her time and for our collaboration.

Also, I would like to thank my committee members Cyrill Muratov, Yassine Boubendir, David Shirokoff, and Rongjie Lai for their time and assistance.

Thanks to the Department of Mathematical Sciences for the Provost Doctoral Award in 2016, support during 2017 and 2018, and various other support throughout the Ph.D. I am grateful for the National Science Foundation Grants DMS 1313861, DMS 1608077, DMS 1751996, and the Graduate Research Fellowship Program (GRFP 1849508).

Thanks to Binan Gu for being a close friend and confidant during the Ph.D. program, and also co-hosting the Optimization & Machine Learning seminars.

Thanks to my family, my parents, brother, and wife for support during the challenges of the doctoral program.

## TABLE OF CONTENTS

**Chapter** **Page**

**TABLE OF CONTENTS**
**(Continued)**

**Chapter** **Page**

**TABLE OF CONTENTS**
**(Continued)**

**Chapter** **Page**

**TABLE OF CONTENTS**
**(Continued)**

**Chapter** **Page**

## LIST OF FIGURES

**Figure** **Page**

## LIST OF FIGURES
## (Continued)

**Figure** **Page**

## LIST OF SYMBOLS

| | |
|---|---|
| $\copyright$ | Copyright |
| $\int_\Omega$ | Integration over domain $\Omega$ |
| $\mathbb{S}^2$ | 2-sphere |
| $\|\cdot\|$ | Euclidean norm in 3-space |
| $\mathbb{R}^d$ | d-dimensional Euclidean space |
| $\nabla_M$ | intrinsic gradient w.r.t. manifold |
| $D_M^n$ | intrinsic $n$-th derivative w.r.t. manifold |
| $C^{k,\alpha}$ | $(k, \alpha)$ Hölder continuous |
| $L^p(\Omega)$ | $p$th-order integrable over a domain $\Omega$ |
| $\langle\cdot\rangle$ | average of function |
| $\|\cdot\|_\infty$ | infinity norm |
| $\partial$ | subdifferential or partial derivative |
| $\operatorname{diam}(\Omega)$ | diameter of domain $\Omega$ |
| $\to$ | limit or maps to |
| $\log$ | natural logarithm |
| $\lvert\cdot\rvert$ | absolute value |
| $\mathcal{O}$ | asymptotic order |
| $\lfloor\cdot\rfloor$ | floor function |
| $\circ$ | composition of maps |
| $\dot{u}(t)$ | time derivative of $u(t)$ |
| $\hookrightarrow$ | embeds into a space |
| div | divergence operator |
| $\inf_\Omega, \sup_\Omega$ | infimum/supremum over a set $\Omega$ |
| $\liminf_n, \limsup_n$ | limit inferior/limit superior in $n$ |

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

This dissertation is driven by two engineering applications: the reflector antenna problem and moving mesh methods. For the applications we have in mind, both of these problems are posed in non-Euclidean geometry and can be solved using Optimal Transport, see Wang (1996, 2004); Weller et al. (2016b). In the course of our exploration, we have also found that there is strong evidence that it is advantageous to utilize Optimal Information Transport to solve the moving mesh methods problem on compact manifolds, see Bauer et al. (2015).

First, we very briefly introduce the partial differential equation (PDE) formulation of the Optimal Transport problem. The original Optimal Transport engineering problem posed by Gaspard Monge (see Monge (1781)) considered the practical problem of moving a pile of sand into a hole, where the height function $f_0(x)$ of the pile and the depth function $f_1(y)$ of the hole were stipulated. The idea was to figure out how to move the sand from the pile to the hole (i.e., find a mapping $T(x)$) in the most "efficient" way possible, see Figure 1.1.

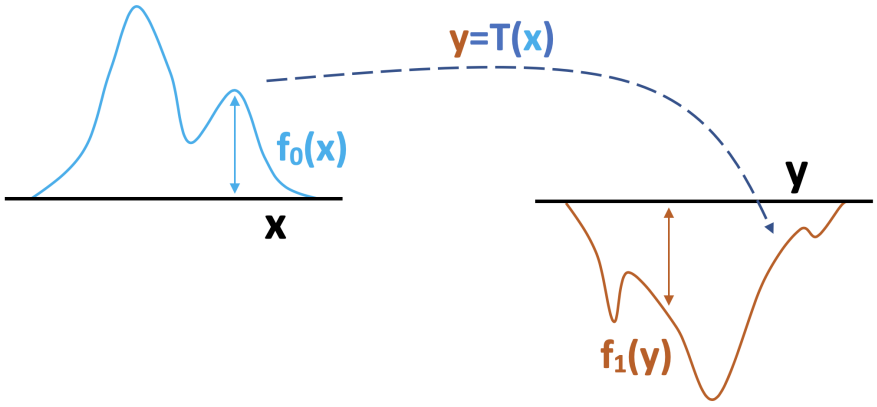



**Figure 1.1** The original engineering problem of shoveling a pile of dirt into a hole. Suppose we have a pile of sand of height $f_0(x)$ and a hole of depth $f_1(y)$. The object is to find a mapping $T(x)$ to prescribe where the mass at a location $x$ gets mapped.

Now, let us consider the source (pile of sand) and target (hole) mass configurations as probability measures (note: they could even be delta-distributions). We do not think of the hole as being "negative" in this case. This less restrictive formulation allows one to, for example, treat discrete, semi-discrete, and continuous formulations at the same time. However, this general formulation does not always lead to an optimal mapping $T(x)$, see Villani (2003). For the vast majority of this dissertation, we will be only treating cases where we end up with an optimal mapping, which is known as the Monge problem of Optimal Transport.

Up to this point, we have not discussed what we mean by "efficient". Usually, efficiency is expressed by defining a cost function $c(x, y)$ of transporting mass from the point $x$ to the point $y$. The most fundamental *a priori* assumptions we make on the cost function, those made in Villani (2003), are that it is measurable and non-negative. In Section 2.1.2, we will introduce the explicit assumptions which will lead to a PDE formulation of the Optimal Transport. Then the total efficiency is computed through a total cost $C(T)$ (which depends on the mapping chosen), integrated over the whole source domain $\Omega$:

$$C(T) \equiv \int_{\Omega} c\left(x, T(x)\right) f_0(x) dx. \tag{1.1}$$

The Monge problem of Optimal Transport on a compact manifold $M$ therefore considers finding a mapping $T$ between two probability measures that minimizes the cost functional in Equation (1.1). After assuming further regularity on the problem, one derives the complicated PDE for the function $u$ (much more information about the conditions and derivation are given in Sections 2.1.2 and 2.1.3):

$$-\det\left(D^2 u(x) + D^2_{xx} c(x,y)|_{y=T(x)}\right) + \frac{\left|D^2_{xy} c(x,y)|_{y=T(x)}\right| f_0(x)}{f_1(T(x))} = 0, \tag{1.2}$$

where the solution $u$ and the mapping $T$ are related via the equation:

$$\nabla u(x) = -\nabla_x c(x, T(x)), \tag{1.3}$$

and all derivatives are defined with respect to the Riemannian metric on the manifold $M$. This PDE formulation of Optimal Transport is a fully nonlinear second-order (degenerate) elliptic PDE. When this PDE is posed on manifolds without boundary, it also lacks a comparison principle, which prevents the use of traditional techniques for elliptic PDE. Furthermore, for many reasonable situations, the solution of this PDE is non-smooth, see Loeper (2009, 2011); Villani (2003) for a summary of the regularity theory in Euclidean space by Caffarelli and the extension to the Riemannian manifold case by many authors. As such, convergence analysis for numerical schemes is particularly challenging. Much more information about the derivation of this PDE and the Optimal Transport problem in general is introduced in Chapter 2.

Here we very briefly introduce the Optimal Information Transport problem, formulated and derived in Bauer et al. (2015). It is also concerned with the transporting of mass between two probability distributions $\mu_0$ and $\mu_1$. However, this is done in such a way that the Fisher-Rao distance between the two probability distributions is minimized. The Fisher-Rao distance arises from the Fisher-Rao metric, which is the second variation of the Kullback-Leibler divergence, which provides a kind of "measure" of how far apart two probability distributions are. Surprisingly, for compact manifolds $M$, the Fisher-Rao distance has an explicit formula. Supposing that $\mu_0 = f_0(x)dx$ and $\mu_1 = f_1(y)dy$, where $dx, dy$ are the standard volume forms on a compact manifold we get

$$d_F(\mu_0, \mu_1) = \sqrt{|M|} \arccos\left(\frac{1}{|M|}\int_M \sqrt{f_0(x)f_1(x)}dx\right), \tag{1.4}$$

where $|M|$ is the volume of the manifold Bauer et al. (2015). Furthermore, and critically for our purposes, one can solve for the mapping $T$ which minimizes the

Fisher-Rao distance between $\mu_0$ and $\mu_1$ by solving the simple equations

$$\begin{cases} \Delta f(t) = \frac{\dot{\mu}(t)}{\mu(t)} \circ \varphi(t)^{-1} \\ \dot{\varphi}(t) = \operatorname{grad}(f(t)) \circ \varphi(t), \quad \varphi(0) = \mathrm{id}, \end{cases} \tag{1.5}$$

where the diffeomorphic mapping $T$ between the probability measures $\mu_0$ and $\mu_1$, is given by $T = \varphi^{-1}(1)$, the terms $\mu(t)$ and $\dot{\mu}(t)$ have explicit forms for compact manifolds $M$, and, of course, $\mu(0) = \mu_0$ and $\mu(1) = \mu_1$ Bauer et al. (2015). The benefit of Optimal Information Transport versus Optimal Transport lies in the simplicity of the PDE and consequently the nicer (compared with the PDE in Equation (1.2)) regularity properties over more general compact surfaces without boundary. Much more information about this derivation will be presented in Chapter 2.

The main contribution of this dissertation is in developing numerical methods for solving fully nonlinear elliptic PDE on manifolds and analyzing and proving their applicability to solving the reflector antenna problem and the moving mesh methods problem. There has been a large body of work on developing numerical methods for such PDE in Euclidean space in the past twenty years or so. Our approach in this dissertation is to develop monotone finite-difference discretizations of such PDE. The definition of monotonicity does not require one to use finite-difference schemes, although most of the constructions of monotones schemes have naturally used finite-difference discretizations. Monotone schemes in Euclidean space were constructed in papers such as Benamou et al. (2016); Benamou and Duval (2017); Benamou et al. (2014); Bonnet and Mirebeau (2021); Chen et al. (2018); Froese (2012, 2018); Froese and Oberman (2011a,b, 2013); Hamfeldt and Salvador (2018); Hamfeldt (2019, 2018); Hamfeldt and Lesniewski (2022a,b); Liu et al. (2017); Oberman (2006, 2008).

## 1.2 Contributions of This Dissertation

In this dissertation, for the Optimal Transport problem on the sphere with two

cost functions, the squared geodesic cost function $c(x, y) = \frac{1}{2}d(x, y)^2$ and the logarithmic cost function $c(x, y) = -\log \|x - y\|$, I have developed a numerical convergence framework for discretizations, designed numerical schemes which satisfy the hypotheses of this framework, and demonstrated the success of computations even in non-smooth cases. It should be mentioned, however, that the convergence framework developed in this dissertation is not specific to these two cost functions, but can be generalized quite easily. The squared geodesic cost computations have direct applications to moving mesh methods and the logarithmic cost function is used in a formulation of the reflector antenna problem of geometric optics.

I have prepared three manuscripts Hamfeldt and Turnquist (2021a,b,c) that develop this avenue of research from theory to full "real-world" implementation and justification. The first manuscript, Hamfeldt and Turnquist (2021a) develops a convergence framework for consistent and monotone numerical discretizations of the Optimal Transport PDE on the sphere. The manuscript Hamfeldt and Turnquist (2021b) develops a finite-difference scheme which satisfies the hypotheses of the convergence framework and implements the scheme on various examples. The manuscript Hamfeldt and Turnquist (2021c) focuses specifically on the reflector antenna problem and proposes adaptations for non-smooth examples arising in optics, demonstrating the success of the implementation and its advantages in efficiency over other provably convergent schemes. In the manuscript Turnquist (2021), I performed a study of Optimal Transport versus Optimal Information Transport as applied to the moving mesh problem on the sphere, with a discussion indicating how one can extend the results to compact 2D surfaces. In the manuscript Hamfeldt and Turnquist (2022), I have also derived explicit convergence rates for monotone discretizations of linear elliptic PDE on compact 2D manifolds without boundary, by constructing smooth barrier functions and then invoking the discrete comparison principle to get bounds. The surprising result, corroborated by empirical evidence, is that the rate of convergence is worse than the formal consistency error. I also develop optimal gradient bounds for such solu-

tions on manifolds without boundary. The idea is to use the results for linear PDE to establish convergence rates for the nonlinear Optimal Transport PDE. A further study on using higher-order schemes that can be adapted into the convergence framework via filtered schemes is also a work in progress.

In Chapter 2, we introduce the Optimal Transport problem. In particular, we focus on the Monge problem and the PDE formulation of Optimal Transport and related regularity results for compact surfaces. We also introduce Optimal Information Transport, which yields an alternative means of defining mapping between probability measures that is not as plagued with regularity issues as Optimal Transport. Then, we introduce the two primary applications of our work on numerical methods, which are the reflector antenna problem and the moving mesh problem. We then discuss the approaches researchers have used in using finite-difference schemes to solve the PDE arising from Optimal Transport. We close this section with a recap of the theoretical and numerical complications arising from our particular geometry.

In Chapter 3, we introduce how the Optimal Transport problem will be solved on the sphere, with the key innovations of a tangent plane interpretation of the PDE and Lipschitz regularization (which allows for compactness arguments to work in the absence of a comparison principle for the underlying PDE). We then show how these innovations, along with the construction of consistent and monotone schemes, allow for a uniform convergence result of the discrete solution to the solution of the PDE using compactness arguments. The work in this chapter is mostly from the publications Hamfeldt and Turnquist (2021a,b).

In Chapter 4, we explicitly show how to construct the discretization by deriving a monotone scheme for the second-order derivatives, deriving an explicit formula for the mixed Hessian term, and by adding regularizing discrete Laplacians to establish monotonicity overall. We also discuss extensions to non-smooth problems and derive the computational complexity of the entire algorithm. We cap off this chapter with computational results showing the generality of the scheme

over different grids, validating the computations with some simple examples, and performing a computation for non-smooth densities. The work in this chapter is mostly from the publications Hamfeldt and Turnquist (2021b,c).

In Chapter 5, we compare the computational complexity of our scheme with other schemes proposed for the reflector antenna problem. We then demonstrate our computations with various examples, including a non-smooth globe example and some other singular examples and then perform validation via ray tracing. The work in this chapter is from the publication Hamfeldt and Turnquist (2021c).

In Chapter 6, we propose using Optimal Information Transport on the sphere for the moving mesh methods problem. We thereby conduct a side-by-side numerical study of the implementation of our Optimal Transport scheme and a monotone and provably convergent implementation for Optimal Information Transport on the sphere (with details on the analysis of the scheme in Appendix E). We demonstrate the computations with a smooth example where both moving mesh computations perform well along with a non-smooth example (the globe) where Optimal Information Transport appears to perform better. The work from this chapter is from the publication Turnquist (2021).

In Chapter 7, for monotone schemes for linear PDE on compact surfaces, we establish convergence rates which depend on the consistency error of the monotone scheme. We show, using a simple 1D example on the torus, that there is empirical evidence that these convergence rates are tight. We also show that a post-processing step can be made to make the numerical gradient of the discrete solution converge to the gradient of the solution. The goal then is to relate the linear convergence rates to the nonlinear case. The work from this chapter is from the publication Hamfeldt and Turnquist (2022).

In Chapter 8 regarding ongoing/future work, we introduce how higher-order schemes for the Optimal Transport problem on the sphere can be incorporated into provably convergent schemes by constructing a "filtered scheme". We then introduce our ideas for extending the Optimal Transport and Optimal Information

Transport methods to compact surfaces. And finally, we also introduce the ideas underlying a PDE-based monotone finite-difference scheme for computing $W_1$.

In Chapter 9, the conclusion, we summarize the various contributions and core ideas of this dissertation. In the Appendices, there are also various technical derivations that are not essential to the telling of the story.

# CHAPTER 2

# BACKGROUND

## 2.1 Optimal Transport

In the past ten years, computational Optimal Transport has garnered much special attention from the optimization community, not least after its gradual introduction to the field of machine learning. The computational implementation in learning problems was greatly popularized after entropically regularized Optimal Transport came into vogue thanks in part to the work of Marco Cuturi who introduced a particularly computationally efficient means to approximate the Optimal Transport distance in Cuturi (2013).

The applications of Optimal Transport extend much beyond machine learning, however. There are applications of Optimal Transport in computer graphics Solomon et al. (2014), image registration Haker et al. (2001, 2004), geophysical inverse problems Engquist and Froese (2014); Engquist et al. (2016); Yang et al. (2018), gene expression Schiebinger et al. (2017), optics Wang (1996, 2004); Yadav (2018), astronomy McCann (2006), dissipative equations Caffarelli et al. (2003); Otto (2001), probability Léonard (2012, 2013), economics and mean-field games Gomes et al. (2015); Lasry and Lions (2007), matching problems Pass (2015), diffeomorphic density matching Bauer et al. (2015), image analysis Gangbo et al. (2019); Wang et al. (2013), shape recognition Gangbo and McCann (2000), generative adversarial networks Arjovsky et al. (2017), barycenter computations Agueh and Carlier (2011); Carlier et al. (2015); Julien et al. (2011), moving mesh methods Budd et al. (2013); Budd and Williams (2009); Weller et al. (2016b), statistics Bigot (2020), and traffic modeling Santambrogio (2015), to name a few.

### 2.1.1 The Optimal Transport Problem in General

Here, we first present a very general formulation of the Optimal Transport problem in $d$-dimensional Euclidean space $\mathbb{R}^d$. The original engineering problem posed by

Gaspard Monge (if you are so inclined to read the original text from the eighteenth century, see the manuscript Monge (1781)) considered the practical problem of moving a pile of sand into a hole. The idea was to figure out how to do it in the most "efficient" way possible, see Figure 2.1. For Gaspard Monge, the cost of moving mass from a point $x$ to a point $y$ was proportional to the distance from $x$ to $y$. One of the ways that Optimal Transport is applicable to a broad range of problems is due to the fact that there are very general conditions on the cost of moving mass from $x$ to $y$ that allow for fruitful interpretations, see Santambrogio (2015); Villani (2003).

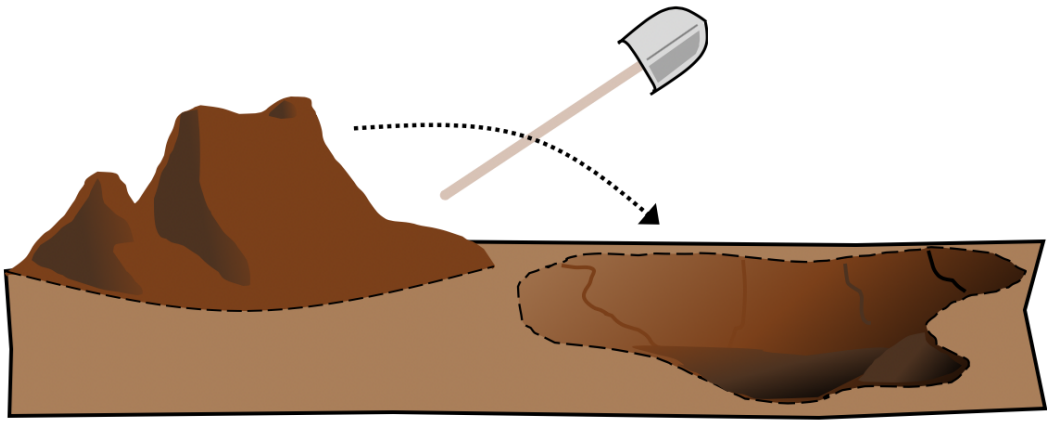


**Figure 2.1** The original engineering problem of shoveling a pile of dirt into a hole.

The problem is formulated more clearly and generally if one considers the source (pile of sand) and target (hole) mass configurations as probability measures. Thus, we dispense with the notion of the depth of the hole being "negative" and simply model both source and target distributions by positive probability measures. There are many advantages to this generalization. The most obvious, perhaps, from an applied perspective, is that it allows one to formulate and treat discrete, semi-discrete, and continuous formulations at the same time. Thus, we start with a source mass distribution denoted by the probability measure $\mu_0$ supported on $X \subset \mathbb{R}^d$, while our target mass distribution will be denoted by the probability measure $\mu_1$ supported on $Y \subset \mathbb{R}^d$. One would like to know how to

"move" the mass from $\mu_0$ to $\mu_1$. One immediately realizes that this cannot be done in general by a mapping, see a simple example given in Figure 2.2. However, one can assign a portion of the source mass to be assigned to different regions of the target mass distribution. The object that will best describe this is mass splitting is known as the transport plan $\pi$, which is a probability distribution supported on $X \times Y$.

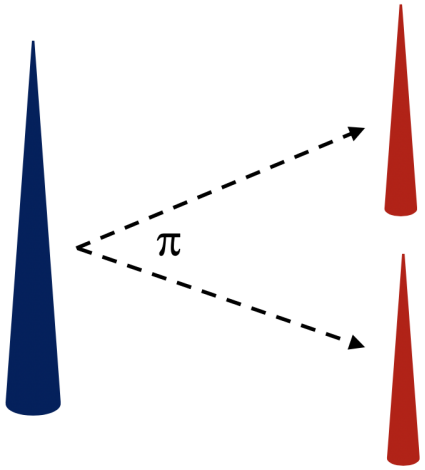



**Figure 2.2** One unit delta source mass (blue) dividing into two half-unit delta target masses (red) requires a true transport plan, which says that the mass must split in half, i.e. half the mass is earmarked for one target delta spike and half is earmarked for the other target delta spike.

The constraint on the problem is that all the source mass must be moved onto the target mass. To close the problem, we would like to add an economic cost function $c(x, y)$ (measurable and non-negative, see Villani (2003)) that when summed up over the source mass, will measure how much transporting the mass from the source distribution to the target distribution will cost. The idea is then to find the transport plan $\pi$ between $\mu_0$ and $\mu_1$ that minimizes the total cost over moving each grain of sand from the pile to the hole. That is, mathematically, we find the minimizer $\tilde{\pi}$:

$$\tilde{\pi} = \text{argmin}_{\pi \in \Pi(\mu_0, \mu_1)} \int_{X \times Y} c(x, y) d\pi(x, y), \tag{2.6}$$

where $\pi \in \Pi(\mu_0, \mu_1)$ denotes the fact that the mass is locally conserved, that is the

simultaneous facts that $\pi(A\times Y) = \mu_0(A)$ and $\pi(X\times B) = \mu_1(B)$ for any Borel sets $A, B \subset X, Y$, respectively. Provided that the cost function is continuous, using the direct method of the calculus of variations immediately yields the existence of such a transport plan $\pi$, since the set $\Pi(X, Y)$ is compact with respect to the weak topology of measures Villani (2003). It is perhaps easier to understand that the minimizers $\pi$ of the discrete formulation are the bistochastic matrices $\pi$, see Peyré and Cuturi (2019); Thorpe (2019) for more information on the discrete formulation.

Another important aspect of this formulation is that when we choose $c(x, y) = \frac{1}{2}\|x - y\|^2$, then the quantity

$$W_2(\mu_0, \mu_1) := \sqrt{\min_{\pi\in\Pi(\mu_0,\mu_1)} \int_{X\times Y} \frac{1}{2}\|x - y\|^2 \, d\pi(x, y)} \tag{2.7}$$

defines a Riemannian distance over the space of square-integrable probability measures $\mathcal{P}_2$, see Villani (2003, 2009) for more details. What happens then if $\pi = \delta_{(x,T(x))}$ for some function $T(x)$? This then reduces to the Monge problem of Optimal Transport, where we end up with a mapping $T(x)$ indicating precisely where the mass located at the point $x \in X$ should be mapped to $T(x) \in Y$.

### 2.1.2 The Monge Problem of Optimal Transport

For the Monge problem we present here, we will make even further assumptions that lead to a PDE formulation. We pose this problem on a $d$-dimensional compact manifold $M$ without boundary. Then, given two probability measures that are absolutely continuous with respect to the canonical volume form on $M$, that is $\mu_0(dx) = f_0(x)dx$ and $\mu_1(dy) = f_1(y)dy$, we seek a Lipschitz and injective mapping $T$, such that $T_\#\mu_0 = \mu_1$. This condition means that

$$\int_{T(E)} f_1(y)dy = \int_E f_0(x)dx \tag{2.8}$$

for every Borel set $E \subset M$. Assuming $T$ is sufficiently differentiable, this condition is equivalent to the following

$$\int_E f_1(T(x))J\left(T(x)\right)dx = \int_E f_0(x)dx, \tag{2.9}$$

where $J\left(T(x)\right)$ is the Jacobian of the transformation $T$ at a point $x \in M$, see Loeper (2009) for more information on the manifold case.

Note that the example given in Figure 2.2, which did not lead to a mapping, was from a delta-measure to a mixture of delta-measures, which are examples of measures which are not absolutely continuous with respect to the volume form. More generally, the condition is that the source mass distribution cannot give mass to "small" sets, see Santambrogio (2015) for more information on the technical details of this condition.

Again, for economic reasons we attach to this problem the auxiliary quantity

$$C(T) = \int_{\mathbb{R}^n} c(x, T(x))d\mu_0(x). \tag{2.10}$$

Our problem is to minimize this total cost:

$$T^* = \operatorname{argmin} C(T). \tag{2.11}$$

This represents a Calculus of Variations problem with a nonlinear equality constraint $T_{\#}\mu_0 = \mu_1$. Since the source measure is absolutely continuous with respect to the volume form, the problem has a solution, see Villani (2003), i.e. there is a unique minimizer $T$ that also satisfies the equality constraint. The distance-squared cost function yields good regularity results in $n$-dimensional Euclidean space and the $n$-sphere, promising that the PDE formulation of the Monge problem is well-suited to applications and computation. After assuming the following technical conditions originating from Ma et al. (2005) on the cost function we get the PDE formulation of the Monge problem:

1. $c(x, y)$ is a $C^4\left(\bar{X} \times \bar{Y}\right)$ function in both variables.

2. For all $x \in \bar{X}$, the mapping $y \mapsto -\nabla_x c(x,y)$ is injective on $\bar{Y}$.

3. The cost function $c$ satisfies $\det D^2_{xy} c \neq 0$.

4. There exists a $C_0 > 0$ such that for all $(x,y) \in X \times Y$, for all $\xi, \nu \in T_x M$ such that $(\xi,\nu)_g$,

$$\mathfrak{G}_c(x,y)(\xi,\nu) \leq C_0 |\nu|^2 |\xi|^2 \tag{2.12}$$

where

$$\mathfrak{G}_c(x,y)(\xi,\nu) = D^4_{p_\nu p_\nu x_\xi x_\xi} \left[(x,p) \to -c(x, c - \exp(p))\right]|_{x, p=-\nabla_x c(x,y)}, \tag{2.13}$$

where $c - \exp(p)$ is the given by the mapping $y \mapsto -\nabla_x c(x,y)$. Usually we will require $C_0 > 0$, however, the case $C_0 \geq 0$ is also special, since it is the necessary condition if one desires regularity from the mapping. More will be explained about this in Section 2.1.4.

The first condition, of course, is necessary for the object $\mathfrak{G}_c(x,y)(\xi,\nu)$ in the fourth condition to exist. The second and third conditions are necessary to assert the existence of maximizers of the dual formulation of the Optimal Transport problem, see Section 2.1.3 for more detail about the dual formulation. The potential functions arising in the dual formulation are related directly to the solution of the PDE formulation. The fourth condition is explained particularly clearly for the manifold case in Loeper (2009) as the cost-sectional curvature condition. As stated above, it is the necessary condition (when $C_0 \geq 0$) that is used in the regularity proofs.

Note: these technical conditions are satisfied by the squared geodesic cost $c(x,y) = \frac{1}{2} d_{\mathbb{S}^2}(x,y)$ and the logarithmic cost $c(x,y) = -\log \|x - y\|$. Note that the underlying manifold **does** have an effect on the squared geodesic cost function via the Riemannian distance function. Given these assumptions on the cost function, one can derive the PDE (whose formal derivation is outlined in Section 2.1.3):

$$F(x, u(x), Du(x), D^2u(x)) := -\det\left(D^2u(x) + D^2_{xx}c(x,y)|_{y=T(x)}\right) + \frac{\left|D^2_{xy}c(x,y)|_{y=T(x)}\right| f_0(x)}{f_1(T(x))}, \tag{2.14}$$

where

$$\nabla u(x) = -\nabla_x c(x, T(x)), \tag{2.15}$$

and $u$ is a $c$-convex function on $\mathbb{S}^2$. The definition for manifolds is presented in Loeper (2009). The support of the source mass will be denoted by $\Omega$ and the support of the target will be denoted as $\Omega'$.

**Definition 2.1.** *A function $\phi : M \to \mathbb{R}$ is c-convex if at each point $x \in \Omega$, there exist $y \in \Omega'$ and a value $\phi^c(y)$ such that*

$$-\phi^c(y) - c(x, y) = \phi(x), \tag{2.16}$$

$$-\phi^c(y) - c(x', y) \leq \phi(x'), \quad \forall x' \in \Omega, \tag{2.17}$$

*where the function $\phi^c(y)$ is defined by:*

$$\phi^c(y) = \sup_{x \in \Omega} \left(-c(x, y) - \phi(x)\right). \tag{2.18}$$

Interestingly, in the dual formulation of Optimal Transport, which is explained in Villani (2003), the function $u$ here is the same as one of the two potential functions arising in the dual formulation. Thus, we will occasionally refer to $u$ as the potential function.

Equation (2.14) is fully nonlinear and degenerate elliptic for $c$-convex functions $u$ (which solve the Optimal Transport problem). Utilizing the notion of viscosity solutions allows one to build convergent numerical schemes even for non-smooth solutions $u$ for this challenging elliptic PDE, see Barles and Souganidis (1991) and the summary in Section 2.5. However, many standard results for elliptic PDE, such as the comparison principle, do not apply due to the lack of boundary on the manifold.

By convention, in the Euclidean case with the squared geodesic cost $c(x, y) = \frac{1}{2}\|x - y\|^2$, Equation (2.14) reduces to

$$\det D^2 u(x) = \frac{f_0(x)}{f_1(T(x))}, \tag{2.19}$$

and this equation is usually referred to as the Monge-Ampère equation. Conscious of our slight abuse of terminology, we will sometimes refer to Equation (2.14) as the Monge-Ampère equation, or more correctly, as an equation of Monge-Ampère type.

### 2.1.3 Formal Derivation of the Optimal Transport PDE

Here we briefly, and formally, derive Equation (2.14). A very clean derivation for the squared distance case is shown in Evans (1997) and the general case is shown in Ma et al. (2005). First, one begins with a dual formulation of the Kantorovich problem, see Equation (2.6), whose supremum is equal to the solution of the Kantorovich problem

$$\min_{\pi \in \Pi(\mu_0, \mu_1)} \int_{X \times Y} c(x, y) d\pi(x, y) = \sup_{(u,v) \in K} \mathcal{I}(u, v), \tag{2.20}$$

where

$$\mathcal{I}(u, v) = \int_M u(x) f_0(x) dx + \int_M v(y) f_1(y) dy, \tag{2.21}$$

and

$$K = \{(u, v) | u(x) + v(y) \leq -c(x, y), \forall x \in M, y \in Y\}. \tag{2.22}$$

Given the MTW conditions on the cost function, see Ma et al. (2005), it turns out that this problem is solved by the pair of maximizers:

$$\begin{cases} u(x) = \inf_y \{-c(x, y) - v(y)\}, \\ v(y) = \inf_x \{-c(x, y) - u(x)\}. \end{cases} \tag{2.23}$$

Fixing a point $x_0$, then, there exists a $T(x_0)$ such that $u(x_0) = -c(x_0, T(x_0)) - v(T(x_0))$. At any other point $x \in M$ we have, $u(x) \leq -c(x, T(x_0)) - v(T(x_0))$,

since $T(x_0)$ is not necessarily an optimal choice corresponding to the point $x$. Therefore, the gradient is zero at the minimum (first-order condition) and the function has non-negative definite Hessian (second-order condition) there as well

$$\begin{cases} \nabla_x \left( u(x) + c\left(x, T(x_0)\right) + v\left(T(x_0)\right)\right)|_{x=x_0} = 0, \\ D_x^2 \left( u(x) + c\left(x, T(x_0)\right) + v\left(T(x_0)\right)\right)|_{x=x_0} \geq 0. \end{cases} \tag{2.24}$$

Therefore, we write

$$\begin{cases} \nabla u(x) + \nabla_x c\left(x, T(x)\right) = 0, \\ D^2 u(x) + D_{xx}^2 c\left(x, T(x)\right) \geq 0. \end{cases} \tag{2.25}$$

We take the first-order condition from Equation (2.25) and differentiate it again with respect to $x$:

$$D^2 u(x) + D_{xx}^2 c\left(x, T(x)\right) = -D_{xy}^2 c\left(x, T(x)\right) \nabla T(x). \tag{2.26}$$

Taking the determinant, we get:

$$\det\left(D^2 u(x) + D_{xx}^2 c\left(x, T(x)\right)\right) = \det\left(-D_{xy}^2 c\left(x, T(x)\right)\right) \det\left(\nabla T(x)\right). \tag{2.27}$$

From here, one proceeds proving that such a $T$ is measure-preserving (i.e. it satisfies Equation (2.9). Thus, we can replace $\det \nabla T(x)$ by $f_0(x)/f_1(T(x))$ to get

$$\det\left(D^2 u(x) + D_{xx}^2 c\left(x, T(x)\right)\right) = \det\left(-D_{xy}^2 c\left(x, T(x)\right)\right) f_0(x)/f_1(T(x)). \tag{2.28}$$

By the second-order condition, see Equation (2.25), we get

$$\det\left(D^2 u(x) + D_{xx}^2 c\left(x, T(x)\right)\right) = \left|\det\left(D_{xy}^2 c\left(x, T(x)\right)\right)\right| f_0(x)/f_1(T(x)). \tag{2.29}$$

### 2.1.4 Regularity

For Equation (2.14) posed on a manifold $M$, one could, of course, try and apply available *a priori* estimates. That is, one could look for the most general results for regularity theory for fully nonlinear degenerate elliptic PDE, see Caffarelli and Cabré (1995) and apply them across the sphere with a patching argument. However, one may run into issues relating to mass transporting (via the mapping $T(x)$) from $x$ to points $y = T(x)$ where the exponential map loses differentiability. Specific work on Equation (2.14) for the squared geodesic and the logarithmic cost functions, however, has been done for the sphere, see Loeper (2011).

The *a priori* regularity of the solution $u$ has a direct effect on the construction of numerical schemes, on applications, and on the convergence rates of the discrete solution to $u$. Here we separate the results on the $n$-sphere into two régimes, smooth and non-smooth (but differentiable).

**Hypothesis 2.2** (Conditions on data (smooth))**.** *We require problem data to satisfy the following conditions:*

*(a) There exists some $m > 0$ such that $f_1(x) \geq m$ for all $x \in \mathbb{S}^n$.*

*(b) The mass balance condition holds, $\int_{\mathbb{S}^2} f_0(x)\,dx = \int_{\mathbb{S}^2} f_1(y)\,dy$.*

*(d) The data satisfies the regularity requirements $f_0, f_1 \in C^{1,1}(\mathbb{S}^n)$.*

*(e) The cost functions are either the squared geodesic cost $c(x,y) = \frac{1}{2} d_{\mathbb{S}^2}(x,y)^2$ or the logarithmic cost $c(x,y) = -\log \|x - y\|$.*

Hypotheses 2.3 will lead to $C^{1,\alpha}$ solutions.

**Hypothesis 2.3** (Conditions on data (non-smooth))**.** *We require problem data to satisfy the following conditions:*

*(a) There exists some $m > 0$ such that $f_1(x) \geq m$ for all $x \in \mathbb{S}^n$.*

*(b) The mass balance condition holds, $\int_{\mathbb{S}^2} f_0(x)\,dx = \int_{\mathbb{S}^2} f_1(y)\,dy$.*

*(d) The data satisfies the regularity requirements $f_0 \in L^p(\mathbb{S}^n)$.*

(e) *The cost functions are either the squared geodesic cost* $c(x,y) = \frac{1}{2} d_{\mathbb{S}^2}(x,y)^2$ *or the logarithmic cost* $c(x,y) = -\log \|x - y\|$.

From these hypotheses, we get the following regularity result Loeper (2011)

**Theorem 2.4** (Regularity)**.** *The Optimal Transport problem in Equation* (2.14) *with data satisfying Hypothesis 2.2 has a classical solution* $u \in C^3(\mathbb{S}^2)$.

The following result, also from Loeper (2011) and following the reasoning in Appendix A.

**Theorem 2.5** (Regularity)**.** *The Optimal Transport problem in Equation* (2.14) *with data satisfying Hypothesis 2.3 has a viscosity solution* $u \in C^1(\mathbb{S}^2)$.

As a final note, the solution to Equation (2.14) is unique only up to additive constants. For uniqueness, in this manuscript, sometimes we will fix a point $x_0 \in M$ and add the additional constraint:

$$u(x_0) = 0. \tag{2.30}$$

At other times, it may be more convenient to choose the mean-zero solution, that is to impose the constraint

$$\int_M u = 0. \tag{2.31}$$

### 2.1.5 Beyond the Sphere

In some cases, manifolds $M$ with non-positive cost-sectional curvature at any point $x \in M$ can have positive measures $\mu_0, \mu_1 \in C^\infty(M)$, but $T$ is not even guaranteed to be continuous Loeper (2009). In order to explain this phenomenon, we introduce the Ma, Trudinger, Wang tensor Loeper (2009); Ma et al. (2005), which is the non-Euclidean generalization of Equation (2.12):

$$\mathfrak{G}_c(x_0, y_0)(\xi, \nu) = D^4_{p_\nu, p_\nu, x_\xi, x_\xi} \left[(x,p) \mapsto -c\left(x, T_{x_0}(p)\right)\right] |_{x_0, p_0 = -\nabla_x x(x_0, y_0)}. \tag{2.32}$$

The cost-sectional curvature is negative at a point $(x, y)$ if there exist $\xi, \nu$ such that $\mathfrak{G}_c(x, y)(\xi, \nu) < C_0 |\xi|^2 |\nu|^2$. The idea is perhaps more transparent if one looks at the squared geodesic cost $c(x, y) = \frac{1}{2} d_M(x, y)$. In this simple case, the non-positivity of the cost-sectional curvature is equivalent to the negativity of the sectional curvature of $M$ at any point $x \in M$ due to the following equality:

$$\frac{\mathfrak{G}_c(x, x)(\nu, \xi)}{|\xi|^2 |\nu|^2 - (\xi \cdot \nu)^2} = \frac{2}{3} \cdot \text{sectional curvature of } M \text{ at } x \text{ in the plane } (\xi, \nu). \quad (2.33)$$

But, even if the underlying manifold has strictly positive curvature everywhere, the Ma, Trudinger, Wang tensor $\mathfrak{G}_c(x, y)$ may be negative on the off-diagonal $(x, y)$ s.t $x \neq y$. This was shown true even for some ellipsoids of revolution in the paper Figalli et al. (2010). This suggests that numerically computing Optimal Transport for such cases is more challenging since one is not guaranteed a diffeomorphic mapping $T$.

## 2.2 Optimal Information Transport

The Monge problem of Optimal Transport, see Section 2.1.2, hints at a deeper relation between the space of probability distributions and the mappings between probability measures. We have seen in Section 2.1 that for the squared geodesic cost $c(x, y) = \frac{1}{2} d_M(x, y)$ there exists a unique mapping $T$ of the Monge problem which is then used in computing the total cost of transporting the probability measure $\mu_0$ to $\mu_1$. Furthermore, for the squared geodesic cost, this total cost defines a Riemannian distance between probability measures, see Villani (2003, 2009), usually referred to as the Wasserstein distance. An interpolation between the source mass $\mu_0$ and the target mass $\mu_1$, denoted by $\mu(t)$ can be uniquely defined from a convex combination of the identity map and the Optimal Transport map from $\mu_0$ to $\mu_1$. That is, by defining the path of maps $T(t) = (1 - t)\mathrm{Id} + tT$ we can construct an interpolation of measures $\mu(t) = T(t)_* \mu_0$, which defines a geodesic in

Wasserstein space, see Figure 2.3 and Villani (2003, 2009) for more detail on the Wasserstein distance and interpolation. This geodesic is minimizing with respect to the Wasserstein metric.

**Figure 2.3** An interpolation of probability measures (left to right) is achieved by traveling along a geodesic in the space of probability measures with respect to a metric. In the case of Optimal Transport, the metric is known as the Wasserstein metric developed by Otto, see the explanation in Villani (2003), and we will refer to the resulting interpolation as the Wasserstein interpolation. In the case of Optimal Information Transport, the metric used is the Fisher-Rao metric.

A different distance between probability measures arising primarily in statistical applications is the Fisher-Rao distance. This distance is defined as the Riemannian distance arising from the Fisher-Rao metric on the space of probability measures. The Fisher-Rao metric, in turn, is the second variation of the Kullback-Leibler divergence between the probability measures $\mu$ and $\nu$, that is the quantity

$$\mathrm{KL}(\mu, \nu) := \int_M \log\left(\frac{d\mu}{d\nu}\right) d\mu, \tag{2.34}$$

which measures the relative entropy between one probability measure $\mu$ and another $\nu$. The Fisher-Rao metric we discuss here is the infinite-dimensional version, first studied in Friedrich (1991). Typically in statistics it is given in the finite-dimensional setting where the probability measures can be parametrized by a finite parameter space which is a subset of $k$-dimensional Euclidean space $\mathbb{R}^k$.

Given the space of probability measures then equipped with the Fisher-Rao metric, one may ask if there exists a path of mapping $T(t)$ that pushes $\mu_0$ forward to the geodesic (with respect to the Fisher-Rao metric) connecting $\mu_0$ and $\mu_1$. The short answer is yes, see Bauer et al. (2015), just like in the way that this can be done for geodesics with respect to the Wasserstein metric. However, it turns out that if we are on compact manifolds $M$ and the source mass is the canonical volume

form, $\mu_0 = \text{vol}$, then the formula for the mapping $T$ ends up being quite simple, as proved in Bauer et al. (2015); Modin (2015). For this situation, a geodesic in the space of diffeomorphisms that pushes $\mu_0$ forward to $\mu_1$ is actually a horizontal geodesic and it descends (via an explicit projection map) onto a geodesic (with respect to the Fisher-Rao metric) on the space of smooth densities. Solving this problem for $T$ is known as Optimal Information Transport and was introduced in Bauer et al. (2015); Modin (2015).

For more information about Optimal Information Transport, see much more background and detail presented in Bauer et al. (2015); Modin (2015) and the highlighted resources therein. For all the discussion in this section, we will assume that the underlying manifold $M$ is compact and connected and is equipped with the standard volume form vol. We introduce the Fréchet Manifold of smooth volume forms over $M$ with total volume $\text{vol}(M)$:

$$\text{Dens}(M) = \left\{ \mu \in \Omega^n(M) : \int_M \mu = \text{vol}(M), \mu > 0 \right\}, \tag{2.35}$$

where the notation $\Omega^n(M)$ denotes the $n$-forms over the manifold $M$, using topology induced by the Sobolev seminorms. The infinite-dimensional Fisher-Rao metric

$$G^F_\mu(\alpha, \beta) = \frac{1}{4} \int_M \frac{d\alpha}{d\mu} \frac{d\beta}{d\mu} d\mu \tag{2.36}$$

yields geodesics that have explicit formulas! Furthermore, there is an explicit formula for the distance which is simply given by the integral

$$d_f(\mu_0, \mu_1) = \sqrt{\text{vol}(M)} \arccos \left( \frac{1}{\text{vol}(M)} \int_M \sqrt{\frac{\mu_0}{\text{vol}} \frac{\mu_1}{\text{vol}}} \text{vol} \right). \tag{2.37}$$

which is a surprising result that was shown in Modin (2015).

We introduce another manifold with rich structure. The set of diffeomorphisms on $M$ is denoted $\text{Diff}(M)$ with topology induced by the Sobolev seminorms. We have assumed $M$ compact, which makes the set $\text{Diff}(M)$ is a Fréchet Lie group

under the composition of maps. The Lie algebra of $\mathrm{Diff}(M)$ is given by the space $\mathfrak{X}(M)$ of smooth vector fields. We equip this manifold with the information metric on $\mathrm{Diff}(M)$, defined by

$$G^I_\varphi(U,V) = \int_M g(\Delta u, v)\mathrm{vol} + \lambda \sum_{i=1}^{k} \int_M g(u,\xi_i)\mathrm{vol} \int_M g(v,\xi_i)\mathrm{vol}, \tag{2.38}$$

where $\lambda > 0$, $u = U \circ \varphi^{-1}$, $v = V \circ \varphi^{-1}$, $g$ is the underlying metric of the manifold, $\Delta$ is the Laplace-de Rham operator on the space of vector fields, defined by $\Delta u = -\left(\delta du^\flat + d\delta u^\flat\right)^\sharp$, where $\sharp$ and $\flat$ are the usual musical isomorphisms of differential geometry, $d : \Omega^k(M) \to \Omega^{k+1}(M)$ is the exterior differential and $\delta : \Omega^k(M) \to \Omega^{k-1}(M)$ is the codifferential, and $\{\xi_i\}_i$ is an orthonormal basis of the harmonic vector fields on $M$, that is those vector fields $\xi$ for which $\Delta\xi = 0$. This metric, Equation (2.38), yields well-posed geodesics, see Bauer et al. (2015).

Now, the diffeomorphic maps define a group action on the space of densities, via the pullback $\varphi^*\mu = \nu$. The volume preserving maps (with respect to $\varphi$): define the isotropy group

$$\mathrm{Diff}_\mu(M) := \{\varphi \in \mathrm{Diff}(M); \varphi^*\mu = \mu\}\,. \tag{2.39}$$

This group action is transitive, that is, given $\mu$ and $\nu$, there exists a diffeomorphism $\varphi$ such that $\varphi^*\mu = \nu$, which was proved in Moser (1965). Now, fix $\mu$. By defining the projection map

$$\pi_\mu : \mathrm{Diff}(M) \ni \varphi \mapsto \varphi^*\mu \in \mathrm{Dens}(M) \tag{2.40}$$

we can define the following principal bundle structure:

$$\begin{array}{ccc} \mathrm{Diff}_\mu(M) & \hookrightarrow & \mathrm{Diff}(M) \\ & & \downarrow{\scriptstyle \pi_\mu} \\ & & \mathrm{Dens}(M). \end{array}$$

Now, choose $\mu = \mathrm{vol}$. We denote $\pi_{\mathrm{vol}} = \pi$. The goal is now to give a Riemannian structure to the principal fiber bundle using Equations (2.36) and (2.38).

The upshot is that although $\pi$ is a submersion it is actually also a *Riemannian* submersion with respect to $G^F$ and $G^I$, i.e. the metric $G^I_\varphi$ descends to $G^F$ as explained in Bauer et al. (2015). That is,

$$G^I_\varphi(U,V) = G^F_{\pi(\varphi)}\left(T_\varphi\pi \cdot U, T_\varphi\pi \cdot V\right). \tag{2.41}$$

If $\varphi(t)$ is a geodesic in $\mathrm{Diff}(M)$, then it is actually a horizontal geodesic in $\mathrm{Diff}(M)$ and furthermore $\mu(t) := \pi(\varphi(t))$ is a geodesic curve in $\mathrm{Dens}(M)$, see Bauer et al. (2015).

The problem of diffeomorphic density matching is that given a path of densities $\mu(t)$, we desire to find the path $\varphi(t)$ which project onto $\mu(t)$, that are also of minimal length with respect to $G^I$. That is, solve the exact density matching problem, that is find a $\varphi(t)$ such that

$$\begin{cases} \varphi(0) = \mathrm{id}, \\ \varphi^*\mu_0 = \mu(t), \\ \text{minimizing } \int_0^1 G^I_{\varphi(t)}\left(\dot\varphi(t), \dot\varphi(t)\right) dt. \end{cases} \tag{2.42}$$

We assume that $\mu_0 = \mathrm{vol}$. We take the equation $\varphi^*(t)\mu_0 = \varphi^*(t)\mathrm{vol} = \mu(t)$ and differentiate with respect to $t$:

$$\dot\mu(t) = \partial_t\left(\varphi(t)^*\mathrm{vol}\right) = \varphi^*\mathrm{div}_{\mathrm{vol}}v(t), \tag{2.43}$$

where $v(t) = \dot\varphi \circ \varphi^{-1}$. This can be rewritten, using the formalism of Lie derivatives (see Bauer et al. (2015)), as

$$\dot\mu(t) = \mathrm{div}\left(v(t)\right) \circ \varphi(t)\mu(t). \tag{2.44}$$

From here, we perform the Hodge-Helmholz decomposition for the vector field $v$, by writing $v = \mathrm{grad}f + w$. It turns out that the Hodge-Helmholtz decomposition is orthogonal with respect to the information metric $G^I$ so the length of

the path $\varphi(t)$ is minimal for $w = 0$, see Bauer et al. (2015). Therefore, in order to solve for the mapping, we can solve the following Poisson equation for the curl-free term $f$ as an intermediate step:

$$\begin{cases} \Delta f(t) = \frac{\dot{\mu}(t)}{\mu(t)} \circ \varphi(t)^{-1} \\ \dot{\varphi}(t) = \operatorname{grad}(f(t)) \circ \varphi(t), \quad \varphi(0) = \mathrm{id}. \end{cases} \tag{2.45}$$

As noted before, the geodesics for the Fisher-Rao metric have *explicit forms*! Thus, we can insert explicit forms for $\mu(t)$ and $\dot{\mu(t)}$ and thus this is a closed problem for $f$ and $\varphi(t)$. The explicit geodesics $\mu(t)$ are given as follows, see Bauer et al. (2015) for more detail. Define $W : \mathrm{Dens}(M) \to C^\infty(M)$ by $\mu \mapsto \sqrt{\frac{\mu}{\mathrm{vol}}}$, then the geodesics are given by

$$[0,1] \ni t \mapsto \left( \frac{\sin((1-t)\theta)}{\sin\theta} f_0 + \frac{\sin(t\theta)}{\sin\theta} f_1 \right)^2 \mathrm{vol}, \tag{2.46}$$

where

$$\theta = \arccos\left( \frac{\langle f_0, f_1 \rangle_{L^2}}{\mathrm{vol}(M)} \right), \tag{2.47}$$

and $f_i = W(\mu_i)$. The Optimal Information Transport problem is then given by

$$\begin{cases} \varphi(0) = \mathrm{id}, \\ \varphi_* \mu_0 = \mu(t), \end{cases} \tag{2.48}$$

where $\varphi_* \mu_0$ denotes the pushforward of $\mu_0$. The solution of this problem, Equation (2.48), is the diffeomorphic mapping $T$ given by $T = \varphi^{-1}(1)$, where $\varphi(1)$ solves Equation (2.45).

## 2.3 The Reflector Antenna Problem

Here we briefly summarize the derivation of the reflector antenna problem and its connection to Optimal Transport on the sphere, which leads to an equation

of Monge-Ampère type that can be solved using techniques from numerical PDEs and Optimal Transport. We begin by following the physical derivation in Wang (1996, 2004) and then show that merely through a change of variables, we can derive a particular instance of Equation (2.14).

We start with a light source or detector $\mu_0$ located at the origin, which is a probability measure indicating directional intensity and is supported on a set $\Omega \subset \mathbb{S}^2$. Next we consider a reflector surface $\Sigma$, which is a radial graph over the domain $\Omega$ and can be represented as

$$\Sigma = \{x\rho(x) \mid x \in \Omega, \quad \rho > 0\}, \tag{2.49}$$

where $\rho : \Omega \to \mathbb{R}$ is a non-negative function indicating the distance between the reflector surface and the origin. The light from the source $\mu_0$ in the direction $x$ bounces off the reflector $\Sigma$ without any refraction or absorption and travels in the direction $T$ following the law of reflection. Over all directions this produces the far-field intensity $\mu_1$, which is also a probability measure indicating directional intensity and is supported on some target domain $\Omega^* \subset \mathbb{S}^2$. See Figure 2.4 for a schematic of the setup.

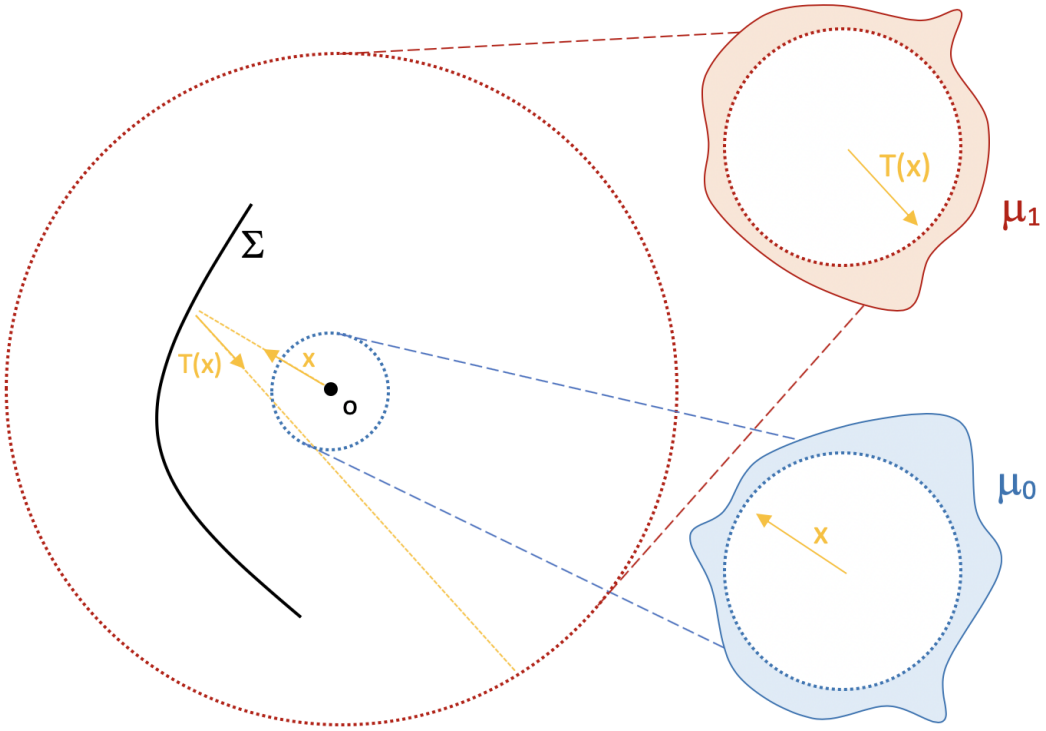



**Figure 2.4** Reflector antenna with source/detector $\mu_0$, reflector $\Sigma$ and target far-field intensity $\mu_1$. The directional vectors $x$ and $T(x)$ are unit vectors.

The reflector antenna problem is thus: given source and target intensity probability distributions $\mu_0$ and $\mu_1$, respectively, find the shape of the reflector $\Sigma$ that transmits the light from the source to the target while satisfying conservation of energy.

We make the assumption that the probability densities $\mu_0$ and $\mu_1$ have density functions $f_0$ and $f_1$ respectively (so that $d\mu_0(x) = f_0(x)dS(x), d\mu_1(y) = f_1(y)dS(y)$). Now we seek a PDE that will allow us to determine the reflector height function $\rho(x)$, which fully determines the reflector surface, in terms of the prescribed intensity functions $f_0$ and $f_1$.

The first of the two physical laws that will be used to derive the governing PDE for this setup is the well known geometric law of reflection, which yields the optical map

$$T(x) = x - 2\langle x, n(x)\rangle n(x), \tag{2.50}$$

where $n(x)$ is the outward normal to $\Sigma$ at the point $z = x\rho(x)$, $x \in \Omega$. See

Figure 2.5. We emphasize that this is the geometric optics limit and as such we ignore all quantum mechanical effects.

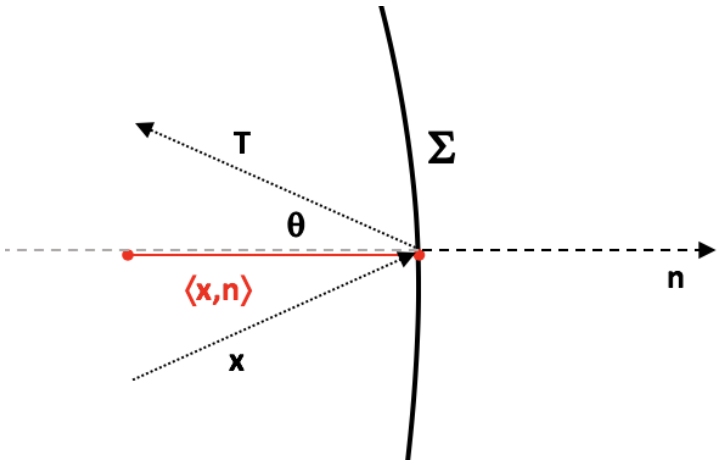



**Figure 2.5** Incident light direction $x$, reflector $\Sigma$, outward normal $n$, and outward light ray $T$.

The second physical law that completes the problem is the law of conservation of energy:

$$\int_{T^{-1}(E)} f_0(x)\,dx = \int_E f_1(y)\,dy, \tag{2.51}$$

for any Borel set $E \subset \Omega^*$.

By introducing local coordinates on the sphere, it was observed in Wang (1996) that the unit normal $n$ can be given by

$$n(x) = \frac{\nabla\rho(x) - x\rho(x)}{\sqrt{\rho(x)^2 + \|\nabla\rho(x)\|^2}}. \tag{2.52}$$

Then the law of reflection, Equation (2.50), yields the mapping

$$T(x) = \frac{2\rho(x)\nabla\rho(x) + \left(-\rho(x)^2 + \|\nabla\rho(x)\|^2\right)x}{\rho(x)^2 + \|\nabla\rho(x)\|^2}. \tag{2.53}$$

Applying the change of variables formula to the conservation of energy constraint, Equation (2.51), and combining these equations yields the PDE

$$\eta^{-2}\det\left(-\nabla_i\nabla_j\rho + 2\rho^{-1}\nabla_i\rho\nabla_j\rho + (\rho-\eta)\delta_{ij}\right) = f_0(x)/f_1(T(x)), \tag{2.54}$$

where $\eta = \left(|\nabla\rho|^2 + \rho^2\right)/2\rho$ and $\delta_{ij}$ is the usual Kronecker delta (in terms of the indices of the local coordinate system), see Wang (1996). We may begin to recognize this PDE as similar to an equation of Monge-Ampère type, with the usual second boundary value condition, see Urbas (1997) for more detail. The second boundary value condition is

$$T(\Omega) = \Omega^*. \tag{2.55}$$

In order to use recently improved regularity results, it is much better to perform the change of variables:

$$\rho = e^{-u}. \tag{2.56}$$

It was shown in Wang (2004) that under an equivalent change of variables (modulo a sign change), the function $u$ solves the dual formulation of the Optimal Transport problem with cost function $\tilde{c}(x, y) = -\log(1 - x \cdot y)$. This then allows one to check the MTW conditions, see Section 2.1.2 for the logarithmic cost function and achieve the desired regularity results, which is what was done in Loeper (2011).

### 2.3.1 Numerical Methods for the Reflector Antenna

Computational approaches to solving optical design problems can be roughly divided into three basic categories: (1) techniques that use a ray-mapping to design the optical surface, (2) methods that approximate the optical surfaces by supporting quadrics, and (3) methods that represent the optical surface through the solution to an Optimal Transportation problem.

The ray-mapping approach generally involves a two-step procedure. In the first step, a ray mapping is produced between the input and output intensities. In the second step, the laws of reflection and/or refraction are employed to construct a surface that achieves this ray mapping as nearly as possible. Several methods based on this general approach are available including Bruneton et al. (2011);

Desnijder et al. (2019); Feng et al. (2016); Fournier et al. (2010); Parkyn and Pelka (2006). A downside to this general approach is that it can be difficult to theoretically justify the existence of an optical surface that exactly produces the desired ray mapping.

Oliker's method of supporting quadrics involves representing the optical surface via supporting ellipsoids or hyperboloids Oliker (2006); Oliker et al. (2015). The simple optical properties of these quadrics is used to produce a pixelated version of the desired target. This approach has the advantage of being theoretically well-founded, but can be costly to implement in practice.

Many optical inverse problems have yielded fruitful interpretations via Optimal Transport by deriving an appropriate cost function $c(x, y)$ Yadav (2018). To give a simple example, a parallel-in, far-field out setup yields the cost function $c(x, y) = \frac{1}{2} \|x - y\|^2$, where $x, y \in \mathbb{R}^2$. The reflector antenna problem considered in this article has a slightly more challenging set-up in that the cost function $c(x, y) = -2 \log \|x - y\|$ is unbounded and the intensity functions $f_0, f_1$ are supported on $\mathbb{S}^2$ (the unit 2-sphere), as opposed to subsets of Euclidean space Gangbo and Oliker (2007); Oliker and Newman (1993); Wang (1996, 2004).

One approach to solving Optimal Transport problems in optical design is to use optimization techniques, including linear assignment Doskolovich et al. (2019) and linear programming Glimm and Oliker (2003). This approach has the advantage of being theoretically well-understood. However, the optimization problems typically involve a very large number of constraints and the resulting methods are computationally complex.

Recently, several methods have been proposed for solving optical design problems involving a point source via the solution of a Monge-Ampère type equation. These methods replace the PDE on the sphere with a corresponding equation on the plane by representing subsets of the unit sphere using spherical coordinates Wu et al. (2013), a vertical projection of coordinates onto the plane Brix et al. (2015), or stereographic projection Romijn et al. (2020). As the numeri-

cal solution of these Monge-Ampère type equations is a very new field, many of the numerical methods used in optical design problems are not yet equipped with theoretical guarantees of convergence.

## 2.4 Moving Mesh Methods

Here we outline the moving mesh problem, a type of adaptive mesh method that has found special usage in computational PDE techniques. The meshes consist of a fixed number of nodes and an edges connecting the nodes which does not change. The idea is to transform an automatically generated mesh (or any given mesh) into a mesh with prescribed node density. For computational reasons, a core requirement is that the target mesh should not be tangled. We desire to complete this adaptive mesh problem with one step and without requiring post-processing.

The resulting adapted mesh is often used in high-resolution PDE computations, such as the Eady problem, which is a 2D vertical cross-section of an incompressible Bousinnesq fluid, which was treated in Budd et al. (2013). Using a fixed number of grid points leads to simpler data structures (than some other adaptive mesh techniques) in the PDE solving step, as was stressed in the manuscripts Budd et al. (2013); Chacón et al. (2011); Weller et al. (2016b). Beyond such computational fluid mechanics applications, the related problem of diffeomorphic density matching has found widespread applications in medical image registration, see, for example, Chen and Öktem Ozan (2017); Gorbunova et al. (2012); Haker et al. (2004); Rottman et al. (2015) and random sampling from a probability measure see, for example, Bauer et al. (2017); Moselhy and Marzouk (2012) among other applications. These applications are necessarily more challenging from a theoretical perspective as well as a computational perspective when the geometry is non-Euclidean. In this dissertation, we focus on applications of moving mesh methods on the sphere with extensions to compact 2D surfaces via diffeomorphic density matching.

Technically speaking, the moving mesh methods presented here can be con-

sidered as an application of the diffeomorphic density matching problem. This problem a long history in the imaging sciences. Classical methods consist of positive scalar image functions composed from the right by transformations Younes (2010). That is, given a source probability density function $f_0$ and a target probability density function $f_1$, classical techniques compute $f_0 \circ T = f_1$. It is not possible in many cases to find diffeomorphic mappings, see Villani (2009). Furthermore, this formulation is not appropriate for moving mesh methods. Non-classical methods, like Optimal Transport and Optimal Information Transport, allow the transformation to act as a pushforward or pullback on the density, that is $f_0 = |DT|\, f_1 \circ T$ or $|DT|\, f_0 \circ T = f_1$, respectively. This generalization has particular benefit in that it allows for proving the existence of a diffeomorphic mapping over a much wider range of densities $f_0$ and $f_1$. Furthermore, it is clear that this is the formulation of diffeomorphic density matching that is appropriate for moving mesh methods, since it allows one to change the local density of mesh points.

In the Euclidean setup, the moving mesh problem setup requires the "physical" target domain $\Omega_p \subset \mathbb{R}^d$, where the PDE is posed, while the "computational" input domain $\Omega_c \subset \mathbb{R}^d$ is usually chosen to be a uniform rectangular grid (in Euclidean space), see Figure 2.6.

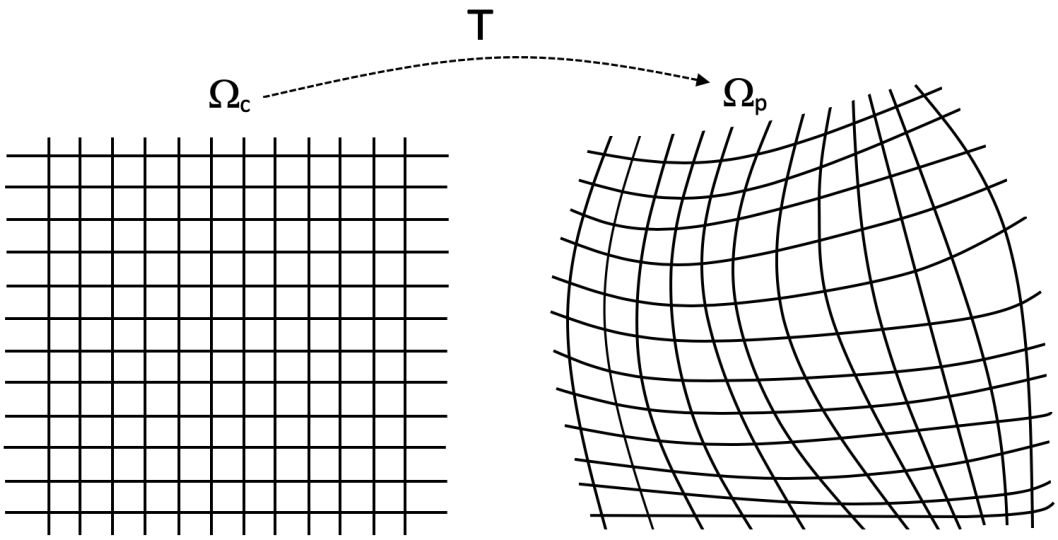



**Figure 2.6** Mapping $T$ from the computation domain $\Omega_c$ to the physical domain $\Omega_p$.

In non-Euclidean geometries, one would like to use any simple off-the-shelf mesh generator for the computational mesh. Then, one would like to find a diffeomorphic mapping $T : \Omega_c \to \Omega_p$. Typically, in applications, the local density of grid points in the physical domain $\Omega_p$ is determined by, for example, the time scales involved in the solution of a fluid mechanics PDE. For example, in an evolving front or shock, it is desirable for the density of points in $\Omega_p$ to be greater than in the relatively unchanging parts of the solution, as in meteorological applications where it is desirable to have high resolution in areas of high precipitation Weller et al. (2016b). This process could be done iteratively to get a very accurate solution of the PDE, by solving the shock on more and more resolved grids as one iterates. The way this is done is by feeding the information from the PDE (desired density) into a scalar monitor function $\mathcal{M}(y,t) > 0$ and then solving the change of variables formula:

$$\mathcal{M}(y,t)J(T(x)) = \theta(t), \tag{2.57}$$

where $J$ is the Jacobian of the diffeomorphic mapping $T$ and $\int_{\Omega_p} \mathcal{M}(y,t)dy = \theta(t)$. Moving mesh methods require that the mapping be a diffeomorphism, which means

that the Jacobian of the mapping $T$ satisfies $0 < J(T) < \infty$ in the strong sense. This condition will prevent the mesh from tangling. An example of a monitor function constructed from information from a function $f(y,t)$ is the scaled arc-length function:

$$\mathcal{M}(y,t) = \sqrt{1 + |S\nabla_y f(y,t)|^2}, \tag{2.58}$$

where $S$ is a normalization factor, see Budd et al. (2013).

The diffeomorphic density matching problem can be solved via the Jacobian equation, which is the change of variables. The existence of a change of variables between densities is actually quite general, see Villani (2009).

**Theorem 2.6** (Jacobian)**.** *Let $M$ be an $n$-dimensional Riemannian manifold. Let $\mu_0$ and $\mu_1$ be two probability measures on $M$, and let $T : M \to M$ be a measurable function such that $T_{\#}\mu_0 = \mu_1$. Let $\nu$ be a reference measure, of the form $\nu(dx) = e^{-V(x)} vol(dx)$, where $V$ is continuous and vol is the volume measure on $M$. Further assume that $\mu_0(dx) = f_0(x)\nu(dx)$ and $\mu_1(dy) = f_1(y)\nu(dy)$, $T$ is injective, and the distributional derivative of the mapping $DT$ is a locally integrable function (this can be relaxed slightly). Then, $\mu_0$-almost surely,*

$$f_0(x) = f_1(T(x))J(T(x)), \tag{2.59}$$

*where $J$ is the Jacobian determinant of $T$ at $x$, defined weakly by:*

$$J(T(x)) := \lim_{\epsilon \to 0} \frac{\nu\left[T(B_\epsilon(x))\right]}{\nu[B_\epsilon(x)]}. \tag{2.60}$$

While Theorem 2.6 stipulates the conditions for which we have a distributional solution of the monitor equation, Equation (2.57), it does not state that such a mapping $T$ is unique or when it is diffeomorphic. One particular unique choice of the map is furnished by the Monge problem of Optimal Transport, which was introduced in Section 2.1.2. The approach of Optimal Information Transport, introduced in Section 2.2 and in the works Bauer et al. (2015); Modin (2015), is to

work directly in the space of diffeomorphisms and thus guarantee a diffeomorphic mapping at the expense of limiting oneself to smooth density functions strictly bounded away from zero.

What we will find is that under fairly general assumptions on the source and target masses, for subsets of Euclidean space and the $n$-sphere we can find diffeomorphic mappings using the techniques of Optimal Transport and Optimal Information Transport, which both satisfy the Jacobian equation and give us diffeomorphic mappings, appropriate for our moving mesh methods, the regularity results are contained in Loeper (2011) for the Optimal Transport problem and in Bauer et al. (2015) for the Optimal Information Transport problem.

The mapping arising from Optimal Transport is elected by further requiring that the map $T$ minimize the following integral:

$$I = \int_{\Omega_c} |T(x) - x|^2 \, d\mu_0(x). \tag{2.61}$$

If the underlying manifold is $\mathbb{R}^d$, it can be shown that such a mapping is the gradient of a convex function $u$, that is: $T(x) = \nabla_x u(x)$, which makes it irrotational (meaning $\nabla \times T = 0$ Budd and Williams (2009)). More specifically, the regularity theory of Caffarelli (see the summary in Villani (2003)) shows us that if $f_0, f_1 \in C^{0,\alpha}(\mathbb{R}^d)$ for $0 < \alpha < 1$ and $0 < f_0, f_1 < \infty$, then we are guaranteed that the mapping is, in fact, a diffeomorphism.

In the more general manifold setting, we have that the mapping solving the Optimal Transport problem with squared geodesic cost is given by $T(x) = \exp_x(\nabla u)$ (as long as the manifold is geodesically complete), see McCann (2001), where $u$ is as a $c$-convex function, see Definition 2.1. For the $n$-sphere, fairly general conditions on the source and target masses are given in Loeper (2011), in which it is shown that under those assumptions (see Section **??**), mass transports to a distance bounded strictly below the injectivity radius and gives us differentiability. Like the case of $\mathbb{R}^d$, we also need $0 < f_0, f_1 < \infty$. More detail on the regularity results of the Optimal Transport problem will be shown in Section 2.1.

The more general manifold case fails quite spectacularly in terms of differentiability, see the result in Figalli et al. (2010) for a simple example where differentiability fails to hold.

Optimal Information Transport provides another option for computing a mapping for moving mesh methods as a change of variables between two probability measures. We provided much more detail on the Optimal Information Transport problem in Section 2.2, but we summarize the relevant details here. The regularity results of Optimal Information Transport is are clearly geometrically more general than the corresponding result for Optimal Transport, since they only depend on the manifold being compact. In Optimal Transport case, in the space of probability measures endowed with the Wasserstein metric between smooth source and target probability measures bounded away from zero, there exists a unique path $T_t : [0, 1] \to \text{Diff}$, where $\text{Diff}(M)$ in the cases $M = \mathbb{S}^d$ and $M = \mathbb{R}^d$. For more general compact manifolds, see Figalli et al. (2010), this does not hold. For the moving mesh problem, then, we expect both Optimal Transport and Optimal Information Transport to be versatile and useful in $\mathbb{R}^d$ and $\mathbb{S}^d$. For more general manifolds $M$, thus, Optimal Information Transport is preferable.

## 2.5 Numerical Analysis and Challenges for Optimal Transport

As far as numerical Optimal Transport is concerned, there exist many fruitful avenues of research which address the issues of convergence proofs and rates, efficiency and stability of solvers, and scalability with dimension. As noted earlier, the difficulty in proving convergence for numerical solutions to the Optimal Transport problem can come from many sources like nonlocal boundary conditions Hamfeldt (2019) and a lack of a comparison principle for the PDE (in the case of the sphere). The underlying geometry may produce discontinuous mapping $T$ for even $C^\infty$ data, see Figalli et al. (2010). The cost function $c$ itself can lead to some problems, such as for $c(x, y) = d(x, y)$ which does not even yield a unique solution Santambrogio (2015); Villani (2003). The point is that there is a menagerie

of problems with clear applications that can be targeted for exploration.

There are many different numerical approaches to the Optimal Transport problem, many of which are the best choice for particular cases. The methods we present here are usually used in Euclidean space for Monge-Ampère-type equations or for computing the Wasserstein distance, the mapping $T$ or the potential function $u$. Most of the competing schemes outlined below were developed for the quadratic cost. The approach in this dissertation is to establish a convergence theorem based on the theory of viscosity solutions and then construct finite-difference methods that satisfy the hypotheses of the convergence theorem. The crux of the type of construction we pursue is usually in the explicit construction of monotone schemes.

There is a long avenue of approach for constructing monotone schemes in Euclidean space and was performed in the papers Benamou et al. (2016); Benamou and Duval (2017); Benamou et al. (2014); Bonnet and Mirebeau (2021); Chen et al. (2018); Froese (2012, 2018); Froese and Oberman (2011a,b, 2013); Hamfeldt and Salvador (2018); Hamfeldt (2019, 2018); Hamfeldt and Lesniewski (2022a,b); Liu et al. (2017); Oberman (2006, 2008). The authors Feng et al. (2013b); Feng and Lewis (2014a) introduced the notion of generalized monotonicity (referred to as $g$-monotonicity in the papers) in for Galerkin methods as well as finite-difference schemes in 1D and extended the results for Galerkin methods to higher dimensions $d \geq 2$ in Feng and Lewis (2014b, 2018). One of the original numerical schemes proposed for solving the Monge-Ampère equation used the notion of generalized solutions of the Monge-Ampère equation to build piecewise convex solutions, see Oliker and Prussner (1988). A discretization based on taking the logarithm or the $n$th root of the Monge-Ampère equation and then solving via standard optimization techniques was the intriguing idea developed in Lindsey and Rubinstein (2017) A linear programming solution of the discrete Kantorovich formulation in Section 2.1.1 that uses a multigrid approach to reduce the computational complexity was pursued in Oberman and Ruan (2020). Linear programming was also used for non-quadratic squared cost in Schmitzer (2016). Schemes for the

semi-discrete case were proposed in Lévy (2015). Semi-Lagrangian schemes were used in Feng and Jensen (2017) and explained more in the review paper Feng et al. (2013a). Galerkin-type and finite-element schemes were investigated extensively by Neilan and others, see Feng et al. (2013a); Feng and Neilan (2007); Neilan (2010, 2014). Inspired by the original notion of viscosity solution, the vanishing viscosity method was developed, see Feng et al. (2013a); Feng and Neilan (2009a,b). It should be noted that the vanishing viscosity method is not a discretization in of itself, but could adapt to use any discretization, be it achieve by finite differences, finite elements, etc. It must be clearly noted that the vanishing viscosity method fortuitously captured the convexity of the problem in Euclidean space. A fixed-point method with finite differences was used in Benamou et al. (2010). A finite-element scheme was produced for the $W_1$ distance after simplification via a Hodge-decomposition and a spectral representation in Solomon et al. (2014) and a primal-dual algorithm was proposed for a regularization version of the $W_1$ distance in Li et al. (2016). Discretization via discrete entropic regularization is an extremely popular approach popularized by the paper Cuturi (2013), which takes advantage of the efficient Sinkhorn algorithm. Augmented Lagrangian methods were used in Dean and Glowinski (2006a,b) and a least-squares and operator splitting methods in Dean and Glowinski (2005, 2006b, 2008); Glowinski et al. (2008); Prins et al. (2015) and a least squares method for the logarithmic cost in Yadav et al. (2019). Some techniques for solving the Optimal Transport problem also are derived from the Benamou-Brenier formulation of Optimal Transport, see the original paper Benamou and Brenier (2000) as well as Benamou and Carlier (2015), for example. We will review existing methods for the reflector antenna problem in Chapter 5 when we compare which methods have convergence proofs and how efficient the discretizations are.

Recently, some progress has been made in the solution of the Optimal Transport problem on the sphere. The work of Weller et al. (2016a) used a geometric interpretation of a Monge-Ampère type equation on the sphere to produce the first

such method, which applies to the squared geodesic cost. A finite element solution of this Monge-Ampère type equation was produced in McRae et al. (2018). For problems posed on a subset of the sphere, the stereographic projection can be used to reframe the problem as an Optimal Transport problem on the plane (with non-quadratic cost); this was the approach of Romijn et al. (2020). For a particular logarithmic cost function, the semi-discrete Optimal Transportation problem on the sphere admits a particularly nice interpretation in terms of generalized (spherical) power diagrams. The work of Cui et al. (2019) recently exploited this interpretation to develop a fast, convergent method using techniques from computational geometry. However, these methods lack convergence guarantees and are limited to specific cost functions.

### 2.5.1 The Convergence Framework of Barles-Souganidis

A powerful contribution to the numerical approximation of elliptic (and parabolic as well) equations was provided by the Barles-Souganidis framework, which states that the solution to a scheme that is consistent, monotone, and $L^\infty$-stable will converge to the viscosity solution, provided the underlying PDE satisfies a comparison principle Barles and Souganidis (1991). The original paper demonstrates the convergence framework posed on an open set $\Omega \subset \mathbb{R}^n$. We first start with the definition of lower and upper semicontinuous envelopes.

**Definition 2.7** (Semi-continuous envelopes)**.** *The* upper and lower semicontinuous envelopes *of a function $u$ are given by*

$$u^*(x) = \limsup_{y \to x} u(y), \quad u_*(x) = \liminf_{y \to x} u(y).$$

Now we can introduce the definition of viscosity solutions, see also Figure 2.7. Consider the PDE:

$$F(x, \nabla\phi(x), D^2\phi(x)) = 0, \ x \in \Omega. \tag{2.62}$$

**Definition 2.8** (Viscosity Solutions)**.** *An upper semi-continuous function $u : \bar{\Omega} \to$*

$\mathbb{R}$ *is a viscosity subsolution (resp. supersolution) of Equation* (7.1)*, if for all* $\phi \in C^2(\bar{\Omega})$ *and all* $x \in \bar{\Omega}$ *such that* $u^* - \phi$ *(resp.* $u_* - \phi$*) has a local maximum (resp. minimum) at* $x$*, we have* $F_*\left(x, u^*(x), D\phi(x), D^2\phi(x)\right) \leq 0$ *(resp.* $\geq 0$*). The function* $u$ *is a viscosity solution if it is a subsolution and a supersolution.*

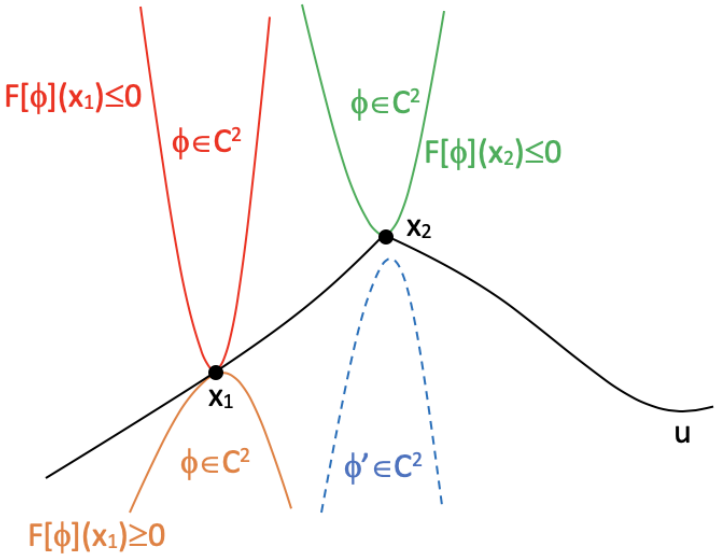



**Figure 2.7** The definition of a viscosity solution $u$ for an elliptic PDE $F$ is perhaps more obvious at points of second differentiability $x_1$ where the $C^2$ test functions above and below form subsolutions and supersolutions, respectively. For the point of non-differentiability $x_2$, the test function defines a subsolution as a bounding paraboloid. However, the test function below $\phi'$ satisfies the supersolution condition vacuously.

We consider finite difference schemes that have the form

$$F^h\left(x, u(x), u(x) - u(\cdot)\right) = 0 \quad x \in \mathcal{G}^h, \tag{2.63}$$

and

$$h = \sup_{x \in \Omega} \min_{y \in \mathcal{G}^h} \|x - y\| \tag{2.64}$$

denotes the grid resolution.

In this setting, the properties required by the Barles-Souganidis framework can be defined as follows.

**Definition 2.9** (Consistency)**.** *The scheme, Equation* (2.63)*, is* consistent *with*

*Equation* (7.1) *if for any smooth function* $\phi$ *and* $x \in \bar{\Omega}$,

$$\limsup_{h\to 0, y\to x, z\in\mathcal{G}^h\to x,\xi\to 0} F^h(z,\phi(y)+\xi,\phi(y)-\phi(\cdot)) \leq F^*(x,\phi(x),\nabla\phi(x),D^2\phi(x)),$$

$$\liminf_{h\to 0, y\to x, z\in\mathcal{G}^h\to x,\xi\to 0} F^h(z,\phi(y)+\xi,\phi(y)-\phi(\cdot)) \geq F_*(x,\phi(x),\nabla\phi(x),D^2\phi(x)).$$

**Definition 2.10** (Monotonicity)**.** *The scheme, Equation* (2.63)*, is* monotone *if* $F^h$ *is a non-decreasing function of its final two arguments.*

**Definition 2.11** (Stability)**.** *The scheme, Equation* (2.63)*, is* stable *if there exists* $M \in \mathbb{R}$ *(independent of* $h$*) such that whenever* $u^h$ *is a solution of Equation* (2.63) *then* $\|u^h\|_\infty \leq M$*.*

Also, we have the comparison principle, which is a very important property that many elliptic PDE have.

**Definition 2.12** (Strong Comparison Principle)**.** *If* $u$ *is an upper semi-continuous solution of Equation* (7.1) *and* $v$ *is a lower semicontinuous solution of Equation* (7.1)*, then* $u \leq v$ *on* $\bar{\Omega}$*.*

Then, we have the strong result that the numerical solution of the discretization converges uniformly to the *a priori* continuous viscosity solution of the underlying PDE, see Barles and Souganidis (1991).

**Theorem 2.13** (Barles-Souganidis)**.** *Assume that the PDE operator* $F$ *is continuous in all its variables. Let the discrete operator* $F^h$ *be consistent, monotone, and* $L^\infty$ *stable. Furthermore, let the PDE operator* $F$ *satisfy the strong comparison principle. Then, we have* $u^h \to u$ *uniformly where* $u$ *is the unique continuous viscosity solution of Equation* (7.1)*.*

The definition of monotonicity presented in Barles and Souganidis (1991) is very important in establishing many convergence results, but monotonicity is equivalent to the much simpler definition of degenerate ellipticity (of a scheme) when using finite-difference schemes. For this definition, we will also now be working on a 2D manifold $M$.

We begin with an unstructured grid $\mathcal{G}$ consisting of $N$ points $x_i \in M, i, \ldots, N$. The discretization operator $S$ is indexed by $i$ where $F_i^h[\phi]$ is used to perform a computation at each point $x_i$ on the function $\phi : M \to \mathbb{R}$. In order to perform a computation at the point $x_i$, there is associated a list of "neighboring" points $N(i)$ used in the computation. Assume that the discrete scheme at each point can be written as

$$F_i^h[\phi] := F_i^h\left(\phi_i, \phi_i - \phi_j|_{j=N(i)}\right), \tag{2.65}$$

then we have the following definition from Oberman (2006).

**Definition 2.14.** *The scheme $F_i^h$ is degenerate elliptic if for each index $i$, we have $F_i^h$ is nondecreasing in each variable.*

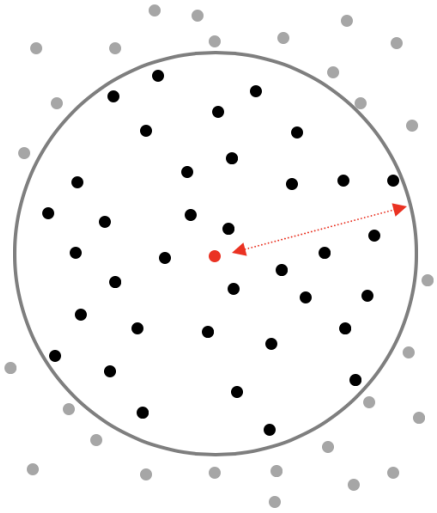
Then, the result from Oberman (2006) is that $F_i^h$ is degenerate elliptic if and only if $F_i^h$ is monotone.

## 2.6 Wide-Stencil Schemes in $\mathbb{R}^2$

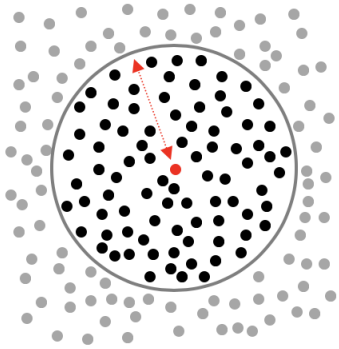
One of the biggest challenges in setting up finite difference schemes for fully nonlinear elliptic PDE is satisfying the monotonicity property, see Definition 2.10. Even for some linear elliptic equations, it is not possible to build a consistent, monotone scheme on a finite stencil Kocan (1995). To resolve this issue, wide-stencil schemes have been introduced for a range of fully nonlinear elliptic PDE. To achieve both consistency and monotonicity, these schemes require the width of finite difference stencils to become unbounded as the grid is refined. A variety of monotone schemes now exist for the Monge-Ampère equation Benamou et al. (2016); Benamou and Duval (2017); Finlay and Oberman (2019); Froese and Oberman (2011a); Oberman (2008), including schemes that can be posed on very general grids Froese (2018); Hamfeldt and Salvador (2018); Nochetto et al. (2018).

The essential idea is that as a discretization parameter $h$ decreases, the number of points used in the neighborhood increases, but the radius of the neigh-

borhood decreases. The grid points must resolve all directions as $h \to 0$. An example, from Froese (2018) and also presented in Chapter 4, is that the radius of the computational neighborhoods $r$ decreases not as $r = \mathcal{O}(h)$, but rather as $r = \mathcal{O}(\sqrt{h})$. This means then that the number of points in the computational neighborhoods increases as $h \to 0$. The idea is that all directions are then resolved as $h \to 0$, fixing the issue noted in Kocan (1995), provided that the grid satisfy some regularity requirements. In Hamfeldt and Turnquist (2022), we propose the construction of wider stencils in post-processing to ensure the convergence of numerical gradients. See Figure 2.8 for a pictorial representation of the wide stencil idea of decreasing the radius of computational neighborhoods while increasing the number of points.



**(a)** Larger discretization parameter $h$.



**(b)** Smaller discretization parameter $h$.

**Figure 2.8** Wide-stencil schemes utilize computational neighborhoods, here denoted by the grey circles. Going from Figure 2.8a to Figure 2.8b, the discretization parameter decreases (i.e. the minimum spacing $h$ between points decreases), but the number of points located in the computational neighborhoods about the red points increases.

## 2.7 The Effect of Non-Euclidean Geometry

What is the effect of non-Euclidean geometry on the solution of our Optimal Transport and Optimal Information Transport problem and for the design of numerical schemes? Here we do a quick rundown of the effects that concern us. The

effect of the non-Euclidean geometry will become more obvious as the schemes and convergence proofs are constructed in the sequel. However, we can highlight some important effects the geometry has on the PDE level and on the level of discretization:

- The geometry can direct affect the regularity of the solution of Monge-Ampère, allowing for the existence of smooth source and target masses, but a mapping $T$ which is not $C^1(M)$, see Figalli et al. (2010).

- Compactness is important for the simplicity of the moving-mesh method, since otherwise the density of mesh nodes would be required to decay at infinity, see Bauer et al. (2015).

- The regularity theory does not restrict one to the dimension $d = 2$, but this makes it simpler to build efficient schemes which use the tangent plane construction.

- Alternative methods (beyond finite-difference schemes) will be necessary to construct efficient discretizations in high dimensions to overcome the curse of dimensionality.

- Lack of a boundary on the manifold $M$ requires the selection of a particular solution and usually leads to slower convergence rates, see Hamfeldt and Turnquist (2022).

- The manifolds we are dealing with are geodesically complete, and therefore geodesics exist connecting any points $x, y \in M$. However, the explicit formulas for such geodesics is hard to generalize beyond the case of the sphere and other simple cases, see Lee (2006).

- The (negative and positive) curvature bounds of the manifold $M$ are the backbone upon which $C^2$ regularity results can be built, see Loeper (2009).

- For our Monge-Ampère equation, Equation (2.14), the geometry results in the PDE having nonlinear first-order derivative terms that are mixed in with second-order derivative terms (unlike the Optimal Transport PDE commonly derived in Euclidean space for the squared cost), see Loeper (2009). Our more general case, in practice, thus makes discrete solvers much more unstable.

# CHAPTER 3

# CONVERGENCE FRAMEWORK

Here, we propose a general convergence framework for the Optimal Transport problem on the sphere, the majority of which is contained in the publication Hamfeldt and Turnquist (2021a), but in this chapter we also include some parts from the publication Hamfeldt and Turnquist (2021b) which are, perhaps, more appropriately placed in this chapter. The key point we wish to stress is that this convergence framework is very easily adapted to PDE beyond the Optimal Transport PDE considered here and to other cost functions, which satisfy the conditions explicitly stated in Theorem 4.1 of Loeper (2011). While not stated explicitly, in our later construction of the scheme, see Chapter 4, we will discretize using finite-difference schemes. The convergence theorem in this chapter will require that the discretization scheme satisfy the key properties of consistency and monotonicity. It is the latter which is perhaps simpler to construct using finite-difference schemes versus other discretizations.

Our convergence framework is inspired by the Barles-Souganidis framework introduced in Section 2.5.1, but requires considerable consideration of the spherical geometry and dealing with the fact that there is no comparison principle for this PDE. Furthermore, our convergence result will apply to discretizations on very general meshes and point clouds on the sphere, which need only satisfy very mild regularity conditions. The convergence framework applies very generally also to any consistent, monotone approximation schemes, so it generalizes beyond the case of Optimal Transport. Although we state here that we will be addressing two specific cost functions: the squared geodesic cost function and the logarithmic cost function, the convergence framework proposed here is not specific to any cost function. We introduce appropriate local coordinates and therefore solve the PDE on local tangent planes, which allows for the use of a wide range of monotone approximation schemes for PDE in $\mathbb{R}^2$. In addition, we introduce Lipschitz control

on the PDE, which introduces sufficient stability to guarantee convergence in the absence of a comparison principle for the PDE.

## 3.1 Background

### 3.1.1 Optimal Transport on the Sphere

We consider the Optimal Transport problem in Equation (2.14) under Hypotheses 2.2 and 2.3. To reiterate, we are interested in two different cost functions $c(x, y)$: the "squared geodesic cost" on the sphere,

$$c(x,y) = \frac{1}{2} d_{\mathbb{S}^2}(x,y)^2 = \frac{1}{2}\left(2\sin^{-1}\left(\frac{\|x-y\|}{2}\right)\right)^2, \tag{3.1}$$

and the "logarithmic cost" arising in the reflector antenna problem,

$$c(x,y) = -\log\|x-y\|. \tag{3.2}$$

While this problem can be interpreted classically under fairly general assumptions, for very general density functions ($f_0, f_1 \in L^p(\mathbb{S}^2)$) or for more general manifolds (including smooth compact manifolds such as certain ellipsoids Figalli et al. (2010) even with $f_0, f_1 \in C^\infty(\mathbb{S}^2)$), $C^1$ solutions $u$ need not exist. Moreover, the type of convergence analysis frequently used for classical solutions of linear equations is not easily adapted to constrained fully nonlinear equations. For these reasons, it is also advantageous to be able to interpret Equation (2.14) in a weak (viscosity) sense.

Classically speaking, however, we denote the elliptic PDE by the second-order elliptic operator $F : \mathbb{S}^2 \times \mathbb{R} \times \mathcal{T}_x \times \mathcal{T}_x \otimes \mathcal{T}_x \to \mathbb{R}$ acting on a $C^2$ function $u$ at the point $x$ as:

$$F\left(x, u(x), \nabla_{\mathbb{S}^2} u(x), D^2_{\mathbb{S}^2} u(x)\right) = 0. \tag{3.3}$$

Let $F$ now denote the PDE operator arising from Equation (2.14). In order to

introduce our notion of viscosity solutions for Equation (2.14), we introduce the notation $\mathcal{E}(F)$ to denote the space of functions on which the PDE operator $F$ is elliptic. Recall the concepts of upper and lower envelopes of a function presented in Section 2.5.1.

**Definition 3.1** (Viscosity Solutions)**.** *An upper (lower) semicontinuous function* $u : \mathbb{S}^2 \to \mathbb{R}$ *is a* viscosity sub (super)-solution *of Equation* (2.14) *if for every* $x_0 \in \mathbb{S}^2$ *and* $\phi \in C^\infty(\mathbb{S}^2) \cap \mathcal{E}(F)$ *such that* $u - \phi$ *has a local maximum (minimum) at* $x_0$ *we have*

$$F_*^{(*)}(x_0, \phi(x_0), \nabla_{\mathbb{S}^2}\phi(x_0), D^2_{\mathbb{S}^2}\phi(x_0)) \leq (\geq) 0.$$

*A continuous function* $u : \Omega \to \mathbb{R}$ *is a* viscosity solution *of Equation* (2.14) *if it is both a sub-solution and a super-solution.*

### 3.1.2 Numerical Methods for Fully Nonlinear Elliptic Equations

The convergence framework of Barles and Souganidis, introduced in Section 2.5.1, does not apply to all elliptic PDEs, including Equation (2.14), which does not have the required comparison principle. Nevertheless, it provides an important starting point for the development of convergent numerical methods. In particular, monotone schemes possess a weak form of a discrete comparison principle even if the limiting PDE does not (Hamfeldt, 2018, Lemma 5.4). If the scheme additionally exhibits an increasing dependence on the function $u$ itself, we obtain a traditional strong form of the discrete comparison principle that guarantees solution uniqueness. The discrete operator $F^h$ being an increasing dependence on $u$ is known as being proper.

**Definition 3.2** (Proper)**.** *The scheme, Equation* (2.63)*, is* proper *if* $F^h$ *is an increasing function of its second argument.*

**Lemma 3.3** (Discrete comparison principle (Oberman, 2006, Theorem 5))**.** *Let* $F^h$ *be a monotone, proper scheme and* $F^h(x, u(x), u(x) - u(\cdot)) \leq F^h(x, v(x), v(x) - v(\cdot))$ *for every* $x \in \mathcal{G}^h$*. Then* $u(x) \leq v(x)$ *for every* $x \in \mathcal{G}^h$*.*

Another property that has recently proved important in establishing convergence of some numerical methods for the Monge-Ampère equation is the concept of underestimation Benamou and Duval (2017); Hamfeldt (2019); Lindsey and Rubinstein (2017). This concept will be important for our efforts to extend our convergence framework to the non-smooth setting.

**Definition 3.4** (Underestimation)**.** *The scheme, Equation* (2.63)*,* underestimates *Equation* (7.1) *if*

$$F^h(x, u(x), u(x) - u(\cdot)) \leq 0$$

*for every (possibly non-smooth) solution* $u$ *of Equation* (7.1)*.*

## 3.2 PDE on the Sphere

We begin by introducing an appropriate characterization of Equation (2.14) on the sphere, which will show how the numerical computations can be performed in local tangent planes. Wide-stencil schemes, see Section 2.6, will be built thus in the tangent planes. We also introduce a modification of the PDE that will allow us to build $c$-convexity and additional Lipschitz stability into our numerical framework.

### 3.2.1 Interpretation of the PDE

Solving Equation (2.14) is unique up to an arbitrary constant. For this reason, we also require that the solution $u$ satisfy

$$\langle u \rangle \equiv \frac{\int_{\mathbb{S}^2} u dV}{\int_{\mathbb{S}^2} dV} = 0. \tag{3.4}$$

With both cost functions, the gradient (an object in the tangent plane) appears in the mapping $T$. Letting $g$ be the standard round metric on the sphere, then the gradient is given by $\nabla u(x) = g^{ij}\partial_i u \partial_j$, where $\partial_j \in \mathcal{T}_x$ and $g^{ij}$ is the inverse of the round metric tensor expressed in local coordinates. The mapping $T$ then can be computed directly by solving Equation (2.15).

For the squared geodesic cost, the optimal mapping $T(x, p)$ has a very simple expression in terms of the exponential map. Given a tangent vector $p$ (which, in particular, would include the gradient defined above) the exponential map is defined as

$$\exp_x(p) = \gamma_{x,p}(\|p\|). \tag{3.5}$$

Here $\gamma_{x,p}(t)$ denotes the point a distance $t$ (parametrized by arclength) along the geodesic beginning from $x \in \mathbb{S}^2$ and oriented in the direction $p$. Then the optimal map corresponding to the squared geodesic cost is given by

$$T(\nabla u(x)) = \exp_x(\nabla u(x)).$$

As in McRae et al. (2018), this map can be found explicitly as

$$T(x, p) = \cos(\|p\|)\, x + \sin(\|p\|)\, \frac{p}{\|p\|}. \tag{3.6}$$

We derive a similar explicit form of the optimal map corresponding to the log cost (see Appendix B):

$$T(x, p) = x\frac{\|p\|^2 - 1/4}{\|p\|^2 + 1/4} - \frac{p}{\|p\|^2 + 1/4}. \tag{3.7}$$

The explicit formulas for the mapping $T$ for both costs demonstrates that they are continuous functions of the gradient. Thus, a smooth gradient $\nabla u(x)$ leads to a smooth mapping $T$, which simplifies the task of obtaining consistent approximations of the mapping.

Computing derivatives of order $n \geq 2$ in the tangent plane introduces some local distortion due to the choice of coordinate system. The Hessian on manifolds usually includes an additional first-order term that is non-zero if the Christoffel symbols are non-zero. In our approach in this article, we will be interested in a choice of local coordinates (geodesic normal coordinates) that cause the Christoffel symbols to vanish. This, in turn, will allow us to compute the spherical Hessian

as a "flat" Hessian on the local tangent plane.

The condition that a solution $u$ must be $c$-convex implies that $u$ can be characterized as the $c$-transform of some function $\psi$, recall the definition in Section 2.1.2. For $u$ and $T(x,p)$ smooth and $c$-convex, this condition implies that

$$D^2u(x) + D^2_{xx}c(x, T(x, \nabla u(x))) \geq 0, \tag{3.8}$$

where the inequality here means that the matrix is positive semidefinite. We remark that Equation (2.14) is elliptic only on the space of functions satisfying this constraint. That is,

$$\mathcal{E}(F) = \{u \in C^2(\mathbb{S}^2) \mid D^2u(x) + D^2_{xx}c(x, T(x, \nabla u(x))) \geq 0\}. \tag{3.9}$$

### 3.2.2 Tangent Plane Characterization

In order to actually approximate Equation (2.14) at a point $x_0 \in \mathbb{S}^2$, we wish to define a set of local coordinates $v_{x_0}(x)$ that will map points on the sphere to points on the tangent plane $\mathcal{T}_{x_0}$. This would then allow us to draw from the discretization schemes that are already available for approximating fully nonlinear elliptic PDE in $\mathbb{R}^2$.

We mention that the determinant of the Hessian, and the magnitude and direction of the gradient, are coordinate-invariant quantities. Our particular choice of normal coordinates is motivated primarily by the desire for computational ease. We reemphasize that the computational challenge here is that local coordinates can distort the Hessian and require the introduction of an additional first-order term. To avoid the need to modify the PDE, we choose to work with geodesic normal coordinates, see Figure 3.1. These retain sufficient local structure of the manifold to cause the Christoffel symbols to vanish, which in turn causes the first-order correction term to vanish.

In particular, this choice of normal coordinates preserves distances from the

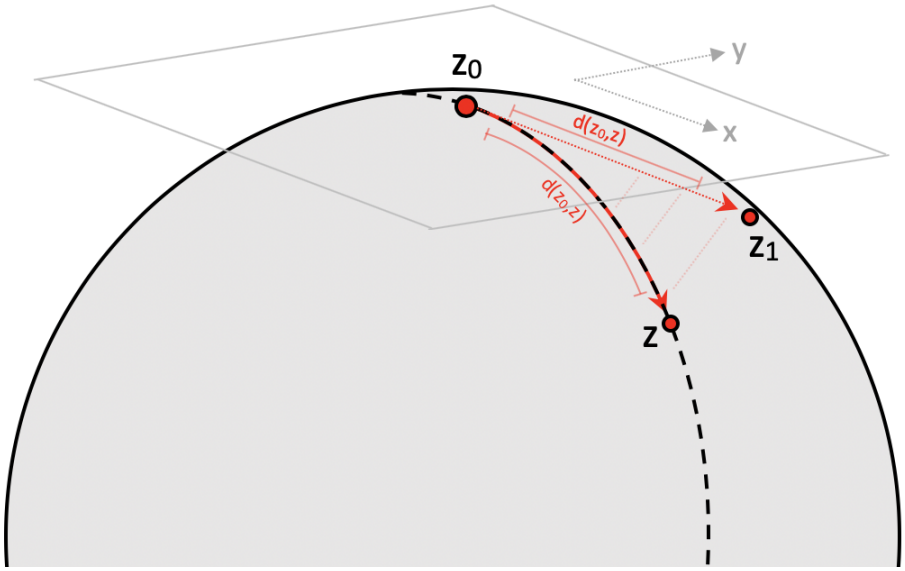



**Figure 3.1** In local normal coordinates about the point $z_0$, the coordinates of a point $z$ are expressed in the local tangent coordinates $(x, y)$ with $z_0$ as the origin. The distance from a point $z$ to the origin in the coordinate system $(x, y)$ is equal to the Euclidean length of the vector $v = z_1 - z_0$ which the exponential map takes to the point $z$. On the sphere, normal coordinates can be made explicit due to the fact that the exponential map has an explicit formula.

reference point $x_0$. That is, if $x \in \mathbb{S}^2$ and $v_{x_0}(x) \in \mathcal{T}_{x_0}$ are sufficiently close to $x_0$, then

$$\|x_0 - v_{x_0}(x)\| = d_{\mathbb{S}^2}(x_0, x).$$

These coordinates also preserve orientation so that the projection of $x - x_0$ into the tangent plane is parallel to $v_{x_0}(x) - x_0$. On the sphere it is possible to construct such coordinates for neighborhoods of uniform size and, in addition, the mapping $v_{x_0}$ is invertible and differentiable. We compute the following explicit representation Appendix C:

$$v_{x_0}(x) = x_0 \left(1 - d_{\mathbb{S}^2}(x_0, x) \cot d_{\mathbb{S}^2}(x_0, x)\right) + x \left(d_{\mathbb{S}^2}(x_0, x) \csc d_{\mathbb{S}^2}(x_0, x)\right). \quad (3.10)$$

For each point $x_0 \in \mathbb{S}^2$ we can now define a function $\tilde{u}_{x_0}(z)$ on the relevant tangent plane $\mathcal{T}_{x_0}$ in a neighborhood of $x_0$ by

$$\tilde{u}_{x_0}(z) = u\left(v_{x_0}^{-1}(z)\right). \quad (3.11)$$

This choice of coordinates allows us to express Equation (2.14) at the point $x_0 \in \mathbb{S}^2$ as a generalized Monge-Ampère equation

$$F\left(x_0, \nabla\tilde{u}(x_0), D^2\tilde{u}(x_0)\right) \equiv -\det\left(D^2\tilde{u}(x_0) + A(x_0, \nabla\tilde{u}(x_0))\right) + H(x_0, \nabla\tilde{u}(x_0)) = 0, \tag{3.12}$$

which is now conveniently posed locally on two-dimensional planes. Thus the problem of approximating the PDE at $x_0$ reduces to the problem of constructing an approximation to the two-dimensional generalized Monge-Ampère equation, Equation (3.12), at $x_0$, posed on the tangent plane containing the points $v_{x_0}(x)$.

We emphasize again that the gradient and Hessian of $\tilde{u}$ on the tangent plane at $x_0$ are equivalent to the surface gradient and Hessian on the original function $u$ on the sphere at $x_0$ (Lee, 2006, Lemma 4.8 and Proposition 5.11). Thus using these local coordinates indeed allows us to interpret our PDE, without modification, on the tangent plane.

**Lemma 3.5.** *Let $u \in C^2(\mathbb{S}^2)$ and $x_0 \in \mathbb{S}^2$, with $\tilde{u} : \mathcal{T}_{x_0} \in \mathbb{R}$ defined in geodesic normal coordinates via Equation* (3.11)*. Then the PDE operator in Equation* (2.14) *applied to $u$ at the point $x_0$ is equivalent to the generalized Monge-Ampère operator in Equation* (3.12) *applied to $\tilde{u}$ at the point $x_0$:*

$$F(x_0, \nabla_{\mathbb{S}^2} u(x_0), D^2_{\mathbb{S}^2} u(x_0)) = F(x_0, \nabla\tilde{u}(x_0), D^2\tilde{u}(x_0)).$$

### 3.2.3 Constraints

We now turn our attention to the problem of incorporating constraints into the PDE. We recall that the PDE operator in Equation (2.14) is elliptic only on the space of functions satisfying the constraint in Equation (3.9). Consequently, this constraint is necessary for the equation to be well-posed. We propose instead to produce a globally elliptic extension of Equation (2.14) that does not require additional constraints. To do so, we introduce a modified determinant operator

satisfying

$$\det^+(M) = \begin{cases} \det(M), & M \geq 0, \\ < 0, & \text{otherwise.} \end{cases} \tag{3.13}$$

Then we can absorb the constraint into Equation (3.12) through the modification

$$F^+(x, \nabla u(x), D^2u(x)) \equiv -\det{}^+(D^2u(x) + A(x, \nabla u(x))) + H(x, \nabla u(x)) = 0. \tag{3.14}$$

Since the function $H > 0$, (sub)solutions of this will automatically satisfy the condition

$$D^2u(x) + A(x, \nabla u(x)) \geq 0.$$

The solution $u$ of Equation (2.31) is also known to satisfy *a priori* bounds on its gradient,

$$\|\nabla u\| \leq R \tag{3.15}$$

for any $R > \pi$ in the case of the squared geodesic cost and $R > C$ in the case of the logarithmic cost. Here $C$ is the bound on $\nabla u$ determined in (Loeper, 2011, Proposition 6.1).

With the goal of constructing Lipschitz stable approximation schemes, we state a modification of the PDE that explicitly includes these constraints on the gradient.

$$G(x, \nabla u(x), D^2u(x)) \equiv \max\left\{F^+(x, \nabla u(x), D^2u(x)), \|\nabla u(x)\| - R\right\} = 0. \tag{3.16}$$

We again emphasize that this new PDE is elliptic on all $C^2$ functions ($\mathcal{E}(G) = C^2(\mathbb{S}^2)$), and does not require any additional constraints. Moreover, as we demonstrate below, the $c$-convex solution of Equation (2.14) is indeed a solution of this modified equation.

**Remark 3.6.** *Under the assumption that the globally elliptic equation, Equation* (3.16)*, has a unique solution, it must automatically coincide with the c-convex*

*solution of the original equation. Comparison principles and uniqueness results for many fully nonlinear elliptic PDEs of this form are available Crandall et al. (1992). However, these calculations are highly technical and need to be specifically adapted to the PDE at hand. This is beyond the scope of the present article.*

It is not *a priori* obvious that solutions of this new PDE operator will automatically satisfy the original PDE. Indeed, because of the action of the maximum operator, they need only be subsolutions. To establish the plausibility of this new operator, we establish that the equivalence of these two equations for smooth, $c$-convex functions.

**Theorem 3.7** (Equivalence of PDE (smooth case))**.** *Under the conditions of Hypothesis 2.2, a c-convex function $u \in C^2$ is a solution of Equation* (2.14) *if and only if it is a solution of Equation* (3.16)*.*

Before completing the proof, we establish a few lemmas relating to the transportation of mass by subsolutions. The following proofs will make use of an abbreviated notation for the transport map:

$$T_u(x) = T(x, \nabla u(x)).$$

**Lemma 3.8.** *If $u \in C^2$ is c-convex then it satisfies the constraint in Equation* (3.8)*: $D^2u(x) + D^2_{xx}c(x, T_u(x)) \geq 0$.*

*Proof.* If $u$ is $c$-convex, then for every $x_0 \in \mathbb{S}^2$ we can fix $y = T_u(x_0)$ and find that the supremum in

$$u^c(y) = \sup_{x \in \mathbb{S}^2} \{-c(x, y) - u(x)\}$$

is attained at $x_0$. The optimality condition for this is precisely Equation (3.8). □

**Lemma 3.9.** *Under the conditions of Hypothesis 2.2, let $u \in C^2$ be a subsolution of Equation* (2.14)*. Then*

$$\int_{T_u(\mathbb{S}^2)} f_1(y)\, dy \leq \int_{\mathbb{S}^2} f_0 s(x)\, dx.$$

*Proof.* By design, the transport maps from Equations (3.6)-(3.7) satisfy $T_u(\mathbb{S}^2) \subset \mathbb{S}^2$. Because of mass balance we conclude that

$$\int_{\mathbb{S}^2} f_0(x)\,dx = \int_{\mathbb{S}^2} f_1(y)\,dy \geq \int_{T_u(\mathbb{S}^2)} f_1(y)\,dy. \qquad \square$$

The preceding lemma will be used to derive a contradiction that shows smooth subsolutions of Equation (2.14) are, in fact, solutions.

**Lemma 3.10.** *Under the conditions of Hypothesis 2.2, let $u \in C^2(\mathbb{S}^2)$ be a subsolution of Equation* (2.14)*. Then $u$ is a solution of Equation* (2.14)*.*

*Proof.* Suppose $u$ is not a solution. Since $u \in C^2(\mathbb{S}^2)$, there exists some open set $E \subset \mathbb{S}^2$ such that

$$F(x, \nabla u(x), D^2u(x)) < 0.$$

We recall that the mapping $T_u$ satisfies the condition given by Equation (2.15):

$$\nabla u(x) = -\nabla_x c(x, T_u(x)).$$

Differentiating yields

$$D^2u(x) = -D^2_{xx}c(x, T_u(x)) - D^2_{xy}c(x, T_u(x))DT_u(x).$$

Since $u$ is a subsolution of Equation (2.14), we know that

$$\begin{aligned}\left|\det(D^2_{xy}c(x, T_u(x)))\right| f_0(x)/f_1(T_u(x)) &\leq \det(D^2u(x) + D^2_{xx}c(x, T_u(x))) \\ &= \left|\det(D^2_{xy}c(x, T_u(x)))\right| \det(DT_u(x)).\end{aligned}$$

Therefore,

$$f_0(x) \leq \det(DT_u(x))f_1(T_u(x))$$

with strict inequality on an open set $E \subset \mathbb{S}^2$.

Integrating, we obtain

$$\int_{\mathbb{S}^2} f_0(x)\,dx < \int_{T_u(\mathbb{S}^2)} f_1(y)\,dy.$$

This contradicts Lemma 3.9 and thus $u$ is a solution of Equation (2.14). □

*Proof of Theorem 3.7.* Let $u$ be a c-convex solution of Equation (2.14). Then it satisfies the gradient bound $\|\nabla u\| - R \leq 0$ from Equation (3.15). Because it is c-convex, it also satisfies the constraint Equation (3.8) (Lemma 3.8) so that

$$F^+(x, \nabla u(x), D^2u(x)) = F(x, \nabla u(x), D^2u(x)) = 0.$$

Then trivially the maximum of these operators also vanishes, and the modified PDE, Equation (3.16), is satisfied.

Now we let $u$ be a solution of the modified PDE, Equation (3.16), so that

$$\max\left\{F^+(x, \nabla u(x), D^2u(x)), \|\nabla u(x)\| - R\right\} = 0.$$

This implies that $u$ is a subsolution of the convexified PDE operator in Equation (3.14) denoted by $F^+$. Subsolutions of this equation automatically satisfy the constraint (Equation (3.8)) (see the definition of $\det^+$) so that

$$F(x, \nabla u(x), D^2u(x)) = F^+(x, \nabla u(x), D^2u(x)) \leq 0.$$

From Lemma 3.10, $u$ is necessarily a solution of Equation (2.14). □

We also partially extend this equivalence result to the non-smooth case for the squared geodesic cost.

**Theorem 3.11** (Equivalence of PDE (non-smooth case))**.** *Under the conditions of Hypothesis 2.3, let $u \in C^{0,1}(\mathbb{S}^2)$ be a c-convex viscosity solution of Equation* (2.14)*. Then $u$ is a viscosity solution of Equation* (3.16)*.*

**Remark 3.12.** *The key to proving this result is the observation that subsolutions of the modified equation satisfy* a priori *Lipschitz bounds. This is fairly straightforward for the squared geodesic cost, but more challenging for the logarithmic cost because of the singularity in the cost function. Below, in Section 3.2.4, an alternative (regularized) version of the logarithmic cost function is used that yields an a priori Lipschitz bound and thus the same reasoning used in this section for the squared geodesic cost applies as well to the unregularized logarithmic cost.*

Once again, we begin with a few lemmas.

**Lemma 3.13** (Local $c$-convexity of test functions)**.** *Let $u \in C^{0,1}(\mathbb{S}^2)$ be $c$-convex with cost function $c(x,y) = \frac{1}{2}d_{\mathbb{S}^2}(x,y)^2$ and $\phi \in C^\infty(\mathbb{S}^2)$. Suppose that $u - \phi$ has a local maximum at $x_0$. Then*

$$D^2\phi(x_0) + D^2_{xx}c(x_0, T_\phi(x_0)) \geq 0.$$

*Proof.* At the maximizer $x_0$ of $u - \phi$, we must have $\nabla\phi(x_0) \subset \partial u(x_0)$.

Since $u$ is $c$-convex, there exists a function $u^c$ such that

$$u(x) + u^c(y) = -c(x,y), \quad y \in \partial u(x).$$

Thus the maximizer $x_0$ of $u - \phi$ will also maximize the function $-u^c(y) - c(x,y) - \phi(x)$, where we can in particular choose $y = T_\phi(x_0)$. The optimality condition for this is

$$-D^2_{xx}c(x_0, y) - D^2\phi(x_0) \leq 0, \quad y = T_\phi(x_0). \qquad \square$$

**Lemma 3.14** (Lipschitz bounds on subsolutions)**.** *Let $u \in C^{0,1}$ be $c$-convex where $c(x,y) = \frac{1}{2}d^2_{\mathbb{S}^2}(x,y)$. Then the Lipschitz constant of $u$ is bounded by $\pi$.*

*Proof.* We first consider $x \in \mathbb{S}^2$ such that $u$ is differentiable at $x$. As in Loeper (2011), we define the set

$$G_u(x) = \{y \in \mathbb{S}^2, u(x) + u^c(y) = -c(x,y)\}.$$

Letting $\partial^c u(x)$ denote the $c$-subdifferential of $u$, defined as

$$\partial^c u(x) = \{-\nabla_x c(x,y), y \in G_u(x)\},$$

from Loeper (2009) Proposition 2.11 we know that for all $c$-convex $u$,

$$\emptyset \neq \partial^c u(x) = \partial u(x).$$

Thus, $\nabla u(x) = \partial^c u(x)$.

To bound $\nabla u$, we need only bound the gradient of the cost function $c(x,y)$:

$$\nabla_x c(x,y) = d_{\mathbb{S}^2}(x,y)\nabla_x d_{\mathbb{S}^2}(x,y).$$

Letting $\hat{n}$ denote a unit tangent vector in the tangent plane $\mathcal{T}(x)$, we compute

$$\nabla_x d_{\mathbb{S}^2}(x,y)\cdot\hat{n} = \lim_{s\to 0}\frac{d_{\mathbb{S}^2}(\exp_x(s\hat{n}),y) - d_{\mathbb{S}^2}(x,y)}{s}.$$

From the triangle inequality we obtain the bounds

$$\nabla_x \cdot \hat{n} d_{\mathbb{S}^2}(x,y) \leq \lim_{\|\Delta x\|\to 0}\frac{d_{\mathbb{S}^2}(\exp_x(s\hat{n}),x) + d_{\mathbb{S}^2}(x,y) - d_{\mathbb{S}^2}(x,y)}{s} = 1$$

and

$$\nabla_x\cdot\hat{n} d_{\mathbb{S}^2}(x,y) \geq \lim_{s\to 0}\frac{d_{\mathbb{S}^2}(\exp_x(s\hat{n}_x),y) - d_{\mathbb{S}^2}(\exp_x(s\hat{n}_x),y) - d_{\mathbb{S}^2}(\exp_x(s\hat{n}_x),x)}{s} = -1.$$

Therefore,

$$\|\nabla u(x)\| \leq \|\nabla_x c(x,y)\| \leq d_{\mathbb{S}^2}(x,y) \leq \pi \tag{3.17}$$

at points $x$ where $u$ is differentiable. Since $u$ is Lipschitz continuous, this gradient bound is also a bound on the Lipschitz constant. □

*Proof of Theorem 3.11.* Suppose that $u$ is a $c$-convex viscosity solution of Equation (2.14). Consider any $x_0 \in \mathbb{S}^2$ and $\phi \in C^\infty(\mathbb{S}^2)$ such that $u - \phi$ has a local

maximum at $x_0$. Then

$$F(x_0, \nabla\phi(x_0), D^2\phi(x_0)) \leq 0.$$

Moreover, since $u - \phi$ is a maximum we know that $\nabla\phi(x_0) \subset \partial u(x_0)$. From Lemma 3.14 we find that $\|\nabla\phi(x_0)\| - R < 0$. Additionally, since $u$ is $c$-convex, $\phi$ must be locally $c$-convex as well near $x_0$ (Lemma 3.13) so that $\phi \in \mathcal{E}(F)$ is a valid test function for the original PDE operator. Thus,

$$F^+(x_0, \nabla\phi(x_0), D^2\phi(x_0)) = F(x_0, \nabla\phi(x_0), D^2\phi(x_0)) \leq 0,$$

and the modified operator will satisfy

$$\max\{F^+(x_0, \nabla\phi(x_0), D^2\phi(x_0)), \|\nabla\phi(x_0)\| - R\} \leq 0.$$

Therefore $u$ is a sub-solution of Equation (3.16).

Next we consider $x_0 \in \mathbb{S}^2$ and $\phi \in C^\infty(\mathbb{S}^2)$ such that $u - \phi$ has a local minimum at $x_0$. If $\phi$ satisfies the constraint (Equation (3.8)) then $\phi \in \mathcal{E}(F)$ is a valid test function for the original PDE operator. Thus, by the fact that $u$ is a supersolution of Equation (2.14), we have

$$\max\{F^+(x_0, \nabla\phi(x_0), D^2\phi(x_0)), \|\nabla\phi(x_0)\| - R\} \geq F(x_0, \nabla\phi(x_0), D^2\phi(x_0)) \geq 0.$$

Otherwise, $D^2\phi(x_0) + D^2_{xx}c(x_0, T_\phi(x_0))$ is not positive semi-definite. From the definition of the modified determinant operator Equation (3.13), this means that

$$F^+(x_0, \nabla\phi(x_0), D^2\phi(x_0)) \geq -\det{}^+(D^2\phi(x_0) + D^2_{xx}c(x_0, T_\phi(x_0))) > 0.$$

This again leads to the inequality

$$\max\{F^+(x_0, \nabla\phi(x_0), D^2\phi(x_0)), \|\nabla\phi(x_0)\| - R\} > 0.$$

In either case, we conclude that $u$ is a super-solution, and therefore also a viscosity solution, of Equation (3.16).

□

### 3.2.4 Regularization of the Logarithmic Cost

One modification of the PDE that we find is necessary to build monotone schemes is to make the logarithmic cost Lipschitz by using a cutoff function. We recall that the solution $u$ to the Optimal Transport satisfies an *a priori* Lipschitz bound Loeper (2011) $\|\nabla u\| < R$, which also yields the following lower bound on the distance mass can be transported:

$$\|x - y\| > \frac{1}{\sqrt{R^2 + 1/4}} \tag{3.18}$$

when $y = T(x, \nabla u(x))$ is the exact transport map.

However, the process of solving a discrete version of Equation (2.14) may evolve through values of $u$ (and consequently $y$) that do not satisfy this bound. This loss of Lipschitz continuity can lead to a breakdown in monotonicity. We thus propose the following $C^3$ regularization of the logarithmic cost function, which agrees with the true logarithmic cost when the bound, Equation (3.2.4), is satisfied:

$$\tilde{c}(x, y) = \begin{cases} -\log \|x - y\|, & \|x - y\| \geq \frac{1}{\sqrt{R^2+1/4}}, \\ \Psi(\|x - y\|), & \|x - y\| < \frac{1}{\sqrt{R^2+1/4}}, \end{cases} \tag{3.19}$$

where $z_* = \dfrac{1}{\sqrt{R^2 + 1/4}}$ and

$$\Psi(z) = -\log(z_*) - \frac{1}{z_*}(z - z_*) + \frac{1}{2z_*^2}(z - z_*)^2 - \frac{1}{3z_*^3}(z - z_*)^3.$$

Because this regularization does not change the solutions of the PDE, the analysis of Section 3.2.3 and the ultimate convergence result (Theorems 3.30 and 3.31) will also apply to discretizations involving this smoother cost function.

**Lemma 3.15** (Equivalence of solution for modified cost function)**.** *Under the assumptions of Hypothesis 2.2, a function $u \in C^3(\mathbb{S}^2)$ is a solution of Equation* (2.14) *with the logarithmic cost (Equation* (3.2)*) if and only if it is a solution of Equation* (2.14) *with the regularized cost (Equation* (3.19)*).*

*Proof.* First let $u$ be a solution using the original cost function $c(x, y) = -\log \|x - y\|$. Because of the *a priori* bounds on the gradient, this automatically satisfies Equation (2.14) with the regularized cost function.

Next let $v$ be any solution of Equation (2.14) using the regularized cost function $\tilde{c}(x, y)$ from Equation (3.19). Notice that for $\|x - y\| \geq \frac{1}{\sqrt{R^2+1/4}}$, we have

$$\tilde{c}(x, y) = -\log\left(2\sin\left(\frac{1}{2}d_{\mathbb{S}^2}(x, y)\right)\right). \tag{3.20}$$

It is easily verified via differentiation that this is convex in $d_{\mathbb{S}^2}(x, y)$ for $\|x - y\| < \frac{1}{\sqrt{R^2 + 1/4}}$, and the original logarithmic cost is also convex in $d_{\mathbb{S}^2}(x, y)$. The modified cost is $C^3$, so verifying that the second derivative in the variable $d_{\mathbb{S}^2}(x, y)$ is everywhere positive is sufficient to guarantee convexity in this variable. Since this new cost function is Lipschitz and convex in the Riemannian distance, the $\tilde{c}$-convex solution of Equation (2.14) for the regularized cost $\tilde{c}(x, y)$ is guaranteed to be unique by McCann as noted in Loeper (2009). Since the solution $u$ of the original equation solves this modified equation, uniqueness requires that $v = u$. □

**Remark 3.16.** *Because this regularization transforms the Optimal Transport problem with a singular cost function into an Optimal Transport problem with a smooth cost function, the techniques of Section 3.2.3 (which were introduced for the squared geodesic cost) can be extended to show that weak ($C^{0,1}$) solutions of this modified problem are also unique. This assumption of uniqueness of $C^{0,1}$ viscosity solutions was needed to prove convergence even in the smooth setting.*

In the development and analysis of our numerical method in the following sections, we will often refer to the cost function $\tilde{c}$. In the case of the logarithmic cost, this refers to the regularized cost (Equation (3.19)). In the case of the squared geodesic

cost, $\tilde{c}$ will refer to the original cost function $c$, which is automatically smooth.

## 3.3 Convergence Framework

### 3.3.1 Discrete Formulation

In order to numerically solve Equation (2.14), we begin with a point cloud $\mathcal{G}^h \subset \mathbb{S}^2$ that discretizes the sphere. We define the discretization parameter $h$ as

$$h = \sup_{x \in \mathbb{S}^2} \min_{y \in \mathcal{G}^h} d_{\mathbb{S}^2}(x, y). \tag{3.21}$$

In particular, this guarantees that any ball of radius $h$ on the sphere will contain at least one discretization point.

We will impose some mild structural regularity on the grid.

**Hypothesis 3.17** (Conditions on point cloud)**.** *There exists a triangulation $T^h$ of $\mathcal{G}^h$ with the following properties:*

*(a) The diameter of the triangulation, defined as*

$$diam(T^h) = \max_{t \in T^h} diam(t), \tag{3.22}$$

*satisfies $diam(T^h) \to 0$ as $h \to 0$.*

*(b) There exists some $\gamma < \pi$ (independent of $h$) such that whenever $\theta$ is an interior angle of any triangle $t \in T^h$ then $\theta \leq \gamma$.*

We remark that these are fairly standard assumptions on a grid: we are simply prohibiting long, thin triangles.

We also associate to each point cloud $\mathcal{G}^h$ a search radius $r(h)$ chosen to satisfy

$$r(h) \to 0, \ \frac{h}{r(h)} \to 0 \text{ as } h \to 0, \quad \text{diam}(T^h) < r(h). \tag{3.23}$$

Now we considering the problem of constructing a discretization of Equation (3.16) at the point $x_0 \in \mathcal{G}^h$. We begin by projecting nearby grid points onto the local tangent plan $\mathcal{T}_{x_0}$, which is spanned by the orthonormal vectors $\left(\hat{\theta}, \hat{\phi}\right)$.

For all points $x_i \in \mathcal{G}^h \cap B(x_0, r(h))$, we define their projection onto the tangent plane through geodesic normal coordinates via

$$z_i = x_0 \left(1 - d_{\mathbb{S}^2}(x_0, x_i) \cot d_{\mathbb{S}^2}(x_0, x_i)\right) + x_i \left(d_{\mathbb{S}^2}(x_0, x_i) \csc d_{\mathbb{S}^2}(x_0, x_i)\right). \tag{3.24}$$

Let $\mathcal{Z}^h(x_0) \subset \mathcal{T}_{x_0}$ be the resulting collection of points. See Figure 3.2.

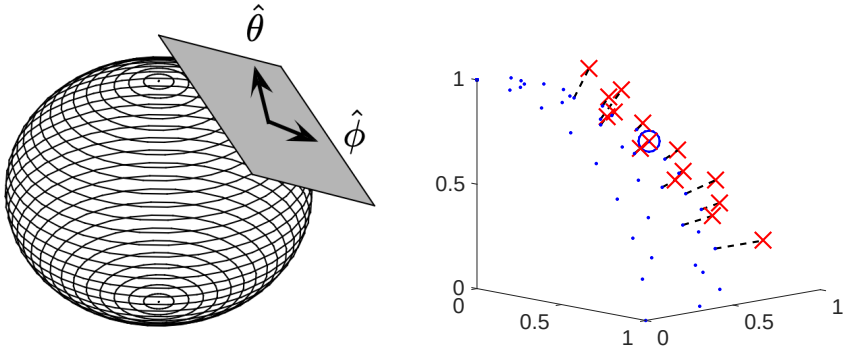



**Figure 3.2** The sphere $\mathbb{S}^2$ (left) and tangent plane $\mathcal{T}_{x_0}$ (right). A point cloud discretizing one octant of the unit sphere ($\cdot$), the point $x_0$ (o), and the projections $z$ of neighboring nodes onto $\mathcal{T}_{x_0}$ ($\times$).

These are now the discretization points available to use for the approximation of Equation (3.16) at $x_0$; recall that this PDE is posed on the two-dimensional tangent plane. There are three components to this discretization: approximation of the Monge-Ampère type operator $F^+(z, \nabla u(z), D^2 u(z))$ (Equation (3.14)), approximation of the Eikonal term $\nabla u(x)$, and approximation of the averaging term $\langle u \rangle$. Let $F^h$, $E^h$, and $A^h$ be suitable discretizations of these three operators.

Our framework will allow for a very general choice of schemes $F^h$ and $A^h$. In particular, many currently available methods for the Monge-Ampère equation can be adapted to fit within our requirements. The specific requirements are:

**Hypothesis 3.18** (Conditions on schemes)**.** *We require the schemes $F^h(x, u(x) - u(\cdot))$ and $A^h(u(\cdot))$ to satisfy:*

*(a) $F^h$ is consistent with Equation* (3.14) *on all $C^2$ smooth functions,*

*(b) $F^h$ is monotone,*

(c) *$A^h$ is consistent with the averaging operator (Equation (3.4)) on all Lipschitz continuous functions,*

(d) *$A^h$ is linear and $A^h(c) = c$ for any constant function $c$.*

If we wish to obtain non-smooth solutions, $F^h$ will also need to be underestimating. We will require additional structure on $E^h$ in order to obtain the strong form of stability needed to guarantee convergence. In particular, we propose

$$E^h(z, u(z) - u(\cdot)) = \max_{y \in \mathcal{Z}^h(z)} \frac{u(z) - u(y)}{\|z - y\|}, \tag{3.25}$$

which is consistent with $\|\nabla u(z)\|$ and monotone (Lemma 3.19).

This allows us to produce the following consistent, monotone approximation of Equation (3.16):

$$G^h(x, u(x) - u(\cdot)) = \max\{F^h(x, u(x) - u(\cdot)), E^h(x, u(x) - u(\cdot)) - R\}. \tag{3.26}$$

Finally, we represent our overall approach through the following two-step approach:

1. Solve the discrete system

$$G^h(x, v^h(x) - v^h(\cdot)) + \tau(h)v^h(x) = 0, \quad x \in \mathcal{G}^h \tag{3.27}$$

for the grid function $v^h$.

2. Define the candidate solution:

$$u^h(x) = v^h(x) - A^h(v^h(\cdot)), \quad x \in \mathcal{G}^h. \tag{3.28}$$

We remark that our candidate solution $u^h$ could also be obtained directly through solution of the non-local approximation scheme

$$G^h(x, u^h(x) - u^h(\cdot)) + \tau(h)u^h(x) + A^h\left(G^h(x, u^h(x) - u^h(\cdot))\right) = 0. \tag{3.29}$$

### 3.3.2 Stability

We now establish some important stability properties of the solutions $v^h$, $u^h$ of the schemes in Equations (4.24)-(3.29). Consistency and monotonicity underpin these results. They are built into our hypotheses on the scheme for the Monge-Ampère type operator in order to allow for great flexibility in the numerical method. However, we also need to establish these properties for our proposed discretization of the Eikonal operator.

**Lemma 3.19** (Approximation of Eikonal operator)**.** *The scheme $E^h$ is consistent with $\|\nabla u\|$ and monotone.*

*Proof.* Monotonicity is immediately evident from the definition of $E^h$ in Equation (3.25).

Now we recall that the magnitude of the gradient can be characterized as a maximal directional derivative,

$$\|\nabla u\| = \max_{\|\nu\|=1} \frac{\partial u}{\partial \nu}.$$

We can obtain an approximation of the first directional derivative in the direction $\nu = \dfrac{z-y}{\|z-y\|}$ via standard backward differencing:

$$\mathcal{D}_{z-y} u(z) = \frac{u(z) - u(y)}{\|z-y\|}. \tag{3.30}$$

Now we consider the set of all such directions that can be resolved using our given set of neighbours $\mathcal{Z}^h(z)$, defined as

$$V^h(z) = \left\{ \frac{z-y}{\|z-y\|} \mid y \in \mathcal{Z}^h(z) \right\}.$$

The discretization $E^h$ can be rewritten as

$$E^h(z, u(z) - u(\cdot)) = \max_{\nu \in V^h(z)} \mathcal{D}_\nu u(z). \tag{3.31}$$

We denote the directional resolution of this approximation by $d\theta$, which can be computed by

$$d\theta = \sup_{\|\nu\|=1} \min_{y \in \mathcal{Z}^h(z)} \cos^{-1}\left(\frac{z-y}{\|z-y\|} \cdot \nu\right).$$

We also remark that projecting the points $x_i \in \mathcal{G}^h \cap B(x_0, r(h))$ onto the plane preserves both the spacing of grid points $h$ and the effective search radius $r(h)$ up to a constant scaling. Since $r(h) \to 0$, the effective grid spacing also goes to zero and thus Equation (3.30) is a consistent differencing operator. Since $\dfrac{h}{r(h)} \to 0$ as $h \to 0$, we will also have $d\theta \to 0$ as $h \to 0$ as in (Froese, 2018, Lemma 11). Thus $E^h$ defined as in Equation (3.31) is consistent. □

An immediate consequence of this is the consistency and monotonicity of our overall scheme in Equation (3.26).

**Lemma 3.20** (Consistency and monotonicity)**.** *Let $\mathcal{G}^h$ and $F^h$ satisfy the conditions of Hypotheses 3.17 and 3.18 respectively. The the approximation $G^h$ given by Equation* (3.26) *is monotone and consistent with Equation* (3.16)*.*

We now use the monotonicity property (and resulting discrete comparison principle) to establish existence and bounds for the solution to our approximation scheme.

**Lemma 3.21** (Existence and stability (smooth case))**.** *Consider the schemes in Equations* (4.24)-(3.28) *under the conditions of Hypotheses 2.2, 3.17, and 3.18. Then solutions $v^h$, $u^h$ exist and are unique. Moreover, there exists some $M > 0$ (independent of $h$) such that $\|v^h\|_\infty, \|u^h\|_\infty \leq M$ for all sufficiently small $h > 0$.*

*Proof.* We remark first of all that the scheme in Equation (4.24) is monotone and proper and therefore has a unique solution $v^h$ (Oberman, 2006, Theorem 8), which immediately yields existence of $u^h$.

Let $u$ be the unique mean-zero solution to Equation (3.16). We know that $u \in C^3(\mathbb{S}^2)$ (Theorem 2.5) and consequently is bounded. From consistency of the

scheme in Equation (3.26) we have that

$$\left|G^h(x, u(x) - u(\cdot))\right| \leq \tau(h)$$

for all $x \in \mathcal{G}^h$ and sufficiently small $h > 0$.

Now we choose some $c > 0$ and substitute $u + c$ into the scheme in Equation (4.24).

$$\begin{aligned} G^h(x, (u(x) + c) - (u(\cdot) + c)) + \tau(h)(u(x) + c) &\geq -\tau(h) + \tau(h)(-\|u\|_\infty + c) \\ &> 0 \\ &= G^h(x, v^h(x) - v^h(\cdot)) + \tau(h)v^h(x) \end{aligned}$$

for $c > \|u\|_\infty + 1$. By the discrete comparison principle (Lemma 3.3), we have that $v^h \leq u + c \leq 2\|u\|_\infty + 1$. A similar argument produces a lower bound for $v^h$.

This allows us to also bound the discrete average of $v^h$ via

$$A^h(v^h(\cdot)) \leq A^h(2\|u\|_\infty + 1) = 2\|u\|_\infty + 1,$$

with a similar lower bound.

Since $v^h$ and $A^h(v^h)$ are bounded uniformly, $u^h = v^h - A^h(v^h)$ is also bounded uniformly. □

With some additional structure on our discretization, we can modify this stability result to also hold in the non-smooth setting.

**Lemma 3.22** (Existence and stability (non-smooth case))**.** *Consider the schemes in Equations* (4.24)-(3.28) *under the conditions of Hypothesis 2.3, 3.17, and 3.18. Suppose also that $F^h$ is an underestimating scheme. Then solutions $v^h$, $u^h$ exist and are unique. Moreover, there exists some $M > 0$ (independent of $h$) such that $\|v^h\|_\infty, \|u^h\|_\infty \leq M$ for all sufficiently small $h > 0$.*

*Proof.* As in Lemma 3.21, $v^h$ and $u^h$ are uniquely defined.

Let $u$ be the exact mean-zero solution of Equation (3.16). Now we know that $u$ is Lipschitz continuous with Lipschitz constant less than $R$. This implies that

$$E^h(x, u(x) - u(\cdot)) = \max_{y \in \mathcal{Z}^h(x)} \frac{u(x) - u(y)}{\|x - y\|} \leq R.$$

Because $F^h$ is an underestimating scheme, we also know that

$$F^h(x, u(x) - u(\cdot)) \leq 0.$$

Choosing any $c > \|u\|_\infty$ we then obtain

$$\begin{aligned} G^h(x, (u(x) - c) - (u(\cdot) - c)) + \tau(h)(u(x) - c) &\leq \tau(h)(\|u\|_\infty - c) \\ &< 0 \\ &= G^h(x, v^h(x) - v^h(\cdot)) + \tau(h)v^h(x) \end{aligned}$$

and by the discrete comparison principle we have the bound $v^h \geq u - \|u\|_\infty \geq -2\|u\|_\infty$.

A simple smooth supersolution of Equation (3.16) is the constant function $\phi(x) = c$. Substituting this into the consistent scheme we find that

$$\begin{aligned} G^h(x, \phi(x) - \phi(\cdot)) + \tau(h)\phi(x) &\geq -\tau(h) + \tau(h)c \\ &> 0 \\ &= G^h(x, v^h(x) - v^h(\cdot)) + \tau(h)v^h(x) \end{aligned}$$

if we choose $c > 1$, which yields the bound $v^h \leq 1$.

As in Lemma 3.21, these uniform bounds on $v^h$ immediately yield uniform bounds on $u^h$. □

An immediate consequence of this is that $u^h$ satisfies a discrete system that is consistent with Equation (3.16).

**Lemma 3.23** (Scheme for $u^h$)**.** *Under the hypotheses of either Lemma 3.21 or*

*Lemma 3.22, $u^h$ satisfies a scheme of the form*

$$G^h(x, u^h(x) - u^h(\cdot)) + \tau(h)u^h(x) + \sigma(h) = 0 \tag{3.32}$$

*where $\sigma(h) \to 0$ as $h \to 0$.*

Another immediate consequence of these lemmas is that $u^h$ satisfies a discrete Lipschitz bound uniformly in $h$.

**Lemma 3.24** (Discrete Lipschitz bounds)**.** *Under the hypotheses of either Lemma 3.21 or Lemma 3.22, $u^h$ satisfies a local discrete Lipschitz bound of the form*

$$\left|u^h(z) - u^h(y)\right| \leq L \left\|z - y\right\| \tag{3.33}$$

*for all $y \in \mathcal{Z}^h(z)$ and sufficiently small $h > 0$ where $L \in \mathbb{R}$ is independent of $h$.*

*Proof.* Note that $u^h$ satisfies Equation (3.32). For small enough $h$, we can assume $\tau(h), |\sigma(h)| < 1$ and $\|u^h\|_\infty \leq M$. By construction,

$$E^h(z, u^h(z) - u^h(\cdot)) \leq G^h(z, u^h(z) - u^h(\cdot)) = -\tau(h)u^h(z) - \sigma(h) \leq M + 1 \equiv L.$$

From the definition of $E^h$, we then have

$$u^h(z) - u^h(y) \leq L \left\|z - y\right\|, \quad y \in \mathcal{Z}^h(z).$$

If $u^h(z) - u^h(y) \geq 0$ we are done. Otherwise, we notice that $z \in \mathcal{Z}^h(y)$ and we can use the fact that

$$0 < u^h(y) - u^h(z) \leq L \left\|y - z\right\|,$$

which establishes the result. □

Because of our choice of geodesic normal coordinates, we can immediately extend this to a discrete Lipschitz bound for the function $u^h$ defined on $\mathcal{G}^h \subset \mathbb{S}^2$

in terms of geodesic distances on the sphere (rather than distances on the tangent plane).

**Lemma 3.25** (Discrete Lipschitz bounds on sphere)**.** *Under the hypotheses of either Lemma 3.21 or Lemma 3.22, $u^h$ satisfies a local discrete Lipschitz bound of the form*

$$\left|u^h(x) - u^h(y)\right| \leq L d_{\mathbb{S}^2}(x, y) \tag{3.34}$$

*for all $x \in \mathcal{G}^h$, $y \in \mathcal{G}^h \cap B(x, r(h))$, and sufficiently small $h > 0$. Here $L \in \mathbb{R}$ is independent of $h$.*

### 3.3.3 Interpolation

In order to establish convergence of the grid function $u^h$ to the solution of Equation (3.16), we will need to construct an appropriate (Lipschitz continuous) extension of it onto the sphere.

We start by considering linear interpolation of a grid function $w : \mathcal{G}^h \to \mathbb{R}$ onto the triangulated surface $T^h$ described in Hypothesis 3.17. In particular, we want to show that the local discrete Lipschitz bounds in Equation (3.34) are inherited by the resulting piecewise linear interpolant.

**Lemma 3.26** (Interpolation onto triangulated surface)**.** *Let $\mathcal{G}^h$ be a point cloud satisfying Hypothesis 3.17 and let $w : T^h \to \mathbb{R}$ be a piecewise linear function, linear on each triangle $t \in T^h$, that satisfies the local discrete Lipschitz bounds in Equation* (3.34)*. Then there exists some $L \in \mathbb{R}$ (independent of $h$) such that for every $t \in T^h$ and $x, y \in T$, $w$ satisfies the Lipschitz bound $|w(x) - w(y)| \leq L \|x - y\|$.*

*Proof.* First we consider the gradient of $w$ on a single triangle $t \in T^h$. Let $t$ have the vertices $x_0, x_1, x_2 \in \mathcal{G}^h$. Without loss of generality, we suppose that the maximal interior angle of $t$ occurs at the vertex $x_0$. Since $\operatorname{diam}(T^h) < r(h) \to 0$

as $h \to 0$, there exists a constant $\tilde{L}$ (independent of $h$) such that

$$|w(x_i) - w(x_j)| \leq L d_{\mathbb{S}^2}(x_i, x_j) = 2L \sin^{-1}\left(\frac{\|x_i - x_j\|}{2}\right) \leq \tilde{L} \|x_i - x_j\|$$

for all $i, j \in \{0, 1, 2\}$. That is, we also have discrete Lipschitz bounds on this triangle.

For $x \in t$, we can express $w$ as

$$w(x) = w(x_0) + q \cdot (x - x_0)$$

where $q$ is in the space spanned by $x_1 - x_0$ and $x_2 - x_0$; that is,

$$q = q_1(x_1 - x_0) + q_2(x_2 - x_0)$$

for some $q_1, q_2 \in \mathbb{R}$. We also denote by $\theta$ the angle between $x_1 - x_0$ and $x_2 - x_0$. Note that $\theta \leq \gamma < \pi$ under Hypothesis 3.17.

Then at the vertices of $t$ we can write

$$w(x_i) = w(x_0) + q_i \|x_i - x_0\|^2 + q_j \|x_i - x_0\| \|x_j - x_0\| \cos\theta, \quad i, j \in \{1, 2\}, i \neq j.$$

Solving this system for the coefficients $q_1, q_2$, we find that

$$q_1 = \frac{(w(x_2) - w(x_0)) \|x_1 - x_0\| \cos\theta - (w(x_1) - w(x_0)) \|x_2 - x_0\|}{\|x_1 - x_0\|^2 \|x_2 - x_0\| (\cos^2\theta - 1)}.$$

Applying the discrete Lipschitz bound and since $\theta$ is the largest interior angle of the triangle $t$, we have $\frac{\pi}{3} \leq \theta \leq \gamma$, so

$$\begin{aligned}
|q_1| &\leq \frac{\tilde{L}\left(\|x_1 - x_0\| \|x_2 - x_0\| |\cos\theta| + \|x_1 - x_0\| \|x_2 - x_0\| |\cos\theta|\right)}{\|x_1 - x_0\|^2 \|x_2 - x_0\| (1 - \cos^2\theta)}, \\
&= \frac{\tilde{L}(\cos\theta + 1)}{\|x_1 - x_0\| (1 - \cos^2\theta)}, \\
&\leq \frac{\tilde{L}}{\|x_1 - x_0\| (1 - \cos\gamma)},
\end{aligned}$$

with a similar bound on $q_2$.

Combining these, we find that

$$|q| \leq |q_1| \, \|x_1 - x_0\| + |q_2| \, \|x_2 - x_0\| \leq \frac{2\tilde{L}}{1 - \cos\gamma}. \qquad \square$$

In particular, we can define $w^h : T^h \to \mathbb{R}$ as the unique piecewise linear interpolant of $u^h : \mathcal{G}^h \to \mathbb{R}$ that is linear on each triangle $t \in T^h$. Notice that $w^h$ satisfies the Lipschitz bounds of Lemma 3.26. This allows us to produce a Lipschitz continuous interpolant of $u^h$ on the sphere by means of the closest point projection $\mathrm{cp} : T^h \to \mathbb{S}^2$,

$$\mathrm{cp}(x) = \frac{x}{\|x\|}. \tag{3.35}$$

We remark that since $\mathrm{diam}(T^h) \to 0$, this is a bijection for small enough $h > 0$.

This leads to the following extension of $u^h$ onto the sphere:

$$u^h(x) = w^h(\mathrm{cp}^{-1}(x)). \tag{3.36}$$

That is, each triangle $t \in T^h$ is distorted to a spherical triangle (Figure 3.3). Importantly, this does not significantly distort the gradient of the underlying function values, and uniform Lipschitz bounds are preserved.

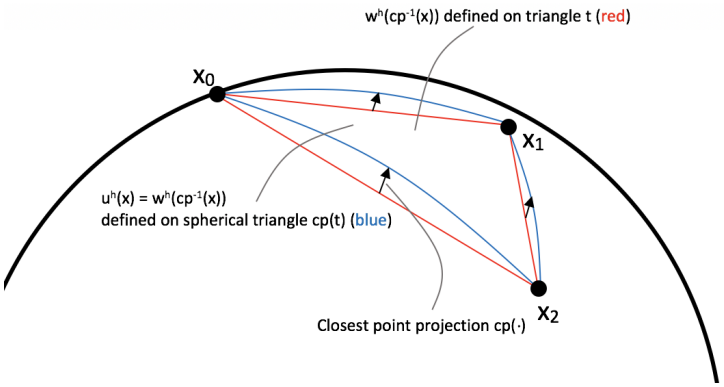



**Figure 3.3** Each triangle $t \in T^h$ is distorted via the inverse closest point map to a corresponding spherical triangle.

**Lemma 3.27** (Lipschitz bounds on the sphere)**.** *Let $u^h : \mathbb{S}^2 \to \mathbb{R}$ be as defined in Equation* (3.36)*. Under the hypotheses of either Lemma 3.21 or Lemma 3.22, there exists some $L > 0$ (independent of $h$) such that*

$$\left|u^h(x) - u^h(y)\right| \leq L d_{\mathbb{S}^2}(x, y)$$

*for all $x, y \in \mathbb{S}^2$.*

*Proof.* Let us first consider a fixed triangle $t \in T^h$ and choose any $x, y \in \mathbb{S}^2$ such that $\mathrm{cp}^{-1}(x), \mathrm{cp}^{-1}(y) \in t$. From Lemma 3.26, we can immediately see that there is some $L > 0$ (independent of $h$ and the particular choice of triangle) such that

$$\left|u^h(x) - u^h(y)\right| = \left|w^h(\mathrm{cp}^{-1}(x)) - w^h(\mathrm{cp}^{-1}(y))\right| \leq L \left\|\mathrm{cp}^{-1}(x) - \mathrm{cp}^{-1}(y)\right\|. \tag{3.37}$$

Now we choose a coordinate system such that the triangle $t$ lies in the plane $x_3 = c$. We recall that $\mathrm{diam}(t) \leq \mathrm{diam}(T^h) \to 0$ and the vertices of $t$ lie on the unit sphere $\mathbb{S}^2$. Thus there exists some $\eta = \mathcal{O}(\mathrm{diam}(T^h))$ such that

$$|z_1|, |z_2| \leq \eta, \quad 0 \leq 1 - z_3 = 1 - c \leq \eta \tag{3.38}$$

for any $z \in t$. (See also Figure 3.3).

In this coordinate system, we can express the closest point function and its inverse as

$$\mathrm{cp}(z) = \frac{(z_1, z_2, c)}{\sqrt{z_1^2 + z_2^2 + c^2}}, \quad \mathrm{cp}^{-1}(x) = \left(\frac{cx_1}{\sqrt{1 - x_1^2 - x_2^2}}, \frac{cx_2}{\sqrt{1 - x_1^2 - x_2^2}}, c\right).$$

Notice that we can interpret the first two components of $\mathrm{cp}^{-1}$ as a transformation from $t \in \mathbb{R}^2$ to $\mathbb{R}^2$. The Jacobian of this transformation is given by

$$\nabla \tilde{\mathrm{cp}}^{-1}(x) = \begin{pmatrix} c(1 - x_2^2)(1 - x_1^2 - x_2^2)^{-3/2} & cx_1x_2(1 - x_1^2 - x_2^2)^{-3/2} \\ cx_1x_2(1 - x_1^2 - x_2^2)^{-3/2} & c(1 - x_1^2)(1 - x_1^2 - x_2^2)^{-3/2} \end{pmatrix},$$

which converges uniformly to the identity matrix as $h \to 0$ given the estimates on the values of $(x_1, x_2) \in t$ from Equation (3.38). Thus for sufficiently small $h > 0$, we have that $\|\nabla \tilde{\text{cp}}^{-1}(x)\| \leq 2$ for all $(x_1, x_2) \in t$.

This leads to a uniform Lipschitz bound on the inverse closest point map, interpreted as a function on $\mathbb{R}^2$. For $x, y \in \mathbb{S}^2$ and sufficiently small $h > 0$ we then obtain the estimates:

$$\begin{aligned}
\left\|\text{cp}^{-1}(x) - \text{cp}^{-1}(y)\right\| &= \left\|\tilde{\text{cp}}^{-1}(x) - \tilde{\text{cp}}^{-1}(y)\right\|, \\
&\leq 2 \left\|(x_1, x_2) - (y_1, y_2)\right\|, \\
&\leq 2 \left\|x - y\right\|, \\
&\leq 4 d_{\mathbb{S}^2}(x, y).
\end{aligned}$$

Substituting this into Equation (3.37) yields the desired uniform Lipschitz bounds on any spherical triangle $\text{cp}(t)$.

Since $u^h$ is continuous, its Lipschitz constant is the maximal Lipchitz constant over any spherical triangle, which can be bounded independent of $h$. □

### 3.3.4 Convergence Theorem

We are now prepared to complete the proof of convergence of the numerical approach outlined in subsubsection 3.3.1. We begin with two lemmas pertaining to uniformly convergent sequences $u^{h_n}$.

**Lemma 3.28.** *Let $u^h$ be defined by the schemes in Equations* (4.24)-(3.28) *and* (3.36) *under the hypotheses of either Lemma 3.21 or 3.22. Suppose that $h_n \to 0$ is any sequence such that $u^{h_n}$ converges uniformly to a continuous function $U$. Then $\langle U \rangle = 0$.*

*Proof.* We recall first that $A^{h_n}(u^{h_n}) = 0$ by design (see Equation (3.28)). Since $A^h$ is consistent on all Lipschitz functions and $u^{h_n}$ enjoy uniform Lipschitz bounds,

we can also say that

$$\left|\langle u^{h_n}\rangle\right| = \left|\langle u^{h_n}\rangle - A^{h_n}(u^{h_n})\right| \leq \tau(h_n).$$

Since convergence is uniform, the Dominated Convergence Theorem yields

$$\langle U\rangle = \lim_{n\to\infty} \langle u^{h_n}\rangle = 0. \qquad \square$$

**Lemma 3.29.** *Let $u^h$ be defined by the schemes in Equations* (4.24)-(3.28) *and* (3.36) *under the hypotheses of either Lemma 3.21 or 3.22. Suppose that $h_n \to 0$ is any sequence such that $u^{h_n}$ converges uniformly to a continuous function $U$. Then $U$ is a viscosity solution of Equation* (3.16).

*Proof.* Here we follow the usual approach of the Barles-Souganidis framework, modified for the setting where the limit function is known to be continuous. Recall that $u^h$ satisfies the scheme

$$G^h(x, u(x) - u(\cdot)) + \tau(h)u^h(x) + \sigma(h) = 0,$$

where $\sigma(h) \to 0$ as $h \to 0$ (Lemma 3.23). Moreover, there exists some $M \in \mathbb{R}$ such that $\|u^h\|_\infty \leq M$ for all sufficiently small $h > 0$ (Lemmas 3.21-3.22).

Consider any $x_0 \in \mathbb{S}^2$ and $\phi \in C^\infty$ such that $U - \phi$ has a strict local maximum at $x_0$ with $U(x_0) = \phi(x_0)$. Because $u^h$ and the limit function $U$ are continuous, strict maxima are stable and there exists a sequence $z_n \in \mathcal{G}^h \cap \mathbb{S}^2$ such that

$$z_n \to x_0, \quad u^{h_n}(z_n) \to U(x_0),$$

where $z_n$ maximizes $u^{h_n}(x) - \phi(x)$ over points $x \in \mathcal{G}^h \cap \mathbb{S}^2$.

From the definition of $z_n$ as a maximizer of $u^{h_n} - \phi$, we also observe that

$$u^{h_n}(z_n) - u^{h_n}(\cdot) \geq \phi(z_n) - \phi(\cdot).$$

We let $G(\nabla u(x), D^2u(x))$ denote the PDE operator in Equation (3.16). Since $u^{h_n}$ is a solution of the scheme, we can use monotonicity to calculate

$$\begin{aligned} 0 &= G^{h_n}(z_n, u^{h_n}(z_n) - u^{h_n}(\cdot)) + \tau(h_n)u^{h_n}(z_n) + \sigma(h_n), \\ &\geq G^{h_n}(z_n, \phi(z_n) - \phi(\cdot)) - M\tau(h_n) + \sigma(h_n). \end{aligned}$$

As the scheme is consistent, we conclude that

$$\begin{aligned} 0 &\geq \liminf_{n\to\infty} \left(G^{h_n}(z_n, \phi(z_n) - \phi(\cdot)) - M\tau(h_n) + \sigma(h_n)\right), \\ &\geq G_*(x_0, \nabla\phi(x_0), D^2\phi(x_0)). \end{aligned}$$

Thus $U$ is a subsolution of Equation (3.16). An identical argument shows that $U$ is a supersolution and therefore a viscosity solution. □

These lemmas lead immediately to our main convergence theorem. The requirements on the schemes for the smooth and non-smooth setting are slightly different, but the proofs of the following two theorems are the same.

**Theorem 3.30** (Convergence (smooth case))**.** *Consider the schemes in Equations* (4.24)-(3.28) *and* (3.36) *under the conditions of Hypotheses 2.2, 3.17, and 3.18. Suppose also that Equation* (3.16) *has a unique mean-zero* $C^{0,1}$ *solution. Then* $u^h$ *converges uniformly to the unique smooth solution of Equation* (2.14)*.*

**Theorem 3.31** (Convergence (non-smooth case))**.** *Consider the schemes in Equations* (4.24)-(3.28) *and* (3.36) *under the conditions of Hypotheses 2.3, 3.17, and 3.18. Suppose also that* $F^h$ *is an underestimating scheme and that Equation* (3.16) *has a unique mean-zero* $C^{0,1}$ *solution. Then* $u^h$ *converges uniformly to the unique Lipschitz continuous solution of Equation* (2.14)*.*

*Proof.* Consider any sequence $h_n \to 0$. Notice that the function $u^{h_n}$ is uniformly bounded (Lemmas 3.21-3.22) and enjoys uniform Lipschitz bounds (Lemma 3.27). Then by the Arzelà-Ascoli theorem there exists a subsequence $h_{n_k}$ and a contin-

uous function $U$ such that $u^{h_{n_k}}$ converges uniformly to $U$, where $U$ has Lipschitz constant $L$.

From Lemmas 3.28-3.29, $U$ is a mean-zero viscosity solution of Equation (3.16). Then by Theorems 3.7 and 3.11, $U$ must agree with the unique mean-zero solution $u$ of Equation (2.14). Since this holds for any sequence $h_n$, we conclude that $u^h$ converges uniformly to $u$. □

# CHAPTER 4

# CONSTRUCTION OF A CONVERGENT SCHEME

## 4.1 Introduction

In this chapter, we produce the first convergent PDE-based finite-difference method for solving the Optimal Transport problem on the sphere. One of the benefits of using finite-difference schemes is that it is relatively straightforward to build higher-order finite-difference schemes by simply Taylor expanding to higher order. In this chapter, however, we do not expand to high order since we are more concerned with producing monotone schemes. Nevertheless, we keep the benefits of flexibility of finite-difference methods in mind as we desire to build schemes that have faster convergence rates Chapter 8. Finally, as emphasized in the introduction of Chapter 3, the philosophy underlying the construction of the finite-difference scheme in this chapter could be easily adapted to other cost functions and other PDE on the sphere.

The finite-difference method presented here is based on an approximation of a Generated Jacobian equation on local tangent planes, using geodesic normal coordinates, as mentioned in Chapter 3 and is largely the method presented in Hamfeldt and Turnquist (2021b), but there are parts of this chapter that are derived from the paper Hamfeldt and Turnquist (2021c). As shown in Theorems 3.30 and 3.31, one vital property in the convergence theorem is monotonicity. A complicating factor for achieving monotonicity is the presence of gradient terms that are mixed with the Hessian terms inside of a nonlinear operator. This requires the introduction of new techniques for approximating both first- and second-order terms in order to preserve both the consistency and the monotonicity of our scheme. We produce an implementation and present computational results that demonstrate the success of this method for both the squared geodesic cost and the logarithmic cost. The method will be shown to handle both structured and unstructured grids, non-smooth data, and validate the constant solution.

In Chapter 5 and Chapter 6, we will show computations that are made to demonstrate the effectiveness of the method specifically for the reflector antenna and moving mesh problems, respectively.

## 4.2 Simple Reformulation of Some Terms in the PDE

### 4.2.1 Variational Formulation of the Determinant of a Hessian

Since our PDE involves computing the determinant of a Hessian, here we show how to reformulate the PDE in order to build a monotone discretization of the eigenvalues of the Hessian. As utilized in Froese and Oberman (2011a), Hadamard's inequality allows the determinant of a positive definite matrix $M$ to be computed via the minimization problem

$$\det M = \min_{v_i \in V} \prod_{i=1}^{d} v_i^T M v_i, \tag{4.1}$$

where $V$ is the set of all orthogonal bases for $\mathbb{R}^d$.

We require a formulation satisfying Equation (3.13), which modifies this formulation to ensure that no negative terms appear when the symmetric matrix $M$ is not positive definite. A simple approach is to use

$$\det{}^+ M = \min_{v_i \in V} \prod_{i=1}^{d} \max\left\{v_i^T M v_i, 0\right\}. \tag{4.2}$$

If our matrix $M = D^2\phi(x)$ is a Hessian matrix, this becomes:

$$\begin{aligned}
\det{}^+(D^2\phi(x)) &= \min_{\nu_1,\nu_2 \in V} \left\{ \prod_{j=1}^{2} \nu_j^T (D^2\phi(x))\nu_j, 0 \right\}, \\
&= \min_{\nu_1,\nu_2 \in V} \prod_{j=1}^{2} \max\left\{ \frac{\partial^2 \phi}{\partial \nu_j^2}, 0 \right\}.
\end{aligned}$$

In particular, this allows us to replace the determinant in Equation (2.14) with

$$\det{}^{+}(D^2u(x)+A(x,\nabla u(x))) = \min_{\nu_1,\nu_2\in V}\prod_{j=1}^{2}\max\left\{\frac{\partial^2 u(x)}{\partial\nu_j^2}+\frac{\partial^2 c(x,y)}{\partial\nu_j^2}\bigg|_{y=T(x,\nabla u(x))},0\right\}. \tag{4.3}$$

### 4.2.2 Mixed Hessian

The framework developed in Hamfeldt and Turnquist (2021a) and shown in Chapter 3 only requires the construction of consistent approximations of derivatives with respect to $x$, expressed in the geodesic normal coordinate system. For the mixed Hessian term, $\det D^2_{xy}c(x,y)$ it is not immediately clear how to do this. In fact, the relative simplicity of the notation obfuscates the actual complexity of the object. In Loeper (2009), a clearer representation of this quantity in curved geometries is presented.

We recall that the optimal map $T(x,p)$ (also known as the $c$-exponential map) satisfies Equation (2.15) and can be constructed explicitly for the cost function of interest to us via Equations (3.6)-(3.7). Then from Loeper (2009), we know that, furthermore, the mixed Hessian satisfies

$$[D^2_{xy}c]^{-1} = -D_pT(x,p)|_{x,p=-\nabla_x c(x,y)}. \tag{4.4}$$

This representation formula from Equation (4.4) shows that the inverse of the mixed Hessian is simply the Jacobian of the map $T(x,p)$ with respect to $p$. Since we will be taking the determinant, this is actually a change of area formula for the transformation $T(x,p)$. See Figure 4.1.

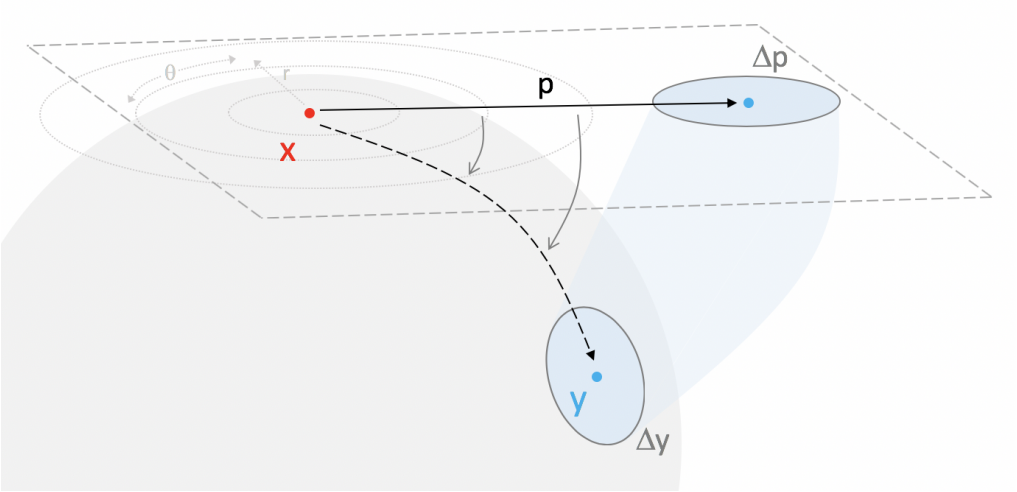



**Figure 4.1** The determinant of the mixed Hessian is the change in area formula from the set $T(x, E)$ on the sphere sphere to the set $E$ on the local tangent plane.

We will compute this change of area by computing the linear differential map $dT$ for both cost functions. That is, the change in area will be computed using orthogonal perturbations $\Delta p_1$, $\Delta p_2$ (so $\Delta p_1 \cdot \Delta p_2 = 0$) in the tangent plane $\mathcal{T}_x$.

$$|\det\ (D_p T(x,p))| = \left| \lim_{\|\Delta p_1\| \to 0} \frac{T(x, p+\Delta p_1) - T(x,p)}{\|\Delta p_1\|} \times \lim_{\|\Delta p_2\| \to 0} \frac{T(x, p+\Delta p_2) - T(x,p)}{\|\Delta p_2\|} \right|.$$

In general, the area element on a manifold is a function of the wedge product of two covectors, which need not be orthogonal. This has the interpretation of the area of a parallelogram on the manifold. However, in the special case where the vectors are orthogonal (or are orthogonal to leading order), this reduces to an ordinary product. This is indeed the case for both the squared geodesic and logarithmic cost functions. Thus the change of area formula reduces to the simpler expression

$$|\det\ (D_p T(x,p))| = \lim_{\|\Delta p_1\|, \|\Delta p_2\| \to 0, \Delta p_1 \cdot \Delta p_2 = 0} \frac{d_{\mathbb{S}^2}\left(T(x,p), T(x, p+\Delta p_1)\right) d_{\mathbb{S}^2}\left(T(x,p), T(x, p+\Delta p_2)\right)}{\|\Delta p_1\|\ \|\Delta p_2\|}$$

and the determinant of the mixed Hessian is given by

$$\left|\det(D^2_{xy}c(x,y))\right| = \left.\frac{1}{|D_pT(x,p)|}\right|_{p=-\nabla_x c(x,y)}. \tag{4.5}$$

This can be computed explicitly for both the squared geodesic cost,

$$\left|\det(D^2_{xy}c(x,y))\right| = \frac{\|p\|}{\sin\|p\|}, \tag{4.6}$$

and for the logarithmic cost,

$$\left|\det(D^2_{xy}c(x,y))\right| = \left(\|p\|^2 + 1/4\right)^2. \tag{4.7}$$

See Appendix D for details.

We note also that these formulas coincide with the formulas that can be derived via standard change of variables formulas by requiring

$$\int_{T(x,E)} dS = \int_E |\det(D_pT(x,p))|\, dp$$

for every measurable $E \in \mathcal{T}(x)$.

## 4.3 Numerical Method

We now explain how we actually construct an approximation scheme for Equation (2.14) at a point $x \in \mathbb{S}^2$.

### 4.3.1 Construction of Finite Difference Stencils

We begin with a point cloud $\mathcal{G} \in \mathbb{S}^2$ that discretizes the sphere; we assume only the minimal regularity required by Hypothesis 3.17. We begin by considering a fixed point $x_i \in \mathcal{G}$ and establishing a computational neighborhood $N(i)$ about this point. For any fixed $C > 0$, we define

$$N(i) = \{j \mid x_j \in \mathcal{G},\, d_{\mathbb{S}^2}(x_i, x_j) \le C\sqrt{h}\}. \tag{4.8}$$

Once $N(i)$ is established, the points $x_j \in N(i)$ are projected on to the local tangent plane $\mathcal{T}_{x_i}$ via the geodesic normal coordinate projection in Equation (3.10):

$$z_j = v_{x_i}(x_j).$$

We denote the resulting point cloud on the tangent plane by

$$\mathcal{N}(i) = \{z_j \mid j \in N(i)\} \subset \mathcal{T}_{x_i}.$$

Figure 4.2 shows an exaggerated example of this tangent plane projection.

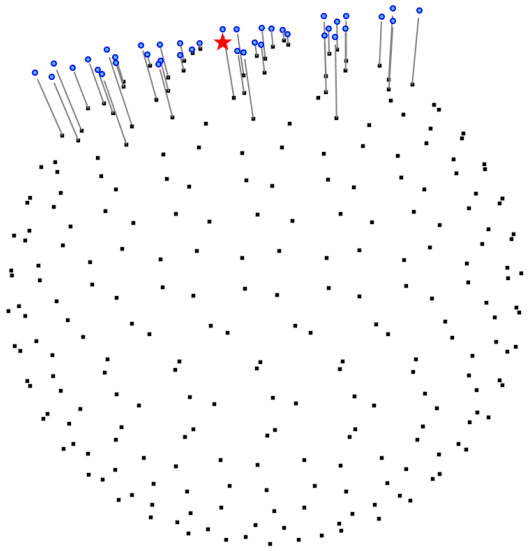


**Figure 4.2** Projection onto the tangent plane via geodesic normal coordinates. The points in $\mathcal{G}$ are indicated with small black squares, the computational point $x_i$ is indicated as a red star, and the projection of the neighborhood $N(i)$ to the local tangent plane is indicated with blue circles.

Next we suppose that we are interested in resolving behavior along some direction $\nu \in \mathbb{R}^2$. Following a slight modification of Froese (2018), we select four points $x_{\nu,j} \in \mathcal{N}(i)$ that are well-aligned with the direction $\nu$. See Figure 4.3. We introduce the following notation:

- $r = C\sqrt{h}$ is the search radius used to define the neighborhood $N(i)$. For all $x_{\nu,j} \in \mathcal{N}(i)$ we have that $\|x_{\nu,j} - x_i\| \leq r$.
- $\theta_{\nu,j}$ is the angle between $x_{\nu,j} - x_i$ and the direction $\nu$.

- $d\theta_{\nu,j}$ is the minimal absolute angle between $x_{\nu,j} - x_i$ and the direction $\nu$.
- $d\theta$ is the overall angular resolution of the stencil, which is related to the search radius through $d\theta = \dfrac{2h}{r} + \mathcal{O}(h) = \mathcal{O}(\sqrt{h})$; see Froese (2018).
- Each $x_{\nu,j}$ can be represented in polar coordinates as $(h_{\nu,j}, \theta_{\nu,j})$ using the coordinate system where $x_i$ is the origin and the coordinate directions $\nu, \nu^\perp$ are orthogonal.
- Components of $x_{\nu,j}$ can be expressed using the shorthand notation

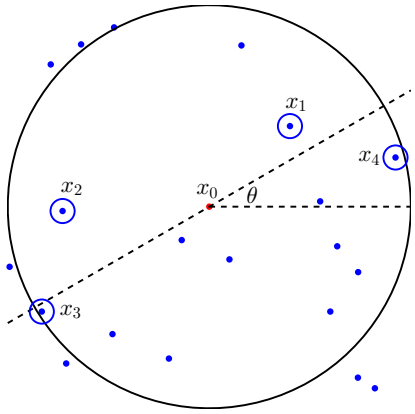
$$C_{\nu,j} = h_{\nu,j} \cos\theta_{\nu,j}, \quad S_{\nu,j} = h_{\nu,j} \sin\theta_{\nu,j}.$$

The following lemma follows immediately from the proof of (Froese, 2018, Lemma 11). Figure 4.3 illustrates the four small balls where each of these four neighbors is required to exist.

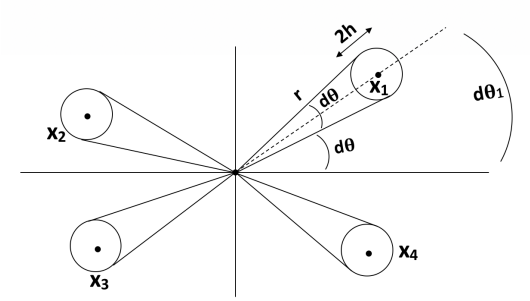
**Lemma 4.1** (Properties of neighbors)**.** *For every $x_i \in \mathcal{G}$ and $\nu \in \mathbb{R}^2$, four neighbors $x_{\nu,j} \in \mathcal{N}(i)$ exist satisfying the following properties:*

- *$x_{\nu,j}$ resides in the $j$th quadrant.*
- *The angular component of $x_{\nu,j}$ satisfies $d\theta \le d\theta_{\nu,j} \le 2d\theta$.*
- *The radial component of $x_{\nu,j}$ satisfies $r - 2h \le h_{\nu,j} \le r$.*

These additional requirements on stencil will be critical to developing monotone approximations of functions of the gradient.




**(a)** The computational points are selected from a neighborhood.




**(b)** The points selected must satisfy an approximate symmetry condition about the computational direction.

**Figure 4.3** Choice of computational points $x_{\nu,j}$ in the local tangent plane. Subfigure 4.3a shows the selection of points aligning with the computational direction from a neighborhood. Subfigure 4.3b shows that the points selected need to satisfy a kind of symmetric balance about the computational direction.

### 4.3.2 Approximation of Second Derivatives

Our overall approximation of Equation (2.14) will hinge on the construction of (negative) monotone schemes for second directional derivatives $\dfrac{\partial^2 \phi}{\partial \nu^2}$. We introduce approximations of the form

$$\mathcal{D}_{\nu\nu}\phi(x_i) = \sum_{j=1}^{4} a_{\nu,j} \left(\phi(x_{\nu,j}) - \phi(x_i)\right). \tag{4.9}$$

As in Froese (2018), consistency and negative monotonicity can be achieved by finding a solution of the system

$$\begin{cases} \sum a_{\nu,j} h_{\nu,j} \cos\theta_{\nu,j} = 0, \\ \sum a_{\nu,j} h_{\nu,j} \sin\theta_{\nu,j} = 0, \\ \sum \frac{1}{2} a_{\nu,j} h_{\nu,j}^2 \cos^2\theta_{\nu,j} = 1, \\ a_{\nu,j} \geq 0. \end{cases} \tag{4.10}$$

An explicit solution is given by

$$\begin{aligned} a_{\nu,1} &= \frac{2S_{\nu,4}(C_{\nu,3}S_{\nu,2} - C_{\nu,2}S_{\nu,3})}{\det(A)}, \\ a_{\nu,2} &= \frac{2S_{\nu,3}(C_{\nu,1}S_{\nu,4} - C_{\nu,4}S_{\nu,1})}{\det(A)}, \\ a_{\nu,3} &= \frac{-2S_{\nu,2}(C_{\nu,1}S_{\nu,4} - C_{\nu,4}S_{\nu,1})}{\det(A)}, \\ a_{\nu,4} &= \frac{-2S_{\nu,1}(C_{\nu,3}S_{\nu,2} - C_{\nu,2}S_{\nu,3})}{\det(A)}, \end{aligned} \tag{4.11}$$

where

$$\begin{aligned} \det(A) =&(C_{\nu,3}S_{\nu,2} - C_{\nu,2}S_{\nu,3})(C_{\nu,1}^2S_{\nu,4} - C_{\nu,4}^2S_{\nu,1}), \\ &- (C_{\nu,1}S_{\nu,4} - C_{\nu,4}S_{\nu,1})(C_{\nu,3}^2S_{\nu,2} - C_{\nu,2}^2S_{\nu,3}). \end{aligned} \tag{4.12}$$

### 4.3.3 Approximation of Functions of the Gradient

Equation (2.14) involves several terms of the form $g(\nabla u)$. Importantly, either automatically or through appropriate regularization (see Section 3.2.4), each of these functions $g$ has a bounded Lipschitz constant $L_g$. This allows us to pursue a generalized Lax-Friedrichs type discretization of the form

$$\begin{aligned} \tilde{g}^{\pm}(\mathcal{D}\phi(x_i)) =& g\left(\sum_{\nu\in\{(1,0),(0,1)\}} \nu \sum_{j=1}^{4} b_{\nu,j}\left(\phi(x_{\nu,j}) - \phi(x_i)\right)\right), \\ &\mp \epsilon_g \sum_{\nu\in\{(1,0),(0,1)\}} \sum_{j=1}^{4} a_{\nu,j}\left(\phi(x_{\nu,j}) - \phi(x_i)\right). \end{aligned} \tag{4.13}$$

Above, the coefficients $a_{\nu,j}$ are identical to the coefficients that arise in the approximation of second directional derivatives. This introduces a Laplacian regularization term, which is carefully chosen to enforce monotonicity (or negative monotonicity) even if the coefficients $b_{\nu,j}$ do not on their own produce a monotone scheme.

We first require coefficients $b_{\nu,j}$ that ensure that

$$\mathcal{D}_\nu \phi(x_i) = \sum_{j=1}^{4} b_{\nu,j} \left( \phi(x_{\nu,j}) - \phi(x_i) \right)$$

is a consistent approximation of the first directional derivative $\dfrac{\partial \phi(x_i)}{\partial \nu}$. Taylor expanding, we obtain

$$\mathcal{D}_\nu \phi(x_i) = \sum_{j=1}^{4} b_{\nu,j} \left( h_{\nu,j} \cos\theta_{\nu,j} \frac{\partial \phi(x_i)}{\partial \nu} + h_{\nu,j} \sin\theta_{\nu,j} \frac{\partial \phi(x_i)}{\partial \nu^\perp} + \mathcal{O}(r^2) \right).$$

Consistency then requires a solution of the system

$$\begin{cases} \sum b_{\nu,j} h_{\nu,j} \cos\theta_{\nu,j} = 1, \\ \sum b_{\nu,j} h_{\nu,j} \sin\theta_{\nu,j} = 0. \end{cases} \tag{4.14}$$

An explicit solution is

$$\begin{aligned} b_{\nu,1} &= \frac{S_{\nu,4}(S_{\nu,3}C_{\nu,2}^2 - S_{\nu,2}C_{\nu,3}^2)}{\det(A)}, \\ b_{\nu,2} &= -\frac{S_{\nu,3}(S_{\nu,4}C_{\nu,1}^2 - S_{\nu,1}C_{\nu,4}^2)}{\det(A)}, \\ b_{\nu,3} &= \frac{S_{\nu,2}(S_{\nu,4}C_{\nu,1}^2 - S_{\nu,1}C_{\nu,4}^2)}{\det(A)}, \\ b_{\nu,4} &= -\frac{S_{\nu,1}(S_{\nu,3}C_{\nu,2}^2 - S_{\nu,2}C_{\nu,3}^2)}{\det(A)}, \end{aligned} \tag{4.15}$$

where $\det(A)$ is again given by Equation (4.12).

We then substitute these coefficients into Equation (4.13) and define a regularization factor satisfying

$$\epsilon_g = \max \left\{ \frac{L_g \left| b_{\nu,j} \right|}{a_{\nu,j}} \mid j \in \{1,2,3,4\},\, \nu \in \{(0,1),(1,0)\} \right\}. \tag{4.16}$$

We will verify that this is finite and bounded in Section 4.4.

### 4.3.4 Approximation of the Nonlinear Operator

We now have the building blocks in place to describe a discretization of the full nonlinear operator (Equation (2.14)) at the point $x_i \in \mathcal{G}$.

The variational formulation of the modified determinant (Equation (4.2)) requires performing a minimization over the set $V$ of orthogonal bases for $\mathbb{R}^2$. In the discrete version, we consider a finite subset of $V$ that ensures that all directions are resolved in the limit $h \to 0$. A simple choice is given by

$$\tilde{V} = \left\{ (\cos\theta, \sin\theta) \mid \theta = jd\theta,\ j = 0, \dots, \frac{\pi}{2d\theta} \right\}, \tag{4.17}$$

where $d\theta = \mathcal{O}\left(\sqrt{h}\right)$ is the angular resolution of the stencil described in Section 4.3.1.

The PDE involves several different functions of the gradient. For compactness, we introduce the shorthand notation

$$g_{1,\nu}(x_i, p) = \mathcal{D}_{\nu\nu}\tilde{c}(x_i, T(x_i, p)), \tag{4.18}$$

where the differencing is performed only in the first argument of $\tilde{c}$. This involves the explicit formulas for the optimal map given in Equation (3.6)-(3.7).

We also define

$$g_2(x_i, p) = \frac{\left|\det D^2_{xy} c\left(x_i, T(x_i, p)\right)\right|}{f_1\left(T(x_i, p)\right)}, \tag{4.19}$$

recalling that the determinant of the mixed Hessian can be replaced with the simple explicit representations obtained in Section 4.2.2, which we here denote by $H(p)$.

The discretization of these functions of the gradient require a regularization parameter (see Equation (4.16)), which involves Lipschitz bounds on these

functions. We select bounds satisfying

$$L_{g_1} > L_{\tilde{c}} L_T \tag{4.20}$$

and

$$L_{g_2} \geq \left\| \frac{1}{f_1} \right\|^2 \left( \|f_1\| L_H + \|H\| L_{f_1} L_T \right), \tag{4.21}$$

where $L_T$ is the Lipschitz constant of the optimal map $T(x,p)$ with respect to the variable $p$, $L_H$ and $L_{f_1}$ are the Lipschitz constants of the functions $H$ and $f_1$ respectively, and

$$L_{\tilde{c}} = \max_{|\nu|=1, x,y \in \mathbb{S}^2} \left\| \nabla_y \left( \nu^T D^2_{xx} \tilde{c}(x,y) \nu \right) \right\|. \tag{4.22}$$

We remark that while these constants $L_{g_1}$ and $L_{g_2}$ depend on the problem data and particular cost function, they are all guaranteed to be bounded under the assumptions of Hypothesis 2.2 and can be computed explicitly using the formulas in Equations (3.6),(3.7), (3.19), (4.6), and (4.7).

We can then write down the full discretization of Equation (2.14) as

$$\begin{aligned} F^h(x, u(x) - u(\cdot)) = - \min_{(\nu_1,\nu_2) \in \tilde{V}} \prod_{j=1}^{2} \max \left\{ \mathcal{D}_{\nu_j \nu_j} u(x_i) + \tilde{g}^-_{1,\nu_j}(\mathcal{D}u(x_i)), 0 \right\} \\ + f_0(x_i) \tilde{g}_2^+ \left( x_i, \mathcal{D}u(x_i) \right). \end{aligned} \tag{4.23}$$

### 4.3.5 Solution Method

In order to efficiently obtain a convergent approximation to Equation (2.14), we will slightly modify the two-step procedure described in Equation (4.24)-(3.28). We propose instead the following solution and verification process, which is equivalent.

1. Solve the discrete system

$$F^h(x, v^h(x) - v^h(\cdot)) + \sqrt{h} v^h(x) = 0, \quad x \in \mathcal{G} \tag{4.24}$$

for the grid function $v^h$.

2. Verify that the grid function $v^h$ satisfies the bounds

$$E^h(x, v^h(x) - v^h(\cdot)) \leq R, \quad x \in \mathcal{G}.$$

3. If the verification step fails, redefine $v^h$ by solving the modified system (Equation (4.24))

$$G^h\left(x, v^h(x) - v^h(\cdot)\right) + \sqrt{h}v^h(x) = 0, \quad x \in \mathcal{G}$$

using the solution obtained in Step 1 as an initial guess.

4. Define the discrete solution

$$u^h(x) = v^h(x) - v^h(x_h^*), \quad x \in \mathcal{G}. \tag{4.25}$$

We note that in practice, we have never found Step 3 above to be necessary. Thus although this procedure appears to be longer than simply performing Steps 3-4, it actually allows us to obtain the same solution by solving a simpler system.

The strong nonlinearity in Equation (2.14), particularly in that it involves nonlinear gradient terms that have very little required structure, makes the construction of a nonlinear Gauss-Jacobi, algebraic multigrid, and/or approximate Newton-type method highly nontrivial. In the present work, we perform all our computations using explicit parabolic schemes of the form

$$v_{n+1}^h(x_i) = v_n^h(x_i) - \Delta t F^h(v_n^h(x_i), v_n^h(x_i) - v_n^h(\cdot)).$$

As discussed in Oberman (2006), $\Delta t$ has to satisfy a nonlinear CFL condition in order to guarantee convergence. In particular, we require $\Delta t < 1/L_{F^h}$, where $L_{F^h}$ is the Lipschitz constant of $F^h$ with respect to the arguments $u_i^h$. This Lipschitz constant scales like $L_{F^h} = \mathcal{O}(h^{-2})$ and can either be determined explicitly *a priori* or adaptively under a requirement that the residual should decrease.

In some cases, acceleration of this process is possible using the approach of Schaeffer and Hou (2016). The accelerated parabolic scheme is as follows. Given $v_0^h, v_1^h = v_0^h - \Delta t F^h\left[v_0^h\right]$ and parameters $\{\gamma_n\}$ where $\gamma_n - n/(n+n_0)$, where typically $n_0 \geq 10$:

**Algorithm 1** Accelerated Parabolic Scheme

1: **while** $\left\|v_n^h - v_{n-1}^h\right\|_\infty > \text{tol}$ **do**
2: $\quad v_{n,E}^h = v_n^h + \gamma_n \left(v_n^h - v_{n-1}^h\right)$
3: $\quad v_{n+1}^h = v_{n,E}^h - \Delta t F^h \left[v_{n,E}^h\right]$
4: **end while**

Faster solvers than the parabolic scheme for these kinds of systems are, of course, desirable and will be explored in future work.

### 4.4 Convergence

We are now prepared to prove that the numerical method defined by Equation (4.23) and the subsequent solution procedure converges.

**Theorem 4.2** (Convergence)**.** *Under the assumptions of Hypothesis 2.2, let* $u \in C^3(\mathbb{S}^2)$ *be the unique solution of Equation* (2.14) *satisfying* $u(x^*) = 0$*. Let* $\mathcal{G}^h$ *be a grid satisfying Hypothesis 3.17 and let* $F^h$ *be defined as in Equation* (4.23)*. Then for each sufficiently small* $h > 0$*, the grid function* $u^h$ *defined in Equation* (4.25) *is uniquely defined. Moreover,* $u^h$ *converges uniformly to* $u$ *as* $h \to 0$*.*

This result follows immediately from the framework described in Theorems 3.30 and 3.31 provided we can verify that our approximation scheme $F^h$ is consistent (Lemma 4.9) and monotone (Lemma 4.11). This will be accomplished in several lemmas throughout the remainder of this section.

#### 4.4.1 Bounds on Coefficients

We begin by demonstrating that the coefficients $a_{\nu,j}, b_{\nu,j}$ appearing in the approximation of the second directional derivatives can be bounded.

**Lemma 4.3** (Bounds on coefficients (second derivatives))**.** *There exists a constant* $C > 0$ *such that for all sufficiently small* $h > 0$ *and* $\nu \in \mathbb{R}^2$*, the coefficients defined by Equation* (4.11) *satisfy*

$$a_{\nu,j} \geq \frac{C}{h}.$$

*Proof.* We establish a bound for the coefficient $a_{\nu,1}$; the remaining coefficients are similar.

Recall the notation $C_{\nu,j} = h_{\nu,j}\cos\theta_{\nu,j}$, $S_{\nu,j} = h_{\nu,j}\sin\theta_{\nu,j}$. Since each $x_{\nu,j}$ lies in the $jth$ quadrant, each of these terms has a definite sign. Based on the requirements on $h_{\nu,j}, d\theta_{\nu,j}$ established in Lemma 4.1, we can record the asymptotic bounds

$$\begin{aligned}(r-2h)\left(1-\mathcal{O}(d\theta^2)\right) &\leq |C_{\nu,j}| \leq r,\\ (r-2h)\left(d\theta-\mathcal{O}(d\theta^3)\right) &\leq |S_{\nu,j}| \leq r\left(2d\theta+\mathcal{O}(d\theta^3)\right).\end{aligned}$$

We recall also that $r, d\theta = \mathcal{O}(\sqrt{h})$.

These observations allow us to establish the following bounds:

$$\begin{aligned}\det(A) =&(C_{\nu,3}S_{\nu,2}-C_{\nu,2}S_{\nu,3})(C_{\nu,1}^2S_{\nu,4}-C_{\nu,4}^2S_{\nu,1})\\ &-(C_{\nu,1}S_{\nu,4}-C_{\nu,4}S_{\nu,1})(C_{\nu,3}^2S_{\nu,2}-C_{\nu,2}^2S_{\nu,3}),\\ &\leq r^5(4d\theta+\mathcal{O}(d\theta^3))(4d\theta+\mathcal{O}(d\theta^3))+r^5(4d\theta+\mathcal{O}(d\theta^3))(4d\theta+\mathcal{O}(d\theta^3)),\\ &=32r^5d\theta^2+\mathcal{O}(h^{9/2})\end{aligned}$$

and

$$\begin{aligned}-2S_{\nu,4}(C_{\nu,2}S_{\nu,3}-C_{\nu,3}S_{\nu,2}) &\geq 2(r-2h)^3(d\theta-\mathcal{O}(d\theta^3))\left(2d\theta-\mathcal{O}(d\theta^3)\right),\\ &=4r^3d\theta^2+\mathcal{O}(h^3).\end{aligned}$$

Combining these bounds, we obtain

$$a_{\nu,1} \geq \frac{4r^3d\theta^2+\mathcal{O}(h^3)}{32r^5d\theta^2+\mathcal{O}(h^{9/2})} = \frac{1}{8r^2}\left(1+\mathcal{O}(\sqrt{h})\right),$$

where $r^2 = \mathcal{O}(h)$. □

**Lemma 4.4** (Bounds on coefficients (first derivatives))**.** *There exists a constant $C > 0$ such that for all sufficiently small $h > 0$ and $\nu \in \mathbb{R}^2$, the coefficients defined*

*by Equation* (4.14) *satisfy*

$$|b_{\nu,j}| \leq \frac{C}{\sqrt{h}}.$$

*Proof.* We proceed as in the proof of Lemma 4.3 and compute the bounds

$$\begin{aligned}
\det(A) =& (C_{\nu,3}S_{\nu,2} - C_{\nu,2}S_{\nu,3})(C_{\nu,1}^2 S_{\nu,4} - C_{\nu,4}^2 S_{\nu,1}) \\
&- (C_{\nu,1}S_{\nu,4} - C_{\nu,4}S_{\nu,1})(C_{\nu,3}^2 S_{\nu,2} - C_{\nu,2}^2 S_{\nu,3}), \\
&\geq 8(r-2h)^5 \left(1 - \mathcal{O}(d\theta^2)\right)^3 \left(d\theta - \mathcal{O}(d\theta^3)\right)^2, \\
&= 8r^5 d\theta^2 + \mathcal{O}(h^4)
\end{aligned}$$

and

$$0 \leq S_{\nu,4}(S_{\nu,3}C_{\nu,2}^2 - S_{\nu,2}C_{\nu,3}^2) \leq 2r^4(2d\theta + \mathcal{O}(d\theta^3))^2 = 8r^4 d\theta^2 + \mathcal{O}(h^4).$$

Combining these, we find that

$$0 \leq b_{\nu,1} \leq \frac{8r^4 d\theta^2 + \mathcal{O}(h^4)}{8r^5 d\theta^2 + \mathcal{O}(h^4)} = \frac{1}{r}\left(1 + \mathcal{O}(\sqrt{h})\right)$$

with $r = \mathcal{O}\left(\sqrt{h}\right)$.
The other coefficients are similar, though some are positive and some are negative.

□

### 4.4.2 Bounds on Lipschitz Constants

We next establish Lipschitz bounds on the functions $g_{1,\nu}, g_2$ defined in Equation (4.18) and Equation (4.19), which play an important role in the discretization of functions of the gradient.

**Lemma 4.5** (Lipschitz bound on $g_{1,\nu}$)**.** *Let $x_i \in \mathcal{G}$ be fixed. Then for $p \in \mathcal{T}_{x_i}$, the function*

$$g_{1,\nu}(p) = \mathcal{D}_{\nu\nu}\tilde{c}(x_i, T(x_i, p)).$$

*has a Lipschitz constant $L$ satisfying*

$$L \leq L_{\tilde{c}} L_T + \mathcal{O}(\sqrt{h}).$$

*Proof.* We first recall that $g_{1,\nu}(p)$ involves a finite difference discretization. By consistency, we have

$$g_{1,\nu}(p) = \nu^T \left( D^2_{xx} \tilde{c}(x_i, T(x_i, p)) \right) \nu + C(p)\sqrt{h},$$

where the coefficient $C(p)$ arising in the discretization error is at least Lipschitz continuous in $p$. Then using regularity, we can calculate

$$\begin{aligned} g_{1,\nu}(p) - g_{1,\nu}(q) &= \nu^T D^2_{xx} \left( \tilde{c}(x_i, T(x_i, p)) - \tilde{c}(x_i, T(x_i, q)) \right) \nu + (C(p) - C(q))\sqrt{h}, \\ &\leq \max_{|\nu|=1, x, y \in \mathbb{S}^2} \left\| \nabla_y \left( \nu^T D^2_{xx} \tilde{c}(x, y) \nu \right) \right\| \| T(x_i, p) - T(x_i, q) \| + \\ &\mathcal{O}(\sqrt{h}) \| p - q \|, \\ &\leq \left( L_{\tilde{c}} L_T + \mathcal{O}(\sqrt{h}) \right) \| p - q \|. \end{aligned}$$

□

**Lemma 4.6** (Lipschitz bound on $g_2$)**.** *Let $x_i \in \mathcal{G}$ be fixed. Then for $p \in \mathcal{T}_{x_i}$, the function*

$$g_2(p) = \frac{\left| \det D^2_{xy} \tilde{c}(x_i, T(x_i, p)) \right|}{f_1(T(x_i, p))}$$

*has a Lipschitz constant $L$ satisfying*

$$L \leq \left\| \frac{1}{f_1} \right\|^2 \left( \| f_1 \| L_H + \| H \| L_{f_1} L_T \right).$$

*Proof.* Using the notation $H(p) = \left|\det D^2_{xy}\tilde{c}(x_i, T(x_i, p))\right|$, we have

$$\begin{aligned}
g_2(p) - g_2(q) &= \frac{H(p)}{f_1(T(x_i,p))} - \frac{H(q)}{f_1(T(x_i,q))}, \\
&= \frac{f_1(T(x_i,q))H(p) - f_1(T(x_i,p))H(q)}{f_1(T(x_i,p))f_1(T(x_i,q))}, \\
&= \frac{f_1(T(x_i,q))(H(p)-H(q)) + H(q)(f_1(T(x_i,q)) - f_1(T(x_i,p)))}{f_1(T(x_i,p))f_1(T(x_i,q))}, \\
&\leq \left\|\frac{1}{f_1}\right\|^2 (\|f_1\|L_H + \|H\|L_{f_1}L_T)\,\|p-q\|. \qquad \square
\end{aligned}$$

### 4.4.3 Lax-Friedrichs Approximations

We now verify that the Lax-Friedrichs type approximations for functions of the gradient defined in Equation (4.13) are both consistent and monotone.

**Lemma 4.7** (Consistency of functions of gradient)**.** *Let $g$ be Lipschitz continuous with Lipschitz constant $L_g$ and $\phi \in C^2$. Then*

$$\begin{aligned}
\tilde{g}^{\pm}(\mathcal{D}\phi(x_i)) =& g\left(\sum_{\nu\in\{(1,0),(0,1)\}} \nu \sum_{j=1}^{4} b_{\nu,j}\left(\phi(x_{\nu,j}) - \phi(x_i)\right)\right) \\
&\mp \epsilon_g \sum_{\nu\in\{(1,0),(0,1)\}} \sum_{j=1}^{4} a_{\nu,j}\left(\phi(x_{\nu,j}) - \phi(x_i)\right).
\end{aligned}$$

*is a consistent approximation of $g(\nabla\phi)$.*

*Proof.* We note that by construction, we have

$$\lim_{h\to 0} \sum_{\nu\in\{(1,0),(0,1)\}} \nu \sum_{j=1}^{4} b_{\nu,j}\left(\phi(x_{\nu,j}) - \phi(x_i)\right) = \nabla\phi(x_i)$$

and

$$\lim_{h\to 0} \sum_{\nu\in\{(1,0),(0,1)\}} \sum_{j=1}^{4} a_{\nu,j}\left(\phi(x_{\nu,j}) - \phi(x_i)\right) = \Delta\phi(x_i).$$

Using Lemmas 4.3-4.4, we can also bound $\epsilon_g$ via

$$\begin{aligned}\epsilon_g &= \max\left\{ \frac{L_g\,|b_{\nu,j}|}{a_{\nu,j}} \mid j \in \{1,2,3,4\},\, \nu \in \{(0,1),(1,0)\} \right\}, \\ &\leq \frac{L_g(C_b/\sqrt{h})}{C_a/h}, \\ &= C\sqrt{h}\end{aligned}$$

where the constant $C$ is independent of $h$.

Combining these results, we obtain

$$\lim_{h\to 0} \tilde{g}^{\pm}(\mathcal{D}\phi(x_i)) = g(\nabla\phi(x_i)),$$

with a truncation error of $\mathcal{O}(\sqrt{h})$. □

**Lemma 4.8** (Monotonicity of functions of the gradient)**.** *Let $g$ be Lipschitz continuous with Lipschitz constant $L_g$ and $\phi \in C^2$. Then the schemes $\tilde{g}^{+}(\mathcal{D}\phi(x_i))$ and $\tilde{g}^{-}(\mathcal{D}\phi(x_i))$ are monotone and negative monotone respectively.*

*Proof.* We verify monotonicity of $\tilde{g}^{+}(\mathcal{D}\phi(x_i))$; the other part of the argument is identical.

Denoting by $d_j$ the differences $\phi(x_i) - \phi(x_{\nu,j})$ allows us to express the scheme more compactly as

$$G(d) = g\left( -\sum_{\nu\in\{(1,0),(0,1)\}} \nu \sum_{j=1}^{4} b_{\nu,j} d_j \right) + \epsilon_g \sum_{\nu\in\{(1,0),(0,1)\}} \sum_{j=1}^{4} a_{\nu,j} d_j.$$

We now introduce a perturbation $\delta > 0$ into the argument $d_k$ to obtain

$$\begin{aligned} G(d + \delta\hat{d}_k) - G(d) &\geq -L_g\,|b_{\nu,k}|\,\delta + \epsilon_g a_{\nu,k}\delta, \\ &\geq 0\end{aligned}$$

since $\epsilon_g \geq L\,|b_{\nu,k}|\,/a_{\nu,k}$. Therefore the scheme is monotone. □

### 4.4.4 Consistency and Monotonicity

We now establish consistency and monotonicity of the approximation Equation (4.23), which in turn establishes the convergence result (Theorem 4.2).

**Lemma 4.9** (Consistency)**.** *Under the assumptions of Hypotheses 2.2,3.17, the scheme $F^h$ defined by Equation* (4.23) *is consistent with Equation* (2.14) *on the space of $C^2$ functions satisfying the c-convexity constraint.*

*Proof.* Let $\phi \in C^2$ satisfy the $c$-convexity constraint. Then the determinant in Equation (2.14) can be equivalently expressed as in Equation (4.3):

$$\det{}^+(D^2\phi(x)+A(x,\nabla\phi(x))) = \min_{\nu_1,\nu_2\in V}\prod_{j=1}^{2}\max\left\{\frac{\partial^2\phi(x)}{\partial\nu_j^2}+\frac{\partial^2 c(x,y)}{\partial\nu_j^2}\bigg|_{y=T(x,\nabla\phi(x))},0\right\}.$$

By design and by Lemma 4.7, the components of the PDE are approximated consistently (with truncation error $\mathcal{O}(\sqrt{h})$). That is,

$$\lim_{h\to 0}\left\{\mathcal{D}_{\nu_j\nu_j}\phi(x_i)+\tilde{g}^-_{1,\nu_j}(\mathcal{D}\phi(x_i))\right\} = \frac{\partial^2\phi(x)}{\partial\nu_j^2}+\frac{\partial^2 c(x,y)}{\partial\nu_j^2}\bigg|_{y=T(x,\nabla\phi(x))}$$

and

$$\lim_{h\to 0}\tilde{g}_2^+\left(x_i,\mathcal{D}\phi(x_i)\right) = \frac{\left|\det D^2_{xy}c(x_i,T(x_i,\nabla\phi(x_i)))\right|}{f_1(T(x_i,\nabla\phi(x_i)))}.$$

We recall that the maximum and minimum operators are continuous, $f_0 \in C^1$, and $\tilde{V}$ is a consistent approximation of the set $V$ with angular resolution $\mathcal{O}(d\theta) = \mathcal{O}(\sqrt{h})$. Thus the combinations of these operators in the scheme $F^h$ satisfy

$$\begin{aligned}&\lim_{h\to 0, x_i\to x} -\min_{(\nu_1,\nu_2)\in\tilde{V}}\prod_{j=1}^{2}\max\left\{\mathcal{D}_{\nu_j\nu_j}\phi(x_i)+\tilde{g}^-_{1,\nu_j}(\mathcal{D}\phi(x_i)),0\right\}+f_0(x_i)\tilde{g}_2^+\left(x_i,\mathcal{D}\phi(x_i)\right)\\ &\quad = -\det{}^+(D^2\phi(x)+A(x,\nabla\phi(x)))+\left|\det D^2_{xy}c(x,T(x,\nabla\phi(x)))\right|\frac{f_0(x)}{f_1(T(x,\nabla\phi(x)))},\end{aligned}$$

which establishes consistency. □

**Corollary 4.10** (Truncation error)**.** *Under the assumptions of Hypotheses 2.2,3.17,*

*the scheme $F^h$ defined by Equation* (4.23) *has local truncation error* $\tau(h) = \mathcal{O}(\sqrt{h})$.

**Lemma 4.11** (Monotonicity)**.** *Under the assumptions of Hypotheses 2.2,3.17, the scheme $F^h$ defined by Equation* (4.23) *is monotone.*

*Proof.* By construction (and see Lemma 4.8), the schemes for $\mathcal{D}_{\nu\nu}\phi$ and $\tilde{g}^-_{1,\nu}(\mathcal{D}u)$ are negative monotone. Addition and the maximum function preserve this so that

$$\max\left\{\mathcal{D}_{\nu\nu}u(x_i) + \tilde{g}^-_{1,\nu}(\mathcal{D}u(x_i)), 0\right\}$$

is also negative monotone for any $\nu \in \mathbb{R}^2$. Since this is also non-negative, products of these terms preserve the negative monotonicity so that

$$\min_{(\nu_1,\nu_2)\in\tilde{V}} \prod_{j=1}^{2} \max\left\{\mathcal{D}_{\nu_j\nu_j}u(x_i) + \tilde{g}^-_{1,\nu_j}(\mathcal{D}u(x_i)), 0\right\}$$

is also negative monotone.

We recall also that $f_0$ is non-negative and $\tilde{g}^+_2$ is monotone (Lemma 4.8). Therefore the full scheme

$$-\min_{(\nu_1,\nu_2)\in\tilde{V}} \prod_{j=1}^{2} \max\left\{\mathcal{D}_{\nu_j\nu_j}u(x_i) + \tilde{g}^-_{1,\nu_j}(\mathcal{D}u(x_i)), 0\right\} + f_0(x_i)\tilde{g}^+_2\left(x_i, \mathcal{D}u(x_i)\right)$$

is monotone. □

### 4.4.5 Extensions to Non-Smooth Problems

The results of Loeper (2011) ensure existence of weak solutions to the Optimal Transport PDE (Equation (2.14)) in the relaxed setting where $f_0, f_1 \in L^1$ with $f_1$ bounded away from zero and $f_0$ bounded away from infinity. In Chapter 3, it was shown that solutions can be computed for the squared geodesic cost as in Theorem 3.31 if the scheme $F^h$ is additionally required to underestimate when applied to the true solution. This result was then extended to the regularized logarithmic cost via Remark 3.16.

Naturally underestimating schemes can be constructed for Monge-Ampère type equations in some cases Benamou and Duval (2017); Hamfeldt (2019). An alternate approach is to utilize a scheme of the form

$$\tilde{F}^h(x, u(x) - u(\cdot)) = F^h(x, u(x) - u(\cdot)) - h^\alpha$$

for a sufficiently small $\alpha > 0$. This preserves the consistency and monotonicity of the original scheme, while decreasing the value of the scheme and forcing it to be negative when applied to the true solution of the PDE.

In addition, definition of consistency (Definition 2.9) in terms of upper and lower semicontinuous envelopes of the PDE operator allows us to accommodate discontinuous data $f_0, f_1$ as in Hamfeldt (2019). This does not require any change in the way $f_0(x)$ is handled. However, functions $f_1 \notin C^{0,1}$ must be carefully regularized to preserve consistency and monotonicity since it takes as its argument terms $T(x, \nabla u(x))$ that involve gradients.

The approach we propose is to introduce a discrete version of the target density function

$$f_1^h(y) = (K_{h^{1/4}} * f_1)(y) \tag{4.26}$$

where $K_{h^{1/4}}$ is a mollifier that ensures that the Lipschitz constant of $f_1^h$ satisfies

$$L_{f_1^h} \leq h^{-1/4}.$$

From here, we use the same discretization of $f_1(T(x, \nabla u(x))$ introduced in Equation (4.13). We note that in this case (following Lemma 4.7), the regularization parameter will satisfy

$$\epsilon_{g_2^h} = \mathcal{O}\left(L_{f_1^h}\sqrt{h}\right) = \mathcal{O}(h^{1/4}).$$

Since this parameter converges to zero as $h \to 0$, the resulting scheme will still be consistent in the sense of Definition 2.9. The monotonicity result (Lemma 4.8) is

unchanged.

## 4.5 Preprocessing of Data

Stability and convergence of the numerical method requires at least one of the densities (denoted by $f_1$) to be strictly positive. This is easily accomplished by choosing $\epsilon > 0$ and letting

$$\tilde{f}_1^\epsilon = (1-\epsilon)f_1 + \frac{\epsilon}{4\pi}. \tag{4.27}$$

As $\epsilon \to 0$, the mapping of the regularized Optimal Transport problem converges in measure to the solution of the given problem Villani (2003), and thus we recover the desired reflector surface.

The numerical method further requires this density function to be smoothed in order to have a (discrete) Lipschitz constant that is at most $\mathcal{O}\left(h^{-1/4}\right)$. We accomplish this via a short-time evolution of the heat equation. That is, we solve

$$\begin{cases} v_t(x,t) = \Delta v(x,t), & (x,t) \in \mathbb{S}^2 \times (0, \sqrt{h}], \\ v(x,0) = \tilde{f}_1^\epsilon(x), & x \in \mathbb{S}^2 \end{cases} \tag{4.28}$$

where $\Delta$ is the Laplace-Beltrami operator. We then set

$$f_1^\epsilon(x) = v(x, \sqrt{h}). \tag{4.29}$$

The Laplace-Beltrami operator can be discretized using the finite difference schemes as

$$\Delta^h = \mathcal{D}_{(1,0),(1,0)} + \mathcal{D}_{(0,1),(0,1)} \tag{4.30}$$

and evolved using forward Euler

$$v^{n+1} = v^n + k\Delta^h v^n. \tag{4.31}$$

The wide stencil nature of the finite difference stencils ($\|x_j - x_0\| = \mathcal{O}(\sqrt{h})$) means that this is stable for a time step $k \leq 1/\sum_j a_j = \mathcal{O}(h)$. Thus a total of $\mathcal{O}\left(h^{-1/2}\right)$ time steps are needed, which leads to an overall cost of $\mathcal{O}\left(N^{5/4}\right)$ that is similar to the cost of discretization.

This regularization procedure can also be applied to unbounded densities, but requires evolving the heat equation to a stopping time of $t = h^{1/6}$ to achieve the required Lipschitz bound.

In the literature, smoothening via the Laplace-Beltrami operator is commonly referred to as heat kernel smoothing, since the Green's function of the Laplace-Beltrami operator is known as the heat kernel. The solution of Equation (4.28) is given by:

$$v(t,x) = \int_{\mathbb{S}^2} K(t;x,y) f_1^\epsilon(y) dy \tag{4.32}$$

where $K(t;x,y)$ is the heat kernel.

Explicit representations of the heat kernel are rather difficult to write down, except, for example, in the asymptotic limit as $t \to 0$. However, Harnack inequality-type bounds have been established for the heat kernel on manifolds with non-negative Ricci curvature Li and Yau (1986) which applies, namely to the 2-sphere:

$$|\nabla_y K(t;x,y)| \leq \frac{C(\epsilon)}{\sqrt{t}\mathcal{L}_{\mathbb{S}^2}\left(B(x,\sqrt{t})\right)} e^{-\frac{d_{\mathbb{S}^2}(x,y)^2}{4(1-\epsilon)t}}, \tag{4.33}$$

where $\mathcal{L}_{\mathbb{S}^2}$ is the Lebesgue measure on the sphere. This inequality is true for any $\epsilon \in (0,1)$. Thus:

$$|\nabla_x v(t,x)| \leq \frac{C(\epsilon)}{\sqrt{t}} \int_{\mathbb{S}^2} \frac{e^{-\frac{d_{\mathbb{S}^2}(x,y)^2}{4(1-\epsilon)t}}}{2\pi\left(1-\cos\left(\sqrt{t}\right)\right)} f_1^\epsilon(y) dy. \tag{4.34}$$

For bounded $f_1^\epsilon$, since we know $\int_{\mathbb{R}^2} \frac{e^{-z^2/at}}{t} dz \leq C$ for some $C > 0$, by standard results on the Euclidean heat kernel, we know that

$$\int_{\mathbb{S}^2} \frac{e^{-d_{\mathbb{S}^2}(x,y)^2/4(1-\epsilon)t}}{2\pi\left(1-\cos\left(\sqrt{t}\right)\right)} dy \leq \int_{\mathbb{R}^2} \frac{e^{-z^2/4(1-\epsilon)t}}{2\pi t} dz + \mathcal{O}(t) \leq C' + \mathcal{O}(t). \tag{4.35}$$

Therefore,

$$|\nabla_x v(t,x)| \leq \frac{C''}{\sqrt{t}} + \mathcal{O}\left(\sqrt{t}\right). \tag{4.36}$$

Thus, having $t \leq \mathcal{O}(h^{1/2})$ ensures that $\|\nabla_x v(t,x)\| \leq \mathcal{O}(h^{-1/4})$ for small enough $h$. For unbounded $f_1^\epsilon$, but under the assumption that $f_1^\epsilon \in L^1(\mathbb{S}^2)$, yields a slightly worse bound:

$$|\nabla_x v(t,x)| \leq \frac{C}{t^{3/2}} \int_{\mathbb{S}^2} e^{-\frac{d(x,y)^2}{4(1-\epsilon)t}} f_1^\epsilon(y) dy + \mathcal{O}\left(\sqrt{t}\right) \leq \\ \frac{C'}{t^{3/2}} \int_{\mathbb{S}^2} f_1^\epsilon(y) dy + \mathcal{O}\left(\sqrt{t}\right) \leq \frac{C''}{t^{3/2}} + \mathcal{O}\left(\sqrt{t}\right), \tag{4.37}$$

which means that $t \geq \mathcal{O}(h^{1/6})$ assures the correct Lipschitz bound. Thus, given an appropriately convergent numerical approximation of the heat equation:

$$f_{n+1}^h = f_n^h + \Delta t \Delta^h f_n^h. \tag{4.38}$$

h Then, the idea is that we can smoothen an $L^1(\mathbb{S}^2)$ function by proceeding $n \approx t/\Delta t$ time steps. The time stepsize $\Delta t$ must satisfy a CFL condition, such as $\Delta t \leq \mathcal{O}(\text{diam}(N_i)^2)$, where $\text{diam}(N_i)$ is the size of the computational neighborhood. Thus, for a fixed $t > 0$, we must take $n = \mathcal{O}(t \cdot \text{diam}(N_i)^{-2})$ time steps. If we use, for example the monotone and consistent schemes for the second directional derivatives in Hamfeldt and Turnquist (2021b), then $\text{diam}(N_i) = \mathcal{O}(\sqrt{h})$, then we get $n = \mathcal{O}(th^{-1})$. Thus, we take:

$$f_1 \in L^\infty(\mathbb{S}^2), \quad n \geq \mathcal{O}(h^{-1/2}),$$
$$f_1 \notin L^\infty(\mathbb{S}^2), \quad n \geq \mathcal{O}(h^{-5/6})$$

to achieve the desired smoothness.

### 4.6 Computational Complexity

We now state the algorithm (Algorithm 2) for finding $u^h$ in a condensed form and then proceed to compute its computational complexity. Note that for the reflector antenna, the final step is to use the solution to compute the reflector antenna shape: $\Sigma^h = x_i e^{-u^h}$.

**Algorithm 2** Computing the discrete solution $u^h$

1: Preprocess data
$$f_1^\epsilon \leftarrow \text{Regularize}(f_1).$$
2: Iterate
$$u^h_{n,E} = u^h_n + \gamma_n \left(u^h_n - u^h_{n-1}\right)$$
$$u^h_{n+1} = u^h_{n,E} - \Delta t \left(F^h(x, u^h_n; f_0, f_1^\epsilon) - \sqrt{h} u^h_n(x)\right)$$
to steady state.
3: Normalize solution
$$u^h(x) \leftarrow u^h(x) - \fint_{\mathbb{S}^2} u^h(x)\, dS(x).$$

Let $N$ be the total number of grid points. The preprocessing step in our algorithm requires evaluating a discrete approximation to the Laplace-Beltrami operator at each grid point, which can be done in $\mathcal{O}(N)$ time. The subsequent evolution, Equation (4.31), requires $\mathcal{O}\left(h^{-1/2}\right) = \mathcal{O}\left(N^{1/4}\right)$ time steps, for an overall cost of $\mathcal{O}\left(N^{5/4}\right)$.

At each point $x_0 \in \mathcal{G}$, evaluating the operator $F^h$ involves computing a minimum over the $\mathcal{O}\left(1/d\theta\right) = \mathcal{O}\left(1/\sqrt{h}\right) = \mathcal{O}\left(N^{1/4}\right)$ pairs of vectors in $V$.

Each pair of vectors $\{\nu_1, \nu_2\} \in V$ requires the construction of two finite

difference operators of the form $\mathcal{D}_{\nu\nu}$. Computing each of these requires identifying the four neighbors $x_1, x_2, x_3, x_4$ in the stencil.

We note that selecting each of these neighboring points $x_j$ as in Equation (3.24) involves searching a region whose area scales like $\mathcal{O}(h^2)$. From the definition of $h$, this is guaranteed to contain at least one point, and expected to contain $\mathcal{O}(1)$ points total. Thus identification of these four neighboring points can be done in $\mathcal{O}(1)$ time.

Thus, given a grid function $u$, the total computational cost of evaluating the operator $F^h$ at all points in the grid is $\mathcal{O}\left(N^{5/4}\right)$. This is the per-iteration cost of the method. The accelerated parabolic solver we use requires approximately $\mathcal{O}(N)$ iterations Schaeffer and Hou (2016) for a total computational complexity of approximately $\mathcal{O}\left(N^{9/4}\right)$.

## 4.7 Computational Results

All computations in this chapter were performed on a 13-inch MacBook Pro, 2.3 GHz Intel Core i5 with 16GB 2133 MHz LPPDDR3 using Matlab R2017b.

### 4.7.1 Structured and Unstructured Grids

The scheme we have built works well on both structured and unstructured grids, provided that they satisfy the mild conditions of Hypothesis 3.17. By structured we mean that there exists a deterministic way of building the grid. Likewise, unstructured here means there is a stochastic element in the construction of the grid.

Here, we describe four types of grids that satisfy these hypotheses and that we make use of in practice: the cube grid (structured), the random grid (fully unstructured), the latitude-longitude grid (unstructured), and the layered grid (structured).

The structured cube grid is constructed as follows. First, a grid of evenly-spaced points is generated on the faces of a cube which contains the sphere. Then,

the points on such a grid on the cube are projected onto the sphere. See Figure 4.4 for an example of the resulting grid.

The semi-unstructured latitude-longitude grid is constructed as follows. We begin with a structured grid composed of points equally spaced in both latitude $\theta$ and longitude $\phi$. However, this produces a highly over-resolved grid near the poles, which does not satisfy the required structure conditions. In order to get rid of this redundancy, we stochastically remove grid points within a geodesic distance $\mathcal{O}(h)$ of both poles. That is, for each grid point $x_i = (\theta_i, \phi_i)$, we compute $\xi_i = \sin\theta_i$ and generate a value using the random variable $\Xi \sim \text{Unif}([0,1])$. If $\Xi > \xi_i$, then we remove the point $(\theta_i, \phi_i)$ from the grid. The removal of these points creates a new unstructured grid that almost surely satisfies Hypothesis 3.17. See Figure 4.4 for an example of such a grid.

To construct the structured layered grid, we take an integer $n$ and $\epsilon = \mathcal{O}(h)$ and define the rows: $\text{row}_j = \epsilon + j\dfrac{\pi - 2\epsilon}{n}$. For each row, we have $\text{col}_{kj} = j\phi_0 + k\dfrac{2\pi}{\lfloor n * \sin\theta_n \rfloor}$, where $\phi_0 = \dfrac{1+\sqrt{5}}{2}$, which is the golden ratio. This creates a nice spread of points inspired by the seed packing of sunflowers. Then, we introduce the grid points $x_{jk} = (\theta, \phi) = (\text{row}_j, \text{col}_{jk})$ for $j = 1, \ldots, n$ and $k = 1, \ldots, \text{floor}\{n * \sin\theta_n\}$. This grid, by construction, will satisfy Hypothesis 3.17. See Figure 4.4.

Finally, we consider a fully unstructured random grid. After defining the set

$$R = \{(x, y) \in [0, \pi] \times [0, 1] : y \leq \sin(x)\}$$

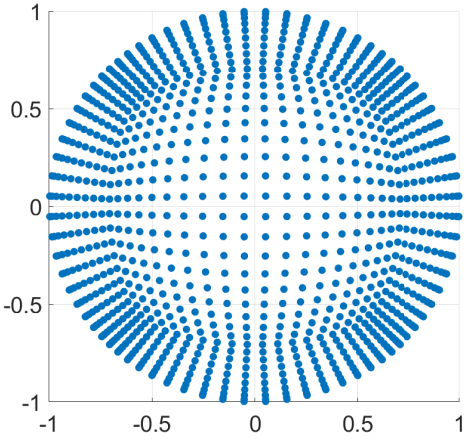
and the projection

$$P(x, y) = x,$$

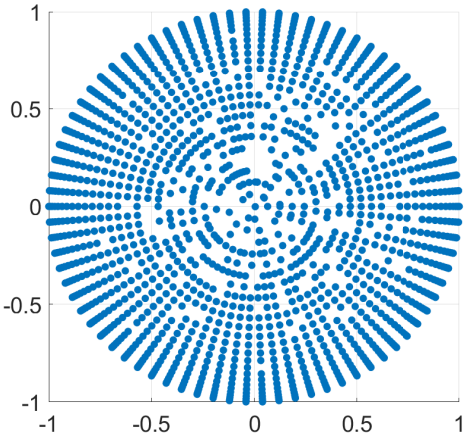
we sample $\Phi \sim \text{Unif}([0, 2\pi])$ and $\tilde{\Theta} \sim \text{Unif}(R)$, then define $\Theta = P(\tilde{\Theta})$.

We take as grid points the random variables

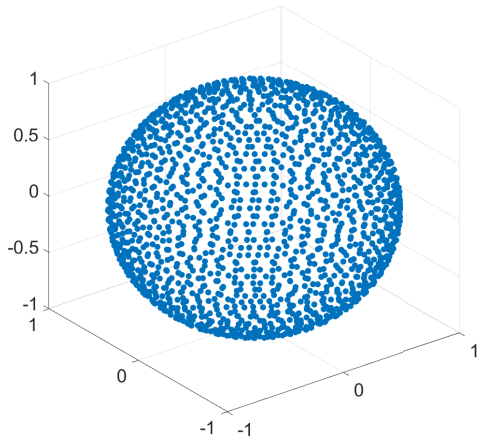
$$(X, Y, Z) = (\sin\Theta\cos\Phi, \sin\Theta\sin\Phi, \cos\Theta),$$



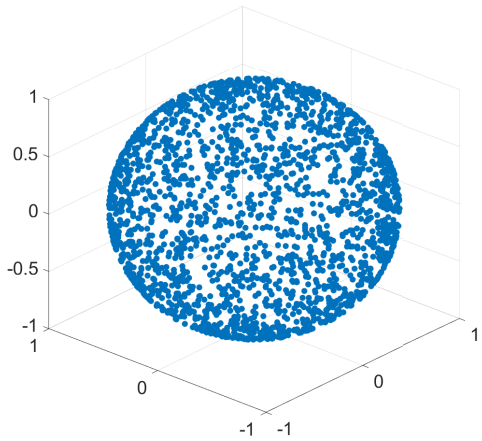

**(a)** Cube grid.




**(b)** Latitude-Longitude grid.




**(c)** Layered grid.




**(d)** Random grid.

**Figure 4.4** Top down views of a cube grid (Figure 4.4a) and latitude-longitude grid (Figure 4.4b). Views of a layered grid (Figure 4.4c) and a random grid (Figure 4.4d).

which satisfy the conditions that $(X, Y, Z) \sim \text{Unif}(\mathbb{S}^2)$. The resulting grid almost surely satisfies Hypothesis 3.17. See Figure 4.4 for an example of such a grid.

### 4.7.2 Recovering Constant Solutions

For both cost functions, in the case where $f_0 = f_1$, the resulting solution will be a constant function ($u(x) = 0$). Figure 4.5 shows the solutions obtained for both the squared geodesic and logarithmic cost functions. Importantly, these are effectively constant to within a tolerance less than the expected consistency error of the method.

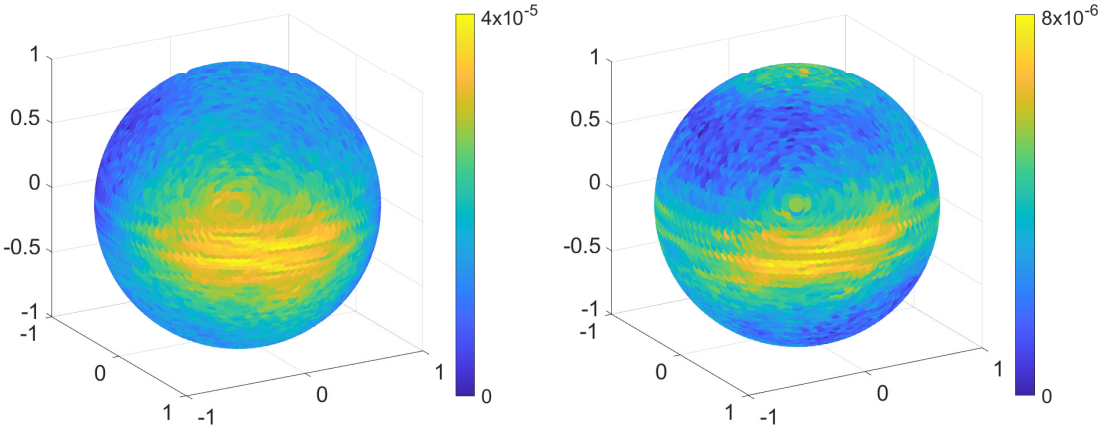



**Figure 4.5** Solutions obtained for the squared geodesic cost (left) and logarithmic cost (right) when $f_0 = f_1$ on a layered grid consisting of $N = 7722$ points.

### 4.7.3 Small Perturbation

Here we demonstrate with computations for the squared geodesic cost what happens when the target mass density $f_1$ is obtained through a slight perturbation (a rotation through an angle $\theta$) of the source mass density $f_0$. In particular, we choose the density functions

$$\begin{cases} f_0(x, y, z) = \frac{1}{4\pi - 4} \left(1 - 0.5 \cos\left(\frac{\pi}{2} x\right)\right), \\ f_1(x, y, z) = \frac{1}{4\pi - 4} \left(1 - 0.5 \cos\left(\frac{\pi}{2} (x \cos\theta + y \sin\theta)\right)\right). \end{cases} \tag{4.39}$$

This problem has the flavor of a translation. In the Euclidean setting, translations are exact solutions of the Optimal Transport problem. However, it is not the case that rotations are exact solutions on the sphere. See Figure 4.6 for a top-view of the computed gradient map. In particular, we observe that the bulk of the mass does undergo a clockwise rotation. However, in order to conserve mass and regularity, there is also a backwards "flow" observed in areas of low density (top and bottom of the figure).

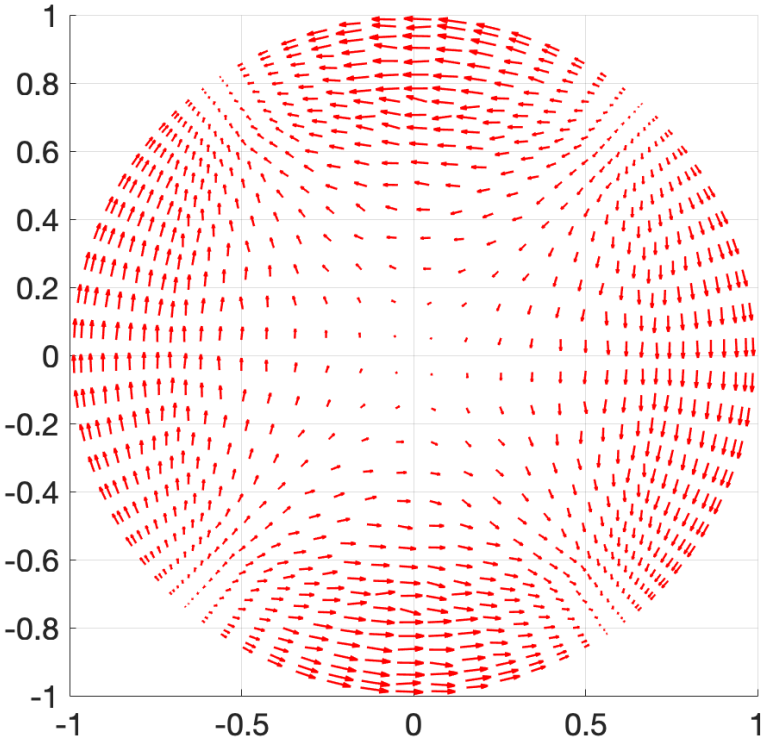



**Figure 4.6** Top view of the local gradient obtained numerically when $f_1$ is obtained from $f_0$ through a small rotation. The solution was computed using a cube grid with $N = 2168$ points.

### 4.7.4 Comparing Structured and Unstructured Grids

To show the robustness of our generalized finite difference scheme with respect to the structure of the grid, we next show a side-by-side comparison of solutions obtained using a fully structured layered grid and a fully unstructured random

grid. We choose a non-smooth source density $f_0$ and constant target density $f_1$:

$$\begin{cases} f_0(\theta,\phi) = (1-\epsilon)\frac{1}{5.8735}\left(\theta - \frac{\pi}{2}\right)^2 + \frac{\epsilon}{4\pi}, \\ f_1(\theta,\phi) = \frac{1}{4\pi} \end{cases} \tag{4.40}$$

for $\epsilon = 0.1$. See Figure 4.7 for the computed solutions, which are identical to within a tolerance on the order of the computed residual.

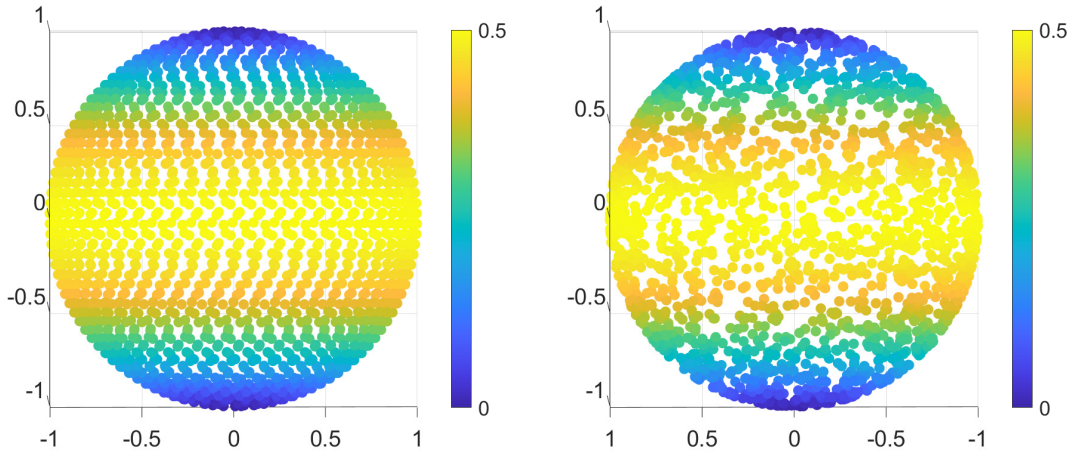



**Figure 4.7** Solution $u$ obtained for densities in Equation (4.40) on $N = 2006$ point layered (left) and random (right) grids.

### 4.7.5 Non-Smooth Examples

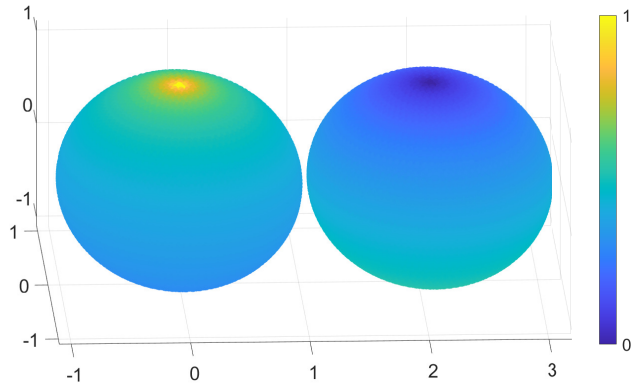
Finally, we present the results of a computation (using the squared geodesic cost) where $f_0$ is unbounded and $f_1$ is not Lipschitz. Recall that this is a situation where the solution $u$ is only guaranteed to be $C^1(\mathbb{S}^2)$ and the non-Lipschitz property of $f_1$ can easily cause issues regarding monotonicity and consistency. However, these issues can be resolved using the ideas in Section 4.4.5. The density functions are given by

$$\begin{cases} f_0(\theta,\phi) = \frac{1}{2\pi\cdot 1.86691}\theta^{-1/4}, \\ f_1(\theta,\phi) = (1-\epsilon)\frac{\theta^{3/4}}{17.2747} + \frac{\epsilon}{4\pi}. \end{cases} \tag{4.41}$$

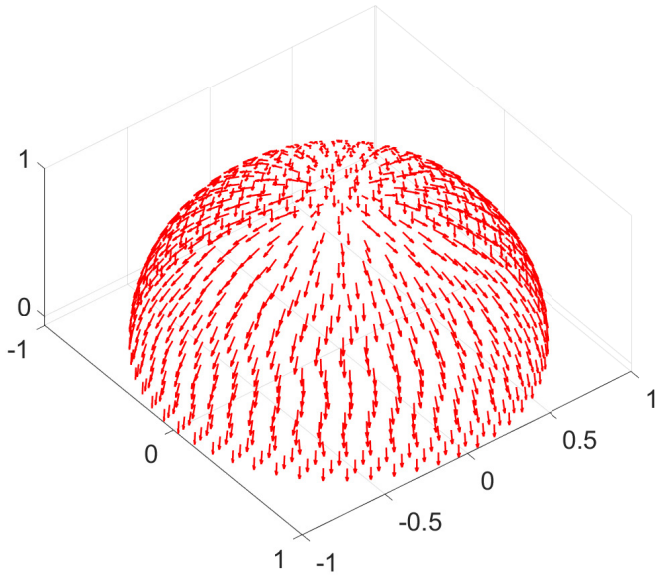
where $\epsilon = 0.5$.

Despite the very strong singularities present in this example, our numer-

ical method has no difficulty computing a solution. The density functions and computed gradient are shown in Figure 4.8. As expected, we observe mass being transported downward away from the singularity.




**(a)** The source and target masses, left is unbounded and right is non-smooth.




**(b)** The movement of mass away from the northern hemisphere and toward the southern hemisphere.

**Figure 4.8** Unbounded source and non-Lipschitz target densities (Figure 4.8a) and resulting gradient (visualized for the northern hemisphere) computed on a $N = 7722$ point layered grid (Figure 4.8b).

# CHAPTER 5

# APPLICATION TO THE REFLECTOR ANTENNA PROBLEM

## 5.1 Introduction

In this chapter, the solution to the reflector antenna problem is obtained by solving a Monge-Ampère type equation directly on the sphere (by the method proposed in Chapter 4). These results are, for the most part, presented in the manuscript Hamfeldt and Turnquist (2021c). One benefit of using a PDE solver for the reflector antenna problem is to allow for intensity distributions supported on complicated subsets of the sphere or even over the entire sphere. Furthermore, PDE solvers allow one to easily design higher-order schemes, an idea which is being explored in current work, see Chapter 8 for more detail. In addition, as shown in Chapter 4, the computational complexity of the method is better than other provably convergent schemes as will be shown in this chapter. Moreover, the approach is intrinsic and thus the solution to the problem will not depend on such details as the choice of the north pole.

We validate this new method through several challenging examples, which include intensities that have complicated discontinuities, that propagate over complicated geometries, or that contain a mix of light and dark regions. It will be shown that the method performs well even in a final example where the physics does not guarantee the existence of a smooth ($C^1$) reflector.

In order to properly compare the scheme to some of those existing in the literature shown in Section 2.3.1, we compare the computational complexity of our scheme as proposed in Chapter 4 with other methods for computing the reflector antenna.

## 5.2 The Reflector Antenna

Here we recapitulate briefly the reflector antenna problem, with more detail provided in Section 2.3. We start with a light source or detector $\mu_0$ located at

the origin, which is a probability measure indicating directional intensity and is supported on a set $\Omega \subset \mathbb{S}^2$. Next we consider a reflector surface $\Sigma$, which is a radial graph over the domain $\Omega$ and can be represented as

$$\Sigma = \{x\rho(x) \mid x \in \Omega, \quad \rho > 0\}, \tag{5.1}$$

where $\rho : \Omega \to \mathbb{R}$ is a non-negative function indicating the distance between the reflector surface and the origin. The light from the source $\mu_0$ in the direction $x$ bounces off the reflector $\Sigma$ without any refraction or absorption and travels in the direction $T$ following the law of reflection. Over every direction this produces the far-field intensity $\mu_1$, which is also a probability measure indicating directional intensity and is supported on some target domain $\Omega^* \subset \mathbb{S}^2$. To recapitulate, the setup is as shown in Figure 5.1

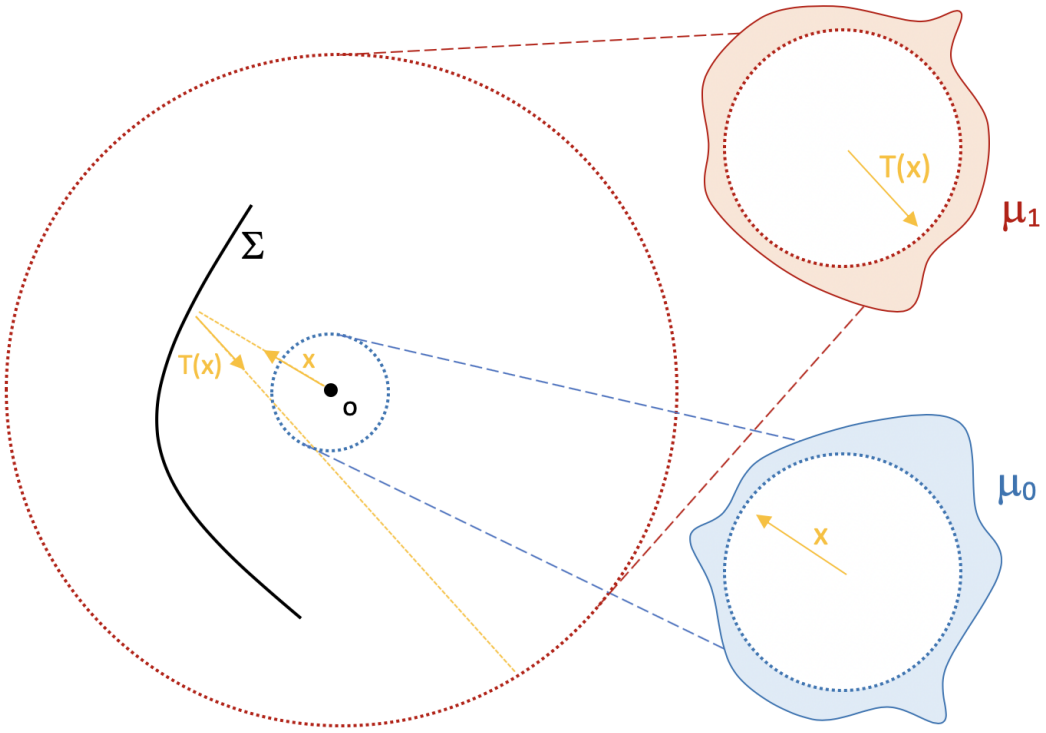



**Figure 5.1** Reflector antenna with source/detector $\mu_0$, reflector $\Sigma$ and target far-field intensity $\mu_1$. The directional vectors $x$ and $T(x)$ are unit vectors.

The reflector antenna problem is thus: given source and target intensity probability distributions $\mu_0$ and $\mu_1$, respectively, find the shape of the reflector $\Sigma$ that

transmits the light from the source to the target while satisfying conservation of energy.

After applying Snell's law of reflection, conservation of energy, and the change of variables $\rho = e^{-u}$, the derivation in Wang (2004) showed that solving for $u$ is equivalent to solving the Optimal Transport problem with cost function $\tilde{c}(x, y) = -\log(1 - x \cdot y)$ on the sphere. Assuming that $d\mu_0 = f_0(x)dx$ and $d\mu_1 = f_1(y)dy$ and under mild conditions on the intensity distributions $f_0$ and $f_1$, the function $u$ can be uniquely obtained as the solution of the following Monge-Ampère type equation:

$$\begin{cases} \det(D^2u + A(x, \nabla u)) = H(x, \nabla u), \quad x \in \mathbb{S}^2, \\ D^2u + A(x, \nabla u) \geq 0. \end{cases} \tag{5.2}$$

Here

$$\begin{aligned} A(x, p) &= D^2_{xx} c\left(x, T(x, p)\right), \\ H(x, p) &= \left|\det D^2_{xy} c\left(x, T(x, p)\right)\right| f_0(x)/f_1\left(T(x, p)\right). \end{aligned} \tag{5.3}$$

and the statement $M \geq 0$ means that $M$ is positive semi-definite. This constraint (related to the so-called $c$-convexity of the optimal map $T$) is needed to ensure that the PDE has a unique solution (up to additive constants) and that this solution corresponds to the desired optical mapping $T$.

## 5.3 Computational Complexity Comparison

Here we present a slight modification of Algorithm 2, which allows for the construction of the reflector, see Algorithm 3. In Section 4.6, it was showed that the computational complexity of our algorithm is $\mathcal{O}\left(N^{9/4}\right)$. Other provably convergent schemes are available for the reflector antenna problem, and we outline them here.

For many of the other numerical methods described in the literature, we

**Algorithm 3** Computing the reflector surface $\Sigma$

1: Preprocess data

$$f_1^\epsilon \leftarrow \text{Regularize}(f_1).$$

2: Iterate

$$u_{n+1}^h = u_n^h + k\left(F^h(x, u_n^h; f_0, f_1^\epsilon) - \sqrt{h}u_n^h(x)\right)$$

to steady state.

3: Normalize solution

$$u^h(x) \leftarrow u^h(x) - \fint_{\mathbb{S}^2} u^h(x)\, dS(x).$$

4: Construct reflector

$$\Sigma^h = \left\{ xe^{-u^h(x)} \mid x \in \Omega \cap \mathcal{G} \right\}.$$

were not able to find a report of computational complexity. However, we make the general observation that there is typically a trade-off between efficiency and convergence guarantees. Our proposed method seeks to find a balance between these issues: it is equipped with a convergence proof while also providing a substantial improvement in efficiency over other provably convergent methods, which typically require at least $\mathcal{O}(N^3)$ time. Other methods based on Monge-Ampère equations or ray-mapping techniques may provide an improved performance, at least for smooth examples, but at the expense of any rigorous convergence guarantees.

Within the category of methods that have a rigorous theoretical foundation we include Oliker's supporting quadrics method Oliker (2006); Oliker et al. (2015) and methods that use optimization techniques to solve the Optimal Transport problem Doskolovich et al. (2019); Glimm and Oliker (2003). The supporting quadrics approach involves successively solving $N$ nonlinear equations for $N$ focal parameters, which are used to construct the $N$ supporting quadrics that approximate the reflector surface. An estimate on the overall complexity of this procedure is $\mathcal{O}\left(N^4 \log N^2 \tau\right)$ where $\tau$ is the computational time to compute a particular integral Kochengin and Oliker (1998). The numerical methods proposed in Doskolovich et al. (2019); Glimm and Oliker (2003) utilize the linear programming and linear assignment formulations for Optimal Transport, which are also theoretically well-founded. The resulting optimization problems are augmented

to involve $\mathcal{O}(N^2)$ unknowns Peyré and Cuturi (2019). The most efficient implementations of the Hungarian algorithm for these problems require approximately $\mathcal{O}(N^3)$ time Jonker and Volgenant (1987).

Several recent methods involve formulating the reflector design problem as a Monge-Ampère type equation Brix et al. (2015); Romijn et al. (2020); Wu et al. (2013). The overall cost of evaluating the discretizations utilized by these methods (i.e. the per-iteration cost) is $\mathcal{O}(N)$, which is lower than our per-iteration cost of $\mathcal{O}\left(N^{5/4}\right)$. However, these efficiency gains come at the expense of any proof of convergence. Solver times vary, and are unreported in many cases. If mesh-independent convergence can be achieved, a total cost of $\mathcal{O}(N)$ may be possible for smooth examples, but it is unclear whether or not any existing methods actually achieve this.

A final class of methods we consider involve producing a ray-mapping between source and target, then using the law of reflection to produce an optical surface that achieves this ray-mapping Bruneton et al. (2011); Desnijder et al. (2019); Feng et al. (2016); Fournier et al. (2010); Parkyn and Pelka (2006). Approaches for accomplishing this vary, and in most cases we were not able to find reports of the overall computational complexity. We again note that while an optimal $\mathcal{O}(N)$ cost may be possible in principle, these computational simplicity comes at the expense of any rigorous guarantees as to the existence of a reflector that produces the desired ray mapping.

## 5.4 Computational Results

Here we demonstrate the effectiveness of our method with several computational examples. These include reflector design problems involving an omnidirectional source, discontinuous intensity distributions, and intensity distributions supported on sets with complicated geometries. In each example, we construct an approximate reflector $\Sigma^h$.

In order to validate our results, we first use the law of reflection in Equa-

tion (2.53) to perform approximate (forward or inverse) ray-tracing. We then construct the resulting intensity patterns via approximation of the conservation of energy (Equation (2.51)) by

$$f_0(x_i)\Delta x_i \approx f_1(y_i)\Delta y_i,$$

where $\Delta x_i$ and $\Delta y_i$ are the areas of the Voronoi regions containing $x_i$ and $y_i = T(x_i)$ respectively.

The areas of these Voronoi regions are estimated as follows. We define the discrete variables $n : \mathcal{G} \to \mathbb{N}$, which defines a histogram whose bins are defined at each grid point. Likewise, we define the discrete variable $m : T(\mathcal{G}) \to \mathbb{N}$, which defines a histogram whose bins are located at the image of the grid points under the mapping $T$. In our computations, we sample $M = 100,000,000$ points $x \sim \text{Unif}(\mathbb{S}^2)$. For each sampled point $z$, we compute

$$x_k = \underset{x_i \in \mathcal{G}}{\text{argmin}}\, d_{\mathbb{S}^2}(x_i, z), \quad y_l = \underset{y_j \in T(\mathcal{G})}{\text{argmin}}\, d_{\mathbb{S}^2}(y_j, z)$$

and increment both of the discrete variables $n(x_k)$ and $m(y_l)$ by 1. After $M$ samples, we are left with histograms $n$ and $m$ that, after appropriate normalization, represent the approximate areas of the Voronoi regions: $n_i \approx \Delta x_i$ and $m_i \approx \Delta y_i$. We emphasize that the approximation of these Voronoi regions introduces additional artifacts into the ray trace which are irrespective of our computed reflector.

After performing ray tracing, the presence of numerical artifacts may require that the data be post-processed to show the results clearly. This is done by rescaling the colorbars to cut off a very small number of the highest values. Any numerical artifacts are presented in plots of the difference between the desired and ray-traced intensities.

All computations were performed on a 13-inch MacBook Pro, 2.3 GHz Intel Core i5 with 16GB 2133 MHz LPPDDR3 using Matlab R2017b. Each computation utilized around $N \approx 20,000$ points on the sphere. Where applicable, regulariza-

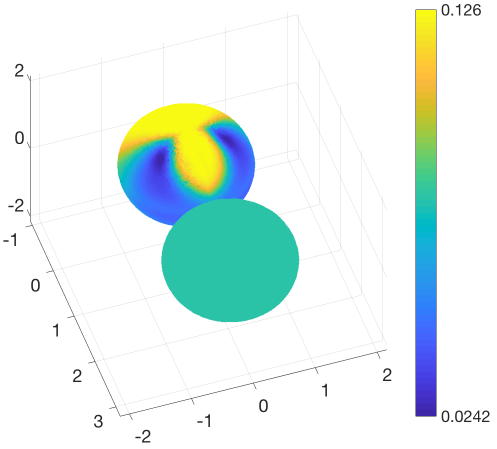
tion was performed using $\epsilon = 0.3$. The precomputation step of approximating all directional derivatives for $N \approx 20,000$ points took about 10 minutes. Solving the parabolic scheme to find the solution took around 30 minutes. Ongoing work will develop faster, more accurate versions of this method. We see therefore that the proposed numerical method can certainly accommodate higher precision computations if necessitated by real-world applications.

### 5.4.1 Peanut Reflector

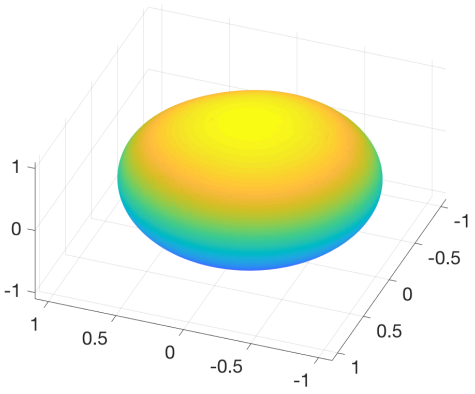
Following the example of Romijn et al. (2020), we consider a source density coming from an ideal headlight intensity emitting from a vehicle's high beams. This headlight intensity pattern is then mapped to the sphere, and inverted, which becomes the source intensity $f_0$. The target density $f_1$ is constant. The computation yields a peanut-shaped oblong reflector lens; see Figure 5.2. Despite the fact that we anticipate error in the reverse ray trace due to the approximate conservation of energy (Equation (2.51)), we see that the absolute error performs quite well in this smooth example. The average error in the reconstruction is 11% of the maximum intensity.

### 5.4.2 Discontinuous Intensities

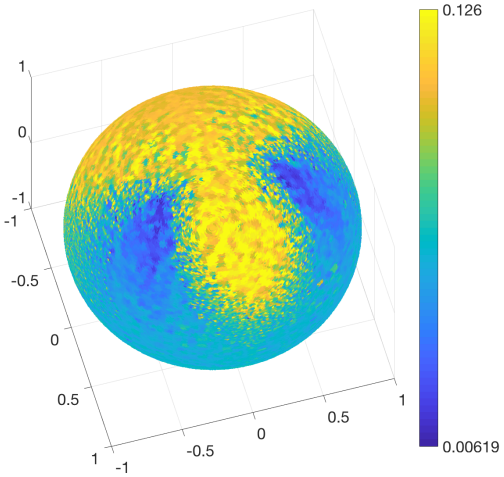
Next, we demonstrate the effectiveness of our method in dealing with discontinuities and complicated densities. In this example, a discontinuous source mass $f_0$ resembling an inverted map of the world is mapped to a constant density $f_1$; see Figure 6.8. This is a particularly challenging example given the very complicated structure of the discontinuities. Nevertheless, we achieve a reconstruction that visually agrees with the world map, with an average error of 19% of the maximum intensity.



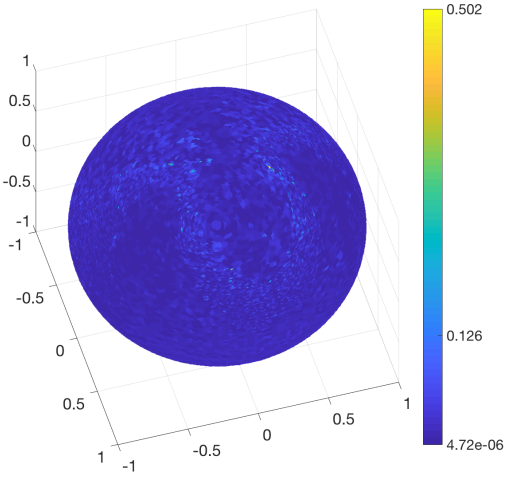

**(a)** The source (headlight intensity) and target masses (constant intensity).



**(b)** The computed reflector shape.




**(c)** The inverse ray trace.



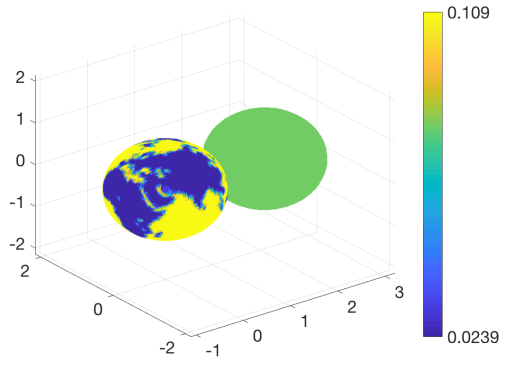

**(d)** The error of the inverse ray trace.

**Figure 5.2** "Peanut" reflector. The source and target intensities (Figure 5.2a), the resulting reflector shape (Figure 5.2b), the ray trace validation (Figure 5.2c), and the absolute error in the validation (Figure 5.2d).

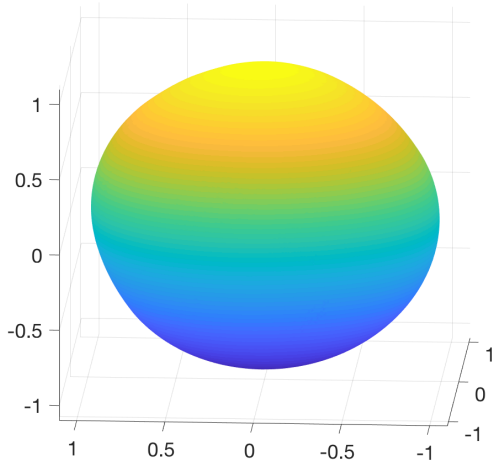
### 5.4.3 Donut Intensities

To further demonstrate the flexibility of our method, we consider the source and target intensities propagating in a donut shape, with a dark region in the center. These are given by

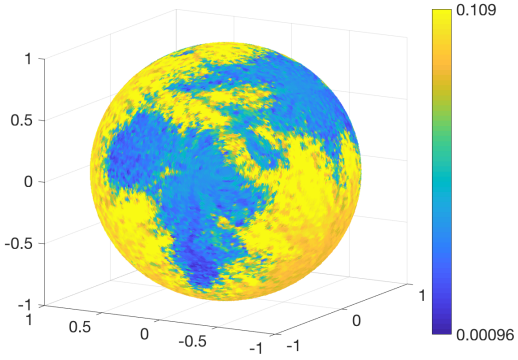
$$f_0(x,y,z) = \begin{cases} \frac{1}{(4\pi/15)(\sqrt{2}+2)} \left(-4\sqrt{x^2+y^2}z^3 + 4(x^2+y^2)^{3/2}z\right), & \sqrt{2}/2 \geq z \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{5.4}$$



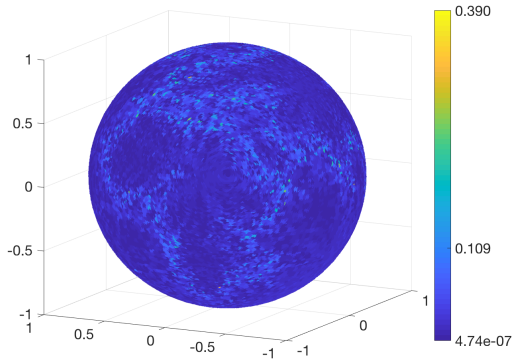

**(a)** Source (globe) and target (constant) densities.



**(b)** The computed reflector shape.




**(c)** The inverse ray trace.




**(d)** The inverse ray trace error.

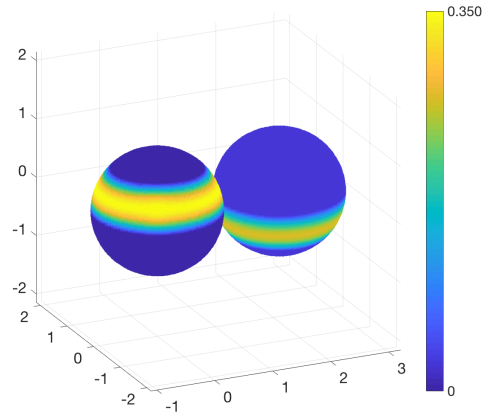
**Figure 5.3** Discontinuous "globe" intensities. The source and target intensities (Figure 5.3a), the resulting reflector shape (Figure 5.3b), the ray trace validation (Figure 5.3c), and the absolute error in the validation (Figure 5.3d).

and

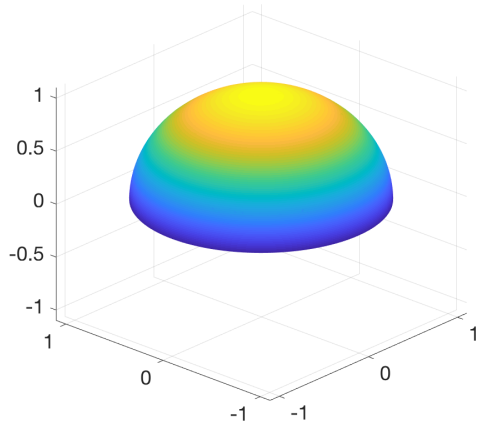
$$f_1(x,y,z) = \begin{cases} \frac{1}{(4\pi/15)(\sqrt{2}+2)}\left(-4\sqrt{x^2+y^2}z^3 + 4(x^2+y^2)^{3/2}z\right), & 0 \geq z \geq -\sqrt{2}/2, \\ 0, & \text{otherwise.} \end{cases} \tag{5.5}$$

These intensities have very complicated support containing holes, which is particularly challenging numerically. Indeed, this challenge is inherent in the theory of the Optimal Transport problem. We note that the $c$-convexity constraint requires the domain $\Omega$ to be $c$-convex in order to guarantee construction of the physically relevant solution of Equation (2.14). Consequently, PDE based methods that are

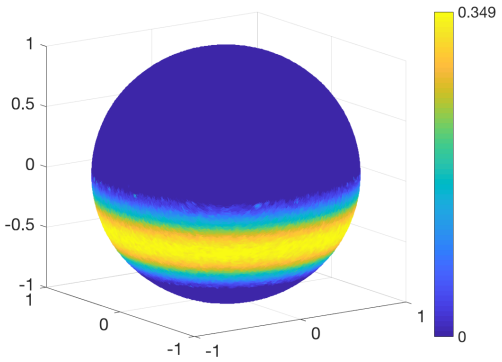
posed only on the support $\Omega$ of the intensity (rather than being extended into the dark regions) will not be assured of producing the correct reflector. This issue is handled naturally by our method, which is posed on the entire sphere. Despite the difficulty of this example, our method performs very well, as evidenced in the results of the ray-tracing. See Figure 5.4. Average error is 9% of the maximal intensity.



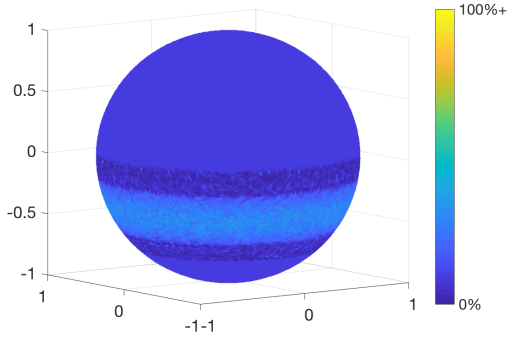

**(a)** The source (left) and target (right) densities.



**(b)** The computed reflector.




**(c)** The forward ray trace.




**(d)** The forward ray trace error.

**Figure 5.4** "Donut" intensities example. The source and target intensities (Figure 5.4a), the resulting reflector shape (Figure 5.4b), the ray trace validation (Figure 5.4c), and the absolute error in the validation (Figure 5.4d).

### 5.4.4 Singular Reflector

We conclude with an example of a hemispheric light source (here designated as $f_1$) that is to be reshaped into a geodesic triangle on the sphere (here designated as

$f_0$). We remark that given the complicated (non $c$-convex) support of this target, we are not even guaranteed the existence of a smooth ($C^1$) reflector; see Loeper (2011).

The intensities are defined as follows. We begin by forming a geodesic triangle $T_\theta \subset \mathbb{S}^2$ from the three vertices $(t_{0,\theta}, t_{1,\theta}, t_{2,\theta})$, where we define $t_{j,\theta} = (\sin\theta\cos(2\pi j/3), \sin\theta\sin(2\pi j/3), \cos\theta)$ for $\pi/2 \leq \theta < \pi$. The geodesic triangle is formed by the small region enclosed by the three vertices $t_i$, which are connected by geodesics on the sphere. That is, a point $x_0 \in T_\theta$ if $x_0$ satisfies the following three inequalities:

$$\begin{aligned} x_0 \cdot (t_{1,\theta} \times t_{2,\theta}) &\leq 0, \\ x_0 \cdot (t_{2,\theta} \times t_{3,\theta}) &\leq 0, \\ x_0 \cdot (t_{3,\theta} \times t_{1,\theta}) &\leq 0. \end{aligned}$$

Then the source intensity is defined by

$$f_0(x,y,z) = \begin{cases} 1/A, & (x,y,z) \in T_\theta, \\ 0, & (x,y,z) \notin T_\theta, \end{cases} \tag{5.6}$$

where $A$ is the area of the geodesic triangle $T_\theta$ and $\theta = 2.1$.

The target intensity is a smoothed version of the identity function on the northern hemisphere:

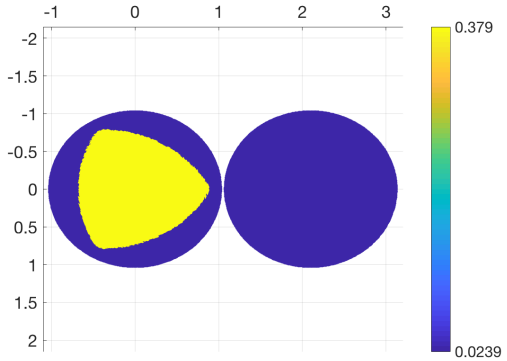
$$f_1(x,y,z) = \begin{cases} \frac{2\pi\log(\cosh(a))}{a}\tanh(az), & z \geq 0, \\ 0, & z < 0, \end{cases} \tag{5.7}$$

where $a = 10$.

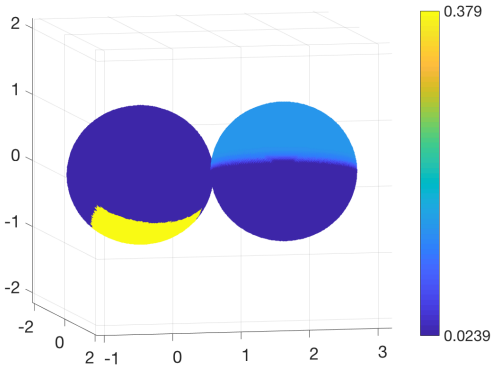
For ease of implementation, we perform pre-processing to bound both $f_0$ and $f_1$ away from zero. Results are presented in Figure 5.5. In the computed reflector, and resulting ray-traced intensity, we observe an approximate triangle shape as expected. In this case, there are notable artifacts present near the boundary of the

triangle. However, to some extent these are a limitation of the physics rather than of our method. We remark that there is no reason to expect the reflector we are approximating to be continuously differentiable, so the accuracy of the ray-tracing verification test is itself rather suspect here. Nevertheless, the absolute error as compared with the ray trace from the approximate conservation of energy equation mostly performs well, with an average error of 16% of the maximal intensity.

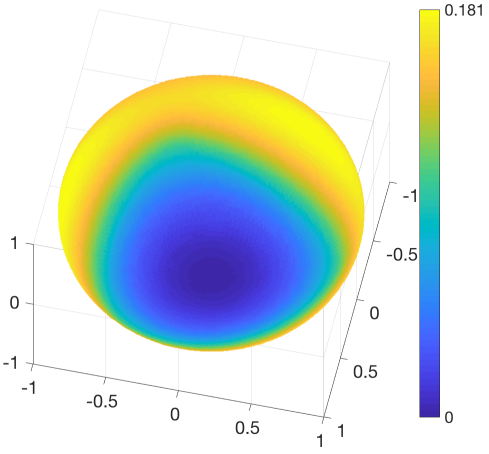
In a challenging problem like this, where the physics itself may not allow for the existence of a reflector with nice properties (from the perspective of manufacturing and outcome), it may also be useful to view our method as a robust way of obtaining a good approximation of the desired reflector. This could then be used to initialize an end-game method, not based on Optimal Transport, that would optimize the reflector surface and enforce any desired smoothness.



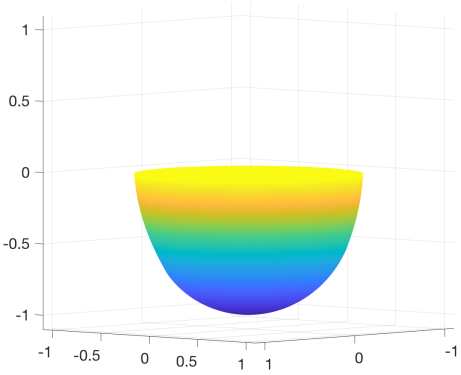

**(a)** The source (left) and target (right) densities from the bottom.



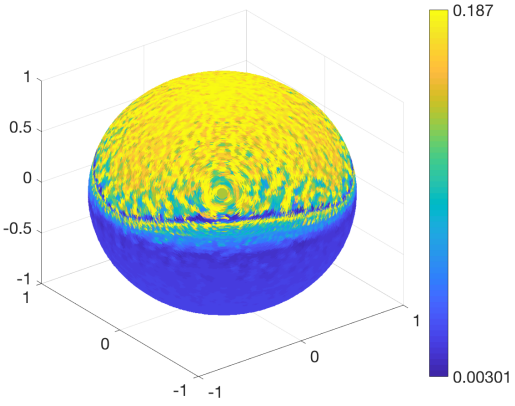

**(b)** The source (left) and target (right) densities from the side.



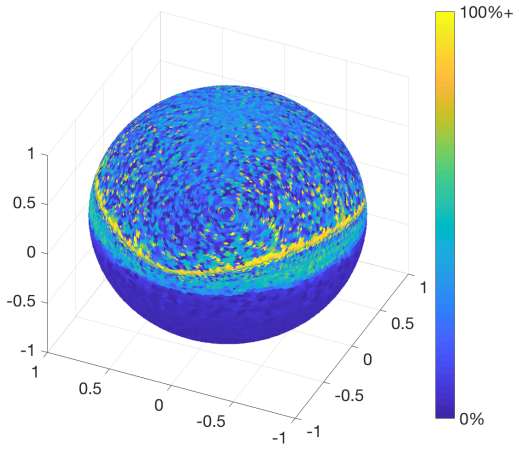

**(c)** The solution $u$.



**(d)** The computed reflector shape.




**(e)** The inverse ray trace.




**(f)** Inverse ray trace error.

**Figure 5.5** Singular reflector. The source and target intensities (Figure 5.5a) and (Figure 5.5b), the solution $u$ (Figure 5.5c), the resulting reflector shape (Figure 5.5d), the inverse ray trace validation (Figure 5.5e), and the absolute error in the inverse ray trace (Figure 5.5f).

# CHAPTER 6

# DIFFEOMORPHIC MAPPING FOR THE MOVING MESH PROBLEM

## 6.1 Introduction

In this chapter, we construct a class of provably convergent methods for computing Optimal Information Transport on smooth, compact, and connected 2D manifolds $M$. We then provide, for the first time, a comparison of the adaptive mesh methods on the sphere via Optimal Transport and Optimal Information Transport. The scheme used for solving the Optimal Transport problem on the sphere was outlined in Chapter 4. We defer the discussion of the issues concerning their generalizations to more general closed, compact surfaces $M$ in Chapter 8. In this more general case, there appear to be more merits to using Optimal Information Transport. The merits include speed of implementation and generalizability in computational terms but also in terms of *a priori* regularity results. Nevertheless, it could be that a particular application requires the computation of Optimal Transport with the squared geodesic cost on the sphere and more general surfaces. Most of the work in this chapter is from the paper Turnquist (2021).

The idea is to use diffeomorphic density matching techniques for moving mesh methods. The theory justifies the use of both Optimal Transport and Optimal Information Transport for doing moving mesh methods on the sphere, at least in the case where both source and target masses are $C^\infty$ and bounded away from zero Bauer et al. (2015); Loeper (2011). Again, the general idea is that we will start with a given or easily generated mesh, complete with $N$ vertices and edges connecting these vertices. The vertices and their edges can be encoded in an adjacency matrix. Then, the local density of vertices $f_0$ is changed to a desired target density $f_1$. The idea is to achieve this without tangling the mesh, that is, without causing the edges to cross each other, see Figure 6.1.

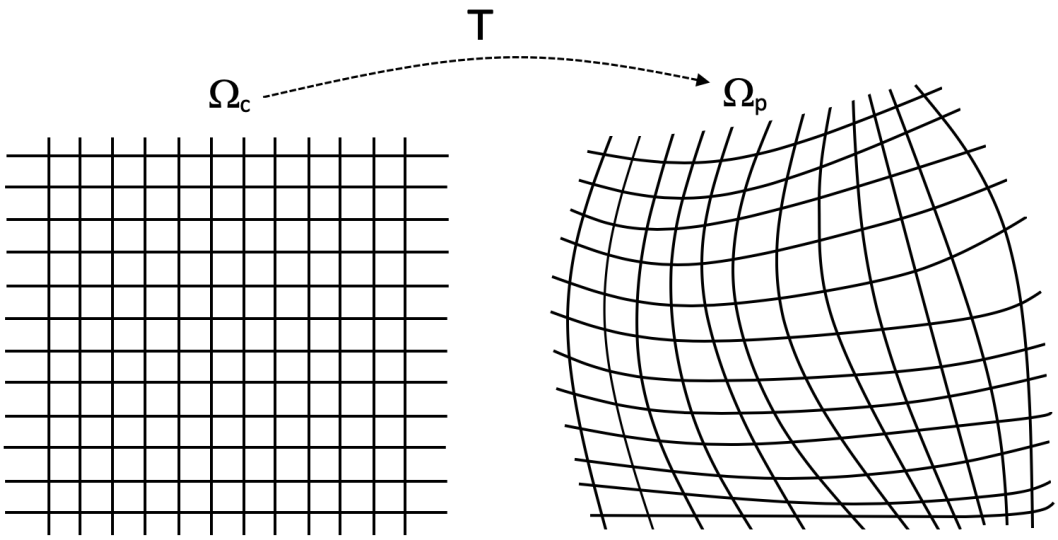



**Figure 6.1** Mapping density of vertices from a computational domain $\Omega_c$ to a physical domain $\Omega_p$.

Here we briefly recapitulate the Optimal Information Transport problem. More detail on the geometry and derivation is shown in Section 2.2 and in the papers Bauer et al. (2015); Modin (2015). The problem of diffeomorphic density matching is that given a path of densities $\mu(t)$, we desire to find the path of diffeomorphisms $\varphi(t)$ which project onto $\mu(t)$, that are also of minimal length with respect to the information metric $G^I$, see Section 2.2 for a definition of this metric. That is, solve the exact density matching problem, that is find a $\varphi(t)$ such that:

$$\begin{cases} \varphi(0) = \mathrm{id}, \\ \varphi^* \mu_0 = \mu(t), \\ \text{minimizing } \int_0^1 G^I_{\varphi(t)} \left(\dot{\varphi}(t), \dot{\varphi}(t)\right) dt. \end{cases} \tag{6.1}$$

Now we assume that $\mu_0 = \mathrm{vol}$ and so $f_0 = 1/\mathrm{vol}(M)$. We take the density matching equation $\varphi^*(t)\mu_0 = \varphi^*(t)\mathrm{vol} = \mu(t)$ and differentiate with respect to $t$, using the formalism of Lie derivatives:

$$\dot{\mu}(t) = \partial_t \left(\varphi(t)^* \mathrm{vol}\right) = \varphi^* \mathrm{div}_{\mathrm{vol}} v(t), \tag{6.2}$$

where $v(t) = \dot{\varphi} \circ \varphi^{-1}$. This can be rewritten as

$$\dot{\mu}(t) = \operatorname{div}(v(t)) \circ \varphi(t)\mu(t). \tag{6.3}$$

From here we perform the Hodge-Helmholz decomposition for the vector field $v$, by writing $v = \text{grad}f + w$. It turns out that the Hodge-Helmholtz decomposition is orthogonal with respect to the information metric $G^I$ and the length of the path $\varphi(t)$ is minimal for $w = 0$ and $\mu(t) = \varphi^*(t)\text{vol}$ is a geodesic path Bauer et al. (2015). Therefore, we can solve the following Poisson equation for the curl-free term $f$:

$$\begin{cases} \Delta f(t) = \frac{\dot{\mu}(t)}{\mu(t)} \circ \varphi(t)^{-1}, \\ \dot{\varphi}(t) = \operatorname{grad}(f(t)) \circ \varphi(t), \quad \varphi(0) = \text{id}. \end{cases} \tag{6.4}$$

The diffeomorphic mapping $T$, is then given by $T = \varphi^{-1}(1)$, where $\varphi(1)$ solves Equation (6.4) Bauer et al. (2015).

## 6.2 Algorithm for Optimal Information Transport

In order to solve the Optimal Information Transport problem on $M$, we must solve Equation (2.45). First, we must perform some pre-computations involving the geometry and the mass densities. Then, we iterate an Euler scheme where each step involves one iteration of a convergent Poisson solver. Such a Poisson solver must be convergent on manifolds without boundary. We implement a monotone Poisson solver whose convergence guarantees are located in Appendix E and are minor adaptations to the convergence framework given in Chapter 3. Here are the steps to compute the approximation to the diffeomorphic mapping $\varphi(1)$. The forward map at a time step $n$ will be denoted by $T_n$ and the corresponding inverse map at a time step $n$ will be denoted by $S_n$. We will outline the algorithm for the sphere, see Algorithm 4, which is inspired by the algorithm shown in Bauer et al. (2015). The algorithm on the sphere requires two important functions: Proj a projection map consistent with the exponential map and Interp, a consistent interpolation

map. For the case of the sphere, with an explicit formula for the exponential map, the function Proj is simple to execute. However, the interpolation function is perhaps more challenging. A robust interpolation is that described in Hamfeldt and Turnquist (2021a), where the values are consistently interpolated over the Delaunay triangles.

**Algorithm 4** Computing the diffeomorphic mapping $\varphi = S(1)$

Initialize $T_0 = \mathrm{id}$;
Initialize $S_0 = \mathrm{id}$;
Fix $\Delta t \ll 1$;
Precompute $\theta$ via quadrature;
**while** $n\Delta t < 1$ **do**
    Compute the density function $\nu_n := \dot{\mu}_n/\mu_n$ using the explicit formulas;
    Interpolate: Interp $\{\nu_n\}$ onto the grid $\{S(x_j)\}_j$;
    Solve $\Delta^h f_n(x_i) = \nu_n\left(S_n(x_i)\right)$ with any monotone, consistent discretization of Equation (E.2);
    Compute $\nabla^h f_n(x_i)$ for all $x_i$;
    Interpolate: $\nabla^h f_n(x_i)$ onto the grid $\{T(x_j)\}_j$;
    Compute $T_{n+1}(x_i) = \mathrm{Proj}\left\{T_n(x_i) + \Delta t \nabla^h f_n\left(T_n(x_i)\right)\right\}$;
    Compute $S_{n+1}(x_i) = S_n\left(\mathrm{Proj}\left\{x - \Delta t \nabla^h f_n(x_i)\right\}\right)$
**end while**

### 6.3 Implementation

The meshes on the sphere are generated as follows. Given $N$ vertices, a cube mesh is generated. That is, vertices are generated on the faces of a cube and an adjacency matrix is fixed for the edge connections between vertices. Each vertex connects to its four nearest neighbors, except for the corner vertices which only connect to the three nearest neighbors. Such a mesh is then projected onto the sphere (and the adjacency matrix is now fixed). After performing the computations for the mapping $T$ given by either Optimal Transport or Optimal Information Transport, the vertices $x_i$ are moved to their new locations $T(x_i)$, the adjacency matrix remains unchanged, and the edges are redrawn, see Figure 6.2 for an example with $N = 2168$ vertices.

All computations were performed on a 13-inch MacBook Pro, 2.3 GHz Intel

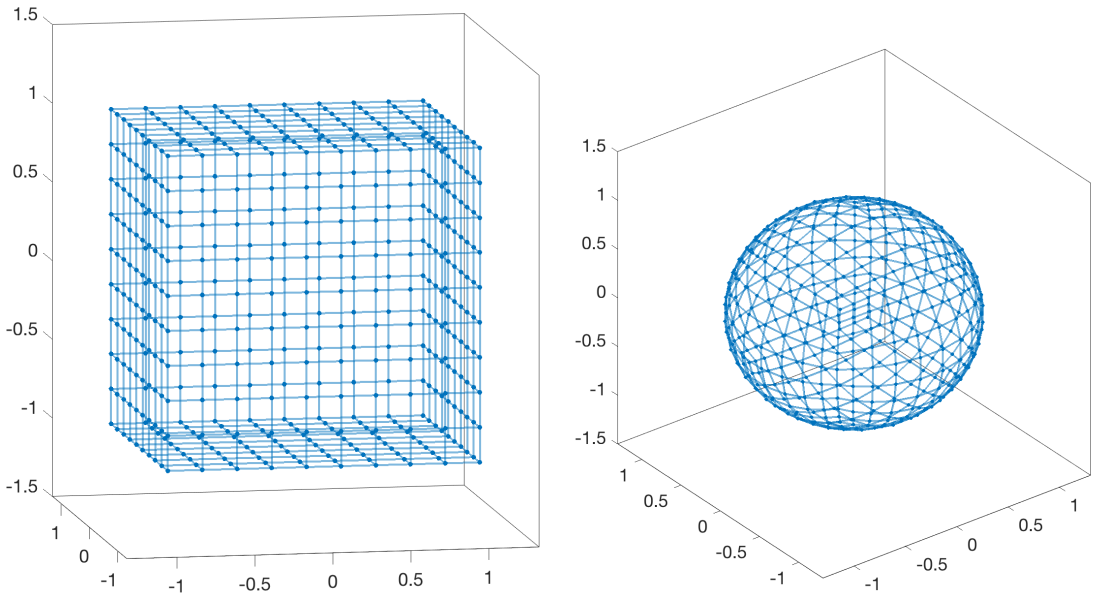



**Figure 6.2** A $N = 2168$-point cube mesh with fixed adjacency matrix (left) then projected to the unit sphere (right).

Core i5 with 16GB 2133 MHz LPPDDR3 using Matlab R2021b.

Our objective is to perform adaptive mesh computations on the sphere. The original mesh will be assumed to be given by some off-the-shelf mesh generator which is simple to implement. Here we will be using a mesh defined on a cube which is then projected onto the sphere. The original mesh, therefore, *does not have constant density*. However, in the moving mesh problem, we will be performing computations that produce diffeomorphic maps from a constant source density to a variable target density. This allows us to redistribute the mesh as desired. That is, we produce an off-the-shelf mesh, have a target mesh in mind, compute the mapping via either Optimal Transport or Optimal Information Transport, and then move the mesh according to the mappings while retaining the same connectivity.

### 6.3.1 Successful Moving Mesh Method Implementations

First, we demonstrate that the computation of mesh redistribution using Optimal Transport and Optimal Information Transport can be both successful in produce meshes which do not exhibit tangling. We select density functions with the goal of producing a transport map $T$ that will concentrate mesh points around the

equator, see Equation (6.5).

$$\begin{cases} f_0(\theta,\phi) = \frac{1}{4\pi}, \\ f_1(\theta,\phi) = \left(1 - \exp\left(-1/30\left(\arccos(z) - \pi/2\right)^2\right)\right)/3.53552. \end{cases} \tag{6.5}$$

The original mesh is generated from the cube mesh that is then projected onto the sphere, see Figure 6.3

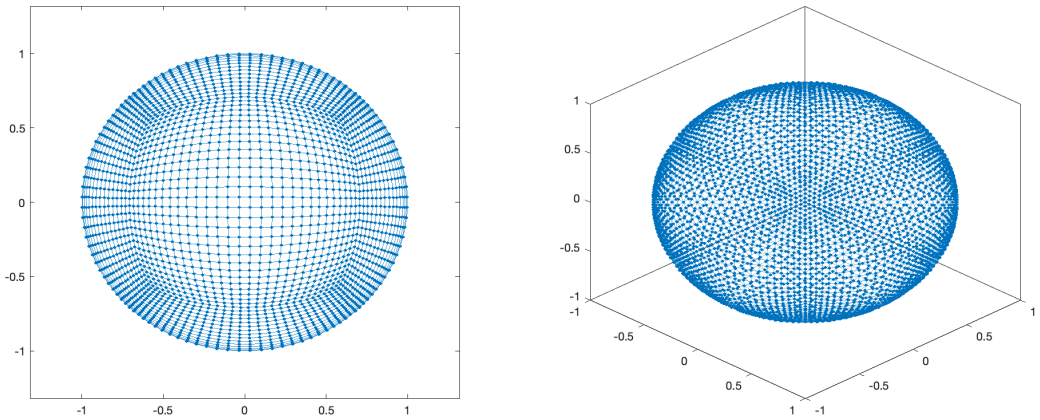



**Figure 6.3** A $N = 5048$-point cube mesh from the top (left) and diagonally above looking down (right).

We then require that the density concentrates about the equator according to Equation (6.5), see the visual representation in Figure 6.4

The resulting mesh restructuring computed using Optimal Transport is pictured in Figure 6.5.

Now we perform the same computation using Optimal Information Transport, see Figure 6.6:

### 6.3.2 Advantage of Optimal Information Transport

Another example serves to demonstrate the versatility and desirability of using Optimal Information Transport for moving mesh problems. We take a constant source density $f_0 = 1/(4\pi)$ and map it to a complicated, discontinuous map of the

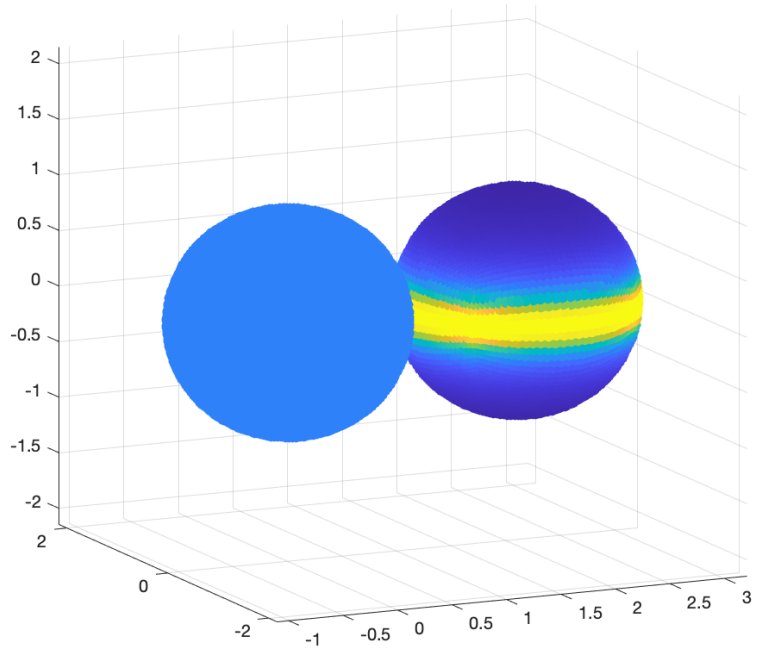


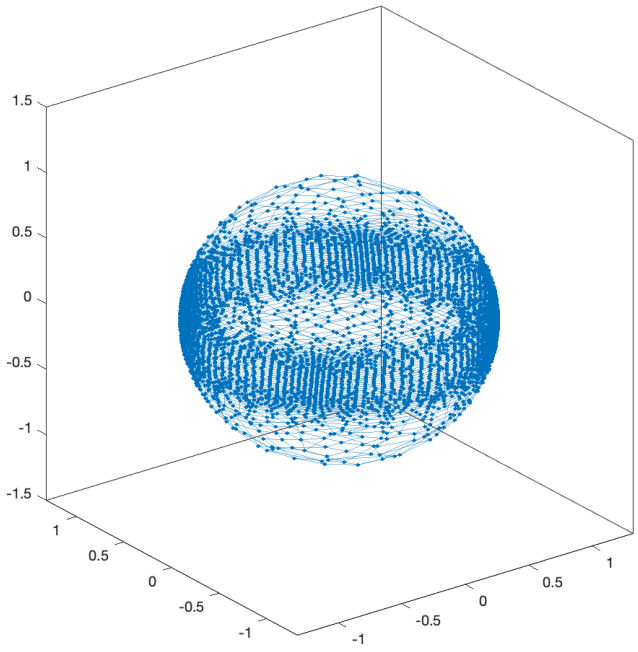

**Figure 6.4** Density in the original mesh (left) and the modified density (right).

world, where a higher density of grid points is required over the oceans and lower density over the continents, see Figure 6.7.

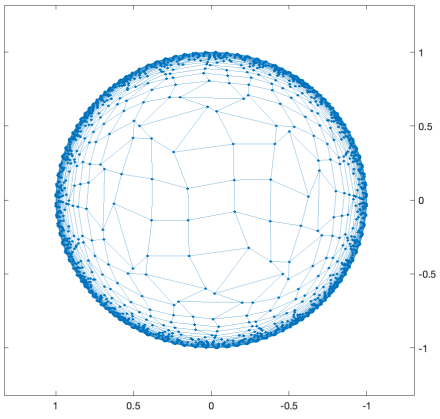
As before, we use the cube mesh in Figure 6.3 and transform it via Optimal Information Transport. The result is shown in Figure 6.8.

Some difficulties are encountered when attempting to perform the same numerical computation via Optimal Transport. First, the necessity of using an interpolation function for the target mass density $f_1$ onto the mesh defined by $T(x)$ leads to instabilities in the scheme due to the very nonlinear nature of the equation we are trying to solve. Compensating for this, by using an inverted image for the source mass $f_0$ and a constant for the target mass $f_1$ will run, but exhibits significant tangling for very complicated images like the world map, see Figure 6.9

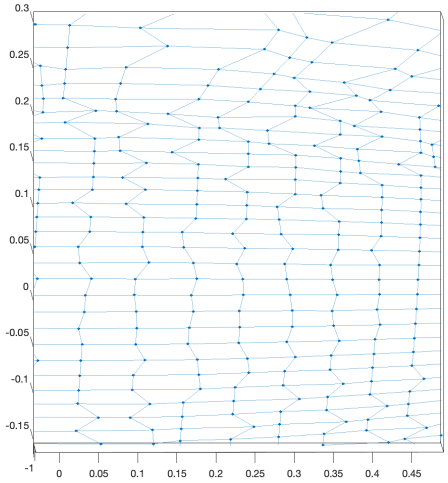
One potential solution to this issue exhibited by the Optimal Transport problem will be explored in further work where the computation of the mapping $T$ is slightly modified to gain convergence guarantees and greater stability.




**(a)** The computed target mesh via Optimal Transport.

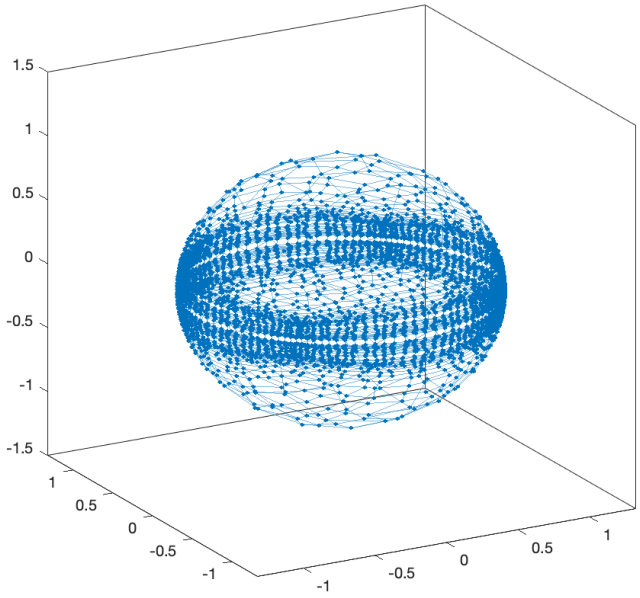



**(b)** The target mesh from above.

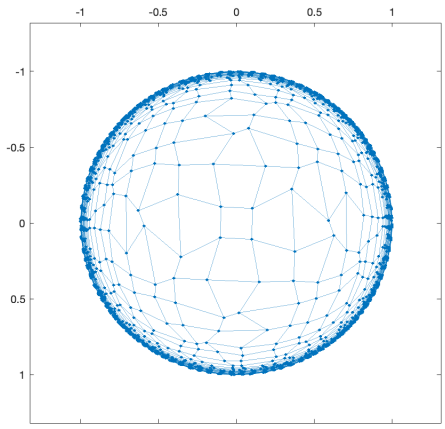



**(c)** Detail showing no tangling of the target mesh.

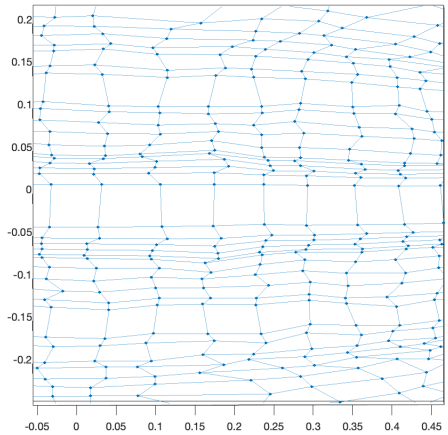
**Figure 6.5** 3D view of the mesh obtained via Optimal Transport map (Figure 6.5a), top view (Figure 6.5b), and a detailed view of the equatorial region (Figure 6.5c) showing that the grid lines do not tangle.




**(a)** The computed mesh via Optimal Information Transport.




**(b)** Top view of the computed mesh.




**(c)** Detail of the equator showing no tangling.

**Figure 6.6** 3D view of the mesh obtained via Optimal Information Transport (Figure 6.6a), top view (Figure 6.6b), and a detailed view of the equator showing that the grid lines do not tangle (Figure 6.6c).

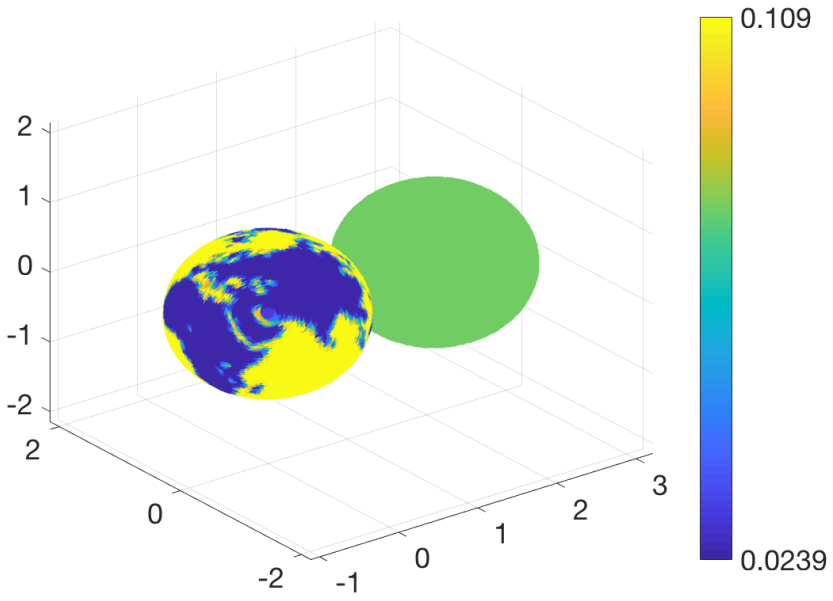



**Figure 6.7** Constant density in the source density (right) and the target globe density (left).

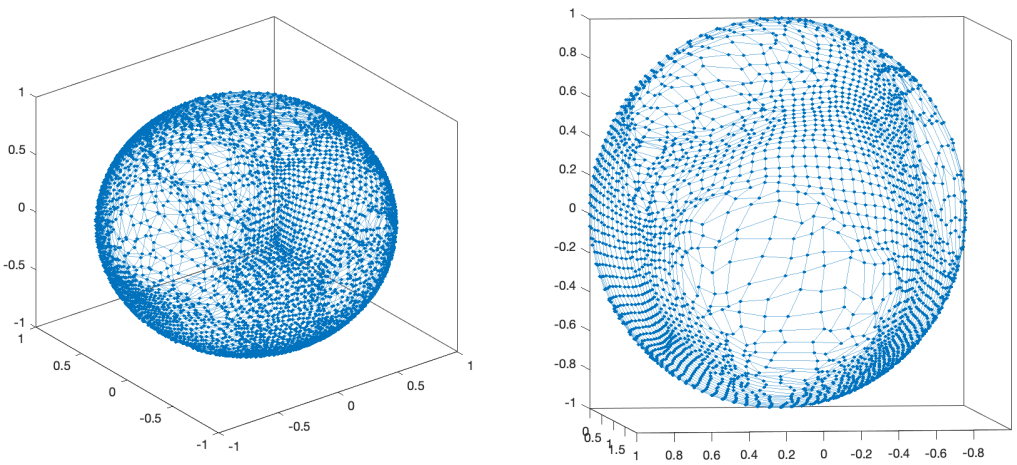


**Figure 6.8** Entire spherical mesh transformed by Optimal Information Transport in 3D view (left). The view is of Africa, Europe and Asia, but the other side of the sphere mesh is also visible. Shown on the right is detail on North America with other side of sphere hidden from view in order to show the mesh more clearly (right). The outlines of North and South America can be discerned and tangling is minimized.

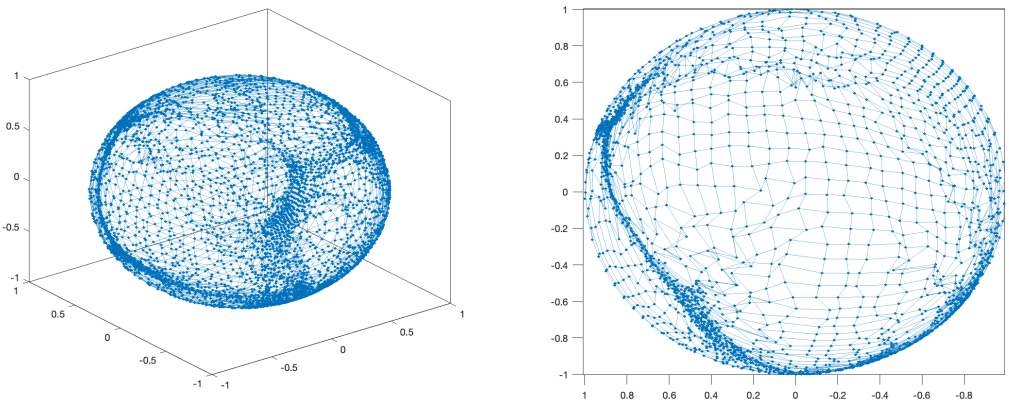



**Figure 6.9** Entire spherical mesh transformed by Optimal Transport in 3D view (left). Shown on the right is detail on Africa, the Middle East, and Asia with other side of sphere hidden from view in order to show the mesh more clearly (right). One can observe some tangling here.

# CHAPTER 7

# TOWARDS CONVERGENCE RATES FOR MONOTONE SCHEMES

## 7.1 Introduction

In this chapter, we develop new convergence rates for monotone numerical schemes for solving linear elliptic partial differential equations (PDEs) posed on compact manifolds. Notably, these rates would apply to the Poisson equation used in the algorithm for the Optimal Information Transport problem in Chapter 6. The surprising result, which is also demonstrated empirically, is that the solution error need not be proportional to the consistency error of the scheme, even in the smoothest problems. For the nonlinear Optimal Transport PDE (Equation (2.14)), the convergence guarantees presented in Chapter 3 only use compactness arguments and thus did not provide error bounds. This chapter represents the first step in establishing convergence rates for the solution of monotone discretizations of Equation (2.14) over the sphere. The majority of the content in this section can also be found in Hamfeldt and Turnquist (2022).

We begin the process of developing convergent rates for numerical schemes on a compact manifold $M$ by considering linear elliptic divergence structure equations of the form

$$L\left(x, u(x), Du(x), D^2u(x)\right) + f(x) = 0, \tag{7.1}$$

where $A(x)$ is a symmetric positive definite matrix and

$$\mathcal{L}[u] \equiv L\left(x, u(x), Du(x), D^2u(x)\right) = -\mathrm{div}_M\left(A(x) D_M u(x)\right). \tag{7.2}$$

which, in local coordinates, has the expression Cabré (2002):

$$\mathcal{L}[u] = \frac{-1}{\sqrt{\det G}} \partial_i \left(\sqrt{\det G} a_{ik} g^{kj} \partial_j u\right), \tag{7.3}$$

where $G$ is the metric tensor and $g^{ij}$ is the $i,j$th entry of the inverse metric tensor $G^{-1}$, and $a_{ij}$ is the $i,j$th entry of $A(x)$.

We notice immediately the nullspace of the PDE operator consists of constants. This in turn requires us to impose an additional condition in order to obtain a unique solution to Equation (7.1). We hereby fix any point $x_0 \in M$ and further impose the condition

$$u(x_0) = 0. \tag{7.4}$$

There exist fairly general conditions upon which there exists a weak $H^1(M)$ solution to Equations (7.1), (7.4) provided the given data satisfies the solvability condition

$$\int_M f(x)dx = 0. \tag{7.5}$$

See Aubin (1998). This solvability condition arises naturally from the fact that $\mathcal{L}$ is self-adjoint and thus

$$\int_M f(x)\,dx = -\int_M \mathcal{L}[u]\,dx = -\int_M u\mathcal{L}^*[1]\,dx = 0.$$

The linearized version of the Monge-Ampère equation arising in Optimal Transport is an example of such a PDE, see Brenner and Neilan (2012). Critically, we will be posing such PDE on compact manifolds $M$ without boundary. Thus, they lack boundary conditions and the usual approaches of establishing convergence rates for numerical schemes do not work.

In this chapter, we investigate the surprising fact that for manifolds without boundary it is possible to construct simple monotone discretizations of linear elliptic PDEs in 1D for which the empirical convergence rate is asymptotically worse than the formal consistency error. In contrast, discrete solutions of the Dirichlet problem are expected to converge on the order of their formal consistency error.

Buttressing this, we derive explicit convergence rates on 2D manifolds without boundary and observe that the bounds are of order $\mathcal{O}\left(h^{\alpha/(d+1)}\right)$ where $h^\alpha$ is the formal consistency error and $d = 2$ is the dimension of the manifold. This

somewhat surprising result demonstrates even more clearly the need to design higher-order (consistency error) schemes for solving such elliptic PDE on manifolds without boundary. Furthermore, we show how this convergence result can immediately be bootstrapped into a convergence result for a discrete gradient. Future work involves relating this convergence result for linear elliptic PDE in divergence form to nonlinear elliptic PDE (whose linearization becomes a linear elliptic PDE in divergence form).

## 7.2 Empirical Evidence of Suboptimal Convergence Rates

For the remainder of this chapter, we will have a manifold $M$ without boundary and $N$ discretization points $\mathcal{G} \subset M$. The parameter $h$ denotes the following spatial resolution of $\mathcal{G}$:

$$h = \sup_{x \in M} \min_{y \in \mathcal{G}} d_M(x, y), \tag{7.6}$$

where $d_M(x, y)$ denotes the Riemannian distance between $x$ and $y$. This choice of $h$ makes sure that there exists a point $y$ on the grid $\mathcal{G}$ within a distance $h$ of any point $x$ on the manifold.

For a monotone numerical discretization $L^h$ of a linear elliptic PDE $L$ on a manifold without boundary that has consistency error $\mathcal{O}\left(h^\alpha\right)$, we may hypothesize that the numerical solution $u^h$ of such a scheme satisfies $\left|u^h - u\right|_\infty = \mathcal{O}\left(h^\alpha\right)$. That is, the rate of convergence of the discrete solution is the same as the consistency error. In this section, we will show empirical evidence that this is not true for manifolds without boundary.

In fact, in Section 7.3, we will derive that the convergence rate of the discrete solution of the monotone discretization satisfies:

$$\left|u^h - u\right| = \mathcal{O}\left(h^{\alpha/(d+1)}\right), \tag{7.7}$$

where $d$ is the dimension of the manifold. In this section, our empirical example

will be given for the 1D torus. Therefore, we will demonstrate in this section that we can construct monotone schemes that have empirical convergence rate:

$$\left|u^h - u\right| = \mathcal{O}\left(h^{\alpha/2}\right). \tag{7.8}$$

### 7.2.1 The Dirichlet Problem

We are trying to solve a PDE on a manifold without boundary (and, therefore, the PDE lacks boundary conditions). Such PDE in general lack a comparison principle. First, we illustrate how we can derive simple convergence bounds when we have a Dirichlet problem. These convergence bounds critically utilize the maximum principle and the bounds are consequently proportional to the consistency error. Let's start with the Dirichlet problem on a subset of Euclidean space $\Omega \subset \mathbb{R}^2$:

$$\begin{cases} -\Delta u(x) + f(x) = 0, & x \in \Omega, \\ u(x) = g(x), & x \in \partial\Omega. \end{cases} \tag{7.9}$$

Suppose, in addition, that we have a consistent monotone discretization scheme

$$\begin{cases} L^h u_i^h + f(x_i), & x_i \in \Omega, \\ u^h = g, & x_i \in \partial\Omega \end{cases} \tag{7.10}$$

with truncation error $L^h\phi_i + f(x_i) = \tau_i(h)$, $|\tau_i(h)| \leq \tau(h)$.

Let $z^h = u - u^h$. We will have $z^h(x_i) = 0$ for $x_i \in \partial\Omega$. Now, choose some $w$ such that $L^h w_i \geq 1$, $\forall i$ with $w(x_i) = 0$ on $\partial\Omega$. Define $v_i^{\pm} \equiv \pm z_i^h - \left\|L^h z^h\right\|_{L^\infty(\Omega)} w(x_i)$. Then,

$$L^h v_i^{\pm} = \pm L^h z_i^h - \left\|L^h z^h\right\|_{L^\infty(\Omega)} L^h w_i \leq \pm L^h z_i^h - \left\|L^h z^h\right\|_{L^\infty(\Omega)} \leq 0. \tag{7.11}$$

Applying the maximum principle, we get:

$$\pm z_i^h - \left\|L^h z^h\right\|_{L^\infty(\Omega)} w_i \leq \left\|z^h\right\|_{L^\infty(\partial\Omega)} - \left\|L^h z^h\right\|_{L^\infty(\Omega)} \|w\|_{L^\infty(\Omega)} = 0. \tag{7.12}$$

Thus, we have

$$\left\|z^h\right\|_{L^\infty(\Omega)} \leq \left\|L^h z^h\right\|_{L^\infty(\Omega)} \|w\|_{L^\infty(\Omega)} = \tau(h) \|w\|_{L^\infty(\Omega)}. \tag{7.13}$$

Thus, the error bounds follow from *a priori* bounds on the solution of, for example,

$$\begin{cases} -\Delta w = 3/2, \quad x \in \Omega, \\ w = 0, \quad x \in \partial\Omega. \end{cases} \tag{7.14}$$

### 7.2.2 The PDE on the Torus without Boundary Conditions

However, our PDE are manifestly different from the Dirichlet problem. To use an even simpler example, suppose we aim to solve:

$$\begin{cases} -u''(x) = 0, \quad x \in \mathbb{T}_1, \\ u(0) = 0 \end{cases} \tag{7.15}$$

on the 1d torus $\mathbb{T}^1$. The condition $u(0) = 0$ is chosen for uniqueness, but it is not a boundary condition. Notice how this differs from the Dirichlet problem. In the Dirichlet problem:

$$\begin{cases} -u''(x) = 0, \quad x \in \mathbb{T}_1 \setminus 0, \\ u(0) = 0. \end{cases} \tag{7.16}$$

We build uniqueness in from the start by designing a proper, monotone scheme:

$$\begin{cases} L_i^h[v] + hv_i = f_i^h, \\ u_i = v_i - v(0). \end{cases} \tag{7.17}$$

Notice that $v$ is well-defined and uniformly bounded. Empirically, we find that $v$ is $\mathcal{O}(1)$ (but not smaller). The equation does $u$ then satisfies

$$L_i^h[u] + hu_i = f_i^h - hv(0). \tag{7.18}$$

From bounds on $v$, this has consistency error $\mathcal{O}(h)$. Note also that by design, we have $u(0) = 0$. Now, let $w$ solve

$$\begin{cases} L_i^h[w] + hw_i = 1, \quad x_i \neq 0, \\ w(0) = 0. \end{cases} \tag{7.19}$$

Empirically, $\|w\|_\infty = C/h$, see Figure 7.1. Applying the maximum principle to $\pm u_i - \left\|L^h u + hu_i\right\|_\infty w_i$, we find that

$$\pm\, u_i - w_i \left\|f^h - hv(0)\right\|_\infty \leq \pm u(0) - w(0) \left\|f^h - hv(0)\right\|_\infty = 0, \tag{7.20}$$

$$\implies \|u\|_\infty \leq \|w\|_\infty \left\|f^h - hv(0)\right\|_\infty = \mathcal{O}(1). \tag{7.21}$$

So this approach does not yield convergence! We can observe, empirically, a convergence rate of $\mathcal{O}\left(\sqrt{h}\right)$ by choosing the following modification of the scheme. Note that the scheme will still be monotone and proper:

$$\begin{cases} L_i^h[v] + hr(x_i)v_i = f_i^h, \\ u_i = v_i - v(0), \end{cases} \tag{7.22}$$

where $r(x) > 0$. E.g., take $f^h = h$, $r(x) = x + 1$. Empirically: we have $\|w\|_\infty = \mathcal{O}(1/h)$, again see Figure 7.1 and $\|v\|_\infty = \mathcal{O}(1)$. If $N$ is chosen to be multiples

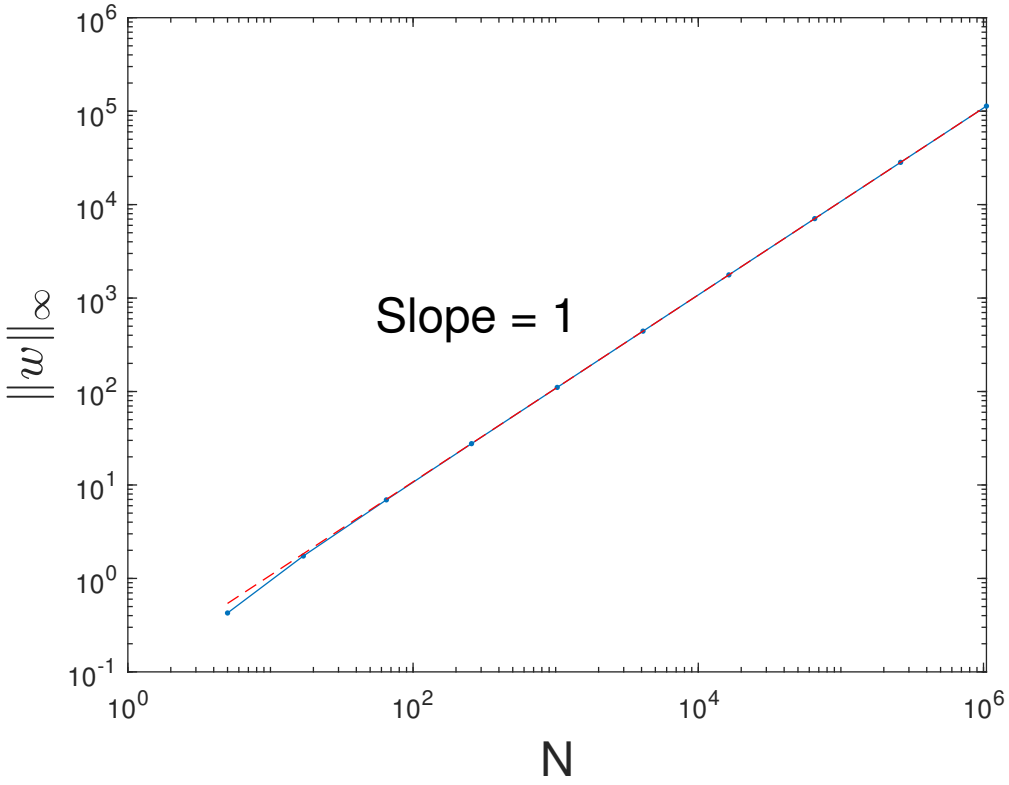



**Figure 7.1** Empirically, $\|w\|_\infty = C/h$ for the problem without boundary conditions.

of 4, then we get a particularly clean plot with observed convergence rate of $\mathcal{O}\left(h^{\alpha/(d+1)}\right)$, where again, for our scheme $\alpha = 1$, see Figure 7.2.

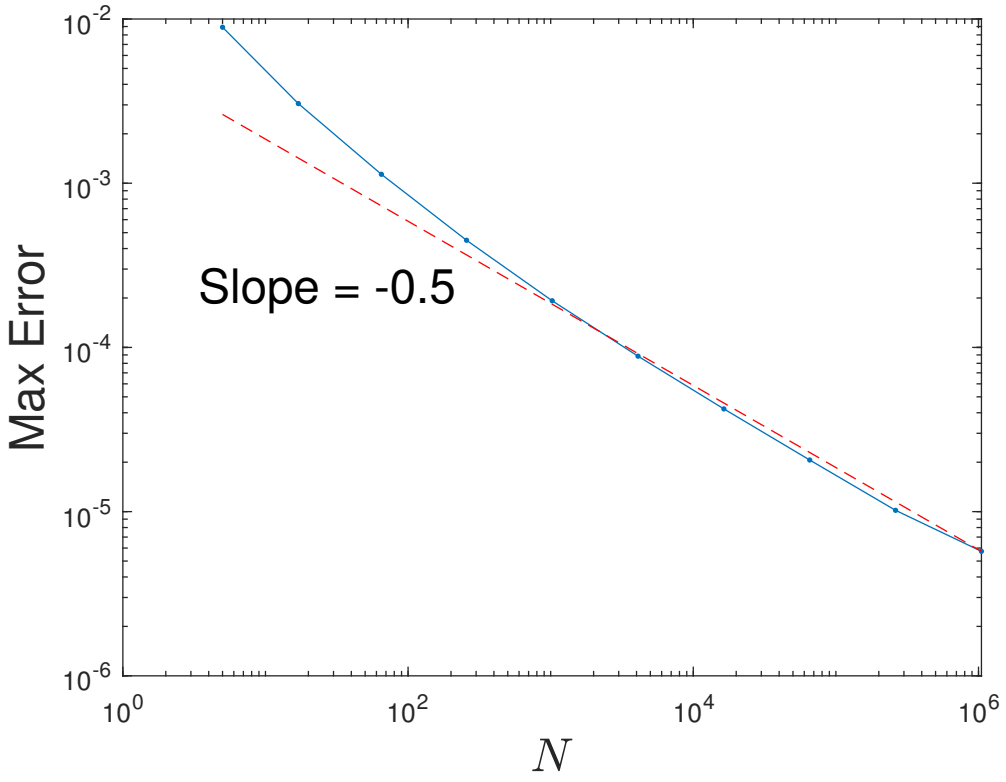



**Figure 7.2** We observe $\mathcal{O}\left(h^{\alpha/(d+1)}\right)$ convergence of the discrete solution $u^h$ to the solution $u = 0$ on $\mathbb{T}_1$, where $\alpha = 1$.

## 7.3 Convergence Rates for Linear Elliptic PDE on Compact Manifolds

We now establish error bounds for a class of consistent, monotone approximations schemes for Equations (7.1), (7.4). The main result is presented in Theorem 7.7. The approach we take here is to construct barrier functions, which are shown to bound the error via the discrete comparison principle. Importantly, the error estimates we obtain are consistent with the empirical convergence rates observed in subsection 7.2.

### 7.3.1 Hypotheses on Geometry and PDE

We begin with the hypotheses on the geometry $M$ and Equation (7.1) that are required by our convergence result.

**Hypothesis 7.1** (Conditions on PDE and manifold)**.** *The manifold $M$ and PDE (Equation (7.1)) satisfy:*

1. *The manifold $M$ is a 2D compact orientable surface without boundary.*
2. *The matrix $A(x) \in C^2(M)$ is symmetric positive definite.*
3. *The function $f(x) \in C^1(M)$ satisfies $\int_M f(x)\, dx = 0$.*

**Remark 7.2.** *The compactness of the 2D manifold $M$ implies that it is geodesically complete, has injectivity radius strictly bounded away from zero, and that the sectional curvature (equivalent to the Gaussian curvature in 2D) is bounded from above and below Lee (2006).*

**Remark 7.3.** *The fact that $M$ is 2-dimensional can quite easily be generalized, since nothing in our convergence theorem fundamentally depends upon the dimension $d$ (though the convergence bound does vary with $d$). We emphasize, however, that the lack of boundary on our manifold $M$ is an essential point in this chapter.*

### 7.3.2 Approximation Scheme

Next, we describe the class of approximation schemes that are covered by our convergence result. The starting point of the scheme is the idea that the uniqueness constraint (Equation (7.4)) should be posed at the point $x_0$, with a reasonable discrete approximation of the PDE posed on other grid points. However, as discussed in subsection 7.2, this approach may not yield a convergent scheme. Instead, we will create a small cap around $x_0$ and fix the values of $u$ at all points in this cap.

To construct an appropriate scheme, we begin with any finite difference approximation $L^h(x, u(x) - u(\cdot))$ of the PDE operator in Equation (7.2) that is defined for $x \in \mathcal{G}^h$ and that satisfies the following hypotheses.

**Hypothesis 7.4** (Conditions on discretization scheme)**.** *We require the scheme $L^h$ to satisfy the following conditions:*

1. *$L^h$ is linear in its final argument.*
2. *$L^h$ is monotone.*
3. *There exist constants $C, \alpha > 0$ such that for every smooth $\phi \in C^2(M)$ the consistency error is bounded by*

$$\left|L^h(x, \phi(x) - \phi(\cdot)) - L(x, \phi(x), D\phi(x), D^2\phi(x))\right| \leq C[\phi]_{C^{2,1}(M)} h^\alpha, \quad x \in \mathcal{G}^h.$$

Next we define some regions in the manifold $M$ that will be used to create "caps" where $u$ is fixed in this scheme, and where additional conditions will be posed on barrier functions. Choose any $0 < \gamma < \alpha$. Define the regions

$$\begin{aligned} b^h &= \{x \in M \mid d_M(x, x_0) < h^\gamma\}, \\ S^h &= \{x \in M \mid h^\gamma < d_M(x, x_0) \leq 2h^\gamma\}, \\ B^h &= M \setminus (b^h \cup S^h). \end{aligned}$$

See Figure 7.3.

We then define the modified scheme $F^h$ as follows:

$$F^h(x, u(x), u(x) - u(\cdot)) \equiv \begin{cases} L^h(x, u(x) - u(\cdot)) + h^\alpha u(x) + f(x), & x \in B^h \cap \mathcal{G}^h \\ u(x), & x \in (S^h \cup b^h) \cap \mathcal{G}^h. \end{cases} \tag{7.23}$$

**Remark 7.5.** *The condition $u(x) = 0$, $x \in S^h \cup b^h$ can be relaxed provided the resulting discrete solution has a uniformly bounded Lipschitz constant in this region and the values of $u$ are close to zero. Pinning the value to zero has the particularly strong effect of setting the local Lipschitz constant to zero.*

Note that the discretization $F^h$ is automatically proper by construction. Therefore, this scheme has a uniformly bounded solution and satisfies the discrete comparison principle as shown in Oberman (2006).

**Lemma 7.6.** *Under the assumptions of Hypotheses 7.1,7.4, the discrete scheme*

$$F^h(x, u^h(x), u^h(x) - u^h(\cdot)) = 0 \tag{7.24}$$

*has a unique solution $u^h$ that is bounded uniformly independent of $h$ for sufficiently small $h > 0$.*

*Proof.* Existence of $u^h$ follows from the fact that $F^h$ is proper Oberman (2006). Now let $u \in C^{2,1}(M)$ be the unique solution of Equations (7.1), (7.4). From the consistency of $L^h$, we have that for any constant $K \in \mathbb{R}$

$$F^h(x, u(x) - K, (u(x) - K) - (u(\cdot) - K)) \leq Ch^\alpha - Kh^\alpha,$$

where $C$ is a constant depending on $\|u\|_{C^{2,1}(M)}$ and $h > 0$ is sufficiently small. Choosing any $K > C$ yields

$$F^h(x, u(x) - K, (u(x) - K) - (u(\cdot) - K)) < 0 = F^h(x, u^h(x), u^h(x) - u^h(\cdot)).$$

By the discrete comparison principle, we have

$$u^h \geq u - K \geq -\|u\|_{L^\infty(M)} - K.$$

Similarily, we obtain

$$u^h \leq \|u\|_{L^\infty(M)} + K.$$

□

## 7.4 Convergence Rates

The idea in this section is to establish the convergence of the discrete solution of a monotone (and proper) scheme to the unique solution of the underlying PDE.

We accomplish this by constructing barrier functions $\phi_{\pm}^h$ such that

$$F^h[\phi_-^h] \leq F^h[u^h - u] \leq F^h[\phi_+^h] \tag{7.25}$$

and then by invoking the discrete comparison principle to conclude that

$$\phi_-^h \leq u^h - u \leq \phi_+^h. \tag{7.26}$$

These barrier functions can be chosen to satisfy $\phi_{\pm}^h = \mathcal{O}\left(h^{\alpha/(d+1)}\right)$. In this article, we explicitly treat the case $d = 2$. In Section 7.2, we saw for $\mathbb{T}_1$ that the empirical convergence rate was $\mathcal{O}\left(h^{\alpha/2}\right)$, which is consistent with our theoretical error bound when $d = 1$. The factor $(d+1)$ appears because there is a contribution of $d$ from the dimension of the underlying manifold (which arises due to the solvability condition (Equation (7.5)), and a contribution of 1 from deriving a Lipschitz bound (also constrained by the solvability condition). Thus, we see that it is the solvability condition on the manifold without boundary that leads to the reduced convergence rate overall of a monotone and proper discretization.

We state the main convergence result:

**Theorem 7.7** (Convergence Rate Bounds for Smooth Case)**.** *Under the assumptions of Hypotheses 7.1 and 7.4, let $u \in C^{2,1}(M)$ be the solution of Equations* (7.1)*,* (7.4)*. Then the discrete solution $u^h$ solving Equation* (7.24) *satisfies*

$$\left\|u^h - u\right\|_{L^\infty(M)} \leq Ch^{\alpha/3}, \tag{7.27}$$

*where $C > 0$ is a constant independent of $h$.*

**Remark 7.8.** *This convergence rate can be extended to more general $d$-dimensional manifolds as follows:*

$$\left\|u^h - u\right\|_{L^\infty(M)} \leq Ch^{\alpha/(d+1)}. \tag{7.28}$$

We will make use of two theorems from differential geometry. The first theorem is the Conjugate Point Comparison Theorem, which tells us that curvature

bounds from above allow us to assert that the conjugate point of a point $x \in M$ is at least a certain distance away, see also Lee (2006):

**Theorem 7.9** (Conjugate Point Comparison Theorem Lee (2006))**.** *Suppose that all sectional curvatures of $M$ are bounded above by a constant $\kappa_+$. If $\kappa_+ \leq 0$, then no point of $M$ has conjugate points along any geodesic. If $\kappa_+ = 1/R^2 > 0$, then the first conjugate point along any geodesic occurs at a distance of at least $\pi R$.*

The Conjugate Comparison Theorem then will allow us to satisfy a condition in the Rauch comparison theorem, which allows us to use curvature bounds in constant curvature spaces to bound in turn the magnitude of the Jacobian on our manifold, see Lee (2006) for more information:

**Theorem 7.10** (Rauch Comparison Theorem Lee (2006))**.** *Let $M$ and $\tilde{M}$ be Riemannian manifolds, let $\gamma : [0,T] \to M$ and $\tilde{\gamma} : [0,T] \to M$ be unit speed geodesic segments such that $\tilde{\gamma}(0)$ has no conjugate points along $\tilde{\gamma}$, and let $J, \tilde{J}$ be normal Jacobi fields along $\gamma$ and $\tilde{\gamma}$ such that $J(0) = \tilde{J}(0) = 0$ and $|D_t J(0)| = \left|\tilde{D}_t \tilde{J}(0)\right|$. Suppose that the sectional curvatures of $M$ and $\tilde{M}$ satisfy $K(\Pi) \leq \tilde{K}(\tilde{\Pi})$ whenever $\Pi \subset T_{\gamma(t)}M$ is a 2-plane containing $\dot{\gamma}(t)$ and $\tilde{\Pi} \subset T_{\tilde{\gamma}(t)}\tilde{M}$ is a 2-plane containing $\dot{\tilde{\gamma}}(t)$. Then $|J(t)| \geq \left|\tilde{J}(t)\right|$ for all $t \in [0,T]$.*

### 7.4.1 Barrier Functions

We now define the barrier functions $\phi_\pm^h$ by solving a linear PDE on the manifold $M$ with an appropriately chosen (small) right-hand side $f^h$ that satisfies the solvability condition (Equation (7.5)). In particular, given a fixed $C_0 > 0$ (which will be fixed later), we let $\phi_\pm^h$ be the solutions of the PDE

$$\begin{cases} \mathcal{L}[\phi_\pm^h] = \pm f^h(x), \quad x \in M, \\ \phi_\pm^h(x_0) = \pm C_0 h^\gamma. \end{cases} \tag{7.29}$$

We emphasize that while the barrier functions $\phi_\pm^h$ depend on the grid parameter $h$, they are solutions of the PDE on the continuous level.

Below, we outline the construction of the upper barrier $\phi_+^h$; the lower barrier is similar. We begin by choosing a function $f^h(x)$ that ensures that $\phi_+^h \in C^{2,1}(M)$:

$$f^h(x) = \begin{cases} \frac{\kappa(h)}{|B^h|}, & x \in B^h, \\ \tilde{\psi}^h(x), & x \in S^h, \\ -\frac{\kappa(h)}{A^h}, & x \in b^h, \end{cases} \tag{7.30}$$

where $|B^h|$ is the area of the region $B^h$ and $\kappa(h) = \mathcal{O}(h^\alpha)$ will be fixed later. Below, we will show how to choose the function $\psi$ and the constant $A^h$. The choice of $f^h$ is constrained by the compatibility condition $\int_M f^h(x)dx = 0$. See Figures 7.3 and 7.4 for two complementary visualizations of the function $f^h(x)$ on the manifold $M$.

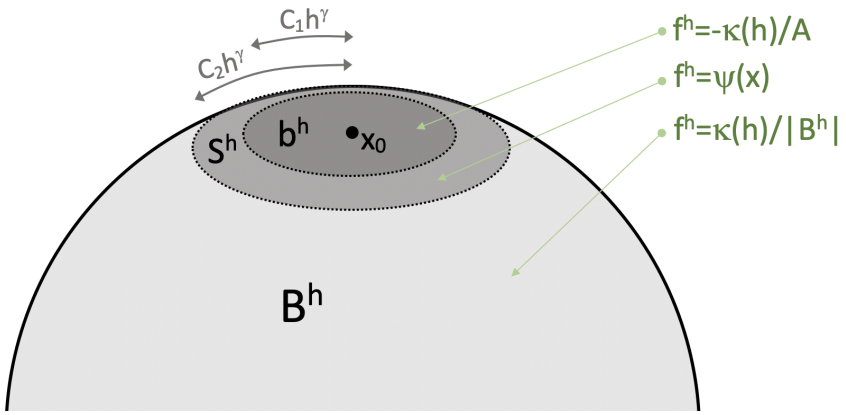



**Figure 7.3** The construction of the barrier functions $\phi_\pm^h$ on the manifold $M$.

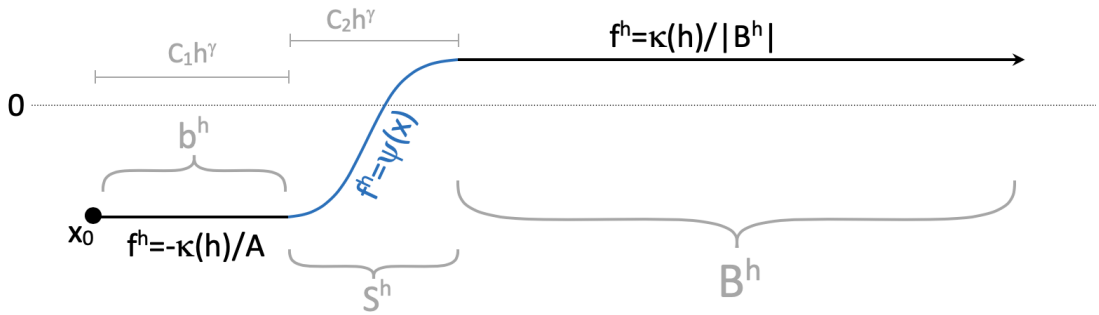



**Figure 7.4** The construction of the function $f_\pm^h$ from a "side profile" parametrized by distance from the point $x_0$.

**Lemma 7.11.** *The function $f^h$ can be chosen to satisfy $f^h \in C^1(M)$.*

*Proof.* We demonstrate this in the case $\phi_+^h$ for simplicity of exposition. We need to show that we can pick a cutoff function $\psi(t)$ such that $\psi \in C^1$ uniformly in $h$ and $\int_M f^h = 0$. Define the constants:

$$\sigma_\pm \equiv \frac{\kappa(h)}{2}\left(\frac{1}{|B^h|} \pm \frac{1}{A^h}\right), \tag{7.31}$$

where we choose

$$A^h = \frac{\frac{|S^h|}{2} + |b^h|}{1 + \frac{|S^h|}{2|B^h|}}, \tag{7.32}$$

where $|S^h|$ is the area of the region $S^h$. Without loss of generality, we choose the function $\psi(t)$ to satisfy:

$$\int_{S^h} \psi dx = \sigma_-. \tag{7.33}$$

The amplitude of the function $\psi(t)$ must be $\mathcal{O}(h^{\alpha/3})$ and the width of $S^h$ is $\mathcal{O}(h^{\alpha/3})$. Thus, we will explicitly fit a one-parameter cosine function with amplitude $\sigma_+$ that is shifted by the value $\sigma_-$ in order to satisfy the requirements of the function $f^h$. In the case of the sphere, the existence of such a smooth cutoff function in the strip $S^h$ boils down to the existence of a one-dimensional smooth cutoff function $\psi(t)$ chosen to satisfy

$$\int_0^1 (\psi(t) - \sigma_-) \sin(tr_1 + (1-t)r_0)\, dt = 0. \tag{7.34}$$

Suppose we want an integral $\int_0^{\phi(1)} \Psi(s)\, ds = 0$. A simple choice for $\Psi$ such that $\Psi'(0) = \Psi'(1) = 0$ is to choose $\Psi(s) = -\sigma_+ \cos\left(\frac{\pi s}{\phi(1)}\right)$. Then, define the change of variables $s = \phi(t)$. In the variable $t$, the integral becomes $\int_0^1 \Psi(\phi(t))J(\phi(t))dt = \int_0^1 \Psi(\phi(t))\phi'(t)dt = 0$. In the case of the sphere, comparing this with Equation (7.34) we see that we should take:

$$\psi(t) = \sigma_- - \sigma_+ \cos\left(\pi \frac{\cos\left(tr_1 + (1-t)r_0\right) - \cos r_0}{\cos r_1 - \cos r_0}.\right) \tag{7.35}$$

From this choice, one can readily see that $\psi'(t) = \mathcal{O}\left(h^\gamma\right)$. We use this choice of $\psi(t)$ to build up the function

$$\tilde{\psi}^h(x) = \psi\left(\frac{d(x, x_0) - h^\gamma}{h^\gamma}\right). \tag{7.36}$$

Since $f^h \in C^1(M)$ and $\int_M f^h = 0$, the result for the case $M = \mathbb{S}^2$ follows.

Beyond the case of the sphere, we must assume both negative and positive curvature bounds. We need a formula for the Jacobian in the strip $S^h$ of the exponential map from $x_0$ as a function of the radius from $x_0$. For small enough $h$, this Jacobian exists (by the positive curvature bound, see Theorem 7.10) and is given by the solution to the Jacobi field equation, see Berger (2003); Carmo (2016), for example. Then, $\phi'(t) = J(t)$ allows us to solve for $\phi(t)$. Since we are concerned with the magnitude of the Jacobian, we will use the Rauch Comparison Principle, see Theorem 7.10. We are allowed to use this since our curvature bound from above allows us to uses the Conjugate Point Comparison Theorem (see Theorem 7.9), which in turn proves that there are no conjugate points in the strip $S^h$. The curvature bound from below allows us to bound the magnitude of the Jacobian from above.

Endowed thus with curvature bounds from above and below: $\kappa_- \leq K \leq \kappa_+$ will allow us to assert $|J| \leq \sinh\left(\sqrt{-\kappa_-}t\right)/\sqrt{-\kappa_-}$, see Berger (2003) for more detail about constant curvature Jacobi fields. Thus,

$$0 \leq \psi'(t) = \sin\left(\frac{\pi\phi(t)}{\phi(1)}\right)\phi'(t) \leq \sin\left(\frac{\pi\phi(t)}{\phi(1)}\right)\sinh\left(\sqrt{-\kappa_-}\left(tr_1 + (1-t)r_0\right)\right)/\sqrt{-\kappa_-} \tag{7.37}$$

and thus again, in the general case $\psi'(t) = \mathcal{O}\left(h^\gamma\right)$. Thus, we have explicitly

constructed a uniformly $C^1$ function $f^h$ on the whole manifold $M$ such that $\int_M f^h = 0$. □

One preliminary result we need is that $\phi^h_\pm$ are bounded.

**Lemma 7.12.** *There exist bounded barrier functions $\phi^h_\pm$ solving Equation* (7.29).

*Proof.* By Theorem 4.7 of Aubin (1998) we have that for a fixed $h$, solutions of $\mathcal{L}\phi^h_\pm = \pm f^h(x)$ are unique up to a constant. The condition $\phi^h_\pm(x_0) = \pm C_0 h^\gamma$ then fixes that constant. Then, from Theorem 4.7 of Aubin (1998) again, we have that if $f^h \in C^k(M)$, then $\phi^h_\pm \in C^{k+2}(M)$. This ensures that for a fixed $h$, the functions $\phi^h_\pm$ are bounded. □

Now, we establish an explicit maximum bound on the magnitude of $\phi^h_\pm$.

**Lemma 7.13** (A Maximum Bound)**.** *For small enough $h$, there exists a constant $C > 0$ such that the barrier functions $\phi^h_\pm$ can be bounded over the entire manifold $M$ as follows:*

$$\left\|\phi^h_\pm\right\|_{L^\infty(M)} \leq C\left(h^\gamma + \frac{\kappa(h)}{h^\gamma}\right). \tag{7.38}$$

*Proof.* We use the fact that we can cover the manifold $M$ with a finite covering $\{B^i_r\}_{i=1,\ldots,n}$, (where the choice of $r$ can be chosen uniformly) established in Theorem G.1 and use the fact that there exists a change of coordinates which expresses the PDE (Equation (7.2)) as a uniformly elliptic PDE in divergence form where the differential operators can be interpreted in the usual Euclidean sense Appendix F. What this allows us to do then, is simply apply Euclidean bounds for uniformly elliptic divergence form PDE in patches over the manifold $M$.

We show the bound just for $\phi^h_+$. Denote

$$\bar{\phi}^h_+ \equiv \phi^h_+ - \inf_M \phi^h_+, \tag{7.39}$$

where $\inf_M \phi^h_+$ exists by Lemma 7.12. This is non-negative, which allows us to apply the de Giorgi-Nash-Moser Harnack inequality, which applies to PDEs in divergence form. The PDE (Equation (7.1)) is in divergence form if expressed in

local coordinates ( Appendix F). See Theorems 8.17 and Theorem 8.18 in Gilbarg and Trudinger (2001), since $\bar{\phi}_+^h$ is a super-solution of Equation (7.29). Then for any ball $B_r$, we have:

$$\sup_{B_r} \bar{\phi}_+^h \leq C \left( \inf_{B_r} \bar{\phi}_+^h + r^{2\delta} \left\| f^h \right\|_{L^{q/2}} \right). \tag{7.40}$$

where $q > d$ and $\delta = 1 - d/q$. Since we can achieve a finite covering of $M$ by balls $B_r$, we now use a chaining argument to get a uniform bound on $\sup_{B_r} \bar{\phi}_+^h$. Denoting the ball $B_\delta^1$ where the minimum is obtained, i.e. $\bar{\phi}_+^h(x) = 0$ for some $x \in B_\delta^1$, we have:

$$\sup_{B_\delta^1} \bar{\phi}_+^h \leq C \left\| f^h \right\|_{L^{q/2}(M)}. \tag{7.41}$$

Now, take a ball $B_\delta^2$ that overlaps with $B_\delta^1$. Then,

$$\sup_{B_\delta^2} \bar{\phi}_+^h \leq C \left( \inf_{B_\delta^2} \bar{\phi}_+^h + \left\| f^h \right\|_{L^{q/2}(M)} \right) \leq$$
$$C \left( \sup_{B_\delta^1} \bar{\phi}_+^h + \left\| f^h \right\|_{L^{q/2}(M)} \right) \leq C' \left\| f^h \right\|_{L^{q/2}(M)}. \tag{7.42}$$

Repeating this chaining argument a finite number of times covers the compact manifold $M$. Thus, we get:

$$\sup_{M} \bar{\phi}_+^h \leq C'' \left\| f^h \right\|_{L^{q/2}(M)}. \tag{7.43}$$

Since, $\phi_+^h(x_0) = C_0 h^\gamma$, this shows that

$$\left| \sup_{M} \phi_+^h \right| \leq C_0 h^\gamma + C'' \left\| f^h \right\|_{L^{q/2}(M)}. \tag{7.44}$$

We have:

$$\left\| f^h \right\|_{L^{q/2}(M)} \leq \left( \int_{S^h \cup b^h} \left( \frac{\kappa(h)}{h^{2\gamma}} \right)^{q/2} dx + \int_{B^h} \kappa(h)^{q/2} dx \right)^{2/q}, \tag{7.45}$$

$$\left\|f^h\right\|_{L^{q/2}(M)} \leq \kappa(h)\left(\frac{C}{h^{(q-2)\gamma}} + \mathrm{vol}(M)\right)^{2/q}. \tag{7.46}$$

Thus,

$$\left\|f^h\right\|_{L^{q/2}(M)} \leq \frac{C'\kappa(h)}{h^{(1-2/q)2\gamma}}\left(1 + \mathrm{vol}(M)h^{(q-2)\gamma}\right)^{2/q}, \tag{7.47}$$

and so

$$\left\|f^h\right\|_{L^{q/2}(M)} \leq \frac{C'\kappa(h)}{h^{(1-2/q)2\gamma}}\left(1 + \mathcal{O}\left(h^{(q-2)\gamma}\right)\right). \tag{7.48}$$

In particular, this holds if we choose $q = 4 > d = 2$. Thus, we get:

$$\left\|f^h\right\|_{L^2(M)} \leq \frac{C'\kappa(h)}{h^{\gamma}} + \mathcal{O}\left(h^{\gamma}\right). \tag{7.49}$$

Hence, we obtain

$$\left|\sup_M \phi_+^h\right| \sim \mathcal{O}\left(h^{\gamma}\right) + \mathcal{O}\left(\kappa(h)/h^{\gamma}\right). \tag{7.50}$$

□

Using the maximum bounds allows one to establish bounds on the Lipschitz constant of the second derivatives of $\phi_{\pm}^h$. Again, for clarity of exposition, we only show the bound for the case $\phi_+^h$.

**Lemma 7.14** (Estimates for Bounding Function)**.** *There exists a constant $C > 0$ such that for small enough $h > 0$, we have the estimate:*

$$\left[\phi_+^h\right]_{C^{2,1}(M)} \leq C\frac{\kappa(h)}{h^{3\gamma}}. \tag{7.51}$$

*Proof.* Just as in Lemma 7.13, we find that we can apply an *a priori* bound on coordinate patches and then this leads to an overall bound. First, we use a classical interior regularity result for uniformly elliptic PDE with bounded first-order coefficients $b$ in a region $\Omega$ from Corollary 6.3 of Gilbarg and Trudinger

(2001):

$$\ell \left\|\phi_+^h\right\|_{C^1(\Omega')} + \ell^2 \left\|\phi_+^h\right\|_{C^2(\Omega')} + \ell^3 \left[\phi_+^h\right]_{C^{2,1}(\Omega')} \leq C \left(\left\|\phi_+^h\right\|_{L^\infty(\Omega)} + \left\|f^h\right\|_{C^{0,1}(\Omega)}\right), \tag{7.52}$$

where $\ell \leq \text{dist}(\Omega, \Omega')$ and $C = C(\lambda, \text{diam}(\Omega))$, and $\lambda$ is the lower bound on the eigenvalues of the matrix $A$.

These coordinate patches (and therefore those in Lemma 7.13) are chosen to overlap by a radius $\delta > 0$ to avoid the fact that the bounds in the estimate (Equation (7.52)) scale with the distance to the boundary of each region $\Omega$. Then, we use the Hölder estimate Equation (7.52) from Gilbarg and Trudinger in each patch $\Omega_{i,\epsilon}$ and bound the overall Hölder estimate by a sum over each patch. Thus, we get:

$$\left[\phi_+^h\right]_{C^{2,1}(M)} \leq C \left(\left\|\phi_+^h\right\|_\infty + \left\|f^h\right\|_{C^{0,1}(M)}\right). \tag{7.53}$$

Thus, from Lemma 7.13 and the fact that the Lipschitz constant of $f^h$, satisfies $\sim \frac{\kappa(h)}{h^{3\gamma}}$ and $\left|f^h\right| \sim \mathcal{O}\frac{\kappa(h)}{h^{2\gamma}}$, we have that there exists a $C > 0$ such that, for small enough $h > 0$, we have that

$$\left[\phi_+^h\right]_{C^{2,1}(M)} \leq C \left(h^\gamma + \frac{\kappa(h)}{h^{3\gamma}}\right). \tag{7.54}$$

□

### 7.4.2 Proof of Convergence Theorem

Now, we prove the main result.

*Proof of Theorem 7.7.* In the region $B^h$ we compute

$$F^h[u^h - u] = L^h\left(u^h - u\right) + h^\alpha\left(u^h - u\right) \leq C'' h^\alpha \tag{7.55}$$

for some $C''$ independent of $h$.

We also substitute $\phi_+^h$ into the discrete operator in the region $B^h$ and exploit consistency to obtain:

$$F^h[\phi_+^h] = L^h\phi_+^h(x_i) + h^\alpha\phi_+^h(x_i) \geq \kappa(h)/\left|B^h\right| - C\left[\phi_+^h\right]_{C^{2,1}(M)} h^\alpha + h^\alpha\phi_+^h(x). \quad (7.56)$$

In order to exploit the discrete comparison principle, we desire to find $\kappa(h)$ such that

$$\frac{\kappa(h)}{|B^h|} - C\left[\phi_+^h\right]_{C^{2,1}(M)} h^\alpha + h^\alpha\phi_+^h(x) \geq C''h^\alpha. \quad (7.57)$$

Applying Lemmas 7.14, 7.13, we require

$$\frac{\kappa(h)}{|B^h|} - Ch^\alpha\left(h^\gamma + \frac{\kappa(h)}{h^{3\gamma}}\right) - C'h^\alpha\left(h^\gamma + \frac{\kappa(h)}{h^\gamma}\right) \geq C''h^\alpha. \quad (7.58)$$

Thus, we must choose:

$$\kappa(h) \geq \frac{h^\alpha\left(C'' + \left(C + C'\right)h^\gamma\right)}{\frac{1}{\left|B^h\right|} - \frac{Ch^\alpha}{h^{3\gamma}} - \frac{C'h^\alpha}{h^\gamma}}. \quad (7.59)$$

This will hold if we choose $\kappa(h) > Kh^\alpha$ for some $K > 0$ large enough and let $\gamma = \alpha/3$. Applying the modified scheme $F^h$ inside $S^h \cup b^h$ to $u^h - u$ and $\phi_+^h$, we get that we require that

$$F_h[\phi_+^h] = \phi_+^h(x) \geq u^h(x) - u(x) = F_h[u^h - u], \quad x \in S^h \cup b^h. \quad (7.60)$$

Recalling that $u^h(x) = 0$ in $S^b \cup b^h$ and $\phi_+^h(x_0) = C_0h^\gamma$ does yield this result provided that we choose the constant $C_0$to satisfy

$$C_0 \geq \left(L_{\phi_\pm^h} + L_u\right), \quad (7.61)$$

where $L_u$ is the Lipschitz constant of $u$ and $L_{\phi_\pm^h}$ is the equi-Lipschitz constant of

the family $\left\{\phi_{\pm}^h\right\}_h$, which is bounded by Equation (7.52):

$$\left|D\phi_+^h\right|_{0;M} \leq C\left(h^\gamma + \frac{\kappa(h)}{h^{3\gamma}}\right). \tag{7.62}$$

Thus, we find that

$$F^h[u^h - u] \leq F^h[\phi_+^h], \quad x \in \mathcal{G}^h. \tag{7.63}$$

Invoking the discrete comparison principle of $F^h$ yields $u^h - u \leq \phi_+^h$.
Doing the same procedure for $\phi_-^h$ we establish the bounds:

$$\phi_-^h(x) \leq u^h - u \leq \phi_+^h(x). \tag{7.64}$$

Thus, applying the maximum bound Equation (7.38) with our choice that $\gamma = \alpha/3$, we obtain

$$\left\|u^h - u\right\| \leq Ch^{\alpha/3}. \tag{7.65}$$

□

## 7.5 Convergent Numerical Gradient

Once the convergence rates are established for the solution $u^h$ of the discrete operator, they can be immediately used to establish a convergence approximate of the gradient of $u^h$ with rates, provided that the *a priori* solution satisfies $u \in C^2(M)$. This is done by modifying the gradient operator by simply taking the finite-difference terms over larger stencils. It must be emphasized that this is done as a post-processing step.

First, we introduce the discrete approximation of the gradient. For every point in the computational grid $x_i \in \mathcal{G}^h$, we have an associated list of neighboring points used in computations, whose indices are denoted by $N(i)$. Below, we will show how this $N(i)$ is chosen. Then, we define the following discrete approximation of the gradient:

$$\tilde{\nabla}^h \phi(x_i) \equiv \max_{j \in N(i)} \frac{\phi(x_j) - \phi(x_i)}{d(x_j, x_i)}. \tag{7.66}$$

Note, this gives an approximation of the magnitude of the gradient. Suppose that the maximum is achieved for some index $j^*$. The approximate direction of the gradient is given by the geodesic connecting $x_i$ and $x_{j^*}$. That is, suppose $\gamma(t) : [0,1] \to M$ is a unit-speed geodesic connecting $x_i$ and $x_{j^*}$, where $\gamma(0) = x_i$ and $\gamma(1) = x_{j^*}$. Then, the approximate direction of the gradient is given by $\gamma'(t)|_{t=0}$. Likewise, given a point $y$, there exists a $\hat{y} \in \mathcal{T}_x$ such that $\exp_x(\hat{y} d(x,y)) = y$ and we will denote the "direction" of $y$ (with respect to $x$) by the notation $\hat{y}$.

**Theorem 7.15.** *Suppose that $u \in C^2(M)$ and the discrete solution satisfies $\left\|u^h - u\right\| = \mathcal{O}(\omega(h))$ where $\omega(h) \to 0$. Then, the gradient operator $\tilde{\nabla}^h$ is such that computing this gradient approximation using the computational points $x_{j \in N(i)}$ such that*

$$N(i) = \left\{ j : C_- \sqrt{\omega(h)} \leq d(x_i, x_j) \leq C_+ \sqrt{\omega(h)} \right\} \tag{7.67}$$

*yields the error:*

$$\tilde{\nabla}^h u^h(x_i) = \|\nabla u(x)\| + \mathcal{O}\left(\sqrt{\omega(h)}\right) + \mathcal{O}(h). \tag{7.68}$$

*Proof.* Use $u^h$ to estimate any directional derivative (e.g. the gradient) as follows:

$$\begin{aligned} \frac{u^h(x_i) - u^h(y)}{d(x_i, y)} &= \frac{u^h(x_i) - u(x_i) - u^h(y) + u(y)}{d(x_i, y)} + \frac{u(x_i) - u(y)}{d(x_i, y)} \\ &= \nabla u(x) \cdot \hat{y} + \mathcal{O}\left(\frac{\omega(h)}{d(x_i, y)}\right) + \mathcal{O}\left(d(x_i, y)\right). \end{aligned} \tag{7.69}$$

Now, on the grid the point $y$ cannot necessarily be taken to be exactly so that $\exp_x(\nabla u(x)) = y$. However, we will search in a strip of width $\mathcal{O}(d(x_i, y))$ for approximate directions. That is, we make the following choice of computational neighborhood:

$$N(i) = \left\{ j : C_- \sqrt{\omega(h)} \leq d(x_i, x_j) \leq C_+ \sqrt{\omega(h)} \right\}. \tag{7.70}$$

Due to the non-zero curvature of $M$, there are necessarily distortions in the search directions. We assume that $d(x_i, y) < 1/\sqrt{\kappa}$, where, again, the Gaussian curvature of $M$ is bounded above by $\kappa$ and below by $-\kappa$. The density of points in the tangent plane $\mathcal{T}_{x_i}$ at a distance $d(x_i, y)$ from the point $x_i$ is distorted by a factor of, at worst, $1 + \mathcal{O}\left(d(x_i, y)^2\right)$. This follows by applying the Rauch comparison principal and the results on constant curvature Jacobi field in Berger (2003).

Since there exists a grid point for every ball of radius $\mathcal{O}(h)$, we have that there exists a sequence of $y^h \in \mathcal{G}^h$ such that

$$\nabla u(x_i) \cdot \hat{y}^h = \|\nabla u(x_i)\| + \mathcal{O}\left(h\right), \tag{7.71}$$

Therefore,

$$\tilde{\nabla}^h u^h(x_i) = \|\nabla u(x_i)\| + \mathcal{O}\left(\sqrt{\omega(h)}\right) + \mathcal{O}\left(h\right). \tag{7.72}$$

We stress here that we are not replacing the old solution, but using the values of $u^h$ to find a post-processing approximation of the gradient. □

This result then naturally extends the results from Section 7.3, by using the fact that the highest order monotone scheme possible for a second-order linear elliptic PDE satisfies $\alpha \leq 2$, see Oberman (2006). Thus, we get:

**Corollary 7.16.** *Suppose we have the PDE (Equation* (7.1)*) on a manifold $M$ with Hypotheses 7.1. Let $u^h$ be the solution of Equation* (7.24) *with the assumptions Equation* (7.4)*. Then, we have:*

$$\tilde{\nabla} u(x_i) = \|\nabla u(x)\| + \mathcal{O}\left(h^{\alpha/6}\right). \tag{7.73}$$

# CHAPTER 8

# CURRENT AND FUTURE WORK

In this chapter, we outline three current and future avenues of research. The first is concerned with establishing higher-order schemes in the hope that the convergence rates for discretizations over the sphere can be improved. Higher-order schemes are not necessarily monotone, but can be worked into monotone schemes via a filter function to build an overall convergent scheme. The second avenue of research is to extend the computations of Optimal Information Transport for compact 2D manifolds, since the moving mesh methods via Optimal Transport are fraught with regularity issues in generality. The final avenue of research is concerned with designing a monotone scheme for the $W_1$ distance based on a PDE formulation.

## 8.1 Higher-Order Schemes for Optimal Transport on the Sphere

While monotone schemes are provably convergent for the Optimal Transport problem on the sphere by the theorem in Chapter 3, the overall consistency error of the discretization is $\mathcal{O}\left(\sqrt{h}\right)$. When we have applications in mind, it would be better if we could design provably convergent schemes that have better consistency error (which would lead to better convergence properties) when the problem is smooth enough. Higher-order (consistency error) schemes can be worked into a convergence framework by using the concept of filtered schemes, which was introduced in the Euclidean case in the paper Froese and Oberman (2013). These filtered schemes, however, lack proofs of convergence in the case of the sphere or more generally on 2D surfaces without boundary. Current work has involved modifying the proof of convergence of filtered schemes to the case of the sphere and demonstrating the improved convergence properties empirically in examples inspired by real-world applications. A further desire is to be able to use explicitly derived convergence rates (in the linear case shown in Chapter 7) for the monotone dis-

cretization of the Optimal Transport problem to get convergence guarantees for the filtered scheme as well.

### 8.1.1 Filtered Schemes

We introduce the concept of filtered schemes by defining what we mean by a perturbation of the discrete operator, see Froese and Oberman (2013).

**Definition 8.1.** *The scheme $F_N^\epsilon$ is a perturbation if there is a nonnegative modulus function $m : \mathbb{R}^+ \to \mathbb{R}^+$ with*

$$\lim_{\epsilon \to 0^+} m(\epsilon) = 0 \tag{8.1}$$

*such that*

$$\sup_{u \in L^\infty(\Omega)} \sup_{x \in \Omega} |F_N^\epsilon[u](x)| \leq m(\epsilon). \tag{8.2}$$

Now, we come to the definition of a scheme which is "close" to monotone, see Froese and Oberman (2013).

**Definition 8.2.** *The scheme $F^\epsilon$ is nearly monotone if it can be written as*

$$F^\epsilon[u] + F_N^\epsilon[u], \tag{8.3}$$

*where $F_M^\epsilon$ is monotone and $F_N^\epsilon$ is a perturbation.*

Now, practically speaking one desires to know that the nearly monotone scheme is simple to construct by somehow using a higher-order scheme in nonsingular regions of the domain. The hybrid scheme will be constructed using a filter function, which we define here, see Froese and Oberman (2013).

**Definition 8.3.** *We define a filter function to be a continuous, bounded function $S$, which is equal to the identity in a neighborhood of the origin and zero outside.*

A particular filtered scheme is then defined via defining the perturbation and fixing $\alpha$, see Froese and Oberman (2013):

$$F_N^\epsilon[u] = \epsilon^\alpha S\left(\frac{F_A^\epsilon[u] - F_M^\epsilon[u]}{\epsilon^\alpha}\right), \tag{8.4}$$

where $F_A^\epsilon$ is the more accurate scheme. Clearly, the choice of $\alpha$ is rather important. It will determine how sensitive the filtered scheme is to detecting singularities. Singularities here are "measured" by how different the values of the monotone and accurate scheme are. If they are approximately the same, then the filtered scheme chooses to use the accurate scheme. But, the advantage lies in the fact that the higher-order scheme may have better convergence properties in smooth situations. We can then develop higher-order schemes by direct Taylor expansion, similarly to what was done in the derivation of monotone schemes in Froese (2018). Then, we discretize the Optimal Transport problem on the sphere with the local tangent plane construction introduced in Chapter 4.

## 8.2 Moving Mesh Methods for 2D Compact Manifolds

The Optimal Transport problem is plagued with regularity issues when one tries to generalize beyond the case of the sphere. Suppose, for example, that we have a mesh generated on an oblate sphere (ellipsoid of revolution with eccentricity $e = 0.25$), see Figure 8.1

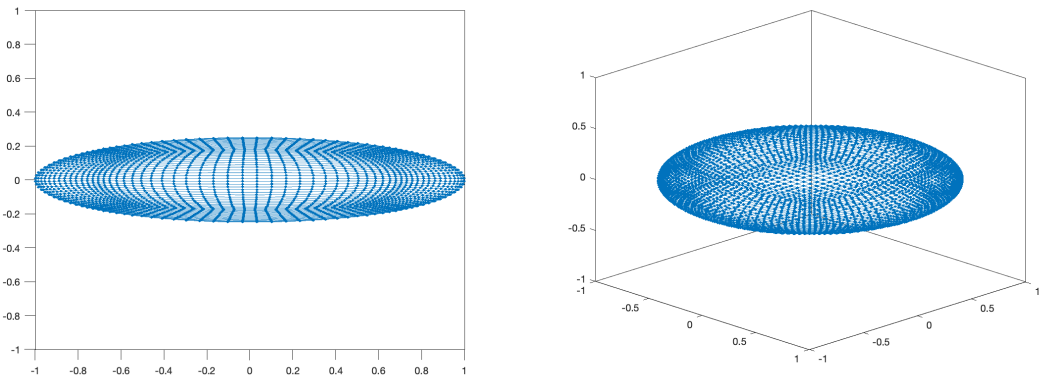



**Figure 8.1** A $N = 5048$-point cube mesh for the ellipsoid from the side (left) and from diagonally above (right).

For the Optimal Transport problem, the result from Figalli et al. (2010) as-

serts that there exist $C^\infty$ source and target mass density functions $f_0$ and $f_1$ such that the mapping $T$ is not guaranteed to be even continuous. Further difficulty in building a discretization comes again from the geometry regarding the computation of geodesics. However, given a local coordinate chart $(x^k)$ in a neighborhood of a point $x$, we can solve the second order ordinary differential equation for the geodesics:

$$\ddot{x}^k(t) + \dot{x}^i(t)\dot{x}^j(t)\Gamma_{ij}^k\left(x(t)\right) = 0, \tag{8.5}$$

where $\Gamma_{ij}^k$ are the Christoffel symbols. The Optimal Transport scheme we have presented, however, exploits an explicit representation formula for the mixed Hessian term, as derived in Hamfeldt and Turnquist (2021b). This required an explicit formula for the exponential map as a surjective map from local tangent planes onto the sphere. This is no longer possible in more general geometries. Thus, a direct discretization of $D_{xy}^2 c(x,y)$ would be necessary. We, similarly, do not have a representation formula for geodesic normal coordinates. This means that they must be approximated. This is possible in practice, however, with any method of finding geodesic normal coordinates that is consistent.

Our Optimal Information Transport algorithm only needs local geodesics. That means that this method is more easily generalizable than that of Optimal Transport. Furthermore, any provably convergent scheme for computing the solution of the Poisson equation on compact manifolds $M$ would lead to the same conclusions as we have presented here. It remains to build a robust numerical method for this computation, however.

## 8.3 A Numerical Scheme for Wasserstein-1 Distance

The Wasserstein-1 distance

$$W_1(\mu,\nu) := \inf \int_\Omega d(x,y) d\pi(x,y) \tag{8.6}$$

is known to have minimizers (maps $T(x)$), but these minimizers are not necessarily unique, see Santambrogio (2015); Villani (2003). Nevertheless, in the computation of the Wasserstein distance, the important quantity of interest is often the distance and not the specific mapping. Supposing that we could "pick out" a particular minimizing mapping $T(x)$, then we could compute the Wasserstein-1 distance and not worry about the other mappings that achieve the same distance. Many current techniques, such as those used in WGAN Arjovsky et al. (2017) rely on performing a minimization over all Lipschitz-1 functions, which is an infinite-dimensional constraint. Usually, this constraint is incorporated into a gradient penalty term. Here we approach the problem directly by using PDE and ODE techniques to capture a particular solution.

In the article Evans and Gangbo (1999), the authors showed that one particular mapping $T$ of the Wasserstein-1 metric, given by:

$$\frac{T(x) - x}{|T(x) - x|} = -\nabla u(x), \tag{8.7}$$

where $|\nabla u(x)| = 1$, could be found explicitly by taking the limit of the following sequence of PDE:

$$\begin{cases} -\nabla \cdot (|\nabla u_p|^{p-2} \nabla u_p) = f, & \text{in } \Omega, \\ u_p = 0, & \text{on } \partial\Omega \end{cases} \tag{8.8}$$

as $p \to \infty$, see Evans and Gangbo (1999). This limit is known as the "infinity" Laplacian. Taking $p \to \infty$, then the above equation can be denoted as:

$$\begin{cases} -\Delta_\infty u = f, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega. \end{cases} \tag{8.9}$$

According to Oberman (2004), such an operator can be interpreted as:

$$\Delta_\infty u = \frac{d^2 u}{d\nu^2}, \tag{8.10}$$

where $\nu = \frac{\nabla u}{|\nabla u|}$. Thus, this is the second directional derivative in the direction of the gradient of $u$. Another way of viewing this operator is to denote the transport density $\sigma$ (a measure) Santambrogio (2015) and write the infinity-Laplacian as

$$-\Delta_\infty u = -\nabla \cdot (\sigma \nabla u), \tag{8.11}$$

where the derivatives must be interpreted in the sense of distributions.

In the manuscript Oberman (2004), the author showed that one could construct a provably convergent scheme for the infinity Laplacian. Then, using Dacorogna-Moser transport Dacorogna and Moser (1990):

$$\begin{cases} \dot{z}(t) = \frac{-a(z(t))\nabla u(z(t))}{t\nu + (1-t)\mu}, \quad (0 \le t \le 1), \\ z(0) = z_0 \end{cases} \tag{8.12}$$

where $a$ satisfies $\sigma = a dx$, we can construct the mapping $T$. Amazingly, the time-1 mapping of the above first-order ODE is the optimal mapping: $z(1) = T$. Using monotone finite-difference schemes, we should be able to construct provably convergent schemes for $W_1$ which do not rely on penalty terms. Furthermore, given that the tangent-plane monotone discretization developed in Chapter 4 was successful, we can explore how to extend such PDE computations to, for example, the $W_1$ distance on the sphere.

# CHAPTER 9

# CONCLUSION

In this dissertation, we have developed numerical methods for solving the PDE formulation of Optimal Transport with applications to the reflector antenna problem and the moving mesh problem. Importantly, we have constructed a convergence framework for such schemes and shown how to adapt the scheme to many non-smooth cases of practical interest. For the reflector antenna problem, it was important to (1) design the scheme to be more computationally efficient than existing provably convergent schemes and (2) be able to deal with non-smooth cases in the source and/or target densities. For the moving mesh method problem, a study of our proposed numerical method for Optimal Transport on the sphere shows its applicability to this problem. Simple computations show that the scheme we proposed avoids tangling of the mesh. However, strong evidence shows that Optimal Information Transport also provides a better computational method for non-smooth cases and also, more importantly, across a wide variety of 2D surfaces.

To solve the Optimal Transport PDE numerically, we first first reformulated the PDE into an equivalent tangent plane formulation in order take advantage of the existing construction of monotone discretizations in the Euclidean plane. Specifically, we used geodesic normal coordinates in our tangent planes in order to reduce the effect of the geometry on our differential operators. We then augmented the numerical method with a Lipschitz control in order to use a compactness argument to achieve the convergence result. This Lipschitz control was designed in lieu of using a comparison-principle-type argument, as the underlying PDE lacks a comparison principle. Monotonicity in the scheme then was explicitly achieved via using existing monotone finite-different schemes for second-order directional derivatives on point clouds in $\mathbb{R}^2$, but, critically, adding discrete Laplacian regularizing terms to deal with the discrete gradient terms. The resulting monotone

discretization was then solved using a parabolic solver. The convergence framework developed here can be very easily adapted to other cost functions and other elliptic PDE on the sphere.

For the reflector antenna problem, the important considerations were allowing for the numerical scheme to solve singular examples. This required the development of the convergence theorem for non-smooth examples by using preprocessing steps for the target mass. The proposed scheme also had to be more computationally efficient than previously proposed provably convergent schemes. Our discretization was achieved via a discretization of computational complexity $\mathcal{O}(N^{9/4})$.

For the moving mesh problem, the important point was to avoid tangling the mesh. For the Optimal Transport problem, this was achievable using the Monge-Ampère PDE numerical scheme on the sphere without any add-ons for smooth enough examples. However, non-smooth examples showed Optimal Information Transport to be a superior method. The lack of a smooth regularity guarantee over a wide variety of compact surfaces further indicates that Optimal Information Transport should be considered in finding diffeomorphic mappings between density functions on general 2D compact manifolds.

Convergence rates for monotone discretizations of linear divergence-form linear elliptic PDE on compact surfaces have been developed, with the object to apply these convergence rates to nonlinear elliptic PDE like the Optimal Transport PDE. What is surprising is that these convergence rates are asymptotically worse than the formal consistency error, a result which is markedly different than that for the Dirichlet problem, for example.

Finally, we are investigating how to incorporate higher-order schemes (with higher formal consistency error) via filtered schemes. It is believed that these will address some issues in the applications, such as numerical artifacts seen in the reflector antenna problem after ray tracing validation and tangling of the mesh for the Optimal Transport implementation of moving mesh methods. Secondly, we

are investigating how to extend the algorithm for Optimal Information Transport to other 2D surfaces without boundary. And, finally, we are investigating how a PDE-based method can be developed for computing the $W_1$ distance based on a monotone discretization of the infinity-Laplacian. Given the development of this PDE method, we can then begin to explore how to make such $W_1$ computations on geometries like the sphere.

# APPENDIX A
# REGULARITY OF THE POTENTIAL FUNCTION

In this appendix, we derive simple conditions to guarantee the solution $u \in C^1(\mathbb{S}^2)$. The results from Loeper (2011) indicate that we have two régimes of regularity: classical and nonsmooth, both encapsulated in Theorem 2.4 of that paper. The classical result, adapted to our notation, is as follows:

**Theorem A.1** (Regularity (smooth))**.** *Given data satisfying the smooth hypothesis, suppose additionally that $f_0$ and $f_1$ are in $C^{1,1}(\mathbb{S}^2)$ (resp. $C^\infty(\mathbb{S}^2)$). Then $u \in C^{3,\alpha}(\mathbb{S}^2)$ for every $\alpha \in [0,1)$ (resp. $u \in C^\infty(\mathbb{S}^2)$).*

The corresponding non-smooth result is:

**Theorem A.2** (Regularity (non-smooth))**.** *Given data satisfying the nonsmooth hypothesis, suppose additionally that there exists some $h : \mathbb{R}^+ \to \mathbb{R}^+$ with $\lim_{\epsilon\to 0} h(\epsilon) = 0$ such that*

$$\int_{B_\epsilon(x)} f_0(y)\, dy \leq h(\epsilon)\epsilon, \quad \text{for all } \epsilon \geq 0, x \in \mathbb{S}^2, \tag{A.1}$$

*then $u \in C^1(\mathbb{S}^2)$.*

As pointed out in Loeper (2011), this condition is automatically satisfied for densities $f_0 \in L^p(\mathbb{S}^2)$ with $p > 2$. In fact, a slightly stronger regularity result is available in this case, and we have $u \in C^{1,\beta}(\mathbb{S}^2)$ with $\beta = \frac{p-2}{7p-2}$. The following Lemma will show that Theorem A.2 also applies to densities $f_0 \in L^1(\mathbb{S}^2)$.

**Lemma A.3** (Integrability condition for $L^1$ densities)**.** *If $\mu_0 \in L^1(\mathbb{S}^2)$, then there exists some $h : \mathbb{R}^+ \to \mathbb{R}^+$ with $\lim_{\epsilon\to 0} h(\epsilon) = 0$ such that*

$$\int_{B_\epsilon(x)} f_0(y)\, dy \leq h(\epsilon)\epsilon, \quad \text{for all } \epsilon \geq 0, x \in \mathbb{S}^2.$$

*Proof.* We use local spherical coordinates $\theta, \phi$ about the point $x$ to compute

$$\int_{B_\epsilon(x)} f_0(y)\, dy = \int_0^{2\pi}\int_0^{\epsilon} f_0(\theta,\phi)\sin\phi\, d\phi\, d\theta \leq \epsilon \int_0^{2\pi}\int_0^{\epsilon} f_0(\theta,\phi)\, d\phi\, d\theta,$$

which holds for sufficiently small $\epsilon$ since then $\sin\phi < \phi \leq \epsilon$. By the Fubini-Tonelli Theorem, we can switch the order of integration and obtain

$$\int_{B_\epsilon(x)} f_0(y)\,dy \leq \epsilon \int_0^\epsilon F(\phi)\,d\phi,$$

where we have defined the partial integral

$$F(\phi) = \int_0^{2\pi} f_0(\theta,\phi)\,d\theta.$$

Since $f_0$ is a non-negative $L^1$ function, the partial integral $F$ is also in $L^1$ and non-negative. We can then define

$$h(\epsilon) = \int_0^\epsilon F(\phi)\,d\phi,$$

which satisfies $\lim_{\epsilon\to 0} h(\epsilon) = 0$ since $F \in L^1$. Thus we obtain the desired result. $\square$

# APPENDIX B

## MAPPING FOR THE LOGARITHMIC COST

In this appendix, we calculate an explicit mapping $T(x,p) \in \mathbb{S}^2$ corresponding to the logarithmic cost $c(x,y) = -\log\|x-y\|$. To accomplish this, we let $x \in \mathbb{S}^2$, $p \in \mathcal{T}(x)$ and solve $\nabla u(x) = \nabla_x c(x,y)|_{y=T(x)}$:

$$\begin{cases} \nabla_{\mathbb{S}^2,x} \log\|x-y\| = p, \\ \|y\| = 1 \end{cases}$$

for $y$.

Let $\hat{\theta}$ and $\hat{\phi}$ be the local orthonormal tangent vectors at the point $x \in \mathbb{S}^2$. Then we can compute this surface gradient in the ambient space in the local tangent coordinates using a simplified formula, which reduces the computational complexity:

$$\nabla_{\mathbb{S}^2} f(x) = \left(\nabla f(x)\cdot\hat{\theta}, \nabla f(x)\cdot\hat{\phi}\right), \tag{B.1}$$

where we emphasize here that the gradient $\nabla$ refers to the usual gradient in $\mathbb{R}^3$. Using this formula, we obtain

$$p = \left(\frac{(x-y)\cdot\hat{\theta}}{\|x-y\|^2}, \frac{(x-y)\cdot\hat{\phi}}{\|x-y\|^2}\right).$$

Note that $x, \hat{\theta}$, and $\hat{\phi}$ form an orthonormal set. Thus we can express the unknown $y$ in the form $y = y_x x + y_\theta\hat{\theta} + y_\phi\hat{\phi}$ and obtain

$$p = \left(\frac{-y_\theta}{2-2y_x}, \frac{-y_\phi}{2-2y_x}\right).$$

Combining this with the requirement that $y_x^2 + y_\theta^2 + y_\phi^2 = \|y\|^2 = 1$ allows us to

solve for the components of $y$:

$$\begin{aligned} y &= (y_x, y_\theta, y_\phi), \\ &= \frac{1}{4\left\|p\right\|^2 + 1}\left(4\left\|p\right\|^2 - 1, -4p_\theta, -4p_\phi\right), \\ &= x\frac{\|p^2\| - 1/4}{\left\|p\right\|^2 + 1/4} - \frac{p}{\left\|p\right\|^2 + 1/4}. \end{aligned}$$

# APPENDIX C

## NORMAL COORDINATES FOR THE SPHERE

In this appendix, we compute an explicit formula for geodesic normal coordinates on the sphere. Consider a point $x_0 \in \mathbb{S}^2$ and the corresponding tangent plane $\mathcal{T}_{x_0}$. Geodesic normal coordinates for points $x \in \mathbb{S}^2$ will take the form

$$v_{x_0}(x) = x_0 + k\text{Proj}_{\mathcal{T}_{x_0}}(x - x_0) \in \mathcal{T}_{x_0},$$

where $k$ is chosen so that $\|x_0 - v_{x_0}(x)\| = d_{\mathbb{S}^2}(x_0, x)$. Recall that the geodesic distance between $x$ and $x_0$ can be expressed as

$$d_{\mathbb{S}^2}(x_0, x) = 2 \arcsin\left(\frac{\|x - x_0\|}{2}\right).$$

Since $x$ and $x_0$ are unit vectors, we can let $\cos\alpha = x \cdot x_0$ and compute

$$\cos d_{\mathbb{S}^2}(x_0, x) = \cos\left(2 \arcsin\left(\frac{\sqrt{2 - 2\cos\alpha}}{2}\right)\right) = \cos\alpha = x \cdot x_0.$$

We will make use of the unit tangent vectors $\hat{\theta}$ and $\hat{\phi}$ at the point $x_0$, which define orthonormal coordinates. The projection of the displacement $x - x_0$ onto the tangent plane can be represented in these coordinates as

$$\text{Proj}_{\mathcal{T}_{x_0}}(x - x_0) = \left[(x - x_0) \cdot \hat{\theta}\right] \hat{\theta} + \left[(x - x_0) \cdot \hat{\phi}\right] \hat{\phi}.$$

By computing a unit vector in this direction and scaling by the geodesic distance $d_{\mathbb{S}^2}(x_0, x)$, we obtain the following expression for the geodesic normal coordinates:

$$v_{x_0}(x) = x_0 + d_{\mathbb{S}^2}(x_0, x) \frac{\left[(x - x_0) \cdot \hat{\theta}\right] \hat{\theta} + \left[(x - x_0) \cdot \hat{\phi}\right] \hat{\phi}}{\sqrt{\left[(x - x_0) \cdot \hat{\theta}\right]^2 + \left[(x - x_0) \cdot \hat{\phi}\right]^2}}.$$

Since $x_0$ is a unit vector orthogonal to both $\hat{\theta}$ and $\hat{\phi}$, the actual displacement between points on the sphere can be expressed as

$$x - x_0 = \left[(x - x_0) \cdot \hat{\theta}\right] \hat{\theta} + \left[(x - x_0) \cdot \hat{\phi}\right] \hat{\phi} + \left[(x - x_0) \cdot x_0\right] x_0,$$

which has squared Euclidean length

$$\|x - x_0\|^2 = \left[(x - x_0) \cdot \hat{\theta}\right]^2 + \left[(x - x_0) \cdot \hat{\phi}\right]^2 + \left[(x - x_0) \cdot x_0\right]^2.$$

These relationships allow us to simplify the expression for geodesic normal coordinates to

$$\begin{aligned}
v_{x_0}(x) &= x_0 + d_{\mathbb{S}^2}(x_0, x) \frac{x - x_0 - \left[(x - x_0) \cdot x_0\right] x_0}{\|x - x_0\|^2 - \left[(x - x_0) \cdot x_0\right]^2}, \\
&= x_0 + d_{\mathbb{S}^2}(x_0, x) \frac{x - x_0 (x \cdot x_0)}{\sqrt{1 - (x \cdot x_0)^2}}, \\
&= x_0 + d_{\mathbb{S}^2}(x_0, x) \frac{x - x_0 \cos d_{\mathbb{S}^2}(x_0, x)}{\sqrt{1 - \cos^2 d_{\mathbb{S}^2}(x_0, x)}}, \\
&= x_0 \left(1 - d_{\mathbb{S}^2}(x_0, x) \cot d_{\mathbb{S}^2}(x_0, x)\right) + x \left(d_{\mathbb{S}^2}(x_0, x) \csc d_{\mathbb{S}^2}(x_0, x)\right).
\end{aligned}$$

# APPENDIX D

# DERIVATION OF THE MIXED HESSIAN

In this appendix, we fill in the details of the derivation of simple expressions for the determinant of the mixed Hessian. For each cost function, we take the following approach:

1. Introduce orthogonal perturbations $\Delta p_1, \Delta p_2 \in \mathcal{T}_x$ such that $\Delta p_1 \cdot p = 0$ and $\Delta p_2 = \hat{p}\|\Delta p_2\|$.

2. Establish that $T(x,p) - T(x,p+\Delta p_1)$ and $T(x,p) - T(x,p+\Delta p_2)$ are orthogonal to leading order.

3. Compute the change of area formula

$$\begin{aligned}
&|\det\ (D_p T(x,p))| \\
&\quad = \lim_{\|\Delta p_1\|,\|\Delta p_2\|\to 0} \frac{d_{\mathbb{S}^2}\left(T(x,p), T(x,p+\Delta p_1)\right) d_{\mathbb{S}^2}\left(T(x,p), T(x,p+\Delta p_2)\right)}{\|\Delta p_1\|\,\|\Delta p_2\|}, \\
&\quad = \lim_{\|\Delta p_1\|,\|\Delta p_2\|\to 0} \frac{\|T(x,p) - T(x,p+\Delta p_1)\|\|T(x,p) - T(x,p+\Delta p_1)\|}{\|\Delta p_1\|\,\|\Delta p_2\|}
\end{aligned}$$

where we can simplify the formulas by using the fact that

$$\|T(x,p) - T(x,p+\Delta p)\| = d_{\mathbb{S}^2}\left(T(x,p), T(x,p+\Delta p)\right) + \mathcal{O}\left(\|\Delta p\|^2\right).$$

## D.1 SQUARED GEODESIC COST

We begin with the squared geodesic cost, recalling that the mapping $T$ has the explicit form

$$T(x,p) = \cos\left(\|p\|\right) x + \sin\left(\|p\|\right) \frac{p}{\|p\|}.$$

First consider a perturbation satisfying $\Delta p_1 \cdot p = 0$.

$$\begin{aligned}
T(x,p) - T(x,p+\Delta p_1) = x\left(\cos\|p\| - \cos\|p+\Delta p_1\|\right) \\
+ \frac{p}{\|p\|}\sin\|p\| - \frac{p+\Delta p_1}{\|p+\Delta p_1\|}\sin\|p+\Delta p_1\|.
\end{aligned}$$

Now, since $p$ and $\Delta p_1$ are orthogonal,

$$\|p + \Delta p_1\| = \sqrt{\|p\|^2 + \|\Delta p_1\|^2} = \|p\| + \mathcal{O}\left(\|\Delta p_1\|^2\right).$$

Thus to leading order we obtain

$$T(x,p) - T(x,p+\Delta p_1) = \Delta p_1 \frac{\sin\|p\|}{\|p\|} + \mathcal{O}\left(\|\Delta p_1\|^2\right).$$

Now, suppose that $\Delta p_2 = \|\Delta p_2\| \frac{p}{\|p\|}$. In this case, $\|p + \Delta p_2\| = \|p\| + \|\Delta p_2\|$. As before, we compute to leading order:

$$\begin{aligned}
&T(x,p) - T(x,p+\Delta p_2) \\
&\quad = x\left(\cos\|p\| - \cos\|p+\Delta p_2\|\right) + \frac{p}{\|p\|}\sin\|p\| - \frac{p+\Delta p_2}{\|p+\Delta p_2\|}\sin\|p+\Delta p_2\|, \\
&\quad = x\|\Delta p_2\|\sin\|p\| - \frac{p}{\|p\|}\|\Delta p_2\|\cos\|p\| + \frac{p}{\|p\|^2}\|\Delta p_2\|\sin\|p\| - \frac{\Delta p_2}{\|p\|}\sin\|p\|, \\
&\qquad + \mathcal{O}\left(\|\Delta p_2\|^2\right), \\
&\quad = x\|\Delta p_2\|\sin\|p\| - \frac{p}{\|p\|}\|\Delta p_2\|\cos\|p\| + \mathcal{O}\left(\|\Delta p_2\|^2\right).
\end{aligned}$$

Now since $x \cdot \Delta p_1 = p \cdot \Delta p_1 = \Delta p_1 \cdot \Delta p_2 = 0$, we can easily verify that

$$\left(T(x,p) - T(x,p+\Delta p_1)\right) \cdot \left(T(x,p) - T(x,p+\Delta p_2)\right) = o\left(\|\Delta p_1\| + \|\Delta p_2\|\right)$$

so that the perturbations in the map are indeed orthogonal to leading order.

Now we can use orthogonality to easily compute the magnitudes of these perturbations via

$$\|T(x,p) - T(x,p+\Delta p_1)\|^2 = \|\Delta p_1\|^2 \frac{\sin^2\|p\|}{\|p\|^2} + o\left(\|\Delta p_1\|^2\right)$$

and

$$\begin{aligned}\|T(x,p) - T(x,p+\Delta p_2)\|^2 &= \|\Delta p_2\|^2 \sin^2 \|p\| + \|\Delta p_2\|^2 \cos^2 \|p\| + o\left(\|\Delta p_2\|^2\right) \\ &= \|\Delta p_2\|^2 + o\left(\|\Delta p_2\|^2\right).\end{aligned}$$

where we have used the fact that $x$ and $p/\|p\|$ are unit vectors. Then we can compute the change of area formula as

$$\begin{aligned}&|\det\ (D_p T(x,p))| \\ &\quad= \lim_{\|\Delta p_1\|,\|\Delta p_2\|\to 0} \frac{\|T(x,p) - T(x,p+\Delta p_1)\|\|T(x,p) - T(x,p+\Delta p_2)\|}{\|\Delta p_1\|\,\|\Delta p_2\|}, \\ &\quad= \lim_{\|\Delta p_1\|,\|\Delta p_2\|\to 0} \frac{\|\Delta p_1\|\|\Delta p_2\| \sin\|p\|/\|p\| + o\left(\|\Delta p_1\|\|\Delta p_2\|\right)}{\|\Delta p_1\|\,\|\Delta p_2\|}, \\ &\quad= \frac{\sin\|p\|}{\|p\|}.\end{aligned}$$

Hence, the determinant of the mixed Hessian for the squared geodesic cost is

$$\left|\det D^2_{xy} c(x,y)\right| = \frac{\|p\|}{\sin\|p\|}. \tag{D.1}$$

## D.2 LOGARITHMIC COST

Now we perform the same procedure for the logarithmic map, which has the explicit form

$$T(x,p) = x\frac{\|p\|^2 - 1/4}{\|p\|^2 + 1/4} - \frac{p}{\|p\|^2 + 1/4}.$$

We again begin with a perturbation satisfying $p \cdot \Delta p_1 = 0$ so that

$$\|p + \Delta p_1\| = \|p\| + \mathcal{O}\left(\|\Delta p_1\|^2\right).$$

To leading order, we can compute

$$\begin{aligned} T(x,p) - T(x,p+\Delta p_1) &= x\frac{\|p\|^2 - 1/4}{\|p\|^2 + 1/4} - x\frac{\|p+\Delta p_1\|^2 - 1/4}{\|p+\Delta p_1\|^2 + 1/4} \\ &\quad - \frac{p}{\|p\|^2 + 1/4} + \frac{p+\Delta p_1}{\|p+\Delta p_1\|^2 + 1/4}, \\ &= \frac{\Delta p_1}{\|p^2\| + 1/4} + \mathcal{O}\left(\|\Delta p_1\|^2\right). \end{aligned}$$

Next we consider an orthogonal perturbation $\Delta p_2 = \|\Delta p_2\| \frac{p}{\|p\|}$ so that

$$\|p + \Delta p_2\|^2 = \|p\|^2 + 2\|p\|\|\Delta p_2\| + \mathcal{O}\left(\|\Delta p_2\|^2\right).$$

Now we can compute to leading order

$$\begin{aligned} \|T(x,p) - T(x,p+\Delta p_2)\| &= x\frac{\|p\|^2 - 1/4}{\|p\|^2 + 1/4} - x\frac{\|p\|^2 + 2\|p\|\|\Delta p_2\| - 1/4}{\|p\| + 2\|p\|\|\Delta p_2\|^2 + 1/4} \\ &\quad - \frac{p}{\|p\|^2 + 1/4} + \frac{p+\Delta p_2}{\|p\|^2 + 2\|p\|\|\Delta p_2\| + 1/4} + \mathcal{O}\left(\|\Delta p_2\|^2\right), \\ &= -x\frac{2\|p\|\|\Delta p_2\|}{\|p\|^2 + 1/4} + x\frac{2\|p\|\|\Delta p_2\|(\|p\|^2 - 1/4)}{(\|p\|^2 + 1/4)^2} \\ &\quad + \frac{\Delta p_2}{\|p\|^2 + 1/4} - p\frac{2\|p\|\|\Delta p_2\|}{(\|p\|^2 + 1/4)^2} + \mathcal{O}\left(\|\Delta p_2\|^2\right), \\ &= -x\frac{\|p\|\|\Delta p_2\|}{(\|p\|^2 + 1/4)^2} + \hat{p}\frac{\|\Delta p_2\|(1/4 - \|p\|^2)}{(\|p\|^2 + 1/4)^2} + \mathcal{O}\left(\|\Delta p_2\|^2\right). \end{aligned}$$

Since $x$, $p$, and $\Delta p_1$ are mutually orthogonal, we can immediately verify that

$$(T(x,p) - T(x,p+\Delta p_1)) \cdot (T(x,p) - T(x,p+\Delta p_2)) = o\left(\|\Delta p_1\| + \|\Delta p_2\|\right)$$

so that the perturbations in the mapping are again orthogonal to leading order. Next we compute the lengths of the perturbations using orthogonality:

$$\|T(x,p) - T(x,p+\Delta p_1)\| = \|\Delta p_1\| \frac{1}{\|p\|^2 + 1/4} + o(\|\Delta p_1\|)$$

and

$$\begin{aligned}\|T(x,p)-T(x,p+\Delta p_2)\|^2 &= \frac{\|p\|^2\|\Delta p_2\|^2+\|\Delta p_2\|^2(1/4-\|p\|^2)^2}{(\|p\|^2+1/4)^4}+o(\|\Delta p_2\|^2),\\ &= \frac{\|\Delta p_2\|^2}{(\|p\|^2+1/4)^4}\left(\|p\|^4+\frac{1}{2}\|p\|^2+\frac{1}{16}\right)+o(\|\Delta p_2\|^2),\\ &= \frac{\|\Delta p_2\|^2}{(\|p\|^2+1/4)^2}+o(\|\Delta p_2\|^2).\end{aligned}$$

Then we can again compute the change of area formula as

$$\begin{aligned}&|\det\ (D_pT(x,p))|\\ &\quad= \lim_{\|\Delta p_1\|,\|\Delta p_2\|\to 0}\frac{\|T(x,p)-T(x,p+\Delta p_1)\|\|T(x,p)-T(x,p+\Delta p_2)\|}{\|\Delta p_1\|\,\|\Delta p_2\|},\\ &\quad= \lim_{\|\Delta p_1\|,\|\Delta p_2\|\to 0}\frac{\|\Delta p_1\|\|\Delta p_2\|/(\|p\|^2+1/4)^2+o\left(\|\Delta p_1\|\|\Delta p_2\|\right)}{\|\Delta p_1\|\,\|\Delta p_2\|},\\ &\quad= \frac{1}{(\|p\|^2+1/4)^2}.\end{aligned}$$

Hence, the determinant of the mixed Hessian for the logarithmic cost is

$$\left|\det D^2_{xy}c(x,y)\right| = \left(\|p\|^2+1/4\right)^2. \tag{D.2}$$

# APPENDIX E
# MODIFIED POISSON EQUATION

In this appendix, we will detail the convergence theorem for the proposed discretization of the Poisson equation. Due to the fact that we are solving on a compact 2D manifold $M$, the convergence framework we present here hinges first on the reformulation of each PDE to include some control on its Lipschitz constant. Given the Poisson equation, which we denote by the operator $F$:

$$F(x, Du(x), D^2u(x)) = 0 \tag{E.1}$$

we instead discretize the following modified PDE:

$$\max\{F(x, \nabla u(x), D^2u(x), \|\nabla u\| - R\} = 0, \tag{E.2}$$

where $R$ is chosen to be strictly greater than the *a priori* Lipschitz bound. After discretization, the Lipschitz bound imparts sufficient stability on the solution to prove overall uniform convergence of the discrete solution $u^h$ to the actual solution $u$, provided that we have constructed a consistent, monotone scheme. Here we present the argument for the Poisson equation.

The function $f(x)$ satisfies the compatibility condition: $\int_M f(x)dx = 0$. We now state the equivalence theorem that shows that adding a term to the modified Poisson equation (Equation (E.2)) does not change the solution.

**Theorem E.1.** *For any sequence $\epsilon_n \to 0$, $\epsilon_n > 0$ and for $R > 0$ large enough, the viscosity solution $w_{\epsilon_n}$ of the PDE*

$$\max\{-\Delta w_{\epsilon_n}(x) + f(x), \|\nabla w_{\epsilon_n}(x)\| - R\} + \epsilon_n w_{\epsilon_n}(x) = 0, \tag{E.3}$$

*and for continuous $f(x)$ satisfying $\int_M f(x)dx = 0$, converges uniformly to the unique mean-zero Lipschitz solution $u$ of the PDE $-\Delta u(x) + f(x) = 0$.*

*Proof.* Equation (E.3) is proper, and therefore even though it is degenerate elliptic,

there exists a comparison principle, provided that $f(x)$ is continuous, by Crandall et al. (1992). Thus, by Perron's method for continuous viscosity solutions, there exists a unique solution $w_\epsilon$ to this equation for any $\epsilon > 0$.

By the same reasoning, the viscosity solution $u_\epsilon$ of the PDE $-\Delta u_\epsilon + f(x) + \epsilon u_\epsilon = 0$ is unique as well and is Lipschitz for all $\epsilon > 0$ provided that it remains uniformly bounded in $\epsilon$. It is mean zero, that is $\int_M u_\epsilon = 0$. The distributional solution $\tilde{u}_\epsilon$ of $-\Delta \tilde{u}_\epsilon + f(x) + \epsilon \tilde{u}_\epsilon = 0$ is the same as the viscosity solution $u_\epsilon = \tilde{u}_\epsilon$, see Ishii (1995). Thus, in the sense of distributions, for any $v \in C^2$ and use the integration by parts formula on the closed compact manifold $M$:

$$\int_M v \Delta u_\epsilon dx - \int_M u_\epsilon \Delta v dx = 0. \tag{E.4}$$

Choosing $v \equiv 1$, we get:

$$\int_M \Delta u_\epsilon dx = 0. \tag{E.5}$$

Hence, we integrate the equation where $u_\epsilon$ is known *a priori* to be continuous:

$$\int_M \Delta u_\epsilon + f(x) + \epsilon u_\epsilon dx = 0. \tag{E.6}$$

Thus, since $f$ is already chosen to integrate to zero, we get $\int_M u_\epsilon dx = 0$.

Fix $\epsilon > 0$. We will use interior regularity estimates for subsets of Euclidean space to show that the Lipschitz constant of $u_\epsilon$ is uniformly bounded. The results in Euclidean space can be used for the compact surface $M$ by using a $C^\infty$-atlas to cover the manifold and applying the Euclidean interior estimates for uniformly elliptic PDE in divergence form in coordinate patches $\Omega_i$ that overlap by a fixed distance $\delta > 0$, as was done in Chapter 7. Interior Schauder estimates of uniformly elliptic linear PDE in non-divergence form with bounded coefficients in subsets of Euclidean space can be found in Corollary 6.3 of Gilbarg and Trudinger (2001):

$$d \left\| Du_\epsilon \right\|_{C^0(\Omega_i')} + d^2 \left\| D^2 u_\epsilon \right\|_{C^0(\Omega_i')} \leq C \left( \left\| u_\epsilon \right\|_{C^0(\Omega_i)} + \left\| f \right\|_{C^0(\Omega_i)} \right), \tag{E.7}$$

where $d \leq \text{dist}(\Omega', \partial\Omega)$. In order to avoid the effect of the terms $d$ on the estimate, the result over the compact manifold $M$ will be effected by choosing overlapping coordinate patches $\Omega_i$ that have a uniform constant of overlap $r_M > \delta > 0$. That is, $d(x, y) > \delta$ for any $x \in M \setminus \Omega_i, y \in M \setminus \Omega_j$ for $i \neq j$. Thus, we will be able to choose $d \geq \delta$. Doing this, we derive an estimate over the whole compact manifold $M$:

$$\delta \left\| Du_\epsilon \right\|_{C^0(M)} + \delta^2 \left\| D^2 u_\epsilon \right\|_{C^0(M)} \leq C \left( \left\| u_\epsilon \right\|_{C^0(M)} + \left\| f \right\|_{C^0(M)} \right). \tag{E.8}$$

The term $\left\| u_\epsilon \right\|_{C^0(\Omega_i)}$ on the right-hand side of (E.7) can be bounded by using the Krylov-Safonov Harnack inequality in Euclidean space, see Cabré (2002). Denote $v_\epsilon = u_\epsilon - \inf_M u_\epsilon$. Then, for any ball $B_r$ we have:

$$\sup_{B_r} v_\epsilon \leq C \left( \inf_{B_r} v_\epsilon + \left\| f \right\|_{L^2(B_{2r})} \right), \tag{E.9}$$

where $B_{2r}$ denotes the concentric ball with twice the radius of $B_r$. Since we can achieve a finite covering of $M$ by balls $B_\delta$, where $\delta < r_M$ as explained above, we now use a chaining argument to get a uniform bound on $\sup_{B_\delta} v_\epsilon$. Denoting the ball $B_\delta^1$ where the minimum is obtained, i.e. $v_\epsilon(x) = 0$ for some $x \in B_\delta^1$, we have:

$$\sup_{B_\delta^1} v_\epsilon \leq C \left\| f \right\|_{L^2(M)}. \tag{E.10}$$

Now take a ball $B_\delta^2$ that overlaps by $\delta$ with $B_\delta^1$. Then,

$$\sup_{B_\delta^1} v_\epsilon \leq C \left( \inf_{B_\delta^1} v_\epsilon + \left\| f \right\|_{L^2(M)} \right) \leq C \left( \sup_{B_\delta^1} v_\epsilon + \left\| f \right\|_{L^2(M)} \right) \leq C' \left\| f \right\|_{L^2(M)}. \tag{E.11}$$

Repeating this chaining argument a finite number of times can cover the compact manifold $M$. Thus, we get:

$$\sup_M v_\epsilon \leq C'' \left\| f \right\|_{L^2(M)}. \tag{E.12}$$

Thus, since $u_\epsilon$ is mean-zero, we have therefore that $u_\epsilon$ is bounded and thus $\|u_\epsilon\|_{C^{2,0}(M)} \leq C$ for a universal constant $C$ that does not depend on $\epsilon$. Hence, we know that since the $u_\epsilon$ is bounded in its Lipschitz constant, so, in fact, the solutions coincide, i.e. $w_\epsilon = u_\epsilon$, for $R$ chosen large enough.

By the stability of viscosity solutions Crandall et al. (1992), that $\overline{U} = \limsup_\epsilon u_\epsilon$ is a viscosity subsolution of $-\Delta u + f(x)$. Likewise, we can define $\underline{U} = \liminf u_\epsilon$ and it is a viscosity supersolution of $-\Delta u + f(x)$.

Take any sequence $\epsilon_n \to 0$ such that $u_\epsilon \to U$ uniformly. Then, $U$ is the mean-zero viscosity solution of $-\Delta u + f(x)$. For any sequence $\epsilon_n$, there exists a subsequence $\epsilon_{n_k}$ such that $u_{\epsilon_{n_k}} \to U$ uniformly, by Arzela-Ascoli. Since every sequence contains a subsequence that converges to $U$, we conclude that, in fact $u_{\epsilon_n} \to U$ for any sequence. □

This now allows us to instead discretize Equation (E.2) instead. Then, the procedure for solving the the discrete modified Poisson equation is by using a parabolic scheme presented in Algorithm 5, with or without the speedup provided by Schaeffer and Hou (2016) as necessary.

**Algorithm 5** Computing the solution to elliptic PDE $F[u] = 0$

```
Initialize u_0^h;
Fix ε > 0;
while |G^h(x_i, u_n^h(x_i))| > ε do
    Compute u_{n+1}^h(x_i) = u_n^h(x_i) + ΔtG^h(x_i, u_n^h(x_i))
end while
```

In the case of the Poisson equation when the Lipschitz constraint is inactive, we solve a linear system via standard linear algebra techniques.

As was explained in Chapter 3, the convergence framework presented there could apply to PDE beyond simply the Optimal Transport PDE. Furthermore, the framework can also apply beyond the case of the sphere, provided that the *a priori* solution of the PDE has sufficient regularity. The key properties in the discretization of the elliptic PDE without boundary are consistency, monotonicity, and Lipschitz stability, as shown in Chapter 3. Techniques are extremely similar

to those used in Chapter 3. Uniform convergence can be proved for non-smooth cases ($C^{0,1}(M)$) by assuming that the modified Poisson equation (Equation (E.2)) has a unique solution and assuming the scheme is underestimating, see Chapter 3 for more details. The interpolation in Section 3.3.3 can be modified, *mutatis mutandis*, to apply the interpolation to the manifold $M$.

# APPENDIX F

## EUCLIDEAN DIVERGENCE FORM PDE

In this appendix, we show that a uniformly elliptic linear PDE in divergence form (with respect to the divergence operator on the Riemannian manifold) can be expressed as a uniformly elliptic linear PDE in divergence form in a local coordinate neighborhood with respect to the divergence operator in local tangent planes.

**Lemma F.1** (PDE on local coordinate patches)**.** *Under the assumptions of Hypothesis 7.1, there exists some $r > 0$ such that for every $x_0 \in M$ there exists a bounded region $\Omega \subset \mathbb{R}^2$ and set of coordinates $y : \Omega \to B(x_0, r)$ corresponding to a metric tensor $G \in C^2(M)$ such that the PDE (Equation (7.2)) can be expressed as*

$$\mathcal{L}[\phi] = -\nabla \cdot \left((\det A)^{1/2} \nabla \phi\right).$$

*Proof.* Let $x_0 \in M$ and fix any $r < r_I$ where $r_I$ is the injectivity radius of the manifold $M$. Then we can consider a bounded set $\Omega \subset \mathbb{R}^2$ and a set of coordinates $y : \Omega \to B(x_0, r)$. In local coordinates Cabré (2002), the PDE (Equation (7.2)) takes the form

$$\mathcal{L}[\phi] = \frac{-1}{\sqrt{\det G}} \nabla \cdot \left(\sqrt{\det G} A G^{-1} \nabla \phi\right), \quad y \in \Omega.$$

Now we choose a local metric such that $G = (\det A)^{-1/2} A$. We note that $G \in C^2(M)$ is strictly positive definite since $A$ has both these properties. We note that $\det(G) = 1$ so that the PDE in local coordinates becomes

$$\mathcal{L}[\phi] = -\nabla \cdot \left((\det A)^{1/2} \nabla \phi\right).$$

This is a uniformly elliptic operator since $A$ is positive definite. □

# APPENDIX G

## FINITE GEODESIC BALL COVERING

In this appendix, we show the construction of local coordinate patches that will allows our regularity "patching" argument to follow. What we need is a finite covering of the manifold using geodesic balls of uniform radius $r > 0$ and allowing for uniform overlap $\delta > 0$.

**Lemma G.1.** *There exists a finite covering of the manifold $M$ with geodesic balls of radius $r > 0$, that is a set $\{B_r^i\}_{i=1,\dots,n}$ such that $M \subset \cup_i B_r^i$ and with the geodesic balls overlapping at least an amount $\delta > 0$. More precisely, for every $i$ and every $x \in \partial B_r^i$, there exists a ball $B_\delta$ such that $x \in B_\delta$ and there exists an index $i' \neq i$ such that also we have $x \in B_r^{i'}$.*

*Proof.* Since the diameter of the manifold is finite and the manifold itself is geodesically complete, we can connect any two points with a geodesic and this geodesic has length strictly upper bounded by $\mathrm{diam}(M)$. We then obtain a covering of the points of the geodesic with geodesic balls of radius $r > 0$. That is, if $\gamma_{xy}$ is the minimal geodesic connecting $x, y \in M$, then for any $z \in \gamma_{xy}$, we can construct $\left\{B_\rho^i\right\}_i$ such that $z \in B_\rho^i$ for some $j$. Since this can be done for all points $x, y$ in $M$, we thus obtain a cover of the manifold $M$ by balls of radius $\rho$. By compactness, we may find a finite subcover. From now on, we will denote the finite subcover by $\left\{B_\rho^i\right\}_i$. Take each member of the subcover $B_\rho^i$ and define $B_r^i = \left\{x \in M : d(x, z) \leq \delta, z \in B_\rho^i\right\}$. Then, choosing $r = \rho + \delta$, we achieve the desired finite covering of geodesic balls of radius $r$ which overlap by $\delta$. □

## REFERENCES


Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *Society for Industrial and Applied Mathematics Journal on Mathematical Analysis*, 43(2):904–924.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223.

Aubin, T. (1998). *Some Nonlinear Problems in Riemannian Geometry*. Springer, Berlin, Germany.

Barles, G. and Souganidis, P. E. (1991). Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4:271–283.

Bauer, M., Joshi, S., and Modin, K. (2015). Diffeomorphic density matching by optimal information transport. *Society for Industrial and Applied Mathematics Journal on Imaging Sciences*, 8(3):1718–1751.

Bauer, M., Joshi, S., and Modin, K. (2017). Diffeomorphic random sampling using optimal information transport. *Geometric Science of Information 2017*, pages 135–142.

Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393.

Benamou, J.-D. and Carlier, G. (2015). Augmented Lagrangian methods for transport optimization, mean-field games and degenerate PDEs. *Journal of Optimization Theory and Applications*, 167:1–26.

Benamou, J.-D., Collino, F., and Mirebeau, J.-M. (2016). Monotone and consistent discretization of the Monge-Ampère operator. *Mathematics of Computation*, 85(302):2743–2775.

Benamou, J.-D. and Duval, V. (2017). Minimal convex extensions and finite difference discretisation of the quadratic Monge-Kantorovich problem. *European Journal of Applied Mathematics*, pages 1–38.

Benamou, J.-D., Froese, B. D., and Oberman, A. M. (2010). Two numerical methods for the elliptic Monge-Ampère. *European Series in Applied and Industrial Mathematics: Mathematical Modelling and Numerical Analysis*, 44:737–758.

Benamou, J.-D., Froese, B. D., and Oberman, A. M. (2014). Numerical solution of the optimal transportation problem using the Monge-Ampère equation. *Journal of Computational Physics*, 260:107–126.

Berger, M. (2003). *A Panoramic View of Riemannian Geometry*. Springer, Berlin, Germany.

Bigot, J. (2020). Statistical data analysis in the Wasserstein space. *European Series in Applied and Industrial Mathematics: Proceedings and Surveys*, 68:1–19.

Bonnet, G. and Mirebeau, J.-M. (2021). Monotone discretization of the Monge-Ampère equation of optimal transport. *Hyper Articles en Ligne Open Science*.

Brenner, S. C. and Neilan, M. (2012). Finite element approximations of the three dimensional Monge-Ampère equation. *European Series in Applied and Industrial Mathematics: Mathematical Modelling and Numerical Analysis*, 46:979–1001.

Brix, K., Hafizogullari, Y., and Platen, A. (2015). Designing illumination lenses and mirrors by the numerical solution of Monge-Ampère equations. *Journal of the Optical Society of America A*, 32(11):2227–2236.

Bruneton, A., Bäuerle, A., Loosen, P., and Wester, R. (2011). Freeform lens for an efficient wall washer. In *Optical Design and Engineering IV*, volume 8167, page 816707, California, United States. International Society for Optics and Photonics.

Budd, C. J., Cullen, M. J. P., and Walsh, E. J. (2013). Monge-Ampère based moving mesh methods for numerical weather prediction, with applications to the Eady problem. *Journal of Computational Physics*, 236:247–270.

Budd, C. J. and Williams, J. (2009). Moving mesh generation using the parabolic Monge-Ampère equation. *Society for Industrial and Applied Mathematics Journal on Scientific Computing*, 31(5):3438–3465.

Cabré, X. (2002). Topics in regularity and qualitative properties of solutions on nonlinear elliptic equations. *Discrete and Continuous Dynamical Systems*, 8(2):331–359.

Caffarelli, L., Buttazzo, G., and Salsa, S. (2003). Optimal transportation, dissipative PDE's and functional inequalities. *Optimal Transport and Applications, Lectures given at the C.I.M.E. Summer School, held in Martina Franca, Italy*, pages 53–89.

Caffarelli, L. A. and Cabré, X. (1995). *Fully Nonlinear Elliptic Equations*. American Mathematical Society, Providence, Rhode Island, United States.

Carlier, G., Oberman, A., and Oudet, E. (2015). Numerical methods for matching for teams and Wasserstein barycenters. *European Series in Applied and Industrial Mathematics: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642.

Carmo, M. P. D. (2016). *Differential Geometry of Curves and Surfaces*. Dover Publications, Inc., Mineola, New York, United States, second edition.

Chacón, L., Delzanno, G. L., and Finn, J. M. (2011). Robust, multidimensional mesh-motion based on Monge-Kantorovich equidistribution. *Journal of Computational Physics*, 230:87–103.

Chen, C. and Öktem Ozan (2017). Indirect image registration with large diffeomorphic deformations. *Society for Industrial and Applied Mathematics Journal on Imaging Sciences*, 11(1).

Chen, Y.-Y., Wan, J., and Lin, J. (2018). Monotone mixed finite differencce scheme for Monge-Ampére equations. *Journal of Scientific Computing*, 76:1839–1867.

Crandall, M. G., Ishii, H., and Lions, P.-L. (1992). User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1):1–67.

Cui, L., Qi, X., Wen, C., Lei, N., Li, X., Zhang, M., and Gu, X. (2019). Spherical optimal transportation. *Computer-Aided Design*, 115:181–193.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*, 26.

Dacorogna, B. and Moser, J. (1990). On a partial differential equation involving the Jacobian determinant. *Annales de l'Institut Henri Poincaré*, 7(1):1–26.

Dean, E. J. and Glowinski, R. (2005). Operator-splitting methods and applications to the direct numerical simulation of particulate flow and to the solution of the elliptic Monge-Ampère equation. *Control and Boundary Analysis, Lecture Notes in Pure and Applied Mathematics*, 240:1–27.

Dean, E. J. and Glowinski, R. (2006a). An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in two dimensions. *Electronic Transactions on Numerical Analysis*, 22:71–96.

Dean, E. J. and Glowinski, R. (2006b). Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type. *Computational Methods in Applied Mechanics and Engineering*, 195:1344–1386.

Dean, E. J. and Glowinski, R. (2008). On the numerical solution of the elliptic Monge-Ampère equation in dimension two: a least squares approach. *Partial Differential Equations, Computational Methods in Applied Sciences*, 16:43–63.

Desnijder, K., Hanselaer, P., and Meuret, Y. (2019). Ray mapping method for off-axis and non-paraxial freeform illumination lens design. *Optics Letters*, 44(4):771–774.

Doskolovich, L. L., Bykov, D. A., Mingazov, A. A., and Bezus, E. A. (2019). Optimal mass transportation and linear assignment problems in the design of freeform refractive optical elements generating far-field irradiance distributions. *Optics Express*, 27(9):13083–13097.

Engquist, B. and Froese, B. D. (2014). Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5):979–988.

Engquist, B., Froese, B. D., and Yang, Y. (2016). Optimal transport for seismic full waveform inversion. *Communications in Mathematical Sciences*, 14(8):2309–2330.

Evans, L. C. (1997). Partial differential equations and Monge-Kantorovich mass transfer. *Current Developments in Mathematics*, pages 65–126.

Evans, L. C. and Gangbo, W. (1999). Differential equations methods for the Monge-Kantorovich mass transfer problem. *Memoirs of the American mathematical Society*, 137(653).

Feng, X., Glowinski, R., and Neilan, M. (2013a). Recent developments in numerical methods for fully nonlinear second order partial differential equations. *Society for Industrial and Applied Mathematics Review*, 55(2):205–267.

Feng, X. and Jensen, M. (2017). Convergent semi-Lagrangian methods for the Monge-Ampère equation on unstructured grids. *Society for Industrial and Applied Mathematics Journal on Numerical Analysis*, 55(2):691–712.

Feng, X., Kao, C.-Y., and Lewis, T. (2013b). Convergent finite difference methods for one-dimensional fully nonlinear second order partial differential equations. *Journal of Computational and Applied Mathematics*, 254:81–98.

Feng, X. and Lewis, T. (2014a). Local discontinuous Galerkin methods for one-dimensional second order fully nonlinear elliptic and parabolic equations. *Journal of Scientific Computing*, 59(1):129–157.

Feng, X. and Lewis, T. (2014b). Mixed interior penalty discontinuous Galerkin methods for fully nonlinear second order elliptic and parabolic equations in high dimensions. *Numerical Methods for Partial Differential Equations*, 30(5):1538–1557.

Feng, X. and Lewis, T. (2018). Nonstandard local discontinuous Galerkin methods for fully nonlinear second order elliptic and parabolic equations in high dimensions. *Journal of Scientific Computing*, 77(3):1534–1565.

Feng, X. and Neilan, M. (2007). Galerkin methods for the fully nonlinear Monge-Ampère equation. *arxiv preprint arXiv:0712.1240*.

Feng, X. and Neilan, M. (2009a). Mixed finite element methods for the fully nonlinear Monge-Ampère equation based on the vanishing moment method. *Society for Industrial and Applied Mathematics Journal on Numerical Analysis*, 47:1226–1250.

Feng, X. and Neilan, M. (2009b). Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. *Journal of Scientific Computing*, 38:74–98.

Feng, Z., Froese, B. D., and Liang, R. (2016). Freeform illumination optics construction following an optimal transport map. *Applied Optics*, 55(16):4301–4306.

Figalli, A., Rifford, L., and Villani, C. (2010). On the Ma-Trudinger-Wang curvature on surfaces. *Calculus of Variations*, 39:307–332.

Finlay, C. and Oberman, A. (2019). Improved accuracy of monotone finite difference schemes on point clouds and regular grids. *Society for Industrial and Applied Mathematics Journal on Scientific Computing*, 41:A3097–A3117.

Fournier, F. R., Cassarly, W. J., and Rolland, J. P. (2010). Fast freeform reflector generation using source-target maps. *Optics Express*, 18(5):5295–5304.

Friedrich, V. T. (1991). Die Fisher-Information und symplektische strukturen. *Mathematische Nachrichten*, 153:273–296.

Froese, B. D. (2012). A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions. *Society for Industrial and Applied Mathematics Journal on Scientific Computing*, 34(3):A1432–A1459.

Froese, B. D. (2018). Meshfree finite difference approximations for functions of the eigenvalues of the Hessian. *Numerische Mathematik*, 138(1):75–99.

Froese, B. D. and Oberman, A. M. (2011a). Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampère equation in dimensions two and higher. *Society for Industrial and Applied Mathematics Journal on Numerical Analysis*, 49(4):1692–1714.

Froese, B. D. and Oberman, A. M. (2011b). Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation. *Journal of Computational Physics*, 230:818–834.

Froese, B. D. and Oberman, A. M. (2013). Convergent filtered schemes for the Monge-Ampère partial differential equation. *Society for Industrial and Applied Mathematics Journal on Numerical Analysis*, 51(1):423–444.

Gangbo, W., Li, W., Osher, S., and Puthawala, M. (2019). Unnormalized optimal transport. *arxiv preprint arXiv:1902:03367*.

Gangbo, W. and McCann, R. J. (2000). Shape recognition via Wasserstein distance. *Quarterly of Applied Mathematics*, 58(4):705–737.

Gangbo, W. and Oliker, V. (2007). Existence of optimal maps in the reflector-type problems. *European Series in Applied and Industrial Mathematics: Control, Optimisation and Calculus of Variations*, 13(1):93–106.

Gilbarg, D. and Trudinger, N. S. (2001). *Elliptic Partial Differential Equations of Second Order*. Springer, Berlin, Germany.

Glimm, T. and Oliker, V. (2003). Optical design of single reflector systems and the Monge-Kantorovich mass transfer problem. *Journal of Mathematical Sciences*, 117(3):4096–4108.

Glowinski, R., Dean, E. J., Guidoboni, G., Juárez, L. H., and Pan, T.-W. (2008). Applications of operator-splitting methods to the direct numerical simulation of particulate and free-surface flows and to the numerical solution of the two-dimensional elliptic Monge-Ampère equations. *Japan Journal of Industrial and Applied Mathematics*, 25:1–63.

Gomes, D. A., Nurbekyan, L., and Pimentel, E. A. (2015). *Economic Models and Mean-field Games Theory*. Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil, publicações matemáticas, 30cbm edition.

Gorbunova, V., Sporring, J., Lo, P., Loeve, M., Tiddens, H. A., Nielson, M., Dirksen, A., and de Bruijne, M. (2012). Mass preserving image registration for lung CT. *Medical Image Analysis*, 16:786–795.

Haker, S., Tannenbaum, A., and Kikinis, R. (2001). Mass preserving mappings and image registration. *Medical Image Computing and Computer-Assisted Intervention 2001*, Lecture Notes in Computer Science 2208:120–127.

Haker, S., Zhu, L., Tannenbaum, A., and Angenent, S. (2004). Optimal mass transport for registration and warping. *International Journal of Computer Vision*, 60(3):225–240.

Hamfeldt, B. and Salvador, T. (2018). Higher-order adaptive finite difference methods for fully nonlinear elliptic equations. *Journal of Scientific Computing*, 75(3):1282–1306.

Hamfeldt, B. D. (2019). Convergence framework for the second boundary value problem for the Monge-Ampère equation. *Society for Industrial and Applied Mathematics Journal on Numerical Analysis*, 57(2):945–971.

Hamfeldt, B. F. (2018). Convergent approximation of non-continuous surfaces of prescribed Gaussian curvature. *Communications on Pure and Applied Analysis*, 17(2):671–707.

Hamfeldt, B. F. and Lesniewski, J. (2022a). A convergent finite difference method for computing minimal Lagrangian graphs. *Communications on Pure and Applied Analysis*, 21(2):393–418.

Hamfeldt, B. F. and Lesniewski, J. (2022b). Convergent finite difference methods for fully nonlinear elliptic equations in three dimensions. *Journal of Scientific Computing*, 90(35).

Hamfeldt, B. F. and Turnquist, A. G. R. (2021a). A convergence framework for optimal transport on the sphere. *arXiv preprint arXiv:2103.05739*.

Hamfeldt, B. F. and Turnquist, A. G. R. (2021b). A convergent finite difference method for optimal transport on the sphere. *Journal of Computational Physics*, 445.

Hamfeldt, B. F. and Turnquist, A. G. R. (2021c). Convergent numerical method for the reflector antenna problem via optimal transport on the sphere. *Journal of the Optical Society of America A*, 38:1704–1713.

Hamfeldt, B. F. and Turnquist, A. G. R. (2022). On the reduction in accuracy of finite difference schemes on manifolds without boundary. *arxiv preprint arXiv:2204.01892*.

Ishii, H. (1995). On the equivalence of two notions of weak solutions, viscosity solutions and distribution solutions. *Funkcialaj Ekvacioj*, 38:101–120.

Jonker, R. and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Julien, R., Peyré, G., Delon, J., and Marc, B. (2011). Wasserstein barycenter and its application to texture mixing. *Scale Space and Variational Methods in Computer Vision 2011*, pages 435–446.

Kocan, M. (1995). Approximation of viscosity solutions of elliptic partial differential equations on minimal grids. *Numerische Mathematik*, 72(1):73–92.

Kochengin, S. A. and Oliker, V. I. (1998). Determination of reflector surfaces from near-field scattering data II. numerical solution. *Numerische Mathematik*, 79:553–568.

Lasry, J.-M. and Lions, P.-L. (2007). Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260.

Lee, J. M. (2006). *Riemannian Manifolds: an Introduction to Curvature*, volume 176. Springer Science & Business Media, Berlin, Germany.

Léonard, C. (2012). From the Schrödinger problem to the Monge-Kantorovich problem. *Journal of Functional Analysis*, 262(4).

Léonard, C. (2013). A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574.

Lévy, B. (2015). A numerical algorithm for L2 semi-discrete optimal transport in 3D. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1693–1715.

Li, P. and Yau, S.-T. (1986). On the parabolic kernel of the Schrödinger operator. *Acta Mathematica*, 156:153–201.

Li, W., Osher, S., and Gangbo, W. (2016). A fast algorithm for earth mover's distance based on optimal trannsport and $l_1$ type regularization. *arxiv preprint arXiv:1609.07092*.

Lindsey, M. and Rubinstein, Y. A. (2017). Optimal transport via a Monge-Ampère optimization problem. *Society for Industrial and Applied Mathematics Journal on Mathematical Analysis*, 49(4):3073–3124.

Liu, J., Froese, B. D., Oberman, A. M., and Xiao, M. (2017). A multigrid scheme for 3d Monge-Ampère equations. *International Journal of Computer Mathematics*, 94(9):1850–1966.

Loeper, G. (2009). On the regularity of solutions of optimal transportation problems. *Acta Mathematica*, 202:241–283.

Loeper, G. (2011). Regularity of optimal maps on the sphere: the quadratic cost and the reflector antenna. *Archive for rational mechanics and analysis*, 199(1):269–289.

Ma, X.-N., Trudinger, N. X., and Wang, X.-J. (2005). Regularity of potential functions of the optimal transportation problem. *Archive for Rational Mechanics and Analysis*, 177(2):151–183.

McCann, R. J. (2001). Polar factorization of maps on Riemannian manifolds. *Geometric and Functional Analysis*, 11:589–608.

McCann, R. J. (2006). Stable rotating binary stars and fluid in a tube. *Houston Journal of Mathematics*, 32(2):603–631.

McRae, A. T., Cotter, C. J., and Budd, C. J. (2018). Optimal-transport-based mesh adaptivity on the plane and sphere using finite elements. *Society for Industrial and Applied Mathematics Journal on Scientific Computing*, 40(2):A1121–A1148.

Modin, K. (2015). Generalised Hunter-Saxton equations, optimal information transport, and factorisation of diffeomorphisms. *The Journal of Geometric Analysis*, 25:1306–1334.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *De l'Imprimerie Royale*.

Moselhy, T. A. E. and Marzouk, Y. M. (2012). Bayesian inference with optimal maps. *Journal of Computational Physics*, 231:7815–7850.

Moser, J. (1965). One the volume elements on a manifold. *Transactions of the American Mathematical Society*, 120:286–294.

Neilan, M. (2010). A nonconforminng Morley finite element method for the fully nonlinear Monge-Ampère equation. *Numerische Mathematik*, 115:371–394.

Neilan, M. (2014). A unified analysis of three finite element methods for the Monge-Ampère equation. *Electronic Transactions on Numerical Analysis*, 41:262–288.

Nochetto, R., Ntogkas, D., and Zhang, W. (2018). Two-scale method for the Monge-Ampère equation: convergence to the viscosity solution. *Mathematics of Computation*.

Oberman, A. M. (2004). A convergent difference scheme for the infinity Laplacian: Construction of absolutely minimizing Lipschitz extensions. *Mathematics of Computation*, 74(251):1217–1230.

Oberman, A. M. (2006). Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton–Jacobi equations and free boundary problems. *Society for Industrial and Applied Mathematics Journal on Numerical Analysis*, 44(2):879–895.

Oberman, A. M. (2008). Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. *Discrete and Continuous Dynamical Systems Series B*, 10(1):221–238.

Oberman, A. M. and Ruan, Y. (2020). Solution of optimal transportation problems using a multigrid linear programming approach. *Journal of Computational Mathematics*, 38:933–951.

Oliker, V. (2006). Freeform optical systems with prescribed irradiance properties in near-field. In *International Optical Design Conference 2006*, volume 6342, page 634211, California, United States. International Society for Optics and Photonics.

Oliker, V. and Newman, E. (1993). The energy conservation equation in the reflector mapping problem. *Applied Mathematics Letters*, 6(1):91–95.

Oliker, V., Rubinstein, J., and Wolansky, G. (2015). Supporting quadric method in optical design of freeform lenses for illumination control of a collimated light. *Advances in Applied Mathematics*, 62:160–183.

Oliker, V. I. and Prussner, L. D. (1988). On the numerical solution of the equation $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - (\partial^2 z/\partial x \partial y)^2 = f$ and its discretizations, i. *Numerische Mathematik*, 54:271–293.

Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2).

Parkyn, B. and Pelka, D. (2006). Free-form illumination lenses designed by a pseudo-rectangular lawnmower algorithm. In *Nonimaging Optics and Efficient Illumination Systems III*, volume 6338, page 633808, California, United States. International Society for Optics and Photonics.

Pass, B. (2015). Multi-marginal optimal transport: theory and applications. *European Series in Applied and Industrial Mathematics: Mathematical Modelling and Numerical Analysis*, 49:1771–1790.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: with applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

Prins, C. R., Beltman, R., ten Thije Boonkkamp, J. H. M., IJzerman, W. L., and Tukker, T. W. (2015). A least-squares method for optimal transport using the Monge-Ampère equation. *Society for Industrial and Applied Mathematics Journal on Scientific Computing*, 37(6):B937–B961.

Romijn, L. B., ten Thije Boonkkamp, J. H. M., and IJzerman, W. L. (2020). Inverse reflector design for a point source and far-field target. *Journal of Computational Physics*, 408:109283.

Rottman, C., Bauer, M., Model, K., and Joshi, S. C. (2015). Weighted diffeomorphic density matching with applications to thoracic image registration. *5th Medical Image Computing and Computer Assisted Intervention Society Workshop on Mathematical Foundations of Computational Anatomy*.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*, volume 55. Birkhäuser, Basel, Switzerland.

Schaeffer, H. and Hou, T. Y. (2016). An accelerated method for nonlinear elliptic PDE. *Journal of Scientific Computing*, 69(2):556–580.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., and Lander, E. S. (2017). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *bioRxiv*.

Schmitzer, B. (2016). A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259.

Solomon, J., Rustamov, R., Guibas, L., and Butscher, A. (2014). Earth mover's distances on discrete surfaces. *Association for Computing Machinery Transactions on Graphics*, 33(4):1–12.

Thorpe, M. (2019). Introduction to optimal transport. Technical report, Centre for Mathematical Sciences University of Cambridge.

Turnquist, A. G. R. (2021). Adaptive mesh methods on compact manifolds via optimal transport and optimal information transport. *arXiv preprint arXiv:2111.14276*.

Urbas, J. (1997). On the second boundary value problem for equations of Monge-Ampère type. *Journal für die reine und angewandte Mathematik*, 487:115–124.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society, Providence, Rhode Island, United States.

Villani, C. (2009). *Optimal Transport: Old and New*, volume 338 of *A Series of Comprehensive Studies in Mathematics*. Springer, Berlin, Germany.

Wang, W., Slepcev, D., Basu, S., Ozolek, J. A., and Rohde, G. K. (2013). A linear optimal transportation framework for quanitfying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269.

Wang, X.-J. (1996). On the design of a reflector antenna. *IOP Science*, 12:351–375.

Wang, X.-J. (2004). On the design of a reflector antenna II. *Calculus of Variations and Partial Differential Equations*, 20(3):329–341.

Weller, H., Browne, P., Budd, C., and Cullen, M. (2016a). Mesh adaptation on the sphere using optimal transport and the numerical solution of a Monge-Ampère type equation. *Journal of Computational Physics*, 308:102–123.

Weller, H., Browne, P., Budd, C., and Cullen, M. (2016b). Mesh generation on the sphere using optimal transport and the numerical solution of a Monge-Ampère type equation. *Journal of Computational Physics*, 308:102–123.

Wu, R., Xu, L., Liu, P., Zhang, Y., Zheng, Z., Li, H., and Liu, X. (2013). Freeform illumination design: a nonlinear boundary problem for the elliptic Monge-Ampére equation. *Optics Letters*, 38(2):229–231.

Yadav, N. K. (2018). *Monge-Ampère Problems with Non-Quadratic Cost Function: Application to Freeform Optics*. PhD thesis, Technische Universiteit Eindhoven, Eindhoven, Netherlands.

Yadav, N. K., ten Thije Boonkkamp, J. H. M., and Ijzerman, W. L. (2019). A Monge-Ampère problem with non-quadratic cost function to compute freeform lens surfaces. *Journal of Scientific Computing*, 80(1):475–499.

Yang, Y., Engquist, B., Sun, J., and Hamfeldt, B. F. (2018). Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion. *Geophysics*, 83(1):R43–R62.

Younes, L. (2010). *Shapes and Diffeomorphisms*, volume 171 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin, Germany.