A FreeForm Optics Application of Entropic Optimal Transport

Giorgi Rukhaia

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Université Paris Dauphine

A FreeForm Optics Application of Entropic Optimal Transport

Soutenue par
Giorgi RUKHAIA
Le 23/11/2021

École doctorale n°543
L'École Doctorale SDOSE

Spécialité
Mathématiques

Composition du jury :

Jean-David BENAMOU Directeur de recherche, UNIVERSITE PARIS DAUPHINE - PSL		*Directeur*
Wilbert IJZERMAN Professor, TECHNISCHE UNIVERSITEIT EINDHOVEN		*Co-Directeur*
Bruno LÉVY Directeur de recherche, INRIA-NANCY		*Rapporteur*
Quentin MÉRIGOT Professeur des universités, UNIVERSITE PARIS-SACLAY		*Président du jury*
Roya MOHAYAEE Directeur de recherche, INSTITUT D'ASTROPHYSIQUE DE PARIS		*Membre du jury*
Bernhard SCHMITZER Professor, GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN		*Rapporteur*

Dauphine | PSL
UNIVERSITÉ PARIS

Acknowledgements

First of all, I would like to dedicate this work to my father, who encouraged me to pursue the path of knowledge and self-improvement but did not live long enough to see this work finished.

I would like to thank my supervisors, Jean-David Benamou and Wilbert IJzerman, who bared with my shortcomings and helped me push forward my capabilities, even at times when I myself didn't believe I could. I would like to give Jean-David special thanks for investing a tremendous amount of time in me, far more than what I think I deserved.

Many thanks to Guillaume Chazareix for an amazing teamwork. It is hard to imagine what the final chapter of this work would look like without our collaboration.

I should also thank my friends and colleagues at INRIA, who supplied many amazing memories, insights, and discussions both during work-time meetings or seminar sessions and after-work beer sessions. I also thank the staff of Signify and TU-Eindhoven for their amazing hospitality during my secondments there.

I also want to thank my friends and family in Georgia, who made me feel we were together for all the important moments of our lives, even though I have been living far from them for more than five years.

Last but not the least, I would like to thank the ROMSOC consortium, which provided funding for my work through European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 765374.

This consortium provided not only funds for the research, but also an amazing combination of great friends, a wide range of colleagues with different backgrounds, and many beautiful memories all around Europe.

Contents

Notations

$\mathbb{R}^d$: d-dimensional euclidean space, with the basis elements $\{e_1, ..., e_d\}$

$\mathbb{S}^{d-1}$: $d-1$-dimensional sphere, parametrized either by an angular parametrization, or as a set of unit vectors in $\mathbb{R}^d$.

$\vec{x}$: x will denote an angular parametrization of $\mathbb{S}^{d-1}$ while $\vec{x}$ (with the $\vec{\cdot}$) will be the corresponding unit vector in $\mathbb{R}^d$ (see notation 1 at page 18).

$\mathbb{S}_+^{d-1}$: $d-1$-dimensional northern hemisphere, that is, subset of $\mathbb{S}^d$ such that, for all $x \in \mathbb{S}_+^{d-1}$, d-th element in vector parametrization is strictly positive.

$\mathbb{S}_-^{d-1}$: $d-1$-dimensional southern hemisphere, that is, subset of $\mathbb{S}^d$ such that, for all $\vec{x} \in \mathbb{S}_+^{d-1}$, d-th element in vector parametrization is strictly negative.

S: The collection of source points from where the light is emitted.

s: Parametrizing variable for S.

O_s: An element of S as a source point in $\mathbb{R}^d$, corresponding to the parameter s.

$\mathcal{R}$: The reflecting surface.

$\mathcal{R}_\rho$: The reflecting surface produced using the general radial function ρ : $\mathcal{R}_\rho := \{\vec{x}\rho(x) | x \in X \subset \mathbb{S}_+^{d-1}\}$ (see page 20)

$\mathcal{R}_f$: The reflecting surface produced using the Kantorovich potential (see theorem 2.3) f: $\mathcal{R}_f := \{\vec{x}e^{f(x)} | x \in X \subset \mathbb{S}_+^{d-1}\}$ (see page 53)

μ_s: A measure on $\mathbb{S}_+^{d-1}$, depending on the parameter s, describing the intensity of light emitted from the source point corresponding to s.

$\mathcal{P}(X)$: A set of Borel probability measures on space X.

$H_{b,\alpha}(X)$: A subset of $\mathcal{P}(X)$, containing measures that have Holder continuous densities with the exponent α, bounded from below by b and above by $\frac{1}{b}$ (see notation 4 on page 4).

$X \times Y$: A product space comprising of pairs (x, y) built from $x \in X$ and $y \in Y$.

$\mu \otimes \nu$: A product measure in $\mathcal{P}(X \times Y)$, built from measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ by $\mu \otimes \nu(A \times B) := \mu(A)\nu(B)$ for all measurable $A \subset X$ and $B \subset Y$.

$f \oplus g$: A function on $X \times Y$, given by $f \oplus g(x, y) := f(x) + g(y)$ for functions $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$.

$T_{\#}\mu$: A push-forward measure of μ by the map $T : X \to Y$.

$\langle f, \mu \rangle_X$: The duality product between the continuous functions $f : X \to R$ and probability measures $\mu \in \mathcal{P}(X)$.

$\mathrm{KL}(\alpha \,|\, \beta)$: Kullback-leibler divergence of a measure α with respect to a measure β (see (2.22) on page 38).

ϵ: A regularization parameter for the entropic regularization (see (2.21) on page 37) of the optimal transport problem (see (2.6) on page 32).

f_{OT_ϵ}: Kantorovich potential obtained from solving an entropy regularized optimal transport problem (see (2.21) on page 37).

f_{S_ϵ}: Debiased Kantorovich potential obtained as a μ gradient of the Sinkhorn divergence functional (2.30) $S_\epsilon(\mu, \nu)$ (see (2.32) on page 41).

$\hat{f}$: A c-concave interpolation of the Kantorovich potential f (see (3.27) on page 62).

$\tilde{f}$: An entropic interpolation of the Kantorovich potential f (see (3.28) on page 63).

Introduction

Since the nineteenth century, electricity has altered human life in many ways through various applications. Nowadays, when we think about electricity, we usually think about devices and appliances like laptops and smartphones or fridges and coffee machines, that were made possible because of electricity. But we usually forget to mention one of the first aspects where electricity radically changed human life: Light!

One of the first massive applications of electricity was street and home lighting, and later also expanded to car lights. This cheap and effective method of producing light drastically improved the quality of life for most of the world's population and made it possible to form "cities that never sleep". However, as everything we humans do on the large scale, it had unintended consequences.

The extensive amount of light altered the ecosystem of the cities. Once romanticized herds of fireflies or the view of the night sky became a luxury of the rural areas. At some point at the end of the twentieth century, an anecdotal story circulated in the press, that during a major blackout in Los-Angeles, USA, some people were calling an emergency line to report a sighting of a "strange blue cloud in the sky", only to find out that they were seeing the Milky Way galaxy for the first time. A good demonstration of how much excess light is created by the street lights alone is the famous photo (Figure 1) from Mia Heikkila, taken in the town of Kauttua, Finland, in January 2016. During a rare atmospheric event in the cold Finnish winter night, the wasted light from the street lights did not scatter and created a reverse map of the town far in the sky.

This problem of having a large amount of extra light around settlements is known as *light pollution*. Tackling this problem would provide a twofold advantage. First, the excess light that we do not use means excess waste of energy we produce. Secondly, light pollution harms both human health and the natural ecosystem, hence resolving it would improve the state of the environment and result in increased quality of life.

One way to tackle this problem is to use lenses and reflectors around the

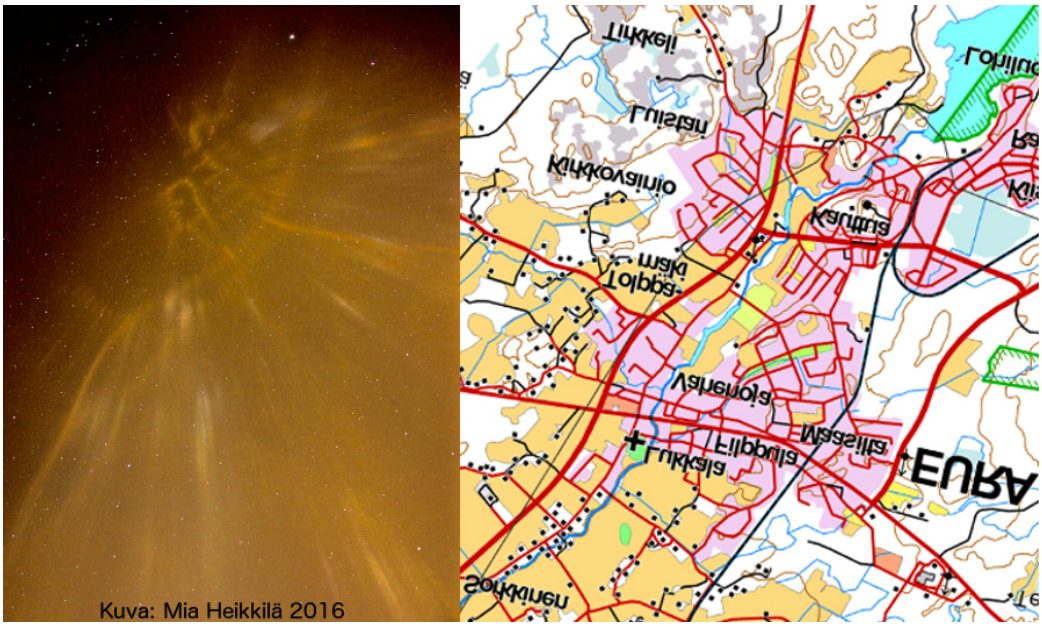


Figure 1: The night sky and the reverse map of the town

source of light, which would send all light only to the directions where we need it. While the main source of light was old-style light bulbs, which produced a high amount of heat, the lenses and reflectors around them had to be built from heat-resisting materials, usually glass, which restricted the design. The development of LED lights (figure 2), which require less energy to produce the same amount of light, while also operating at a lower temperature, removed this restriction and allowed the use of more flexible materials, such as plastic.

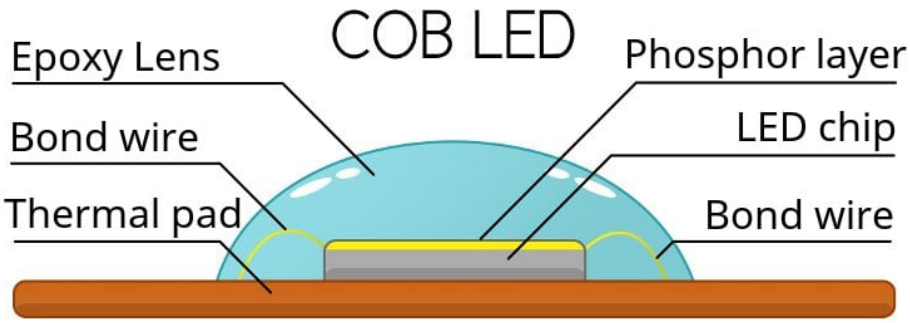


Figure 2: Structure of an LED light source (picture from lamphq.com)

Therefore, the solution boils down to designing optical systems that send all the emitted light to only the desired directions. This is a task of illumina-

tion optics, a branch of optics that deals with constructing optical systems for lighting. However, soon it became apparent, that simple shapes that are produced using classical techniques and enjoy rotational or translational symmetries are not enough to address such problems. This is where freeform optics, a branch of optics that focuses on the construction of optical systems that do not necessarily have rotational or translational symmetries, came into play.

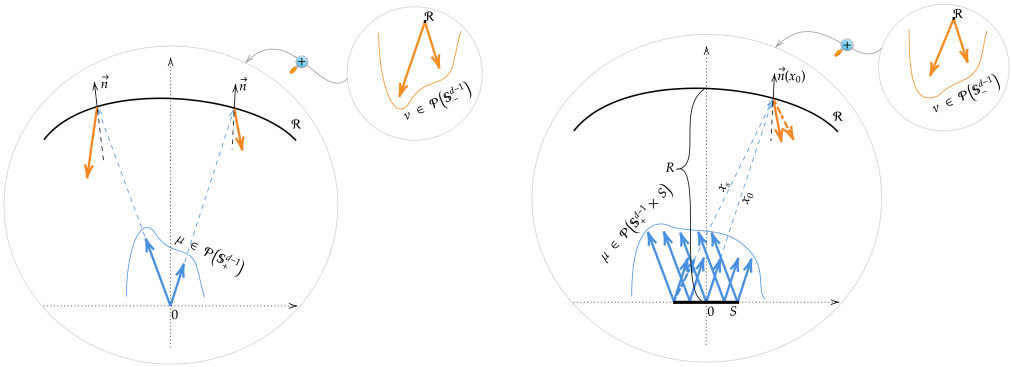
Through this work, we will address a problem at the intersection of illumination and freeform optics: The far-field reflector problem. This is a problem of designing a reflector such that it reflects all the light emitted from some source into the desired distribution. The far-field assumption means that the reflector is so small compared to both the location that needs to be illuminated and the distance to this location, that it can be regarded as a point with respect to the location, therefore only the directions of the reflected rays matter. This assumption usually holds in the applications of lighting, where the light source and the reflector with a dimension of millimeter to several centimeters should illuminate areas of tens of meters wide.

Mathematical model for the far-field reflector problem

The far-field reflector problem is modeled under several idealization assumptions. The most fundamental is that we work under the setup of geometrical optics. This means that we work under the following assumptions:

- Light propagates as a set of rays: straight lines described by their source of origin and propagation angle.
- When a ray hits the reflecting surface, it is reflected into the new direction which is determined by the reflection map T, given by the geometrical reflection law (see definition 1).
- The light source S is a set of points in $\mathbb{R}^d$ and each point emits light with some given intensity distribution over the propagation angles.
- For each source point $s \in S$, the reflection map T satisfies the measure-preserving property (see definition 2) between the emitted and the reflected intensities.

The reflector is modeled as a $d-1$ dimensional surface in $\mathbb{R}^d$. The decision in how to model the source of light is an important choice. When the source



(a) The point source to far-field reflector problem

(b) The extended source to far-field reflector problem

Figure 3: The point source to far-field and the extended source to far-field reflector problem schemes (see chapter 1 for the notations)

is supposed to be drastically smaller than the reflector, it can be modeled as a point in $\mathbb{R}^d$ (usually taken at the origin O for convenience) and emitted light is therefore modeled as a distribution in directions only (A measure defined on $\mathbb{S}^{d-1}$). This choice is known as a point source to far-field reflector problem.

However in real-life applications, such cases are quite rare, and the source should be modeled as a subset of $\mathbb{R}^d$. As we have LED applications in mind, which are rectangular surfaces in $\mathbb{R}^3$ where each point emits light in all directions of the upper hemisphere (see figure 2), we model the source S as a finite diameter subset of $\mathbb{R}^{d-1}$ (usually containing origin for convenience). This is known as an extended source to far-field reflector problem. As we already mentioned, far-field assumption assumes that we are only interested in the direction of the reflected rays, hence the reflected light is modeled as a distribution in the directions only (again a measure on $\mathbb{S}^{d-1}$). For convenience we assume that the source emits light in the direction of the northern hemisphere $\mathbb{S}^{d-1}_+$ and the reflections are directed to the southern hemisphere $\mathbb{S}^{d-1}_-$. In the figure 3 we present the geometrical setup of those models.

Solving the far-field reflector problem

In the case of the point source problem, each point on the reflector has no more than one incoming ray. Therefore for each incoming ray, there is one "control" dedicated specifically for that ray, that is, a normal at the reflector, which defines the reflecting direction. In contrast, for the extended

source problem, each point on the reflector may receive several or even infinite amount of rays (up to the size of S). Nevertheless, each point on the reflector still has only one "control", the normal at that point, which now defines reflecting direction for several rays. This indicates, that the extended source problem might be an over-determined problem. The point source problem is mathematically well-posed and it can be solved numerically, but those techniques do not apply to the extended source problem.

Solving the point source to far-field reflector problem

In [Wan96][Wan04](see also [GO03]), Wang showed that the point source to far-field reflector problem can be formulated as an optimal transport problem.

The optimal transport problem, first posed by Gaspard Monge in "Mémoire sur la théorie des déblais et des remblais" ([Mon81]), is a problem of allocating the mass optimally, with respect to some given cost of transporting a unit of mass. Mathematically it can be formulated as finding some "transfers" from the mass distributed according to the measure μ defined on the space X to another space Y with another mass distribution of measure ν, in a way, that the total cost of this transfer, with respect to some cost function $c(x, y)$ is minimal. Here $c(x, y)$ gives information about what it costs to transfer a unit of mass from a position $x \in X$ to $y \in Y$.

Wang constructed a cost function, that incorporates the information about the reflection, and solving an optimal transport problem with this cost provides a solution of the point source problem. Optimal transport theory provides a range of efficient solvers to tackle this problem. We discuss their applications to the reflector problem in chapter 3.

Solving the extended source to far-field reflector problem

Solvers for the extended source problem can roughly be split into two types: heuristic constructive methods, usually based on SMS (Simultaneous Multy-Surface method, [GBMB$^+$04]) and iterative improvement methods, which create some parametrization of the reflector, compute the reflection from this reflector, and then iteratively modify the reflector in order to produce more accurate reflection (for a recent review of the approaches for the extended source problem, see [WFZ$^+$18]).

There are various choices through iterative approaches, e.g. how is the reflector parametrized, how is the reflected distribution obtained, how is this reflection compared to the desired one, and finally, how are the modifications made to improve the reflector:

Some approaches (e.g. [BB19] [BKM+20]) parametrize reflectors as a general surface in $\mathbb{R}^3$ (e.g. by using spline parametrization, or just a set of points). This approach allows flexibility of deforming the reflecting surface when trying to produce the desired reflection. However, this might lead to the design of irregular surfaces, which might not be feasible for production or even develop "blind spots" during modifications, meaning that some parts of the reflector get in the way of light that was supposed to arrive at the different portion of the reflector.

In some works (see e.g [FCR10] [WZLM21], [LFHL10]), a point source problem is used to parametrize the reflector. In this approach, a reflector is built as a solution of some point source problem, which is then illuminated using the extended source. This approach has more limitations on what kind of reflection it can produce but has the advantage of making it possible to guarantee the regularity properties of the reflector, e.g. convexity or concavity, differentiability under some assumptions on the source and target measures, etc.

Obtaining the reflection from a given reflector and the source measure is done using ray-tracing. It is a widely used technique in various optical applications, based on a discrete sampling of measures and point-wise computation of reflection maps (see e.g. [Gla89] for a good review).

The modifications of the reflector at each iteration are usually based on the point-wise comparison of the desired and obtained reflections. Applying the modification depends on the way the reflector was parametrized. For example, in [BKM+20], spline parametrization is used, which is modified by doing quasi-newton minimization step on the L_2 norm between the desired and obtained reflections. In [FCR10], where the reflector is parametrized by the target distribution of the point source problem, modification is applied by scaling the parametrizing target distribution of the point source problem with the fraction of the desired and obtained reflections.

Our approach to the far-field reflector problem

In order to solve the point source optimal transport problem, we use the entropic regularization approach. This approach was introduced for optimal transport computations in [Cut13] (see [PC18] for a comprehensive review). Regularization is based on the penalization of the total cost of transfer by KullBack-Leibler divergence (2.22) (also known as "relative entropy"), mul-

tiplied by some small parameter ϵ.

Entropic regularization adds substantial regularity to the optimal transport problem and allows using the Sinkhorn algorithm, an efficient solver based on iterative projections. But it also introduces the "entropic bias", an error introduced by solving the altered problem. This is discussed in [FSV+18] where the Sinkhorn divergence correction (2.30) is used to compensate the bias. In chapter 3 we demonstrate that entropic regularization with this correction can be used to accurately solve the point source problem.

For the extended source, we start by analyzing the relationship between the point source and extended source problems. In particular, we consider reflectors generated as a solution of the point source problem between the fixed source measure μ_0 and a target measure ν_0, and compute the reflection ν of the extended source μ from this reflector. We study the relationship between ν_0 and ν, denoted by the functional $\mathcal{F}(\nu_0) = \nu$. We demonstrate that this functional can be expressed as a non-linear convolution, with a kernel involving the jacobian of the reflection map and its inverse, while also depending on ν_0.

Note that if we can invert this map, then we can solve the extended source problem for the desired target ν by finding $\nu_0 = \mathcal{F}^{-1}(\nu)$ and building the corresponding point source reflector. This parametrization fixes the overdetermined feature of the initial extended source problem. But $\mathcal{F}$ has a very complicated non-linear nature, which makes it hard to invert. Moreover, we are not even guaranteed that the inverse will be well-defined or even exist for a given desired target ν. Therefore, we resort to the minimization of the residual $\nu_0 \mapsto \mathcal{L}(\mathcal{F}(\nu_0), \nu)$ with an ad-hoc misfit/loss function $\mathcal{L}$.

Overall, this approach fits in the framework of a regularisation approach as is customary for ill-posed non-linear inverse problems (see for instance [5] for a recent review). It is a concept of trying to find a "best approximation" of the solution within some class, with respect to some loss/misfit function. In terms of the reflector problem for the fixed source distribution μ and the desired target ν, it is based on the following ingredients:

1. *A parametric set* of admissible reflectors. This is a regularization part, where parameter set should be chosen so that the problem is guaranteed to become well-posed, even if the original problem was not. We take parameter set to be the set of target distributions ν_0 and corresponding reflectors are obtained by solving the point source optimal transport problem.

2. *A forward map*, that for a given source distribution and an element

of the parameter set (corresponding to the reflector), produces the reflected distribution. We take the above-mentioned $\mathcal{F}$ as the forward map.

3. *A loss/merit function* that gives information about the "closeness" of the reflected and desired distributions. Ideally, this should be a distance, or at least convex, positive and reaching the minimum value of 0 only when the reflected distribution is equal to the desired one. For this, we use Sinkhorn divergence functional S_ϵ (see 2.30), which approximates the W_2^2, squared Wasserstein distance (see definition 6) between the desired target ν and the obtained reflection $\mathcal{F}(\nu)$.

Then the regularized solution is the reflector corresponding to the parameter ν_0, the best approximation of $\mathcal{F}^{-1}(\nu)$ in the sense that it minimizes the loss function $S_\epsilon(\mathcal{F}(\nu_0), \nu)$. The value of the loss function can be seen as a measure of the quality of the reflector.

In chapter 4 we prove that there exists a minimizer for this loss. We also use gradient-based optimization techniques, relying on the automatic differentiation tools, in order to find a minimizer.

Outline of the structure of this work

In chapter 1, we sum up the required notions from optics. First, we discuss the geometrical reflection law and related concepts used throughout the whole work. Then we present the geometrical optics setup of point source and extended source to far-field problems. Finally, we discuss ray-tracing, a method used to calculate the light distribution of the reflection from the reflector with the given light source distribution.

In chapter 2, we review the foundations of optimal transport and introduce some fundamental results which will be used later in chapters 3 and 4. We will not present the proofs, but rather focus on explaining the results and their role. First, we go through the basic concepts such as the dual formulation of optimal transport problem and the definition of Wasserstein distance. We then proceed to present the entropic regularization of optimal transport and its numerical resolution, which will play a crucial role in chapters 3-4. The material covered in this chapter can be found in [Vil08],[San15] and [PC18].

In chapter 3, we concentrate on the point source problem and its resolution using the optimal transport theory. We start by presenting the construction from [KO03], which leads to the optimal transport formulation of the point source problem from [Wan04]. We then proceed to present the

methods for finding such solutions and discuss the necessary adaptations for applying entropic regularization and the Sinkhorn algorithm to the point source problem. We also discuss ways of evaluating the obtained solution using ray-tracing. We then present the corresponding numerical results. Some of the material covered in this chapter is given in [BIR20].

In chapter 4, we present the study of the extended source problem. We restrict our attention to the reflectors that are generated as the optimal transport solutions of a point source problem and analyze the relation between the extended source and point source problems. Then we propose a regularization approach to find the "best approximation" of the desired reflector with respect to an optimal transport based loss. In the final section, we present our numerical simulations, analyzing the choice of different optimization strategies, choice of ray-tracing, etc. Material covered in this chapter is given in [BCIR21]. Through this chapter, we will work with the 2-dimensional case in order to emphasize the effect of the extended source problem and how to tackle it, clearly seen in this simple case as well, while avoiding non-uniform or weighted grids, discussed in Chapter 3, necessary for the 3-dimensional case.

Chapter 1

Optics

Introduction

In this chapter, we sum up the required notions from optics. First, we discuss the geometrical reflection law and related concepts used throughout the whole work. Then we present the geometrical optics setup of point source and extended source to far-field problems. Finally, we discuss ray-tracing, a method used to calculate the light distribution of the reflection from the reflector with the given light.

1.1 Geometrical Reflection law and Conservation of Energy

Throughout this document, we will work within the framework of Geometrical optics. This means that we work under the following assumptions:

- Light propagates as a set of rays: straight lines described by their source of origin and propagation angle.
- When a ray hits the reflecting surface, it is reflected into the new direction which is determined by the reflection map T, given by the geometrical reflection law (see definition 1 below).
- The light source S is a set of points in $\mathbb{R}^d$ and each point emits light with some given intensity distribution over the propagation angles.
- For each source point $s \in S$, the reflection map T satisfies the measure-preserving property (see definition 2) between the emitted and the reflected intensities.

Notation 1. *A ray will be described by its point of origin in $\mathbb{R}^d$ and its direction as an element of $\mathbb{S}^{d-1}$. By convention, x will denote an angular parametrization of $\mathbb{S}^{d-1}$ while $\vec{x}$ (with the $\vec{\cdot}$) will be the corresponding unit vector in $\mathbb{R}^d$.*

Definition 1 (Snell–Descartes law of geometrical reflection)**.** *When a ray of light, traveling in the direction $\vec{x} \in \mathbb{S}^{d-1}$, hits a reflecting surface at a point with an outward unit normal $\vec{n}$, the direction of the reflected ray $y = T(x)$ is given by*

$$\vec{y} = \vec{T}(x) = \vec{x} - 2\langle \vec{x}, \vec{n} \rangle \vec{n} \tag{1.1}$$

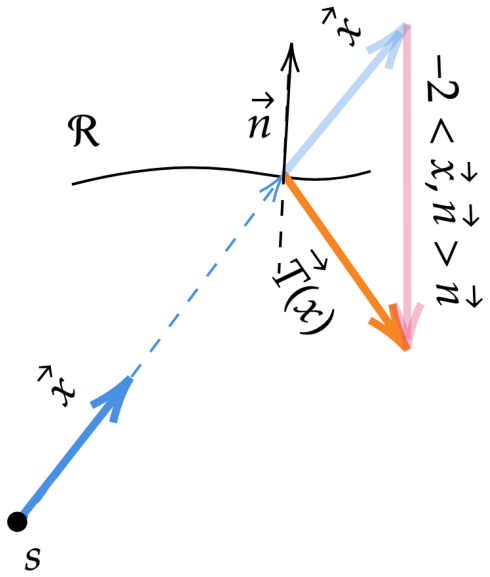


Figure 1.1: Reflection law for the ray x originating from the source $s \in S$, reflected on the surface $\mathcal{R}$ with a normal $\vec{n}$. The angle between the normal and the reflected ray is equal to the angle between the normal and the incoming ray.

Definition 2 (Measure preserving property)**.** *The map $T : X \to Y$ satisfies the measure-preserving property with respect to the measures μ and ν respectively on X and Y, if for any measurable set $E \subset Y$, $\nu(E) = \mu(T^{-1}(E))$. This relation is denoted by*

$$\nu = T_{\#}\mu \tag{1.2}$$

read as T "pushes forward" μ into ν.

Assuming that the measures μ and ν have densities (abusively denoted by μ and ν again), we can express the measure-theoretical notion of preserving the measure as an identity for the functions:

$$\forall\, E \subset Y \text{ measurable}, \qquad \int_{T_\rho^{-1}(E)} \mu(x)dx = \int_E \nu(y)dy \tag{1.3}$$

We will denote a set of points from which the light is emitted by $S \subset \mathbb{R}^d$. Then the intensity of all the emitted light can be modeled as some measure μ on the space $S \times \mathbb{S}^{d-1}$. In the same manner, if we denote the reflecting surface by $\mathcal{R}$, then the intensity of all the reflected light can also be modeled as a measure ν on the space $\mathcal{R} \times \mathbb{S}^{d-1}$. However, throughout this document, we will work with a far-field assumption (definition 2), which we will introduce in the section 1.2, where we will disregard the dependence of the reflected light intensity on $\mathcal{R}$ and consider only the directions of the reflected rays.

To sum up, through this work we always assume that the reflection on the surface $\mathcal{R}$ is given by the map $T : S \times X \subset \mathbb{S}^{d-1} \to Y \subset \mathbb{S}^{d-1}$ satisfying the reflection law (1.1), and for each $s \in S$, T_s is a measure-preserving map between the emitted intensity μ_s and the reflected intensity $\nu_s = T_{s\#}\mu_s$. The total reflected intensity $T_\#\mu$ is then given by:

$$\nu := \int_S \nu_s \, ds = \int_S T_{s\#}\mu_s \, ds \tag{1.4}$$

The angular parametrization of the reflecting direction will be denoted by $y = T_s(x)$. When the source S contains only one point, we will drop the subscript s.

1.2 Point source to far-field reflector problem model

In the point source to far-field reflector problem two "idealization" assumptions are made, the point source assumption and the far-field assumption.

Assumption 1 (The point source assumption)**.** *The point source assumption means that the light source is so small compared to the reflector, that it can be regarded as a point, usually taken at the origin $O \in \mathbb{R}^d$. More formally, we assume that $S = \{O\}$. Hence light emitted from it will have a density μ in directions only (in a subset of $\mathbb{S}^{d-1}$).*

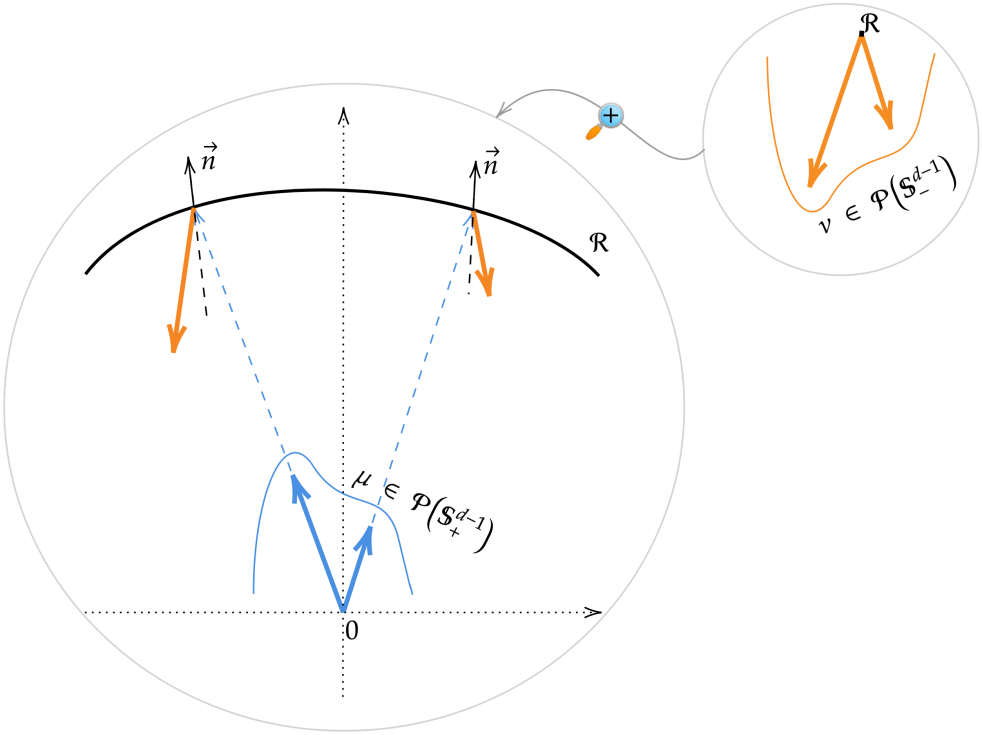


Figure 1.2: Visual description of point source to far-field problem

Assumption 2 (The far-field assumption)**.** *The far-field assumption means that the reflector $\mathcal{R}$ is so small compared to both the illumination scene and the distance to it, that it can be regarded as a point source. Therefore only the directions of the reflected rays matter, that is, the reflected light distribution ν is defined on a subset of $\mathbb{S}^{d-1}$.*

Problem 1 (The point source to far-field reflector problem)**.** *The point source to far-field reflector problem (from now on referred to as point source problem) is to find a reflector $\mathcal{R}$, that will reflect the given source distribution μ into the given desired target distribution ν, in the sense of (1.2)*

We will always restrict the supports of the distributions μ and ν to respectively $\mathbb{S}_+^{d-1}$ and $\mathbb{S}_-^{d-1}$.

Following [Wan04], we parametrize the reflector $\mathcal{R}$ using $x \in X$ a fixed subset of the upper hemisphere $\mathbb{S}_+^{d-1}$ and a given positive "radius" function $\rho \in \mathcal{C}^1(X, \mathbb{R})$. With these notations the reflector is modeled as:

$$\mathcal{R}_\rho = \{\vec{x}\rho(x) | \, x \in X \subset \mathbb{S}_+^{d-1}\} \tag{1.5}$$

Remark 1.1. *Note that in this parametrization, reflectors $\mathcal{R}_{c\rho}$ for any $c > 0$ induce the same angular distribution ν.*

In this work, we use the theory of optimal transport in order to solve the point source problem. This approach was developed in [Wan96][Wan04] and will be discussed in Chapter 3. The optimal transport theory gives access to new efficient solvers to tackle the reflector problem. We discuss those solvers in Chapter 2 and their adaptations to the reflector problem in Chapter 3.

1.3 Extended source to far-field reflector problem model

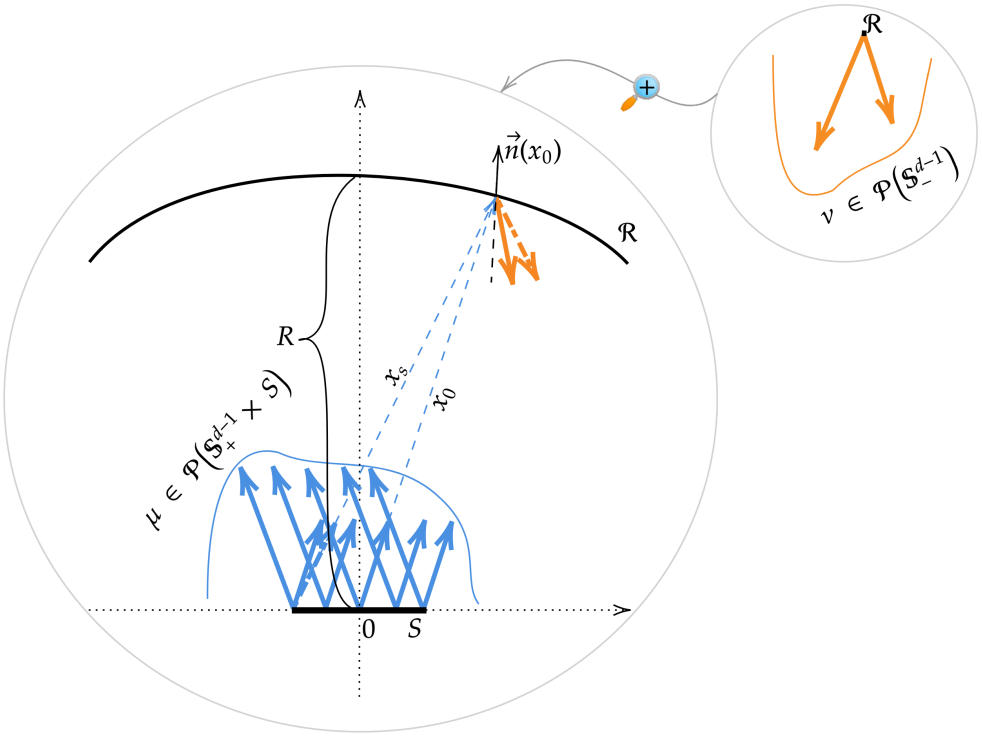


Figure 1.3: Visual description of extended source to far-field problem

The extended source problem follows the same far-field setup as the point source, but drops the point source assumption 1 on the light source. Hence, the light source is not a point, but a finite diameter subset S of $\mathbb{R}^{d-1}$, the subspace orthogonal to e_d, the d-th basis element of R^d. For convenience of notation, we assume that the origin $O := O_{\mathbb{R}^d}$ is always in S.

Each point $s \in S$ emits light in directions only, with the probability distribution μ_s supported on some subset of $\mathbb{S}_+^{d-1}$. The total light source distribution μ is defined on the product of spatial and directional spaces, that is, on $S \times \mathbb{S}_+^{d-1}$. In order to avoid the normalization constant when treating both μ and all μ_s as the probability measures ($\mu(S \times \mathbb{S}_+) = 1$ and $\mu_s(\mathbb{S}_+) = 1$), we will always take $d-1$ dimensional Lebesgue measure of S to be 1 (except of course when in the point source regime, where S contains only one point).

Note that, unlike the point source case, the reflector is not invariant under scaling (remark 1.1). When the reflector is parametrized with the angular parametrization from the origin O, different scalings of that parametrization will have different intersections with a ray originating from another source point. Therefore, it is important to quantify the distance between the reflector $\mathcal{R}$ and the source S. For this, we define a "reflector height", a value $h_{\mathcal{R}}$, corresponding to the distance between the origin O and the point on the reflector, corresponding to the vector e_d in the angular parametrization from the origin. As we use the far-field assumption 2, the reflected target distribution ν is again defined in directions only.

Problem 2 (The extended source to far-field reflector problem)**.** *The extended source to far-field reflector problem (from now on referred to as extended source problem) is to find a reflector $\mathcal{R}$, with a given height $h_{\mathcal{R}}$, that will reflect the given extended source distribution μ into the given desired target distribution ν in the sense of (1.4)*

Remark 1.2 (Ill-posedness)**.** *Existence of a reflector for the point source problem is known and will be detailed in chapter 3. The extended source problem can be understood as a collection of point source problem, one for each point in S, sharing the same unknown reflector. The target illumination constraint is unchanged but there are as many source constraints as points in S. The problem is very likely to be over-determined and therefore ill-posed except maybe for very specific data. We are not aware of a mathematical theory (or partial theory) of existence.*

1.4 Ray-tracing

Problems 1 and 2 are the inverse problems of the following "forward" problem:

Problem 3. *Given the light source distribution μ_s and a reflector $\mathcal{R}$, what is the distribution ν of the reflection from this reflector?*

Formally, the solution to this problem is to compute the push-forward of μ_s by the reflection maps T_s and then integrate over s. In practice, computing those push-forwards analytically is rarely possible (except for some very specific simple reflectors).

On the other hand, there is a simple numerical solution to this problem, known as ray-tracing. It is based on a discrete sampling of the measures, and point-wise computation of reflection maps.

Ray-tracing is a widely used technique in various optical applications. For a good review we suggest [Gla89]. Here we present two approaches to ray-tracing, designed for the extended source problem. The point source problem can be considered as a particular case in which the source set S contains only one point.

1.4.1 Forward ray-tracing

The first, referred to as "forward ray-tracing ", is commonly used in the optics community.

We assume that the $\{\mu_s\}_{s\in S}$ are given as discrete measures $\mu_s := \sum_{i=1}^{N_s} \mu_{s,i}\delta_{x_{s,i}}$, each family $\{x_{s,i}\}_{i=1..N_s}$ being a discretisation of $\mathbb{S}^1_+$. We also assume S is a discrete set and the positive coefficients $\mu_{s,i}$ are normalized:

$$\sum_{s\in S}\sum_{i=1}^{N_s} \mu_{s,i} = 1 \tag{1.6}$$

Then for a given reflector $\mathcal{R}$ and the corresponding family of the reflection maps $\{T_s\}$, the discrete reflected distribution $\bar{\nu} := \sum_s T_{s\#}\mu_s$ is given by:

$$\bar{\nu} = \sum_{s\in S}\sum_{i} \mu_{s,i}\,\delta_{y_{s,i}}, \text{ where } y_{s,i} = T_s(x_{s,i}) \tag{1.7}$$

Remark 1.3. *This form of the reflected distribution is a consequence of the measure preserving property (definition 2) of the reflection maps T_s applied to the discrete measures.*

This approach is not limited by the number of rays that can be traced, except for the computational time. Assuming enough resources, it allows to accurately approximate the continuous illumination.

The computation of $x \mapsto T_s(x)$ requires the normal n_s to $\mathcal{R}$ at the intersection between the ray traced from O_s in the direction $\vec{x}$ and the reflector.

When the source S contains only one point, this intersection is easier to find, since the reflector is parametrized using the same variable. Even when $\mathcal{R}$ is defined on a different discretization $\{x'_j\}_{j\leq M}$, it is usually possible to

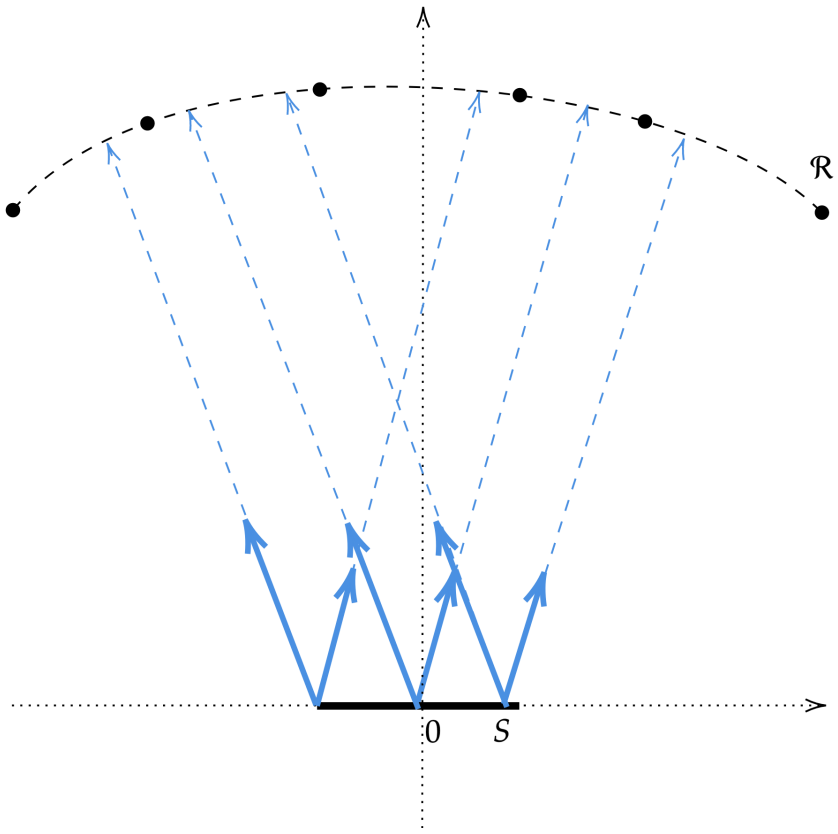


Figure 1.4: When the reflector $\mathcal{R}$ is only computed on the discretization points (black dots), one needs to find the intersection of the sampled rays (blue arrows) with the interpolation of the surface (black dashed curve). This intersection happens at different positions for the rays with the same direction, depending on the starting point on the source S

find the value of the interpolation of $\mathcal{R}$ at $\vec{x}$ and corresponding normal in "constant" number of operations $O(1)$.

On the other hand, when working with the extended source, so that S contains more points, finding the intersection becomes more complicated, since either reflector should be completely reparametrized for each source point, or a search procedure is necessary for pinpointing the intersection (see figure 1.4). Because of this, computation for a large number of rays becomes time-consuming even for the simplest, linear (bi-linear for d=3) interpolations.

Note also that the quality of the ray-tracing depends on the quality of the samplings $\{x_{s,i}\}$. It is well-known (see e.g. [Dav92]) that monte-carlo samplings have an error of order $1/\sqrt{N}$ where N is the number of samples.

To be more precise, $\int_X \mu - \frac{1}{N}\sum_{i\leq N}\mu(x_i) \sim \frac{1}{\sqrt{N}}$ for the monte-carlo samplings $\{x_i\}_{i\leq N}$. This is known as the "curse of diminishing returns", since a linear improvement of the quality of the ray-tracing requires quadratic increase in the number of points used. This can be slightly improved in various ways, by using a structured grid (see e.g. [Dav92] for the details). We will use a quasi monte-carlo grid (see e.g. [Wom17]), which will be discussed in Chapter 3.2.1.

1.4.2 Backward ray-tracing

In order to speed up the computations, we can also use another method, which we will refer to as "backward ray-tracing ".

In this approach, we do not fix the sampling of the source, but instead, we assume that we have access to a continuous analytic density for μ_s and we construct a sampling corresponding to a prescribed discrete support in angular parametrization.

In other words, we look at the discretization points on the reflector, and for each source point $s \in S$, we only consider the rays originating from this source that will intersect the reflector at the discretization points (see figure 1.5). Of course, in order to produce a good sampling of the source measure, some correction of weights will also be necessary. We outline this construction below.

The angular discretization $\{x_{0,i}\}_i$ of $\mathbb{S}^{d-1}_+$ for the O center source point is taken to be the same as the support of the discrete reflector. Normals at those points can be computed by computing the gradient of ρ (e.g by finite differences, computing the gradient of the interpolation, etc.).

Then, for all $s \in S$ we can induce a new (possibly non-uniform) discretization $\{x_{s,i}\}_i$ on $\mathbb{S}^{d-1}_+$ for the source point O_s by taking the angular representation of vectors connecting O_s to points $\{\vec{x}_{0,i}\rho(x_{0,i})\}_i$ on the discrete reflector (see figure 1.5).

In order to account for the non-uniformity of this discretization, we perform a piecewise constant approximation of the density on this grid and use the corresponding correction term $\Delta_{s,i}$. For d=2, following the idea of trapezoidal rule, this correction can be $\Delta_{s,i} := (x_{s,i+1} - x_{s,i-1})/2$. For d=3, the formula for $\Delta_{s,i}$ can be more involved, as it involves computations of the areas of quadrilaterals, but it follows the same concept.

Finally, we use the following empirical approximation

$$\mu \simeq \sum_s \sum_i \Delta_{s,i}\mu_s(x_{s,i})\delta_{x_{s,i}} \tag{1.8}$$

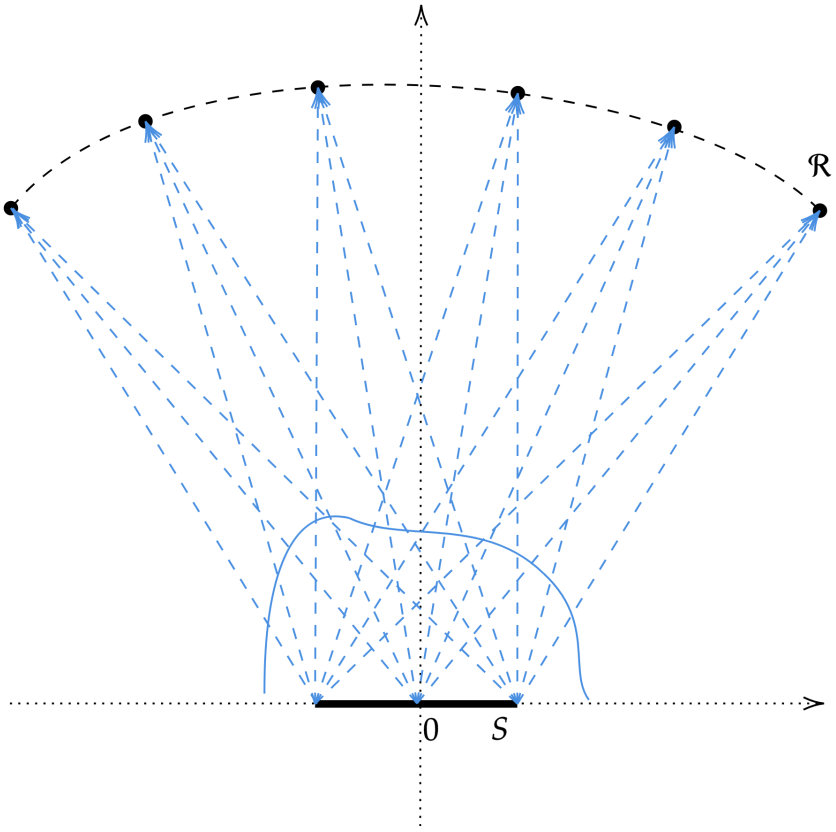


Figure 1.5: When the reflector $\mathcal{R}$ is only computed on the discretization points (black dots), one can induce a discretization on the directions (blue dashed arrows), which will be different for each point on the source S. Such construction guarantees that the intersection points will coincide with the discretization points of the reflector

Which provides the following reflected distribution:

$$\bar{\nu} = \sum_{s \in S} \sum_{i} \Delta_{s,i} \mu_s(x_{s,i})\, \delta_{y_{s,i}}, \text{ where } y_{s,i} = T_s(x_{s,i}) \tag{1.9}$$

There are various ways to improve this strategy by optimizing the weights $\Delta_{s,i}$ using different estimators. The number of rays is fixed and is the same as the discretization of the problem and there is no intersection with the reflector point to compute. Also note that the normal of $\mathcal{R}$, which is used to compute the reflection maps $\{T_s\}$, need to be computed only at the discretization points of $\mathcal{R}$ and not for each ray as in forward ray-tracing.

The drawback of this approach lies in the quality of the discretizations $\{x_{s,i}\}$, which cannot be controlled directly. However, in practice (see figure

4.8) it does not seem to be a problem.

1.4.3 The "Binning" technique

The above discussed ray-tracing methods generate discrete point cloud $\{y_{s,i} = T_s(x_{s,i})\}_{s,i}$ distributions in the angle space $\mathbb{S}_-^{d-1}$ with weights $w_{s,i}$. It could be desirable to have this distribution on a grid or another set of points denoted here $\{z_k\}$ (e.g to have the reflected distribution in the form of a pixelized picture, or because the desired target density is given on such grid: $\nu = \sum_k \nu_k \delta_{z_k}$ and one wishes to do a point-wise comparison).

To achieve this, we define "bins" $\{B_k\}$, that is, a disjoint cells covering of Y with centers z_k. For $d = 2$ and assuming the z_k are ordered, we use $B_k = [\frac{z_k+z_{k-1}}{2}, \frac{z_k+z_{k+1}}{2})$. For $d = 3$ the shape of the bins can be more complicated, but if z_k are induced by some structured grid, this structure will dictate what the shape should be.

The "binned" approximation is constructed by summing the weights of all rays falling into an each bin B_k to obtain $\bar{\nu}(z_k) := \sum_{\{i:T(x_i)\in B_k\}} w_i$.

However, the final discrete distribution $\bar{\nu}$ is usually noisy, in the sense that neighbouring bins might have different number of rays, resulting in oscillating values from bin to bin, even when the desired target is supposed to be smooth. To resolve this problem, we do a convolution with the Gaussian kernel, with $\sigma = 5/N$ (where N is the number of bins), which averages the values of neighboring bins and results in a smooth distribution. The choice $\sigma = 5/N$ was made empirically, governing the width of the smoothing window (the main contribution in convolution comes from 5 closest bins).

In figure 1.6 we demonstrate the "raw" binning and its smoothed counterpart. For this, we use a Gaussian distribution $\mathcal{N}_{\frac{3\pi}{2},\frac{\pi}{21}}$, with a mean $3\pi/2$ and deviation $\pi/21$. We obtain a point cloud sampling of this distribution (all weights are equal) with 100000 points using `pytorch`. We then bin this sampling into 1000 equal bins of the interval $[\frac{9\pi}{8}, \frac{15\pi}{8}]$.

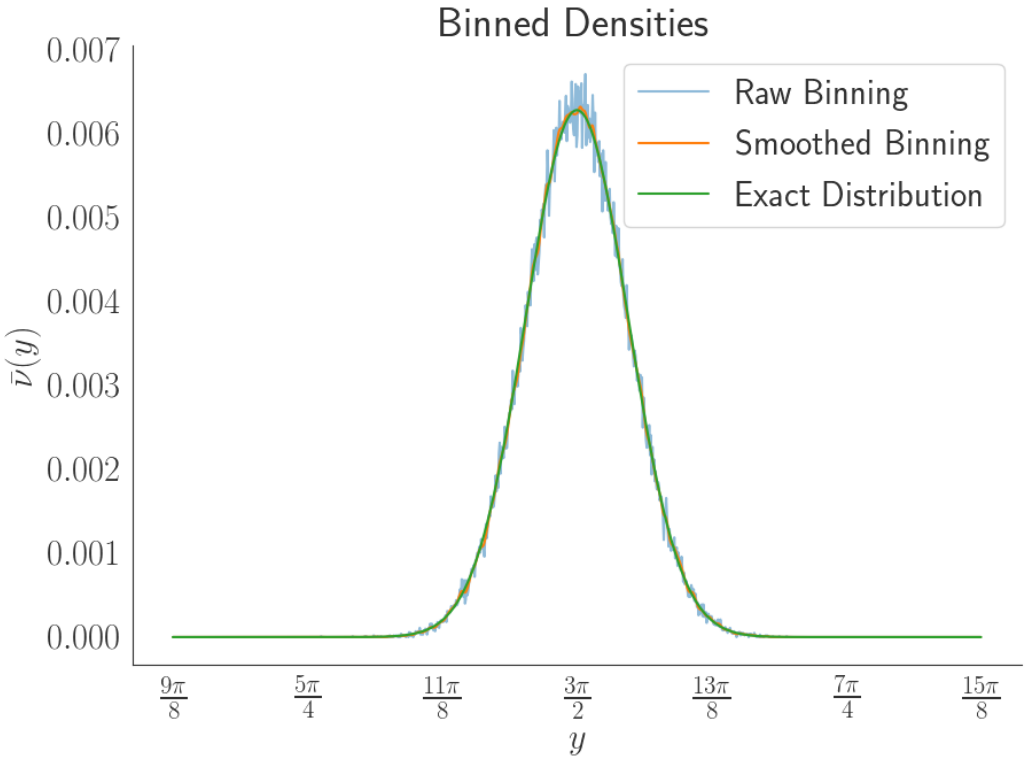


Figure 1.6: Binning of a point cloud (sampled from $\mathcal{N}_{\frac{3\pi}{2},\frac{\pi}{21}}$) with(orange) and without(blue) smoothing.

Chapter 2

Optimal transport

Introduction

In this chapter, we review the foundations of optimal transport and introduce some fundamental results which will be used later in chapters 3 and 4. We will not present the proofs, but rather focus on explaining the results and their role. First, we go through the basic concepts such as the dual formulation of optimal transport problem and the definition of Wasserstein distance. We then proceed to present the entropic regularization of optimal transport and its numerical resolution, which will play a crucial role in chapters 3-4. The material covered in this chapter can be found in [Vil08],[San15] and [PC18].

2.1 Basics from optimal transport

2.1.1 Dual formulation and stability of optimal transport

The optimal transport problem, first posed by Gaspard Monge in "Mémoire sur la théorie des déblais et des remblais" ([Mon81]), is an optimal mass allocation problem, with respect to some given cost of transporting a unit of mass. Mathematically it can be formulated as finding some "transfers" from the mass distributed according to the measure μ defined on the (possibly discrete) space X to another (possibly discrete) space Y with another mass distribution of measure ν, in a way, that the total cost of this transfer, with respect to some cost function $c(x, y)$ is minimal. Here $c(x, y)$ gives information about what it costs to transfer a unit of mass from a position $x \in X$ to $y \in Y$.

As we will apply optimal transport theory to the problems where the conservation of mass is assumed, we will always consider $\mu(X) = \nu(Y) = 1$. Intuitively this means that no mass is "lost" during the transport, and no extra mass appears at the destination. The case $\mu(X) \neq \nu(Y)$ is referred to as "Unbalanced optimal transport problem", which we will not discuss here, and instead refer the interested readers to [CPSV15] [CPSV18] [KMV16] [LMS18].

In the original formulation of Monge, the set of "transfers" was a set of measure-preserving maps.

Problem 4 (Monge Problem)**.** *Given two complete separable metric spaces with borel probability measures, (X, μ) and (Y, ν), and a continuous function $c(x, y) : X \times Y \to \mathbb{R}$ bounded from below. Find a minimizer of the following functional:*

$$\inf_T \int_X c(x, T(x))\, d\mu \tag{2.1}$$

Where the infimum is taken over all measure-preserving maps T from (X, μ) to (Y, ν)

This choice of transfers has a major drawback. First, in general, one can not guarantee that the infimum can be achieved by a map (There are various examples of such cases, for the details see e.g. [Vil08]). Even in the cases where this can be guaranteed, it is still difficult to analyze this formulation of the problem, since the set of measure-preserving maps is not closed under any reasonable convergence, appropriate to this problem.

Monge problem remained mostly unstudied till the middle of the 20th century when Leonid Kantorovich proposed a relaxed formulation of this problem in [Kan42]. Under this relaxation, the "transfers" are not maps anymore, but transport plans, that is, coupling measures γ on the product space $X \times Y$. For each couple of measurable sets $A \in X$, $B \in Y$, this measure $\gamma(A \times B)$ tells us how much mass from A should be transported to B. The measure γ should also satisfy the mass conservation property: all the mass from X should go to Y, and also, no new mass should be created. This translates into the following marginal constraints for the transport plan: For all measurable $A \in X$ and $B \in Y$, $\gamma(A \times Y) = \mu(A)$ and $\gamma(X \times B) = \nu(B)$. Clearly this implies that γ should also be a probability measure ($\gamma(X \times Y) = \mu(X) = \nu(Y) = 1$). For a given metric space Z, we will denote the set of all Borel probability measures on Z by $\mathcal{P}(Z)$.

Problem 5 (Kantorovich relaxation)**.** *Given two complete separable metric spaces with probability measures, (X, μ) and (Y, ν), and a continuous function*

$c(x, y) : X \times Y \to \mathbb{R}$ bounded from below. Find a minimizer of the following functional:

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, d\gamma \tag{2.2}$$

Where the infimum is taken over all coupling measures between μ and ν:

$$\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) | \gamma(\cdot \times Y) = \mu(\cdot), \gamma(X \times \cdot) = \nu(\cdot)\} \tag{2.3}$$

Kantorovich formulation has several advantages. First of all, $\Pi(\mu, \nu)$ is never empty, since it always contains $\mu \otimes \nu$. Second, the set of transport plans $\Pi(\mu, \nu)$ is convex and compact with respect to the weak convergence (see e.g. [Vil08],[AG09])

Definition 3 (Weak Convergence of Measures)**.** *Given a metric space X and a set of all probability measures $\mathcal{P}(X)$ on it, we say that the sequence of measures $\mu_k \in P(X)$ converge weakly to a measure $\mu \in \mathcal{P}(X)$, iff the following holds:*

$$\int_X \phi\, d\mu_k \to \int_X \phi\, d\mu \qquad \text{for all } \phi \in C_b(X) \tag{2.4}$$

Where $C_b(X)$ denotes a set of bounded continuous functions.

Remark 2.1. *It is well-known (as a Portmanteau theorem, see e.g. [Bil99]) that the weak convergence still holds if (2.4) is verified only for bounded Lipschitz functions.*

Convexity and compactness of $\Pi(\mu, \nu)$ provides existence of an optimal plan (the minimizer of the Kantorovich functional), under very mild assumptions on the spaces X, Y and the cost c. For example, if X, Y are complete separable metric spaces and c is bounded from below, then the optimal plan exists for any $\mu \in \mathcal{P}(X)$ $\nu \in \mathcal{P}(Y)$.

Also note that every measure-preserving map T can induce a transport plan $\gamma_T := (Id, T)_{\#}\mu$, a measure concentrated on the graph of T, that for the sets $A \times B$ measures the portion of $T^{-1}(B)$ contained in A:

$$\gamma_T(A \times B) = \mu((Id, T)^{-1}(A \times B)) = \mu(A \cap T^{-1}(B))$$

Hence it is clear, that the minimal value of Kantorovich functional can not be larger than the value of Monge functional. Conversely, however, the minimizer transport plan γ might not be induced by a map. But the value obtained by this minimizer is actually equal to the infimum of the Monge functional, under mild assumptions on the spaces and the cost function. More precisely, the following theorem holds (even if the optimal map T does not exist):

Theorem 2.2 (see e.g. [Pra07]). *Given two complete separable metric spaces with probability measures, (X, μ) and (Y, ν), and a continuous function $c(x, y) : X \times Y \to \mathbb{R}$ bounded from below. Assume that the measure μ has no atoms, then*

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{X\times Y} c(x, y)\, d\gamma = \inf_{T} \int_{X} c(x, T(x))\, d\mu \tag{2.5}$$

Note that the requirement for the cost c to be bounded from below is always satisfied when working with compact spaces X and Y, as we take c to be continuous.

The above-mentioned advantages allow a far larger set of mathematical tools to be applied to the relaxed problem. In particular, Kantorovich proved the duality-type theorem from linear programming for the relaxed problem.

Notation 2. *From now on, in order to keep a unified representation for the discrete and continuous cases, we will switch from integral notation to a "duality product" notation: $\langle f, \mu \rangle_X := \int_X f d\mu$ denoting the duality product between continuous functions and probability measures over X. This notation will also allow us to denote measures and their densities with the same letter, avoiding the further complications of the notations.*

Theorem 2.3 (Kantorovich Duality [Kan42]). *Given two complete separable metric spaces with probability measures, (X, μ) and (Y, ν), and a continuous function $c(x, y) : X \times Y \to \mathbb{R}$ bounded from below. Then the following duality holds*

$$OT(\mu, \nu) := \min_{\gamma \in \Pi(\mu,\nu)} \langle c, \gamma \rangle_{X\times Y} = \sup_{(f,g)\in\mathcal{D}} \langle f, \mu \rangle_X + \langle g, \nu \rangle_Y \tag{2.6}$$

Where the dual constraint set is:

$$\mathcal{D} := \{(f, g) \in C(X) \times C(Y) | f(x) + g(y) \leq c(x, y)\} \tag{2.7}$$

The solutions of the dual problem (f, g) are called Kantorovich potentials, and they play a crucial role in the optimal transport theory, as well as in this work. For any (x, y) in the support of an optimal plan γ_{opt}, f and g saturate the dual constraint, that is, $f(x) + g(y) = c(x, y)$ for all $(x, y) \in supp(\gamma_{opt})$. Moreover, they are related to each other by the relation known as c-transformation: $f = g^c$ and $g = f^c$.

Definition 4 (c-transform). *Given spaces X and Y, and a continuous function $c(x, y) : X \times Y \to \mathbb{R}$ bounded from below, c-transform of the function $f : X \to \mathbb{R}$ is given by:*

$$f^c(y) = \inf_{x\in X} c(x, y) - f(x) \tag{2.8}$$

And similarly, for the function $g : Y \to \mathbb{R}$, the c-transform is given by:

$$g^c(x) = \inf_{y \in Y} c(x, y) - g(y) \tag{2.9}$$

Functions that can be expressed as a c-transform of some function, are known as c-concave functions.

Assuming that X, Y are smooth manifolds, support of ν is connected, μ does not give mass to the sets with Hausdorff dimension less or equal to $d-1$ (where d is the dimension of X) and the cost function $c(x, y)$ satisfies the regularity assumptions A1-A3 (see appendix A), the solution for the Kantorovich problem γ is actually concentrated on an optimal map T, which is a solution of the Monge problem, $\gamma = (Id, T)_{\#}\mu$, and T saturates the dual constraint for all $x \in supp(\mu)$ (see e.g. [Loe09]):

$$f(x) + g(T(x)) = c(x, T(x)) \tag{2.10}$$

Also, T can be computed using the potential f by the following relation:

$$T : x \to y := \{y \to \nabla_x c(x, y)\}^{-1} (\nabla_x f(x)) \tag{2.11}$$

The necessary regularity assumptions mentioned above, are technical and always satisfied for all the applications presented through this work. Therefore we proceed without discussing them here and postpone their presentation to the appendix A.

One useful property of the optimal transport is that it is "stable" with respect to the changes in marginals or the cost. This means that if we approximate the marginals using discretizations, or we don't know exact marginals or exact values of the cost for the given real-world problem and hence we use approximations, we can still obtain solutions that are close to the solution of the original problem.

Theorem 2.4 (Stability of optimal transport plans (see e.g [Vil08]))**.** *Given two complete separable metric spaces with probability measures, (X, μ) and (Y, ν), and a continuous function $c(x, y) : X \times Y \to \mathbb{R}$ bounded from below. Let $\{c_k\}_{k \in \mathbb{N}}$ be a sequence of continuous cost functions converging uniformly to c, $\{\mu_k\}_{k \in \mathbb{N}}$ and $\{\nu_k\}_{k \in \mathbb{N}}$ converging weakly to respectively μ and ν, and for each k let γ_k be an optimal transport plan between μ_k and ν_k with respect to the cost c_k.*

If

$$\forall k \in \mathbb{N} \qquad \langle c_k, \gamma_k \rangle_{X \times Y} < +\infty \qquad \textit{and} \qquad \liminf_{k \in \mathbb{N}} \langle c_k, \gamma_k \rangle_{X \times Y} < +\infty$$

Then γ_k converges weakly (up to the extraction of a subsequence) towards a plan γ, the total cost $\langle c, \gamma \rangle_{X \times Y}$ is finite and γ is an optimal plan between μ and ν

This result also translates into the stability of the Kantorovich potentials, where the uniform convergence can be established (again, up to the extraction of a subsequence, if the limiting optimal plan is not unique).

Theorem 2.5 (Stability of Kantorovich potentials (see e.g [San15]))**.** *Given two complete separable metric spaces with probability measures, (X, μ) and (Y, ν), and a continuous function $c(x, y) : X \times Y \to \mathbb{R}$ bounded from below. Let $\{c_k\}_{k \in \mathbb{N}}$ be a sequence of continuous cost functions converging uniformly to c, $\{\mu_k\}_{k \in \mathbb{N}}$ and $\{\nu_k\}_{k \in \mathbb{N}}$ converging weakly to respectively μ and ν, and for each k let (f_k, g_k) be pair of (normalized by $f(x_0) = 0$ for some fixed $x_0 \in X$) Kantorovich potentials for the optimal transport problem between μ_k and ν_k with respect to the cost c_k.*

If

$$\forall k \in \mathbb{N} \qquad \langle c_k, \gamma_k \rangle_{X \times Y} < +\infty \qquad \text{and} \qquad \liminf_{k \in \mathbb{N}} \langle c_k, \gamma_k \rangle_{X \times Y} < +\infty$$

Then (f_k, g_k) converges uniformly (up to extraction of a subsequence) towards a pair of Kantorovich potentials (f, g), for the optimal transport problem between μ and ν with a cost c.

It is worth noting that the similar stability result for optimal maps is more restrictive in general. It requires the source measure μ to be stable and to assume the existence of optimal transport map in the limiting case.

Theorem 2.6 (Stability of optimal transport maps (see e.g [Vil08]))**.** *Given two complete separable spaces with probability measures, (X, μ) locally compact and (Y, ν) general, and a continuous function $c(x, y) : X \times Y \to \mathbb{R}$ bounded from below. Let $\{c_k\}_{k \in \mathbb{N}}$ be a sequence of continuous cost functions converging uniformly to c, $\{\nu_k\}_{k \in \mathbb{N}}$ converging weakly to ν.*

Assume that for each k there exists a T_k an optimal transport map between μ and ν_k with respect to the cost c_k such that total cost is finite. Also, assume that there exists an optimal map T from μ to ν with respect to c and the total cost is finite.

Then T_k converges to T in probability:

$$\forall \epsilon > 0 \qquad \mu\left(\{x \in X | d_X(T_k(x), T(x)) \geq \epsilon\}\right) \xrightarrow[k \to \infty]{} 0 \tag{2.12}$$

On one hand, those stability results are a useful tool for studying the optimal transport problem and finding approximate solutions. On the other hand, it also allows to study (2.6) as a functional on the set of probability measures.

2.1.2 $\mu \mapsto OT(\mu, \nu)$ and $\nu \mapsto OT(\mu, \nu)$ functionals

In order to avoid extra technicalities, through this subsection X and Y will be compact subsets of $\mathbb{R}^d$. We consider the functionals $\mu \mapsto OT(\mu, \nu)$ for some fixed $\nu \in \mathcal{P}(Y)$ and $\nu \mapsto OT(\mu, \nu)$ for some fixed $\mu \in \mathcal{P}(X)$, defined respectively on $\mathcal{P}(X)$ and $\mathcal{P}(Y)$, and discuss their continuity, convexity and differentiability.

The continuity of those functionals with respect to the weak convergence of measures can be deduced using the stability results and the classical compactness arguments. Moreover, compactness is essential for continuity, and without this assumption, in general, one can only hope for lower semi-continuity.

Theorem 2.7 (Continuity of $\mu \mapsto OT(\mu, \nu)$, $\nu \mapsto OT(\mu, \nu)$ (see e.g [San15]))**.** *Given X and Y compact, a continuous cost function $c : X \times Y \to R$ and a fixed distribution $\nu \in \mathcal{P}(Y)$ (resp. $\mu \in \mathcal{P}(X)$), the functional $\mu \mapsto OT(\mu, \nu)$ (resp. $\nu \mapsto OT(\mu, \nu)$) is continuous with respect to the weak convergence of measures in $\mathcal{P}(X)$ (resp. $\mathcal{P}(Y)$).*

If X and Y are not compact, then in general only lower semi-continuity holds, and in this case, c can also be assumed to be only lower semi-continuous.

The convexity of those functionals is a direct consequence of the Kantorovich duality (2.6), as they are expressed as a supremum of linear functions. However a-priory one does not have a strict convexity, which holds only for some special cases. In contrast, the entropic regularization of optimal transport problem, that we discuss in section 2.2.1, is always strictly convex.

In order to discuss differentiability, one should first define what it means for a functional on $\mathcal{P}(X)$ to be differentiable, or in other words, define its first variation.

Definition 5. *Given a functional $F : \mathcal{P}(X) \to \mathbb{R}$, we denote by $\nabla_\mu F(\mu)$ (if it exists) a measurable function, such that for any $\chi := \mu_1 - \mu$ with $\mu_1 \in \mathcal{P}(X)$, the following equality holds*

$$\lim_{\epsilon \to 0} \frac{F(\mu + \epsilon\chi) - F(\mu)}{\epsilon} = \langle \nabla_\mu F(\mu), \chi \rangle_X \tag{2.13}$$

Note that since χ is a difference of probability measures, its total mass is 0, hence $\langle C, \chi \rangle_X = 0$ for any constant C and therefore, $\nabla_\mu F(\mu)$ is defined only up to an additive constant.

Theorem 2.8 (Differentiability of $\mu \mapsto OT(\mu, \nu)$, $\nu \mapsto OT(\mu, \nu)$ (see e.g [San15]))**.** *Given X and Y compact, a continuous cost function $c : X \times Y \to R$*

and a fixed distribution $\nu \in \mathcal{P}(Y)$, the functional $\mu \mapsto OT(\mu, \nu)$ is convex and its subdifferential at some $\mu_0 \in \mathcal{P}(X)$ is given by the set of Kantorovich potentials of the optimal transport problem:

$$\left\{ f \in C^0(X) | \langle f, \mu_0 \rangle_X + \langle f^c, \nu \rangle_Y = OT(\mu_0, \nu) \right\} \tag{2.14}$$

Moreover, if the solution of optimal transport problem is unique, or equivalently, the Kantorovich potential is also unique up to an additive constant, then the gradient"(first variation) of $\mu \mapsto OT(\mu, \nu)$ exists at μ_0, and is given by f:

$$\nabla_\mu OT(\mu_0, \nu) = f \tag{2.15}$$

Same result holds for $\nu \mapsto OT(\mu, \nu)$, with the second Kantorovich potential g:

$$\nabla_\nu OT(\mu, \nu_0) = g \quad for \quad \langle g^c, \mu \rangle_X + \langle g, \nu_0 \rangle_Y = OT(\mu, \nu_0) \tag{2.16}$$

2.1.3 Wasserstein Metrics

One useful application of optimal transport is the case where the spaces X and Y are the same complete separable metric space with a metric d and the cost function $c(x, y)$ is some p-th $(1 \leq p < \infty)$power of that metric: $c(x, y) = \frac{1}{p} d^p(x, y)$. In this setting, optimal transport can define a metric on the space of probability measures $\mathcal{P}(X)$ on X with a finite p-th moments:

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) | \int_X d^p(x, x_0)\, d\mu < \infty \right\} \quad \text{for some } x_0 \in X \tag{2.17}$$

Definition 6 (Wasserstein Metric)**.** *Given a complete separable metric space X with a metric d, for a given $1 \leq p < \infty$, one can define a metric on $\mathcal{P}_p(X)$ using the following optimal transport Problem:*

$$W_p(\mu, \nu) := \left(\min_{\gamma \in \Pi(\mu, \nu)} \left\langle \frac{1}{p} d^p, \gamma \right\rangle_{X \times X} \right)^{\frac{1}{p}} = (OT(\mu, \nu))^{\frac{1}{p}} \tag{2.18}$$

Theorem 2.9 (see e.g [Vil08])**.** *For a complete separable metric space X with a bounded metric d, W_p for any $1 \leq p < \infty$ defines a metric on the space $\mathcal{P}_p(X)$, which metrizes the weak convergence.*

Remark 2.10. *The requirement of a bounded metric can be relaxed by several means. One way is to require the existence of another bounded metric $\tilde{d}$ inducing the same topology (e.g. $\tilde{d} = d/(d+1)$). Another is to require weakly convergent measures, to also maintain convergence of p-th moments.*

As we plan to use the Wasserstein metric in the framework of the geometric optics, and as in this case the reflection maps are measure-preserving maps, it is worth to note that this metric allows estimating the distance between push-forward measures.

Lemma 2.11. *Given two Polish spaces (X, d_Y) and (Y, d_Y), with Borel probability measures μ_1, μ_2 on X and Lipschitz continuous map $T : X \to Y$ with Lipschitz constant Lip(T), Then*

$$W_p(T_{\#}\mu_1, T_{\#}\mu_2) \leq Lip(T) W_p(\mu_1, \mu_2) \tag{2.19}$$

This lemma is not hard to prove and most likely many experts are aware that it holds, but as we could not find any references in the literature, we include the proof in the appendix B.

Remark 2.12. *The above lemma can be applied to the Monge maps $T : X \to Y$ under mild regularity assumptions: When the target support Y is convex and the source and target densities μ and ν smooth enough ($C^{1,\alpha}$ is sufficient) and bounded below and above by positive constants, a classic regularity result by Caffarelli [Caf92] gives sufficient regularity to get*

$$Lip(T) \leq K \tag{2.20}$$

where the constant K only depends on the dimension d and the data μ and ν. See [PF14] for further refinements and discussions.

2.2 Entropic optimal transport and Sinkhorn algorithm

2.2.1 Entropic regularization of optimal transport

The entropic regularization approach was introduced for optimal transport computations in [Cut13] (see [PC18] for a comprehensive review). Regularization of the Kantorovich problem (2.6) is based on the following KullBack-Leibler divergence or "relative entropy" (KL) penalization :

$$\begin{aligned} OT_\epsilon(\mu, \nu) := \min_{\gamma_\epsilon \in \Pi(\mu,\nu)} \langle c, \gamma_\epsilon \rangle_{X\times Y} + \epsilon \, \mathrm{KL}(\gamma_\epsilon \,|\, \mu \otimes \nu) = \\ \max_{f_\epsilon, g_\epsilon} \langle f_\epsilon, \mu \rangle_X + \langle g_\epsilon, \nu \rangle_Y - \epsilon \left\langle \exp\left(\frac{1}{\epsilon}(f_\epsilon \oplus g_\epsilon - c)\right) - 1, \mu \otimes \nu \right\rangle_{X\times Y} \end{aligned} \tag{2.21}$$

where $\epsilon > 0$ is a small regularization parameter and

$$\mathrm{KL}(\gamma \,|\, \mu \otimes \nu) := \left\langle \log\left(\frac{\gamma}{\mu \otimes \nu}\right), \gamma \right\rangle_{X \times Y} + \langle 1, \mu \otimes \nu \rangle_{X \times Y} - \langle 1, \gamma \rangle_{X \times Y} \tag{2.22}$$

$$\text{if } \gamma << \mu \otimes \nu \text{ and } +\infty \text{ otherwise.}$$

The primal-dual optimality condition is given by

$$\gamma_\epsilon = \exp\left(\frac{1}{\epsilon}(f_\epsilon \oplus g_\epsilon - c)\right) \mu \otimes \nu. \tag{2.23}$$

The optimal entropic plan is therefore a "scaling" of a fixed kernel $\exp(-\frac{1}{\epsilon}c)$ by the regularized Kantorovich potentials f_ϵ and g_ϵ. Note that, as discussed in section 2.1, the optimal plan γ of an unregularized problem (2.6) is "sparse" in the sense that it is supported on the graph of a monge map T (or eqivalently, on the points (x, y) that saturate the dual constraint: $f(x) + g(y) = c(x, y)$). In contrast, γ_ϵ is diffuse, that is, it is supported on the whole $supp(\mu) \times supp(\nu)$. However, note that when ϵ goes to 0, the values $\exp\left(\frac{1}{\epsilon}(f_\epsilon \oplus g_\epsilon - c)\right)$ become exponentially small for the points (x, y) for which $f_\epsilon(x) + g_\epsilon(x)$ is strictly smaller then $c(x, y)$.

Entropic regularization adds substantial regularity to the optimal transport problem. In particular, the functional $\langle c, \gamma \rangle_{X \times Y} + \epsilon \,\mathrm{KL}(\gamma \,|\, \mu \otimes \nu)$ is strictly convex with respect to γ for any $\epsilon > 0$, which results in the existence of unique minimizer γ_ϵ (see e.g [PC18]).

The differentiability of $\mu \to OT_\epsilon(\mu, \nu)$ and $\nu \to OT_\epsilon(\mu, \nu)$ follow the same path as for the unregularized functional $OT(\mu, \nu)$.

Theorem 2.13 ([FSV+18])**.** *Under the assumptions of theorem 2.8, the first variations of $\mu \to OT_\epsilon(\mu, \nu)$ and $\nu \to OT_\epsilon(\mu, \nu)$ at respectively μ_0 and ν_0 are given by:*

$$\nabla_\mu OT_\epsilon(\mu_0, \nu) = f_\epsilon \qquad \nabla_\nu OT_\epsilon(\mu, \nu_0) = g_\epsilon \tag{2.24}$$

Where f_ϵ and g_ϵ are corresponding first and second regularized Kantorovich potentials.

Moreover, this regularization adds extra smoothness to the discrete problem as well. In fact, in [LRPC18] it is demonstrated that the discrete version of the regularized functional is C^∞ smooth with respect to the input data (the discretization vectors of the source and target measures).

Theorem 2.14 ([LRPC18])**.** *Let Δ_n denote a (probability) simplex $\Delta_n := \{v \in \mathbb{R}^n \,|\, v_i > 0, \sum_{i=1}^n v_i = 1\}$. One can identify the measure on the discrete space with N points, with an element of Δ_N. Under this identification, the functional $OT_\epsilon(\mu^N, \nu^M) : \Delta_N \times \Delta_M \to \mathbb{R}$ is C^∞ smooth in the interior of its domain.*

2.2.2 Sinkhorn Algorithm

Numerical solutions of the entropic optimal transport problem are produced under the discretization of this problem, replacing $(X,\, Y\,, c\,, \mu,\, \nu)$ by $(X_N,\, Y_N\,, c^N\,, \mu^N\, \nu^N)$:

$X_N = \{x_i\}_{i\leq N}$, $Y_N = \{y_j\}_{j\leq N}$, $c^N := \{c(x_i, y_j)\}_{i,j\leq N}$ and

$$\mu^N = \sum_{i=1}^{N} \mu^N(x_i)\delta_{x_i}, \quad \nu^N = \sum_{j=1}^{N} \nu^N(y_j)\delta_{y_j}, \quad \sum_{i=1}^{N} \mu^N(x_i) = \sum_{j=1}^{N} \nu^N(y_j) = 1 \tag{2.25}$$

Using this discretization we get the following discrete non-linear system of optimality:

$$\begin{aligned} OT_{\epsilon,N} := &\max_{f_\epsilon, g_\epsilon} \left\langle f_\epsilon, \mu^N \right\rangle_{X_N} + \left\langle g_\epsilon, \nu^N \right\rangle_{Y_N} \\ &- \epsilon \left\langle \exp\left(\frac{1}{\epsilon}(f_\epsilon \oplus g_\epsilon - c^N)\right) - 1, \mu^N \otimes \nu^N \right\rangle_{X_N \times Y_N}. \end{aligned} \tag{2.26}$$

where we use the same notation (f_ϵ, g_ϵ) for discrete vectors in $\mathbb{R}^N$.

This system is solvable using Sinkhorn algorithm. It corresponds to a block coordinate (f_ϵ and g_ϵ) ascent :

Initialize with $g_\epsilon^0 = 0_Y$ and then iterate (in k):

$$\begin{aligned} f_\epsilon^{k+1} &= -\,\epsilon \log\left(\left\langle \exp\left(\frac{1}{\epsilon}(g_\epsilon^k - c^N)\right), \nu^N \right\rangle_{Y_N}\right) \\ g_\epsilon^{k+1} &= -\,\epsilon \log\left(\left\langle \exp\left(\frac{1}{\epsilon}(f_\epsilon^{k+1} - c^N)\right), \mu^N \right\rangle_{X_N}\right) \end{aligned} \tag{2.27}$$

For a fixed ϵ, those iterations converge in k towards a solution (f_ϵ, g_ϵ) of the regularized problem $OT_{\epsilon,N}$. The rate of this convergence is linear for simple costs and asymptotic to $1-\epsilon$ (see [PC18]) again), meaning that the following estimate holds for large enough k: $||f_\epsilon^k - f_\epsilon|| = O(1-\epsilon)^k$.

The exponential convergence of the value $OT_\epsilon(\mu, \nu)$ toward $OT(\mu, \nu)$ when $\epsilon \to 0$ is established and studied in the continuous [Lé13] and discrete setting [CM94].

The numerical stability of Sinkhorn iterations depends on the transport scale of the data $\tau = \sup_{x\in X} c(x, T(x))$ and ϵ. The variable memory in computers overflows or underflows when the ratio τ/ϵ is too large. Decreasing

ϵ below a certain threshold is, therefore, a numerical difficulty. A good review of existing hacks and methods to mitigate this problem can be found in [Sch16] and we re-discuss this in section 2.2.4.

The potentials computed using Sinkhorn iterations (2.27) are defined on the discrete sets X_N and Y_N but they admit a canonical extension on the whole space by replacing $c^N(x_i, y_j)$ respectively by $c(x, y_j), x \in X$ and $c(x_i, y), y \in Y$. Omitting the iteration index :

$$f[g_\epsilon](x) = -\epsilon \log \left(\sum_{j=1\dots N} \exp\left(\frac{1}{\epsilon}(g_\epsilon(y_j) - c(x, y_j))\right) \nu^N(y_j) \right), \quad \forall x \in X. \tag{2.28}$$

We will be interested in the convergence of f_ϵ as $\epsilon \to 0$ and $N \to \infty$. To the best of our knowledge, the joint convergence in N and ϵ has only been studied in [Ber17], we reproduce partially his results :

Theorem 2.15 (Berman joint convergence [Ber17])**.** *Assume μ and ν have $\mathcal{C}^{2,\alpha}$ and positive densities, and that N and ϵ are dependent parameters : $N \approx (1/\epsilon)^d$ where d is the dimension of the problem. A technical condition on the sequence of discretization $(X_N, Y_N, c^N, \mu^N \nu^N)$ called "density property" (see Remark 2.16 below) is also necessary. Then there exists a positive constant A_0 such that for any $A > A_0$ the folowing holds : setting $m_\epsilon = [-A\log(\epsilon)/\epsilon]$ the continuous interpolation provided by $f[g_\epsilon^{m_\epsilon}]$, built using the cannonical extension (2.28) from the discrete Sinkhorn iterate at $k = m_\epsilon$, satisfies the estimate*

$$\sup_X |f[g_\epsilon^{m_\epsilon}] - f| \leq -C\epsilon\log(\epsilon), \tag{2.29}$$

for some constant C (depending on A) and f is an optimal potential for (2.6).

Remark 2.16 (Density property Lemma 3.1 [Ber17])**.** *The "density property" in theorem 2.15 is defined as follows. For any given open set U intersecting the support X of μ (same for Y and ν)*

$$\liminf_{\epsilon \to 0} \epsilon \log(\mu^N(U)) = 0.$$

For the flat space $X \subset \mathbb{R}^d$, this condition is enough. For curved surfaces, a technical generalization is required. But in both cases, this density property ensures that the discretization of X and μ (2.25) is such that, for any $U \subset X$ the sequence of approximations $\mu^N(U)$ never converges faster to 0 than ϵ (remember that $N \approx (1/\epsilon)^d$). For $X = \mathbb{R}^d$, uniform grids are fine. For curved spaces, extra precautions need to be made to make sure that this property is satisfied. We postpone the presentation of the general form of this property to the appendix C.

2.2.3 Entropic bias and Sinkhorn Divergence

From now on, the potentials computed using Sinkhorn iterations (2.27) are denoted by f_{OT_ϵ} and g_{OT_ϵ}.

Entropic optimal transport is popular to approximate transport distances between probability measures, since, as already mentioned, the Entropic cost $OT_\epsilon(\mu,\nu)$ is known to converge exponentially fast to $OT(\mu,\nu) = OT_0(\mu,\nu)$ with ϵ [CM94].

However, in many applications, the transport plan and the Kantorovich potentials are more important than the cost value itself. The error estimate for the potentials, (2.29) in theorem 2.15 has an infinite slope at $\epsilon = 0$. And due to the numerical stability limit, imposed when decreasing ϵ in (2.27), there is a substantial problem when trying to approximate the potentials with high precision.

For the case where $X = Y$ and the cost $c(x,y) = \frac{1}{p}d^p(x,y)$, this problem is discussed in [FSV$^+$18] where it is proposed, in order to correct the bias without decreasing ϵ under the numerical stability limit, to subtract the "diagonal terms" to correct the entropic cost :

$$S_\epsilon(\mu,\nu) = OT_\epsilon(\mu,\nu) - \frac{1}{2}\Big(OT_\epsilon(\mu,\mu) + OT_\epsilon(\nu,\nu)\Big). \tag{2.30}$$

Quite remarkably, the authors show that this quantity, called Sinkhorn divergence, remains positive and is strictly convex. It also obviously vanishes for $\mu = \nu$ which is not the case for OT_ϵ. Thanks to the symmetry, there is only one dual potential for each of the diagonal problems. We denote them f^μ_ϵ and f^ν_ϵ. They can be computed using independent Sinkhorn iterations :

$$\begin{aligned} f^{\mu,k+1}_{OT_\epsilon} &= -\epsilon \log\left(\left\langle \exp\left(\frac{1}{\epsilon}(f^{\mu,k}_{OT_\epsilon} - c^N)\right), \mu\right\rangle_X\right) \\ f^{\nu,k+1}_{OT_\epsilon} &= -\epsilon \log\left(\left\langle \exp\left(\frac{1}{\epsilon}(f^{\nu,k}_{OT_\epsilon} - c^N)\right), \nu\right\rangle_Y\right). \end{aligned} \tag{2.31}$$

The μ gradient of S_ϵ, denoted by f_{S_ϵ} may be formed by a simple subtraction. An open question is whether this correction :

$$f_{S_\epsilon} = f_{OT_\epsilon} - f^\mu_{OT_\epsilon}, \tag{2.32}$$

is a better approximation to f than f_{OT_ϵ}. Numerical simulations of W_2 gradient flows in [FSV$^+$18] do indicate this is the case.

Note finally that solving S_ϵ has the same complexity as OT_ϵ as the diagonal problems typically converge faster in terms of Sinkhorn iterations ([FSV$^+$18]).

Remark 2.17 (Asymptotics of OT_ϵ)**.** *The entropic bias may be related formally to asymptotic results on the difference between OT and OT_ϵ, see [CT19] [Pal19] for recent publications on this subject.*

Here we give the formula for the d^2 cost on a smooth, connected, and closed d-dimensional Riemannian manifold $X = Y$, with a volume measure vol re-scaled to be a probability measure. [CT19] gives result for more general reference measures $m = e^u vol$ and theorem 1 in [Pal19] also treats more costs in the form $c := g(x - y)$, g convex.

If μ and ν have smooth densities ρ^0 and ρ^1 with respect to vol, then the following asymptotic behavior for small ϵ holds :

$$OT_\epsilon(\mu,\nu) - OT(\mu,\nu) = d\epsilon \log\left(\sqrt{2\pi\epsilon}\right) + \frac{\epsilon}{2}\left(KL(\mu|vol) + KL(\nu,vol)\right) + \frac{\epsilon^2}{8} I(\mu,\nu) + O(\epsilon^2). \tag{2.33}$$

Where $I(\mu,\nu)$ is a certain "energy" term involving the W_2 geodesic $\rho_{t\#}vol$ connecting μ to ν.

$$I(\mu,\nu) := \int_X \int_0^1 |\nabla log(\rho_t)|^2 \rho_t \, dt \, vol \tag{2.34}$$

Using (2.33) and that $OT(\mu,\mu) = OT(\nu,\nu) = 0$, $I(\mu,\mu) = I(\nu,\nu) = 0$ we first note that the Sinkhorn Divergence correction removes at least the leading terms :

$$\frac{1}{2}\left(OT_\epsilon(\mu,\mu) + OT_\epsilon(\nu,\nu)\right) = d\epsilon \log\left(\sqrt{2\pi\epsilon}\right) + \frac{\epsilon}{2}\left(KL(\mu|vol) + KL(\nu,vol)\right) + O(\epsilon^2) \tag{2.35}$$

Formally taking the gradient in μ of equation (2.33) (except for the terms of order ϵ^2) we get for small ϵ :

$$f_{OT_\epsilon} \simeq f - \frac{\epsilon}{2}\log(\mu). \tag{2.36}$$

It gives some indication at the leading order of the entropic bias in the potential.

Finally, note that the density property discussed in Remark 2.16 requires for the first order bias term to converge to 0 when discrete mesures μ^N converge to original measure μ.

2.2.4 Computational efficiency of the Sinkhorn algorithm

The simplest implementation of Sinkhorn algorithm requires $O(N^2)$ operations per iteration. If we use the relation between discretization and entropic parameters in theorem 2.15, we need at least $O(N^{\frac{1}{d}}\log(N^{\frac{1}{d}}))$ iterations to reach $O(\epsilon\, log(\epsilon))$ precision for the Kantorovich potential. This takes us to a pessimistic $O(N^{\frac{2d+1}{d}}\log(N^{\frac{1}{d}}))$, far from the optimal complexity of semi-discrete optimal transport solvers for example.

This can be largely improved in practice by using a multi-scale method in ϵ and N. In the case of small ϵ, the limit of the entropic plan concentrates on the graph of the optimal transport map. The bandwidth of (2.23) decreases with ϵ and coarser scales in this parameter can be used to restrict the relevant support on which to perform the summation in the Sinkhorn iterates. This approach has been proposed and tested in [Sch16] and [Fey19] and experimentally yields a $O(N\, logN)$ operations cost. Here we present those concepts, which are used later in simulations of chapters 3 and 4.

ϵ scaling

As discussed in section 2.2.2, decreasing ϵ would result in a more accurate solution for (2.6). On the other hand, the convergence rate $1-\epsilon$ suggests that smaller ϵ we take, higher number of iterations will be required for the Sinkhorn algorithm to converge. Also, taking ϵ too small would result in numerical overflows due to the exponential terms of order $e^{\frac{1}{\epsilon}}$ in (2.27)

As discussed in [Sch16], the problem of numerical stability can be tackled by working with the increments of the potentials rather than full potentials during the iterative steps.

That is, if we look at the updates f_ϵ^{k+1} and g_ϵ^{k+1} in (2.27) as $f_\epsilon^{k+1} = f_\epsilon^k + \bar{f}_\epsilon^{k+1}$ and $g_\epsilon^{k+1} = g_\epsilon^k + \bar{g}_\epsilon^{k+1}$, then by moving previous approximations to the right hand side, we will get the following new iterative scheme for the increments:

$$
\begin{aligned}
\bar{f}_\epsilon^{k+1} &= -\epsilon \log\left(\left\langle \exp\left(\frac{1}{\epsilon}(g_\epsilon^k + f_\epsilon^k - c^N)\right), \nu^N\right\rangle_{Y_N}\right) \\
f_\epsilon^{k+1} &= f_\epsilon^k + \bar{f}_\epsilon^{k+1} \\
\bar{g}_\epsilon^{k+1} &= -\epsilon \log\left(\left\langle \exp\left(\frac{1}{\epsilon}(f_\epsilon^{k+1} + g_\epsilon^k - c^N)\right), \mu^N\right\rangle_{X_N}\right) \\
g_\epsilon^{k+1} &= g_\epsilon^k + \bar{g}_\epsilon^{k+1}
\end{aligned}
\tag{2.37}
$$

Those iterations will be more stable due to the saturation property of the optimizing potentials. This property tells us that quantity $f(x_i) + g(y_j) - c(x_i, y_j)$ is zero for exact potentials and optimal pairs (x_i, y_j) while being strictly negative for non-optimal pairs. Therefore, when the iterates f_ϵ^k and g_ϵ^k are close to the true potentials, new updating steps would not cause a numerical overflow.

However, this approach alone would not help at the first steps of the algorithm, since we have no guarantees that initial approximations would be close to the exact potentials, and for small ϵ we could get an overflow at the first iteration. In order to avoid this, a possible approach would be to start with higher values of ϵ and gradually decrease it to the desired final value ϵ_{final} (see [Sch16] [OR15]).

More formally, one can define a sequence of regularization parameters $\epsilon_k \to \epsilon_{final}$ and use ϵ_k at k-th iteration in (2.37). A common choice is to start with $\epsilon_0 = 1$ and use a scaling parameter $\lambda \in (0, 1)$ to define $\epsilon_k := \max\{\epsilon_{final}, \lambda^k \epsilon_0\}$.

Remark 2.18. *It has been empirically established (see [Sch16] and references therein), that the above-discussed approach of gradually decreasing ϵ_k at each iteration, not only provides a more numerically stable scheme but also increases the convergence speed. In other words, a smaller number of iterations is required for achieving a given error threshold with decreasing ϵ_k at each iteration, then while using fixed ϵ_{final} for all iterations.*

Discretization scaling

The entropic regularization with ϵ acts as a smoothing filter on the source and target densities, which smoothes out any details that are on the finer scale then ϵ [Sch16] (see also [Ber17]). This means that using Sinkhorn iterations with discretizations such that $\min_{i,j} d(x_i, x_j) << \epsilon$ does not provide any valuable improvement over working with discretizations that are on the scale of ϵ.

Therefore, it would be more efficient to also use a sequence of discretizations $(X_{N_k}, Y_{N_k}, c^{N_k}, \mu^{N_k}, \nu^{N_k})$ where $N_k = O(\frac{1}{\epsilon_k})^d$ (where d is the dimension of the problem). In order to implement this approach, one would need to find a way to interpolate approximations $f_{\epsilon_k}^k$, $g_{\epsilon_k}^k$ on the discretization $X_{N_{k+1}}$, $Y_{N_{k+1}}$ respectively, while they are computed on the grids X_{N_k}, Y_{N_k}. For this

one can use the canoncial extension formula (2.28) for both potentials:

$$\tilde{f}^k_{\epsilon_k}(x) := -\epsilon_k \log\left(\sum_{j=1..N_k} \exp\left(\frac{1}{\epsilon_k}(g^k_{\epsilon_k}(y_j) - c(x,y_j))\right)\nu^{N_k}(y_j)\right),\ \forall x \in X. \tag{2.38}$$

$$\tilde{g}^k_{\epsilon_k}(y) := -\epsilon_k \log\left(\sum_{i=1..N_k} \exp\left(\frac{1}{\epsilon_k}(f^k_{\epsilon_k}(x_i) - c(x_i,y))\right)\mu^{N_k}(x_i)\right),\ \forall y \in Y. \tag{2.39}$$

Therefore, at k-th iteration, we can take $k-1$-th approximations to be restrictions of $\tilde{f}^{k-1}_{\epsilon_k}(x)$ and $\tilde{g}^{k-1}_{\epsilon_k}(y)$ on the spaces X_{N_k} and Y_{N_k} respectively.

Putting it all together, we obtain the following iterative procedure in k:

$$\begin{aligned}
f^{k-1}_{\epsilon_k} &= \tilde{f}^{k-1}_{\epsilon_{k-1}}|_{X_{N_k}} \qquad g^{k-1}_{\epsilon_k} = \tilde{g}^{k-1}_{\epsilon_{k-1}}|_{Y_{N_k}} \\
\bar{f}^k_{\epsilon_k} &= -\epsilon_k \log\left(\left\langle \exp\left(\frac{1}{\epsilon_k}(g^{k-1}_{\epsilon_k} + f^{k-1}_{\epsilon_k} - c^{N_k})\right), \nu^{N_k}\right\rangle_{Y_{N_k}}\right) \\
f^k_{\epsilon_k} &= f^{k-1}_{\epsilon_k} + \bar{f}^k_{\epsilon_k} \\
\bar{g}^k_{\epsilon_k} &= -\epsilon_k \log\left(\left\langle \exp\left(\frac{1}{\epsilon_k}(f^k_{\epsilon_k} + g^{k-1}_{\epsilon_k} - c^{N_k})\right), \mu^{N_k}\right\rangle_{X_{N_k}}\right) \\
g^k_{\epsilon_k} &= g^{k-1}_{\epsilon_k} + \bar{g}^k_{\epsilon_k}
\end{aligned} \tag{2.40}$$

In this setting, taking ϵ_{final} to 0 means also refining the discretization. This fits into the setup of the Theorem 2.15, but as already mentioned in remark 2.18, in practice, by the effect of ϵ scaling, the convergence rate is much faster.

Cutoff of iterations

The combination of the ϵ scaling and discretization scalings give us a threefold advantage: Firstly, we avoid the stability issues of overflow, and can decrease epsilon further than naive implementation would allow. Secondly, ϵ scaling provides better convergence, reducing the number of iterations necessary to achieve the desired accuracy. Lastly, discretization scaling allows us to perform a portion of those iterations on the coarser discretizations than desired final discretization, reducing the time necessary for computing those iterations.

But even with all this benefits, final iterations that happen on the finest discretization still require N^2 operations. This can be reduced using again

the saturation property $f(x_i) + g_(y_j) - c(x_i, y_j) = 0$ for optimal pairs (x_i, y_j): When ϵ_k is small, the values $e^{\frac{1}{\epsilon_k}(f^k_{\epsilon_k}(x_i)+g^{k-1}_{\epsilon_k}(y_j)-c^{N_k}(x_i,y_j))}$ become extremely small when the pairs (x_i, y_j) are far from the graph of optimal pairs, since $f(x_i) + g_(y_j) - c(x_i, y_j)$ is strictly smaller then 0.

In other words, during the iterations, many of the summands will be arbitrarily small and will not contribute to the total sum. Therefore, it is possible to exclude those summands from the iteration, without affecting the final result.

Given some tolerance threshold η, define:

$$J_k(i) := \left\{ j \,|\, e^{\frac{1}{\epsilon_k}(f^k_{\epsilon_k}(x_i)+g^{k-1}_{\epsilon_k}(y_j)-c^{N_k}(x_i,y_j))} > \eta \right\} \tag{2.41}$$

$$I_k(j) := \left\{ i \,|\, e^{\frac{1}{\epsilon_k}(f^k_{\epsilon_k}(x_i)+g^{k-1}_{\epsilon_k}(y_j)-c^{N_k}(x_i,y_j))} > \eta \right\} \tag{2.42}$$

And commence the following iterations in k:

$$f^{k-1}_{\epsilon_k} = \tilde{f}^{k-1}_{\epsilon_{k-1}}|_{X_{N_k}} \qquad g^{k-1}_{\epsilon_k} = \tilde{g}^{k-1}_{\epsilon_{k-1}}|_{Y_{N_k}}$$

$$\bar{f}^k_{\epsilon_k}(x_i) = -\epsilon_k \log \left(\sum_{j \in J_k(i)} \exp\left(\frac{1}{\epsilon_k}(g^{k-1}_{\epsilon_k}(y_j) + f^{k-1}_{\epsilon_{k'};}(x_i) - c^{N_k}) \right) \nu^{N_k}(y_j) \right)$$

$$f^k_{\epsilon_k} = f^{k-1}_{\epsilon_k} + \bar{f}^k_{\epsilon_k}$$

$$\bar{g}^k_{\epsilon_k}(y_j) = -\epsilon_k \log \left(\sum_{i \in I_k(j)} \exp\left(\frac{1}{\epsilon_k}(f^k_{\epsilon_k}(x_i) + g^{k-1}_{\epsilon_k}(y_j) - c^{N_k}) \right) \mu^{N_k}(x_i) \right)$$

$$g^k_{\epsilon_k} = g^{k-1}_{\epsilon_k} + \bar{g}^k_{\epsilon_k}$$

Note that as ϵ scaling speeds up convergence, it is easy to estimate the index sets J_k and I_k when ϵ_k becomes small. They can be estimated in a way that they don't exceed $\log(N)$ in size, hence reducing the iterations cost to $N \log(N)$, without evaluating the value at every pair (i, j) individually (which would still require N^2 operations and defeat the purpose of using this approach).

The true complication of this approach comes from memory management. It is common knowledge in software engineering, that arranging computations in a way that memory is accessed in a continuous way, so that processor doesn't have to wait for the delivery of necessary memory components, produces better practical computational time even when the theoretical count of operations is far larger.

Therefore, combining all three approaches described in this section is a delicate software development task. Implementations for CPU usually re-

quire different solutions than for GPU. For the CPU implementation of those approaches see [Sch16], and for the GPU implementation see [Fey19].

Chapter 3

Point source problem

Introduction

In this chapter, we start by presenting the construction from [KO03], which leads to the optimal transport formulation of the point source problem from [Wan04]. We present the methods for finding such solutions. We discuss the necessary adaptations for applying entropic regularization and Sinkhorn algorithm to the point source problem. We also discuss ways of evaluating the obtained solution using ray-tracing. We then present the numerical results.

3.1 An optimal transport model for the point source problem

3.1.1 Constructing a solution of point source problem using paraboloids

In [Wan96] [Wan04], Wang showed that one can solve the point source problem using optimal transport. Here we first present the intuition which provides the optimal transport formulation and then formalize the way such solution is built.

Intuitive construction of the Reflector

First consider the case when the desired target $\nu = \delta_y$ for some $y \in Y$, that is, we want to send all the rays coming from O to the same angle y. It is a well-known property of paraboloids, that they reflect all the rays coming from the focus in the parabola axial direction. Hence we can use a paraboloid with a focus O and axial angle y to achieve the desired target $\nu = \delta_y$.

Such paraboloid can be expressed in the radial parametrization as

$$\Gamma_p := \{\vec{x}p(x) | x \in \mathbb{S}^{d-1}_+\} \quad \text{where} \quad p(x) := p_{y,C}(x) = \frac{C}{1 - \langle \vec{x}, \vec{y} \rangle} \tag{3.1}$$

where and C is the distance between the focus and the directrix (or in other words, $C/2$ is the focal length, aka the distance between the focus and the "tip" of the paraboloid).

Now consider the case where ν is supported on two angles y_1 and y_2. Then, we can build two paraboloids, parametrized using functions p_1 and p_2 (with constants C_1 and C_2), both focused at O and with axial angles respectively y_1 and y_2. Take as a reflector their "envelope", that is,

$$\mathcal{R} = \{\vec{x} \min\{p_1(x), p_2(x)\}\} \tag{3.2}$$

Then all the light will be reflected in the two desired directions. However, in this case we also have to take care how much of the light intensity goes into each direction. Assume that the source measure μ has a continuous density, and the desired target ν has a form $\nu = \alpha_1 \delta_{y_1} + \alpha_2 \delta_{y_2}$. Then, according to the measure-preserving property 1.2, we need to have $\mu(T^{-1}(y_1)) = \alpha_1$ and $\mu(T^{-1}(y_2)) = \alpha_2$ for the reflection map T corresponding to the given reflector. To achieve this, we can adjust the distances from the focus C_1 and C_2 relative to each other.

As shown on figure 3.1, increasing C_2 would take the paraboloid Γ_{p_2} away from focus, reducing $T^{-1}(y_2)$ and conversely, decreasing C_2 expands $T^{-1}(y_2)$. Therefore, $\mu(T^{-1}(y_2))$ depends continuously on C_2, so it can be adjusted in a way that $T_{\#}\mu$ gives the desired ν.

This construction can be generalized for arbitrary N points as a support of $\nu = \sum_{i \le N} \alpha_i \delta_{y_i}$ (see [KO03]), that is, we can find the set of paraboloids $\{p_i\}_{i \le N}$ with coefficients $\{C_i\}_{i \le N}$, such that the reflection map T, corresponding to the reflector obtained by the following radius function pushes forward μ into ν:

$$\rho(x) := \min_{i \le N} \frac{C_i}{1 - \langle \vec{x}, \vec{y_i} \rangle}. \tag{3.3}$$

Taking the logarithm of this relation, and defining $f(x) := \log(\rho(x))$ and $g(y_i) := -\log(C_i)$, we recover the following:

$$f(x) = \min_{i \le N} -\log(1 - \langle \vec{x}, y_i \rangle) - g(y_i) \tag{3.4}$$

Here one can recognize the semi-discrete version of the c-transform relation (2.8) with the cost function

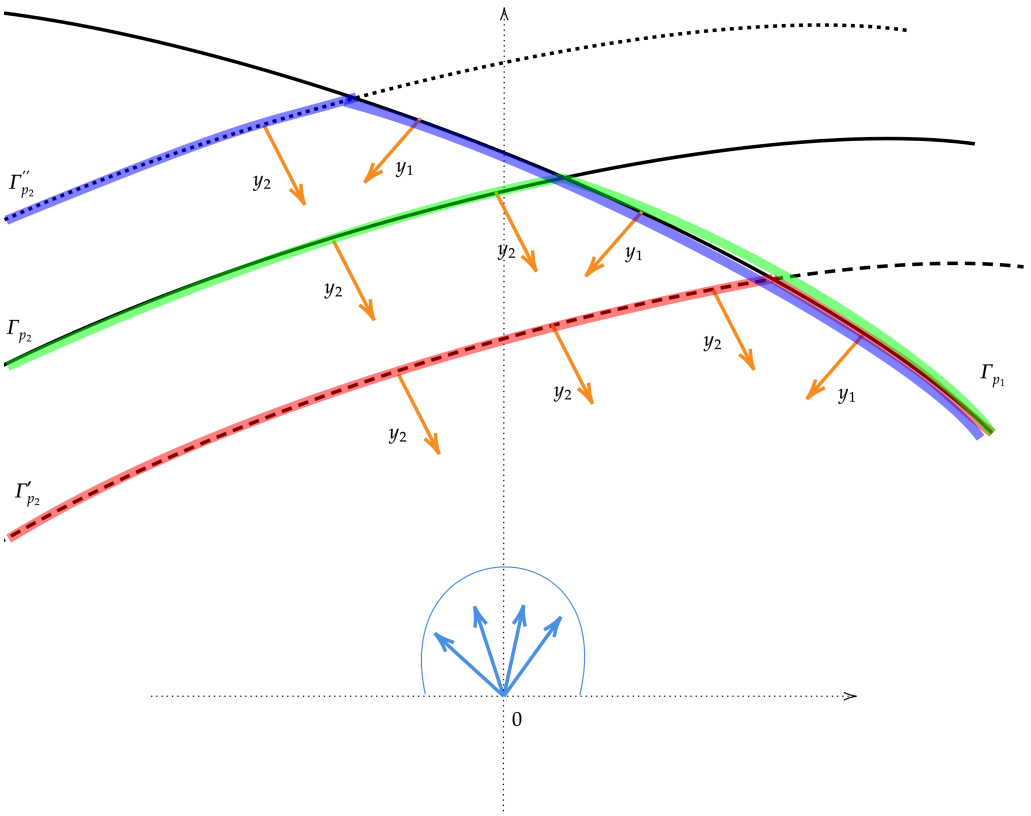


Figure 3.1: When the paraboloid $\Gamma_{p_1} := \{\hat{x}p_1(x)\}$ with a focal direction y_1 is fixed, taking paraboloid Γ_{p_2} with a focal direction y_2 induced by p_2 closer (Γ'_{p_2}, dashed) to the focus O results in reflecting surface changing from the green to red, resulting in increasing the amount of light sent to y_2 (and correspondingly decreasing the amount sent to y_1). Similarly, taking Γ_{p_2} further (Γ''_{p_2}, dotted) from the focus O results in reflecting surface changing from green to blue, resulting in decreased amount of light sent to y_2 (and correspondingly increasing the amount sent to y_1)

$$c(x, y) := -\log(1 - \langle \vec{x}, \vec{y} \rangle) \tag{3.5}$$

This suggests that the solution (at least for the discrete target) can be found among the c-concave functions. Moreover, the measure-preserving requirement for the reflection map T is the same as the measure-preserving constraint for the optimal transport problem (2.1). This intuitively suggests that the solution for the reflector problem could be found by solving an optimal transport problem with the above cost (3.5).

As mentioned above, this intuition was formalized by Wang in [Wan96] and [Wan04]. The solution of the reflector problem was constructed based on the supporting paraboloids. Also, the relation between the radius function ρ and the Kantorovich potential of the optimal transport problem with the cost (3.5) was established. We present this construction and the result below.

Envelope of paraboloid solution of the reflector problem

For the reflector, $\mathcal{R}_\rho$, the supporting paraboloid at point $\vec{x}_*$ is a paraboloid that touches the reflector at $\vec{x}_*$, and is above it for any other point. More formally, we say that Γ_p is a supporting paraboloid for $\mathcal{R}_\rho$ at a point $\vec{x}_*$, if

$$\begin{cases} \rho(x_*) = p_{y,C}(x_*) \\ \rho(x) \leq p_{y,C}(x) \quad \forall x \in X \end{cases} \tag{3.6}$$

Definition 7 (Admissible reflector)**.** *The function ρ and the corresponding reflector $\mathcal{R}_\rho$ are called admissible if $\mathcal{R}_\rho$ has a supporting paraboloid at every point.*

In [Wan04] Wang proved that solving an optimal transport problem with the reflector cost (3.5) provides the reflection map from an admissible reflector, which is a solution to the point source problem.

Theorem 3.1 ([Wan04])**.** *If $X \subset \mathbb{S}^{d-1}_+$ and $Y \subset \mathbb{S}^{d-1}_-$ are connected domains, μ and ν have positive bounded densities on X and Y respectively, then the Kantorovich potential soltions (f, g) of the optimal transport problem (2.6) with the reflector cost $c(x, y) := -\log(1 - \langle \vec{x}, \vec{y} \rangle)$ solve the point source reflector problem in the sense that the radius function $\rho = e^f$ is admissible and the reflection from the corresponding reflector is given by the optimal map T.*

Note that the Kantorovich potential f is a c-transform of g with the reflector cost:

$$f(x) = \min_y c(x, y) - g(y) = \min_y -log\left(1 - \langle \vec{x}, \vec{y} \rangle\right) - g(y) \tag{3.7}$$

Taking the exponent in e of this constraint gives a continuous version of (3.3):

$$e^{f(x)} = \min_y \frac{e^{-g(y)}}{1 - \langle \vec{x}, \vec{y} \rangle} \tag{3.8}$$

On the other hand, as an optimal map, T saturates the dual constraint.:

$$f(x) + g(T(x)) = c(x, T(x)) = -log\left(1 - \left\langle \vec{x}, \vec{T(x)} \right\rangle\right) \tag{3.9}$$

Taking g to the other side and taking the exponent of both sides, gives:

$$e^{f(x)} = \frac{e^{-g(T(x))}}{1 - \left\langle \vec{x}, \vec{T(x)} \right\rangle} \tag{3.10}$$

Combining all this and setting $\rho(x) = e^{f(x)}$, we can see that the supporting paraboloid for the reflector $\mathcal{R}_\rho$ at a fixed x_* is given by a paraboloid with $C = e^{-g(T(x_*))}$ and axial direction $T(x_*)$.

As the reflector has a supporting paraboloid at every point, it is a strictly convex surface, hence differentiable almost everywhere. At the differentiability points, the tangent and normal of the surface coincide with the tangent and normal of the supporting paraboloid at that point. Therefore, as the normal determines the reflection (see (1.1)), at differentiability points the reflecting direction from the surface $\mathcal{R}_\rho$ is given by $T(x)$.

Finally note that as an optimal map, T pushes forward μ into ν, therefore it automatically satisfies the measure-preserving property (1.2) of the reflection map.

Notation 3. *From now on, we will denote reflectors induced by the Kantorovich potential f by*

$$\mathcal{R}_f := \left\{\vec{x}e^{f(x)} | x \in \mathbb{S}^{d-1}_+\right\} \tag{3.11}$$

Note that this is a slight abuse of notation, as for general ρ we denoted by $\mathcal{R}_\rho$ reflector built from the radial function ρ, while for the Kantorovich potential f, $\mathcal{R}_f$ is built from the radial function e^f.

Remark 3.2 (Adaptation of the cost (3.5) to other optical setups)**.** *Constructions similar to the above exist for the problem of designing the lens surfaces as well. In that setup, light travels into a medium with some refractive index k_1 until it hits the surface of this medium, outside of which is another medium with the refractive index k_2. Defining $\kappa := \frac{k_2}{k_1}$, lens cost has the form:*

$$c(x, y) = \log(1 - \kappa\langle \vec{x}, \vec{y} \rangle) \tag{3.12}$$

Depending on whether κ is larger or smaller than 1, different constructions of the surface are necessary, e.g. using ellipsoids or hyperboloids instead of paraboloids, taking max-envelope instead of min-envelope. But not all of those constructions satisfy the A3 condition from appendix A. A good review on which constructions satisfy this condition and how to tackle the cases which do not, can be found in [Mey18].

3.1.2 Reflector cost and the corresponding numerical approaches

The reflector cost satisfies the regularity assumptions necessary for deriving the results presented in Chapter 2 (see Appendix A for the statement of the assumptions). Therefore, as long as the source measure does not have atoms, the transport map T always exists, is computable using the formula (2.11) and corresponds to the reflection map. Moreover, although the cost function c is not a distance on the sphere $\mathbb{S}^{d-1}$, it is a function of a distance: $c(x,y) = -log(|x-y|^2/2)$.

Intuitively, this means that reflecting direction $T(x)$ produces the least transport cost, when it is farthest from x, which in the case of the sphere $\mathbb{S}^{d-1}$ means $-x$. When the distance between $T(x)$ and x becomes smaller, the cost of transportation becomes larger. In other words, the cost function penalizes deviation of the reflected ray from the opposite direction of the incoming ray, as $\operatorname{argmin}_y c(x,y) = -x$.

In proposition 3.4 we summarize some of the results and properties of the reflector cost $c(x,y) = -\log(1 - \langle \vec{x}, \vec{y} \rangle)$, the Kantorovich potential f obtained by solving the corresponding optimal transport problem and the reflector constructed from it. The regularity of the Kantorovich potential f is established for source and target measures with Holder continuous densities, that are bounded away from 0 and infinity.

Notation 4. *For the Riemannian manifold X, $H_{b,\alpha}(X)$ with $0 < \alpha < 1$ and $b > 0$, will denote the subset of $\mathcal{P}(X)$, containing measures that are absolutely continuous with respect to the volume measure vol of X with densities that have Holder continuous derivatives with the exponent α, and are bounded from below by b and above by $\frac{1}{b}$:*

$$H_{b,\alpha}(X) := \left\{ \mu \in \mathcal{P}(X) \,|\, \mu(x) \in C^{1,\alpha}, \forall_{x \in X}\, b < \mu(x) < \frac{1}{b} \right\} \tag{3.13}$$

Remark 3.3. *Note that, as a consequence of Arzela-Ascoli theorem, if X (and hence $\mathcal{P}(X)$) is compact, then $H_{b,\alpha}(X)$ is a compact subset of $\mathcal{P}(X)$.*

Proposition 3.4.

(i) The reflector cost defined on $\mathbb{S}^{d-1}_+ \times \mathbb{S}^{d-1}_-$ satisfies the regularity assumptions A1-A3 from the Appendix A (see e.g. [Loe13]).

(ii) If μ and ν are from $H_{b,1}$ for some $b > 0$ and $0 < \alpha < 1$, then f is bounded in $C^{3,\alpha}$ (see [Loe13] and references therein).

(iii) The reflector $\mathcal{R}_f$, a $d-1$ dimensional surface in $\mathbb{R}^d$ produced using the Kantorovich potential f, is strictly convex. (A direct consequence of the supporting paraboloids construction)

There have been various works about solving the point source reflector problem using optimal transport. Here we briefly summarize the general classes of numerical methods for solving the optimal transport problem and their comparative advantages and disadvantages. For a more detailed presentation we suggest e.g. [PC18] [San15] [LS18]. We also discuss their applications for the point source problem problem.

Linear Programming approach

The first is the linear programming approach. This assumes that the data in discrete form as in (2.25), which provides a natural discretization of the OT problem (2.6):

$$OT_N(\mu^N, \nu^N) := \min_{\gamma^N \in \Pi(p,q)} \left\langle c^N, \gamma^N \right\rangle_{X_N \times Y_N} \tag{3.14}$$

where

$$\Pi(\mu^N, \nu^N) := \left\{ \gamma^N \in \mathbb{R}_+^{N \times N} | \left\langle 1_{X_N}, \gamma^N \right\rangle_{Y_N} = \mu^N,\ \left\langle 1_{Y_N}, \gamma^N \right\rangle_{X_N} = \nu^N \right\} \tag{3.15}$$

Problem (3.14) is a discrete linear programming problem that can be solved numerically using standard linear programming solvers. This approach was suggested for the point source problem in [Wan04]. The main drawback of this method is its high dimensionality. It is a linear problem with $N \times N$ unknowns and $2\,N$ constraints. Numerical resolution with linear solvers which have cubic complexity in the number of unknowns is therefore out of reach for reasonable discretizations (typically $N > 100$).

Partial Differential approach

Assuming sufficient smoothness to interpret the measure-preserving property of the transport map T in a pointwise sense gives us the Jacobian equation:

$$\mu(x) = \nu(T(x)) det(J_T(x)) \tag{3.16}$$

Plugging the expression (2.11) in the above, gives a Monge-Ampère type PDE:

$$det(J_{\{y \to \nabla_x c(x,y)\}^{-1}(\nabla_x f(x))}(x)) = \frac{\mu(x)}{\nu\left(\{y \to \nabla_x c(x,y)\}^{-1}(\nabla_x f(x))\right)} \tag{3.17}$$

Remark 3.5. *The Monge-Ampere type equation can be derived for the point source problem using the measure-preserving property of the reflection map.*

Let T_ρ denote a reflection map from the reflector $\mathcal{R}_\rho$. Let $D \subset X$ be the set of differentiability points of T. On this set we can understand (1.3) pointwise, as an equality of the densities, which gives us the Jacobi equation for T_ρ:

$$det(J_{T_\rho}(x)) = \frac{\mu(x)}{\nu(T_\rho(x))} \tag{3.18}$$

Computing the value of $det(J_{T_\rho}(x))$ in terms of ρ and its covariant derivative ∇ on $\mathbb{S}^{d-1}$, yields the following Monge-Ampere type differential equation for ρ :

$$\eta^{-2} \det\left(-\nabla_i \nabla_j \rho + 2\frac{1}{\rho}\nabla_i \rho \nabla_j \rho + (\rho - \eta)\delta_{i,j}\right) = \frac{\mu(x)}{\nu(T_\rho(x))} \tag{3.19}$$

Where $\eta := \frac{|\nabla\rho|^2+\rho^2}{2\rho}$ and δ_{ij} is a Kronecker symbol, being 1 when $i = j$ and 0 otherwise. Derivation of this equation is not hard but it involves somewhat lengthy computations. We refer the interested reader to [Wan96].

A natural boundary condition for this equation is that T_ρ should map support of μ to the support of ν:

$$T_\rho(X) = Y \tag{3.20}$$

Therefore, solving the Monge-Ampere type equation (3.19) with a boundary condition (3.20), would provide the solution of the reflector problem.

The formulation (3.19) is equivalent to the (3.17), when substituting c with the reflector cost (3.5).

In order to find the solution of the point source problem using the Monge-Ampere type PDE, a B-spline collocation method was proposed and implemented in [BHP14], [BHP15] with convincing numerical results and a numerical complexity in $O(N^{\frac{3}{2}})$, although with a large hidden constant. Another approach of building monotone discretizations was developed in [BM21]. Wu [WXL+13] derives the Monge-Ampere equation for a lens surface and solves the equations using standard finite differences and Newton iterations. Also, the heuristic solution method of finding the solution of the Monge-Ampere PDE based on the least-squares approach is presented in [RtI20] [RtTBI19] for the reflector cost and the numerical study also demonstrates a $O(N^{\frac{3}{2}})$ complexity.

Semi-discrete approach

Semi-discrete optimal transport takes advantadge of the special case where one measure μ has continuous density and the other ν^N has discrete support Y_N as in (2.25). In this case, the dual formulation in (2.6) can be reduced to a semi-dual discrete optimisation problem through the elimination of the constraints replacing f by its c-transform:

$$f(x) = g^c(x) = \min_{y \in Y_N} (c(x,y) - g(y)) = \min_j \big(c(x,y_j) - g(y_j)\big) \tag{3.21}$$

and optimizing over the finite vector $\{g(y_j)\}_{j \leq N}$.

It is now well understood that a Newton's method can be used to solve the Semi-discrete optimal transport problem under classical regularity hypothesis for the cost $c = \frac{1}{2}d^2$ [Mér11]. The implementation relies on fast (linear time) computations of a tessellation of the target domain called Laguerre cells :

$$Lag_j = \{x,\ s.t.\, c(x,y_j) - g(y_j) < c(x,y_i) - g(y_i),\ \forall i \leq N\}\ \forall j \leq N.$$

There have been numerous works on using this method for the reflector problem. Starting from the pioneering work of [KO03], all the way to [MMDCMT16]. Semi-discrete optimal transport has been adapted to some other optical setups as well, such as designing lens surface instead of the reflector, or collimated sources instead of point source (see e.g. [Mey18]).

However, there is an extra technical difficulty that the efficient Laguerre cell computations are only available in $\mathbb{R}^d$. Therefore, in order to compute those on the sphere, one needs to lift the dimension of the computations by one. Then, one needs to compute Laguerre cells on the sphere as an intersection of the Laguerre cells in $\mathbb{R}^d$ intersected with a triangulation of $\mathbb{S}^{d-1}$. The resulting method still has linear complexity, and efficient implementations in dimensions 2 and 3 are available in [Lec19] and [Lév15].

In this work, we will instead explore the use of entropic optimal transport approach, discussed in chapter 2.2.1, which is also applicable for the reflector cost.

3.1.3 Adaptation of Sinkhorn divergences to the reflector cost

Using the Sinkhorn algorithm for the reflector cost seems straightforward at first glance since one can simply use the iterations (2.27) to obtain the potentials. But as the reflector is built using the function $e^{f_{OT_\epsilon}}$, the entropic bias discussed in Section 2.2.3 is observed on numerical solutions of the problem.

We can see the effect in the simplest "identity" reflector case for example, where each ray is reflected in its opposite direction, i.e. when $\nu(y) = \mu(y - \pi)$, $y \in \mathbb{S}^1_-$. The exact potentials f and g are constant and the reflector $\mathcal{R} = \{\vec{x}\, e^{f(x)} | x \in X\}$ is a portion of circle.

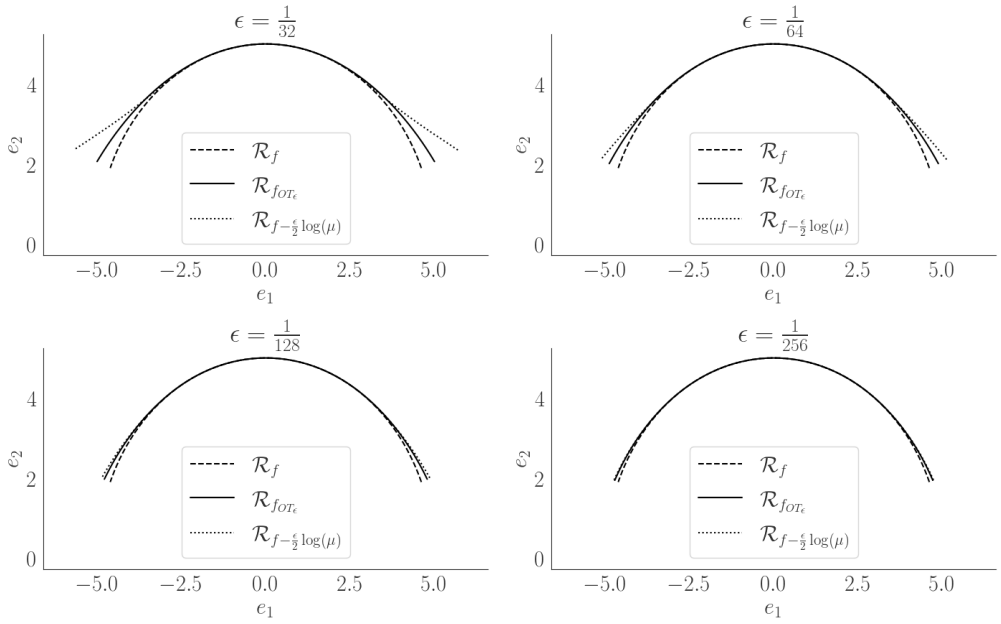


Figure 3.2: As ϵ decreases, reflectors induced from f_{OT_ϵ} and $f - \frac{\epsilon}{2}\log(\mu)$ become closer to each other and to the true reflector $\mathcal{R}_f$, but still remain above it.

Recall from remark 2.17, that the asymptotics of the difference between the entropic and exact potentials can be expressed by $f_\epsilon = f - \frac{\epsilon}{2}\log(\mu)$. For the radial function used for constructing the reflector it translates into $e^{f_\epsilon} = \frac{e^f}{\mu^{\frac{\epsilon}{2}}}$. When μ decreases, the denominator decreases and the radius increases, pushing the reflector further from the light source.

We can see this in figure 3.2, for the plane problem ($d = 2$) and $N = 128$ discretization points, where we plot the exact reflector induced from $f = const$, entropic reflector induced from f_{OT_ϵ}, and the reflector obtained by adding theoretical asymptotic bias to the exact potential: $f - \frac{\epsilon}{2}\log(\mu)$, for 4 different values of ϵ: $\frac{1}{16}$, $\frac{1}{64}$, $\frac{1}{128}$ and $\frac{1}{256}$. Remember that the potentials are defined up to a constant, we therefore adjust the constants so that the reflectors superimpose at $x = \pi/2$.

The source distribution μ is built in a way that it is close to uniform between the angles $\pi/4$ and $3\pi/4$ and decays exponentially fast outside of it. This is achieved by by summing 16 Gaussian distributions with deviation

$\pi/32$ and means taken within the interval $[9\pi/32, 23\pi/32]$. it is plotted in figure 3.3. As already mentioned, The target distribution $\nu(y) = \mu(y - \pi)$.

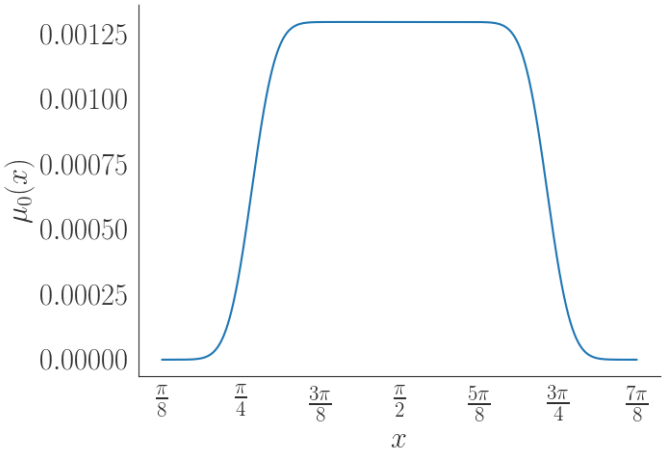


Figure 3.3: Source distribution μ for the simulations in figure 3.2

.

As ϵ decreases, reflectors induced from f_{OT_ϵ} and $f - \frac{\epsilon}{2}\log(\mu)$ become closer to each other and to the true reflector, but still remain above it. This produces a shrinking effect on the reflected distribution (observed e.g in figure 3.10), since the rays reflecting on the biased entropic reflector are reflected strictly inside the support of ν (see figure 3.4).

In order to deal with this bias, we use the notion of Sinkhorn divergences (2.30). When the spaces X and Y are different, we can extend the notion of Sinkhorn divergence as follows :

$$S_\epsilon(\mu, \nu) = OT_\epsilon(\mu, \nu) - \frac{1}{2}\Big(OT_\epsilon(\mu, \mu') + OT_\epsilon(\nu', \nu)\Big). \tag{3.22}$$

where $\mu' = argminOT(\mu, \cdot)$ and $\nu' = argminOT(\cdot, \nu)$.

Note that this is indeed a consistent extension: For the distance costs $c(x, y) = d_p^p(x, y)$, when the spaces X and Y are the same, $\mu' = \mu$ and $\nu' = \nu$ as $arg\min_{y\in Y} d(x, y) = x$.

Also in the reflector cost case, μ' and ν' correspond to the reflections of respectively μ and ν on the circular reflector mentioned above. In other words, that least total cost is induced by reflecting all the rays in the exact opposite direction. Indeed $arg\min_{y\in Y} -log(1- < x, y >) = -x$.

The Kantorovich potential f_{S_ϵ} which we need in order to construct the reflector, can be computed using the same formula as for the distance cost

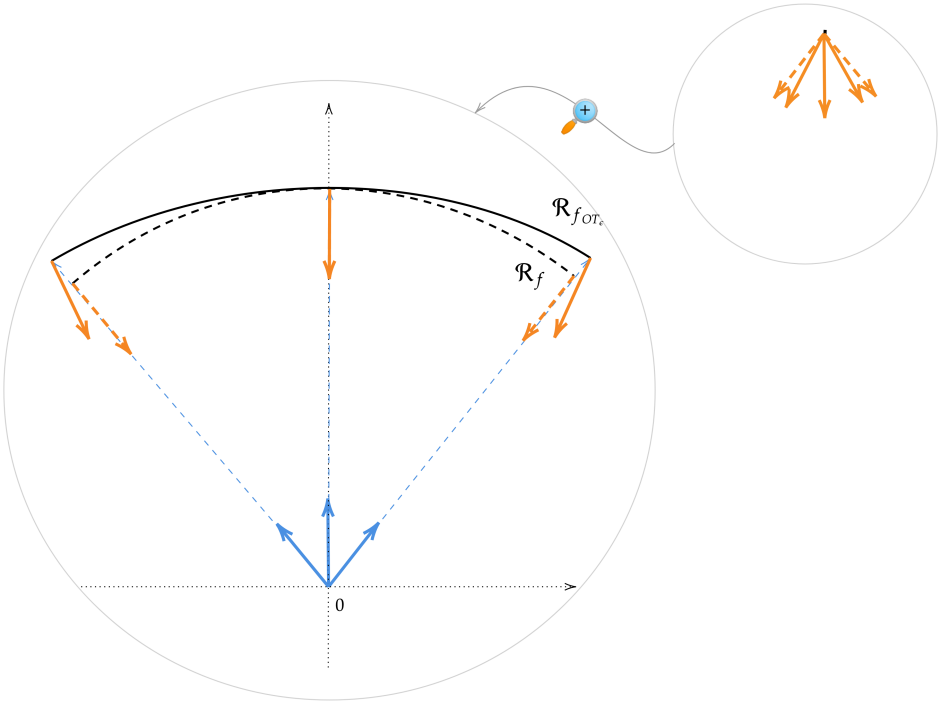


Figure 3.4: In blue and red are rays from μ and ν. Solid and dashed black curves are $\mathcal{R}_{f_{OT_\epsilon}}$ and $\mathcal{R}_f$. The dashed rays are reflected in the exact opposite "identity" direction. The solid rays, reflecting on the biased entropic reflector are reflected strictly inside the support of ν.

Sinkhorn divergence (formula 2.32) :

$$f_{S_\epsilon} = -\epsilon \log\left(\left\langle \exp\left(\frac{1}{\epsilon}(g_{OT_\epsilon} - c_N)\right), \nu_N \right\rangle_{Y_N}\right) + \epsilon \log\left(\left\langle \exp\left(\frac{1}{\epsilon}(f^{\mu}_{OT_\epsilon} - c_N)\right), \mu_N \right\rangle_{X_N}\right) \tag{3.23}$$

3.2 Numerical results

3.2.1 Choice of the Discretization

In Theorem 2.15, the requirement that ϵ is of order $(1/N)^{\frac{1}{d}}$ and that μ_N, ν_N satisfy the density condition (Remark 2.16 and the Appendix C) are closely related. Intuitively this condition means that while choosing the discretizations μ_N, ν_N with N points, it is important to make sure that they approximate the corresponding distributions μ, ν with integration error of order $(1/N)^{\frac{1}{d}}$ for functions which do not oscillate on finer scale then $(1/N)^{\frac{1}{d}}$ (see local density condition in the Appendix C).

The general version of the density property can be satisfied on the sphere S^{d-1} using different discretizations. For $d = 2$, the uniform discretization of angular parametrization creates a uniform grid that is convenient for computations. However, the angular grid in $d = 3$ is not uniform with respect to the area element. This can be corrected by adding the corresponding weights to the discretization points when building the discrete measure μ^N, although the scale of this discretization will be worse than $(1/N)^d$, since the grid will be finer at the center and become gradually coarser at the sides. This can be a disadvantage for some applications where a substantial portion of the mass is away from the center. Different types of projected grids (stereographic projection from the tangent plane, or projections from the equatorial plane), suffer similar disadvantages.

One alternative is to use the Quasi Monte-Carlo grids (see [Ber17]), which are built to minimize the "worst-case error" coming from the non-uniformity of the grids. This discretization is defined by bounding the worst-case error of integration over the desired function space W.

Definition 8. *Given a Riemmanian manifold X of dimension d, a corresponding normalized volume form dV, and a discretization $X_N = x_{i\{i\leq N\}}$, the worst case error of X_N with respect to function space W is:*

$$WCE(X_N, W) = \sup_{f\in W}\left\{\int_X f dV - \sum_{i=0}^{N} f(x_i)\right\}. \tag{3.24}$$

Then, $X_N = x_{i\{i\leq N\}}$ is a Quasi Monte-Carlo discretization of X on lengthscale $(1/N)^{\frac{1}{d}}$ if for any $p \in [1, \infty)$ and $s > \frac{d}{p}$:

$$WCE(X_N, W_p^s) \leq \frac{C_{s,p}}{\left(N^{\frac{1}{d}}\right)^s}, \tag{3.25}$$

where W_p^s is a Sobolev space of functions f such that all derivatives of order s are in $L_p(X)$ and $C_{s,p}$ is a uniform constant depending only on s and p.

[Ber17] shows that for a density μ which is absolutely continuous with respect to the volume form dV with density ρ_μ, the discrete approximation μ_N which satisfies the density requirements of theorem 2.15 can be constructed using Quasi Monte-Carlo discretization X_N of X as the empirical measure :

$$\mu_N = \frac{1}{\sum_{x_i \in X_N} \rho_\mu(x_i)} \sum_{x_i \in X_N} \rho_\mu(x_i)\delta_{x_i}. \tag{3.26}$$

When the space X is linear, standard uniform square grids with step-size h are Quasi Monte-Carlo systems on the lengthscale h. But for curved spaces, such as a sphere in the case of the reflector problem, constructing Quasi Monte-Carlo systems is not straightforward and usually they do not have such a simple structure.

One way of constructing such a discretization on the sphere, which was used for the simulations in this chapter, is given in [Wom17].

3.2.2 Interpolation for ray-tracing

As discussed in Chapter 1.4.1, forward ray-tracing is a convenient method for the point source problem. In order to use this method, the continuous interpolation of the obtained discrete reflector is required. Since we construct the reflector using the potential f, interpolating the potential also induces an interpolation on the reflector (and vice versa).

Apart from all the classical interpolation approaches (linear, bi-linear, spline, etc.), it is also possible to use the structure of the optimal transport problem in order to obtain an interpolation.

Recall that the Kantorovich potentials are saturating the constraint 2.7. Moreover, they are c-transforms of each other: $f = g^c := \min_y c(x, y) - g(y)$ and vice versa. Of course for the discrete solutions f^N, g^N this relation also holds in the discrete setting: $f^N(x_i) = \min_{y_j} c(x_i, y_j) - g^N(y_j)$.

This provides a way to obtain a c-concave interpolation $\hat{f}^N$ for f^N:

$$\hat{f}^N(x) := \min_{y_j} c(x, y_j) - g^N(y_j) \tag{3.27}$$

This interpolation provides a reflector $\hat{\mathcal{R}}^N$, which is a true "envelope of paraboloids" discussed in Chapter 3.1, meaning that it consists of N patches of paraboloids. Such a surface, apart from the fact that it is non-smooth, also sends all light into the discrete target consisting of N axial directions of the paraboloids. This could be an advantage when one wants to produce a pixelized target (with sharp changes from section to section) but is a disadvantage when one wants to approximate a continuous target density.

In order to provide a smooth reflector that would be able to produce continuous target densities, one can use the entropic optimal transport structure. In particular, we can use the canonical extension formula (2.28) for the entropic iterations:

$$\tilde{f}^N(x) = -\epsilon \log \left(\sum_{j=1\dots N} e^{\left(\frac{1}{\epsilon}(g^N(y_j) - c(x,y_j))\right)} \nu_N(y_j) \right), \quad \forall x \in X. \tag{3.28}$$

As the above formula is C^∞ smooth, this interpolation creates a smooth reflecting surface and is a better choice for producing continuous target distributions.

From now on, we will denote potentials and their corresponding reflectors obtained using c-concave interpolations by $\hat{\cdot}$ and entropic interpolations by $\tilde{\cdot}$.

3.2.3 Numerical setup

Our inputs are analytical descriptions of the illumination/illuminance μ and ν described in the test cases section below. All test cases presented in this Chapter will have the same source and target domains X and Y. The source domain $X \subset \mathbb{S}^2$ will be the inverse stereographic projection in the northern hemisphere of the square domain centered at the origin $\{(x_1, x_2) \in \mathbb{R}^2 | -0.6 \leq x_1 \leq 0.6,\ -0.6 \leq x_2 \leq 0.6\}$. Similarly, $Y \subset \mathbb{S}^2$ will be the inverse stereographic projection in the southern hemisphere of the same domain.

The outputs are ray-tracing computed according to the following procedure :

1. Computation of f^N : The discrete Kantorovich potentials are computed for the discretizations (μ_N, ν_N) induced by the Quasi Monte-Carlo grids X_N and Y_N respectively according to the formula (3.26).

2. Interpolation of f^N: Interpolation will happen either using c-concave interpolation $\hat{f}^N$ or entropic interpolation $\tilde{f}^N$ as discussed in the previous section.

3. Ray-tracing: We will use the forward ray-tracing with a Quasi Monte-Carlo sampling of the source domain of size M, with the sampling produced in [Wom17]. We will not use the binning technique discussed in 1.4.3 and instead display directly the point clouds.

Remark 3.6. *The ray-traced distribution of reflected directions (obtained as a distribution of directions in $\mathbb{R}^3$) is then projected using the stereographic projection from the south pole to the equator plane:*

$$(x, y, z) \in \mathbb{R}^3 \to (\frac{x}{1+z}, \frac{y}{1+z}) \in \mathbb{R}^2. \tag{3.29}$$

Definition 9 (Parameters and notation for the forward map)**.** *In order to discuss the numerics, we need to introduce a notation for the forward map induced by the above procedure. This is cumbersome as many parameters are involved : ϵ the entropic regularization, N the discretization of the distributions μ and ν used in the Sinkhorn algorithm. The choice of using or not the Sinkhorn divergence correction. The choice and notation of the interpolation for the reflector.*

We will use the notation $\hat{f}^N_{OT_\epsilon}$, $\tilde{f}^N_{OT_\epsilon}$, $\hat{f}^N_{S_\epsilon}$, $\tilde{f}^N_{S_\epsilon}$ for different potentials, and $\hat{\mathcal{R}}^N_{OT_\epsilon}$, $\tilde{\mathcal{R}}^N_{OT_\epsilon}$, $\hat{\mathcal{R}}^N_{S_\epsilon}$, $\tilde{\mathcal{R}}^N_{S_\epsilon}$ for the respective reflectors. The superscript $\cdot^N$ will denote the discretization number which was used in order to obtain the solution of the optimal transport problem (for simplicity, we will use same number for the source and target). The subscript will identify the regularization parameter ϵ and also weither or not the Sinkhorn divergence correction was used: $\cdot_{OT_\epsilon}$ denoting the use of the "pure" entropic potential and $\cdot_{S_\epsilon}$ denoting the use of the corrected one using the Sinkhorn divergence. Finally the header $\hat{\cdot}$ will denote the use of c-concave interpolation (3.27) while $\tilde{\cdot}$ will denote the use of the entropic interpolation (3.28)

We will denote by $\bar{\nu}^M = \hat{\mathcal{R}}^N_{OT_\epsilon}[\mu^M]$ the ray-traced sampling with M rays (and a similar notation for the other versions of the reflector). Of course, $\bar{\nu}^M$ also depends on the choice of the method and other parameters, which will always be clear from the context. Finally, the N and M discretizations of the domain are QMC discretizations as explained in section 3.2.1.

3.2.4 Test cases and illustrations

We give here several ray-traced images (produced using the QMC sampling of size $M = 128^2$) as an illustration of the different approximations of the reflectors. The code together with the data is available on the repository : `https://github.com/ROMSOC`

The numerical procedure is introduced in section 3.2.3. The test cases we used are described below.

Test Case 1: Square To Circle. The source distribution μ is the uniform distribution over the inverse stereographic projection of a square :

$$StP(supp(\mu)) = \{(x_1, x_2) \in \mathbb{R}^2 | -0.5 \leq x_1 \leq 0.5, -0.5 \leq x_2 \leq 0.5\}$$

The target distribution ν is the uniform distribution over the inverse stereographic projection of a disk : $StP(supp(\nu)) = \{(x_1, x_2) \in \mathbb{R}^2 | x_1^2 + x_2^2 \leq 0.5^2\}$. Even though the densities are constant, mapping from a non-smooth support geometry of the square to the smooth geometry of a circle is not a trivial task.

We show in figure 3.10, from left to right and from top to bottom the ray-traced images using the reflectors $\hat{\mathcal{R}}_{OT_\epsilon}$, $\tilde{\mathcal{R}}_{OT_\epsilon}$, $\hat{\mathcal{R}}_{S_\epsilon}$, $\tilde{\mathcal{R}}_{S_\epsilon}$ and finally also the QMC discretization used for ν_N, $N = 64 * 64$ points. The regularization parameter was taken to be $\epsilon = \frac{1}{2*64}$.

Figures 3.10 (a) and (c) correspond to the c-concave interpolations. This builds the minimal envelope of the family of parabolae with the focal axis given by the ν discretization. All rays hitting one parabola end up at the same position. So up to numerical errors we indeed recover (e), that is ν_N even though $M >> N$. Also, the S_ϵ solution performs better at the boundary.

Figures 3.10 (b) and (d) correspond to the entropic interpolations. This is a smoothing of the c-concave interpolation. The M rays are therefore distributed more evenly over the support of the target. We observe, for (b), at the boundary of the support, an important shrinking effect. It is caused by the entropic bias that was discussed in chapter 2.2.3 and also in section 3.1.3. The figure 3.4 also shows a similar effect in $d = 2$. As expected, the S_ϵ (d) solution is effective to de-bias the solution. Except, at the corners of the square where the map is singular, we see extra artifacts, induced by the smoothed interpolation.

Test Case 2 : Square to Gaussian. The target distribution ν has a Gaussian density $\rho(x_1, x_2) = e^{-\frac{x_1^2 + x_2^2}{2}}$ over the whole domain of computation. The source distribution will be the same as in Test Case 1.

Test Case 3 : Square to Square. The source and target distribution are uniform distribution over the inverse stereographic projection of a square : $StP(supp(\nu)) = \{(x_1, x_2) \in \mathbb{R}^2 | -0.5 \leq x_1 \leq 0.5, -0.5 \leq x_2 \leq 0.5\}$.

This test case corresponds to the "identity reflector" that should be a portion of a sphere and send every ray into its exact opposite direction. As demonstrated in section 3.1.3, such a task can be challenging for entropy-based methods.

Test Case 4 : Circle To Steep Gaussian. The source distribution μ is the uniform distribution over the inverse stereographic projection of a disk $StP(supp(\mu)) = \{(x_1, x_2) \in \mathbb{R}^2 | x_1^2 + x_2^2 \leq 0.5^2\}$. The target distribution ν has a Gaussian density $\rho(x_1, x_2) = e^{-16*(x_1^2+x_2^2)}$ over the whole domain of computation.

Test Case 5 : Circle To Two Steep Gaussians. The source distribution μ is the same as in the fourth. The target distribution ν has a Gaussian density $\rho(x_1, x_2) = e^{-16*((x_1-0.25)^2+(x_2-0.25)^2)} + e^{-16*((x_1-0.25)^2+(x_2+0.25)^2)}$ over the whole domain of computation.

Figures 3.5- 3.9 show the ray-traced images for the 5 test cases with the reflector $\tilde{\mathcal{R}}^N_{S_\epsilon}$ with $N = 128^2$ and $\epsilon = \frac{1}{8*128}$. Those figures show that the point cloud distribution resembles the desired distribution (shown as a color graph). However, precise evaluation and more importantly, quantification of the quality of the approximation is clearly required. We address this in the next section 3.2.5.

3.2.5 Wasserstein metrics as an error estimator

From the optimal transport point of view, a straightforward evaluation of our numerical approach would be to build an approximate transport map T^{app} obtained by plugging the different approximations of the potential $\hat{f}^N_{OT_\epsilon}$, $\tilde{f}^N_{OT_\epsilon}$, $\hat{f}^N_{S_\epsilon}$, $\tilde{f}^N_{S_\epsilon}$ into formula (2.11) and compare it against the exact solution T. For the reflector problem, however, only trivial analytical solutions such as the circle (identity) or a parabola (Dirac target) are known. Also, in general, it is desirable to be able to evaluate the quality of the obtained solutions of real-world problems, where the exact solution is not known.

On the other hand and from the optics application point of view, the main concern is not the shape of the reflector itself (except for designing constraints), but rather the quality of the produced illumination, which is the push-forward ${T^{app}}_{\#}\mu$. Therefore, in order to build an error estimator, we will consider the difference between the push-forward ${T^{app}}_{\#}\mu$ and the desired distribution ν.

For this, we will use the standard L^2 Wasserstein distance $W_2({T^{app}}_{\#}\mu, \nu)$, or its S_ϵ entropic counterpart, which is known to provide a smooth distance between empirical distributions [CRL+20]. In order to build a numerically computable error estimate, we will use the ray-tracing of a QMC-sample μ^M of the source measure μ.

First, using the triangle inequality for the Wasserstein metrics we get :

$$W_2({T^{app}}_{\#}\mu, \nu) \leq W_2({T^{app}}_{\#}\mu, {T^{app}}_{\#}\mu^M) + W_2({T^{app}}_{\#}\mu^M, \nu). \tag{3.30}$$

The first term on the right-hand side ($W_2(T^{app}{}_{\#}\mu, T^{app}{}_{\#}\mu^M)$) can be estimated using Lemma 2.11 and inequality (2.20) giving :

$$W_2(T^{app}{}_{\#}\mu, T^{app}{}_{\#}\mu^M) \leq K\, W_2(\mu, \mu^M), \tag{3.31}$$

where K depends on the data of the problem. The convergence in W^2 norm of such a sampling has been studied in [FG15] and is known to behave asymptotically as $M^{-\frac{1}{2}}$.

The second term on the right hand side of (3.30) can be approximated using $\bar{\nu}^M$ the point-cloud obtained by ray-tracing the computed reflector :

$$W_2(T^{app}{}_{\#}\mu^M, \nu) \simeq W_2(\bar{\nu}^M, \nu). \tag{3.32}$$

The continuous densities μ and ν still appear in estimates (3.31-3.32). The W_2 distance can be computed either using semi-discrete optimal transport (see section 3.1.2) which relies on a $P1$ discretization of the continuous densities or again using Sinkhorn divergence. This is the choice we made in this work and μ and ν are discretized on a finer $N_\infty = 512^2$ grid.

The error estimates are therefore computed using the Sinkhorn divergence approximation of the L^2 Wasserstein distance on the projection plane with a smaller ϵ : $W_2(.,.) \simeq \sqrt{S_{\epsilon=1e-06}(.,.)}$, but we will keep the W_2 notation below.

To sum up, we estimate the error of the computed reflector, by considering the quantity

$$W_2(\bar{\nu}^M, \nu) \tag{3.33}$$

Computed using the Sinkhorn divergences. In order to interpret this quantity, we compare it to

$$W_2(\mu^M, \mu) \tag{3.34}$$

while also taking into account the corresponding supports of the source and the target measures.

3.2.6 Numerical convergence Study in N

In this section $M = 128^2$ the number of rays for the ray-tracing is fixed. The densities μ and ν are discretized with a finer $N_\infty = 512^2$ points orthogonal grid on the plane.

In (3.31), for μ as in Test Cases 1,2 and 3 , we obtained with the above parameters :

$$W_2(\mu^M, \mu^{N_\infty}) = 2.355e-03.. \tag{3.35}$$

For Test cases 4 and 5, we obtained

$$W_2(\mu^M, \mu^{N_\infty}) = 2.310e-03. \tag{3.36}$$

Regarding $W_2(\bar{\nu}^M, \nu^{N_\infty})$ in (3.32), we plot for different tests cases the convergence curves in N for the different reflector approximation methods explained in definition 9 and two values of ϵ : $1/2\sqrt{N}$, $1/8\sqrt{N}$.

From the figures 3.11- 3.15 for the 5 test cases we observe :

1. Convergence to error levels comparable to (3.35-3.36) which makes sense as $\bar{\nu}^M$ is at best a QMC sampling of ν.

2. Decreasing ϵ improves the OT_ϵ reflectors.

3. Sinkhorn divergence S_ϵ de-biasing is effective, both in the sense of obtaining lower errors and being less dependent on the choice of ϵ.

4. As the target is smooth, the entropic interpolation of S_ϵ solutions are less dependent on the choice of the discretization, and moderate values of N are enough to achieve the same error as when using the highest value of N used here.

3.2.7 Numerical convergence Study in M

In this section, we use the Entropic interpolation method and generate the potential with the Sinkhorn divergence method as they seemed to performed the best. We then fix $N = 128^2$ and $\epsilon = 1/(128 * 8)$ and study the dependence of the error term $W_2(\bar{\nu}^M, \nu)$ with M where $\bar{\nu}^M = \tilde{\mathcal{R}}^N_{S_\epsilon}[\mu^M]$.

In figures 3.16-3.18 we plot the error curves for Test Cases 1,3 and 5 in original and log scales. The curves demonstrate that the computed continuous numerical approximation of the reflector(obtained by entropic interpolation of the discrete potential) preserves after ray-tracing the quality of the illumination/source ray sampling. In particular :

1. The convergence curves are similar to the convergence curves in N obtained using the Sinkhorn divergence method with $\epsilon = 1/(8\sqrt{N})$ and the c-concave interpolation. Indeed, with this interpolation method and a good approximation of the potential, the reflector will send all the rays onto ν_N which is also discretized using a QMC system. So increasing N there and increasing M here results in the same empirical measure $\bar{\nu}^M = \tilde{\mathcal{R}}^N_{S_\epsilon}[\mu^M]$.

2. In logarithmic scales the curves agree with the $M^{-\frac{1}{2}}$ rate predicted by the theory [FG15].

3.3 Figures

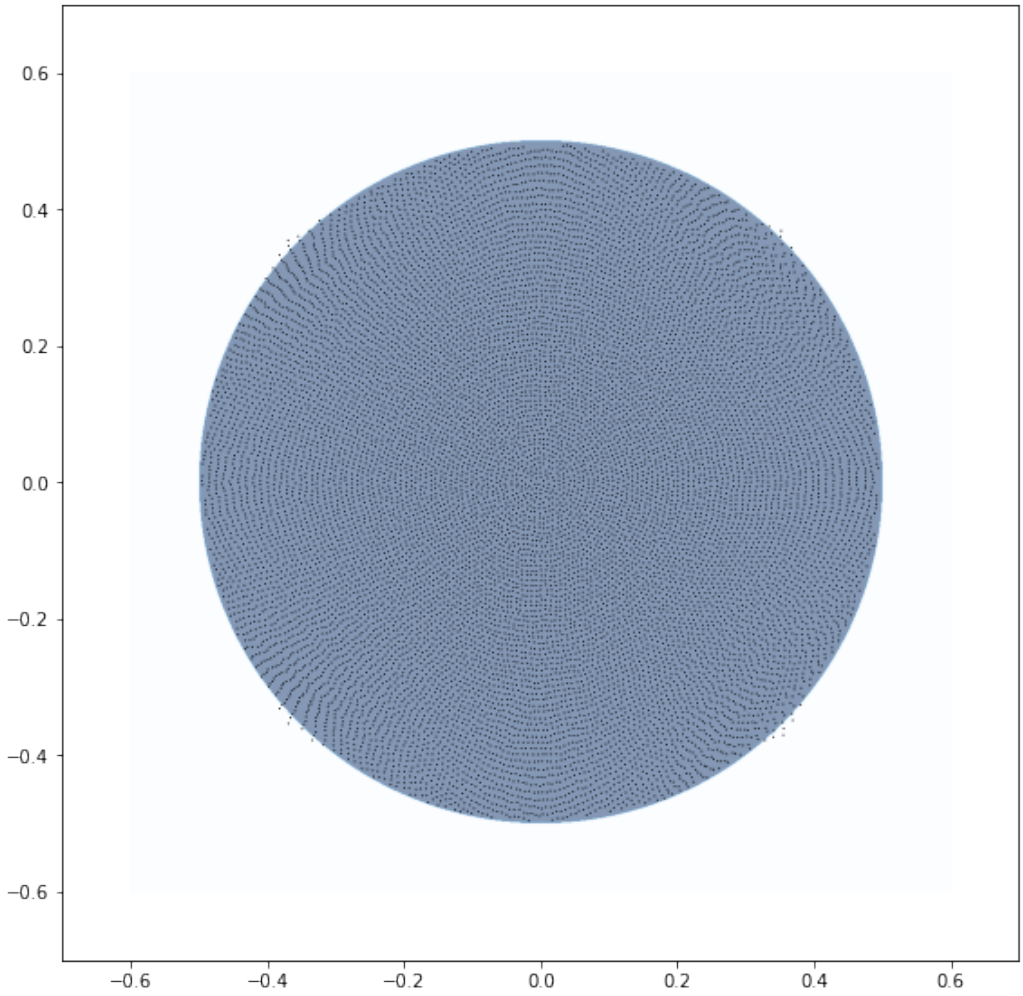


Figure 3.5: Test Case 1 : Ray-traced reflection $\bar{\nu}^M$ from $\tilde{R}_{S_{1/8\sqrt{N}}}$(point cloud) together with the exact desired distribution (color graph)

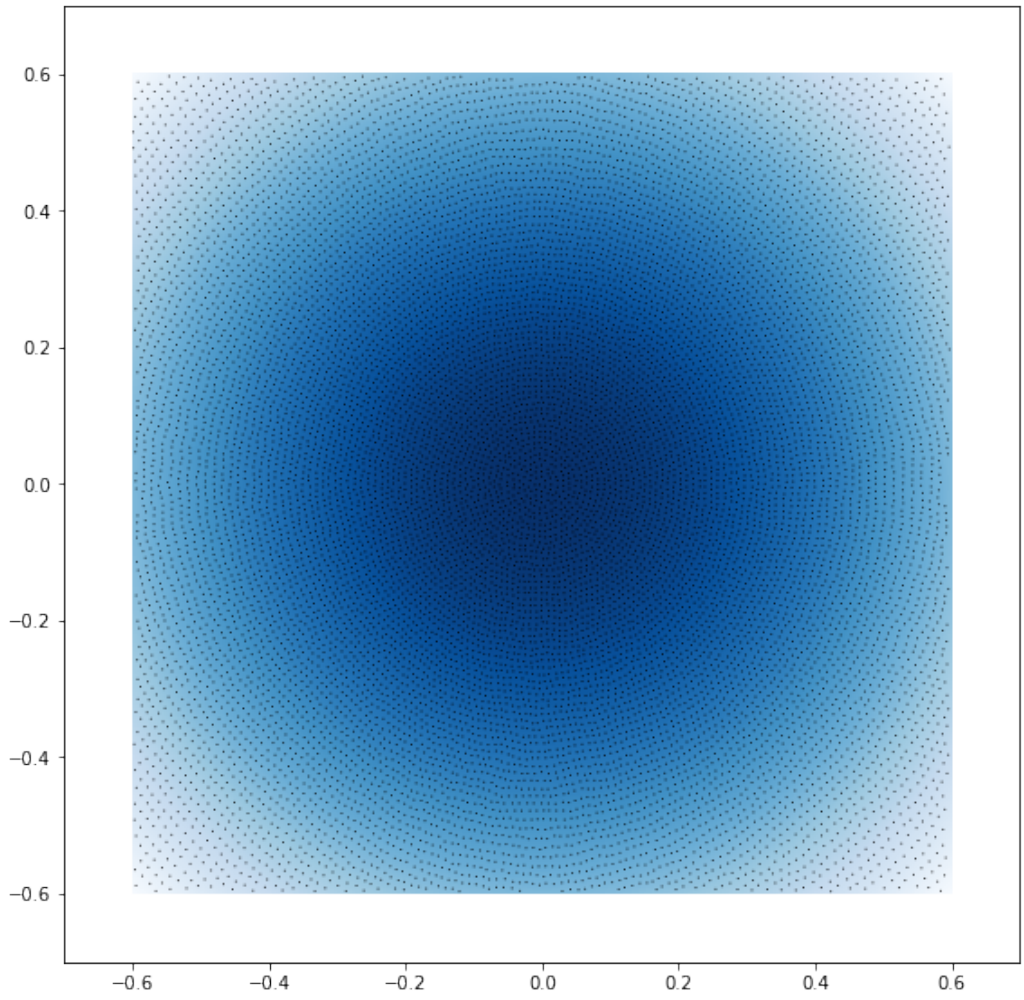


Figure 3.6: Test Case 2 : Ray-traced reflection $\bar{\nu}^M$ from $\tilde{R}_{S_{1/8\sqrt{N}}}$(point cloud) together with the exact desired distribution (color graph)

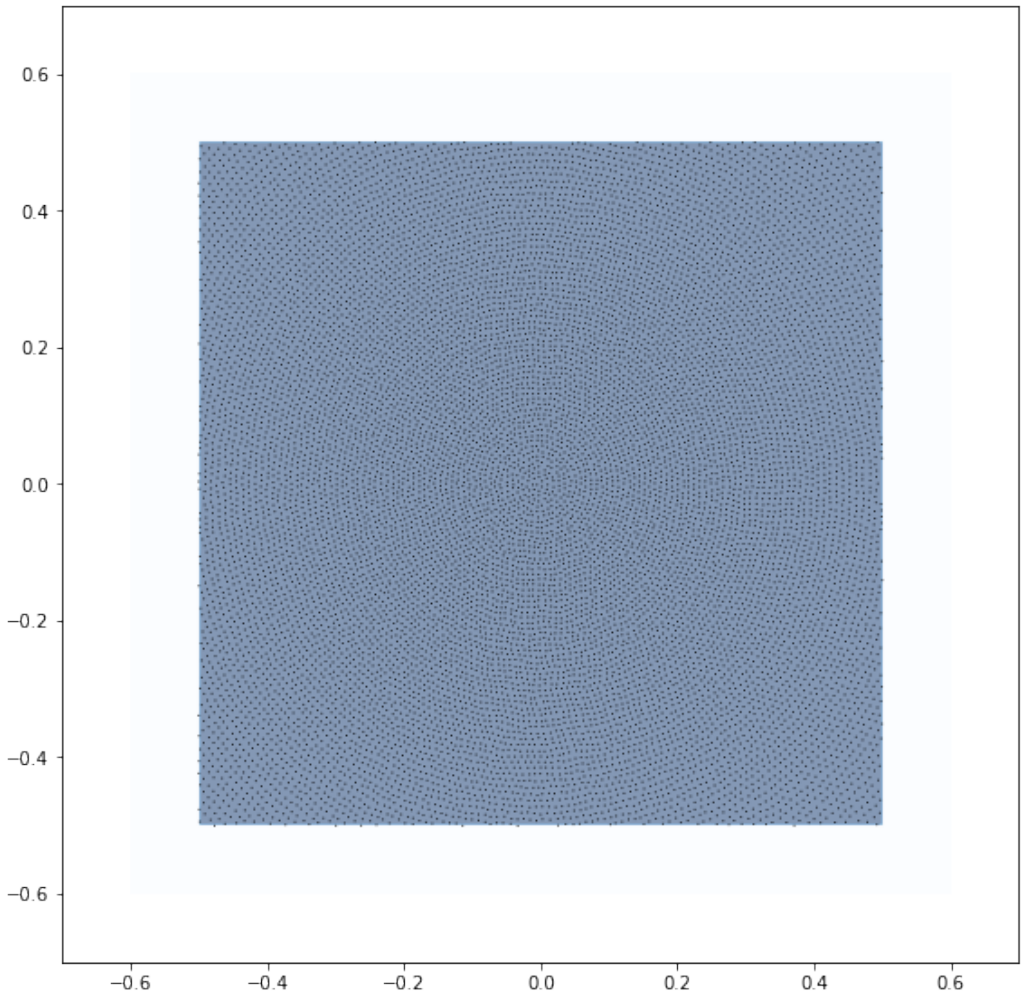


Figure 3.7: Test Case 3 : Ray-traced reflection $\bar{\nu}^M$ from $\tilde{R}_{S_{1/8\sqrt{N}}}$(point cloud) together with the exact desired distribution (color graph)

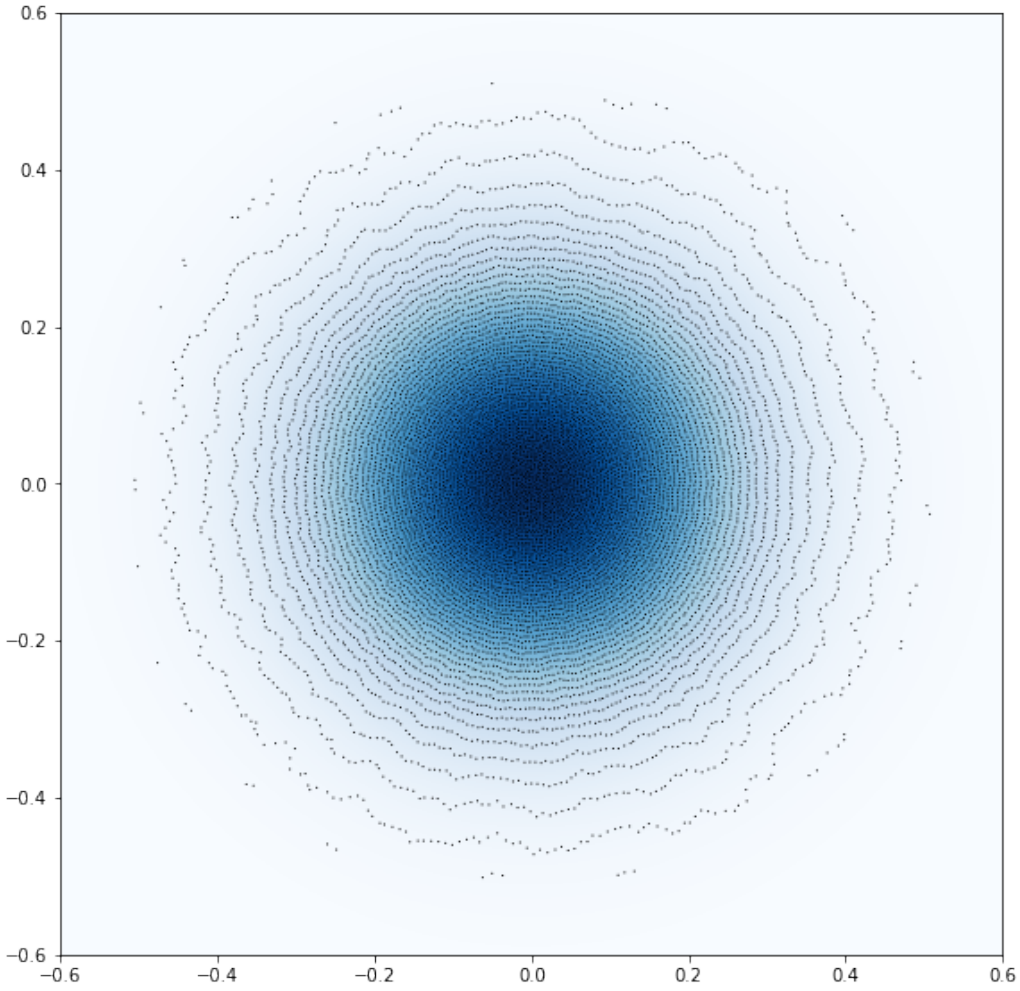


Figure 3.8: Test Case 4 : Ray-traced reflection $\bar{\nu}^M$ from $\tilde{R}_{S_{1/8\sqrt{N}}}$(point cloud) together with the exact desired distribution (color graph)

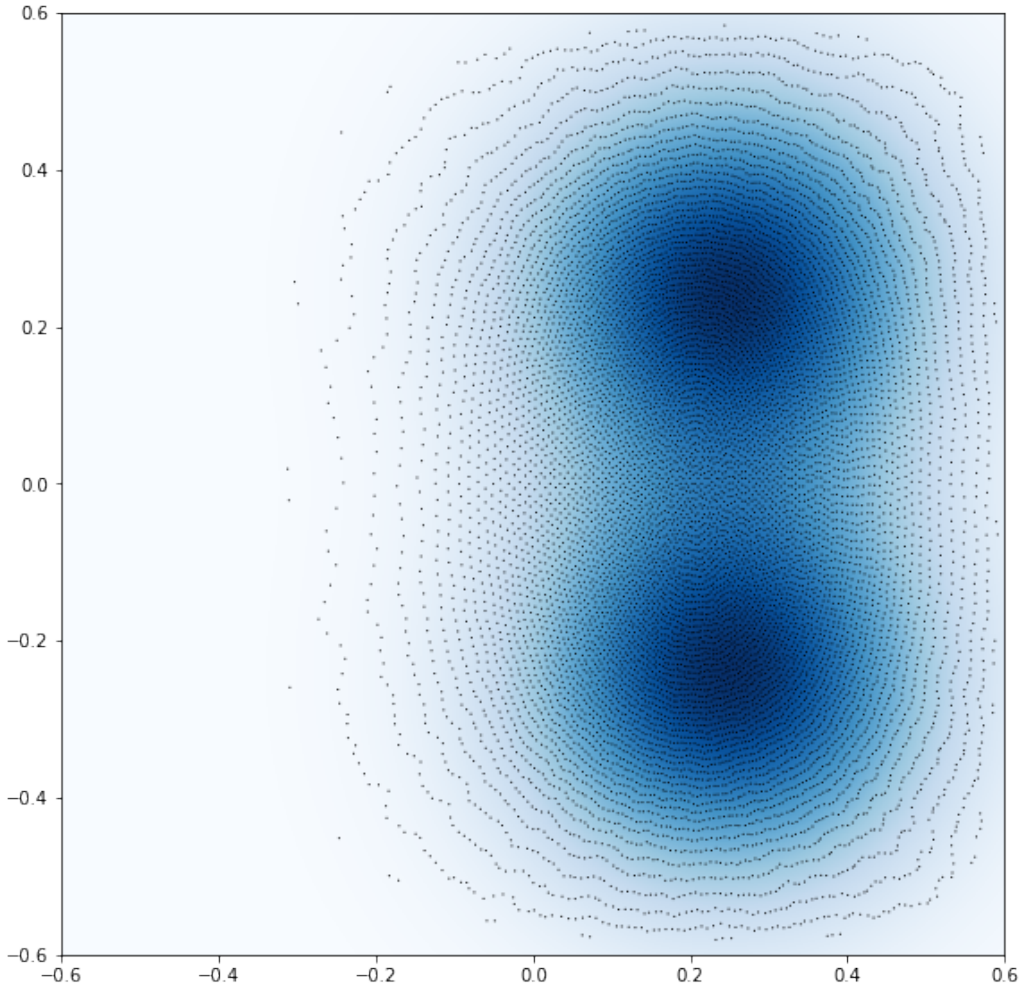
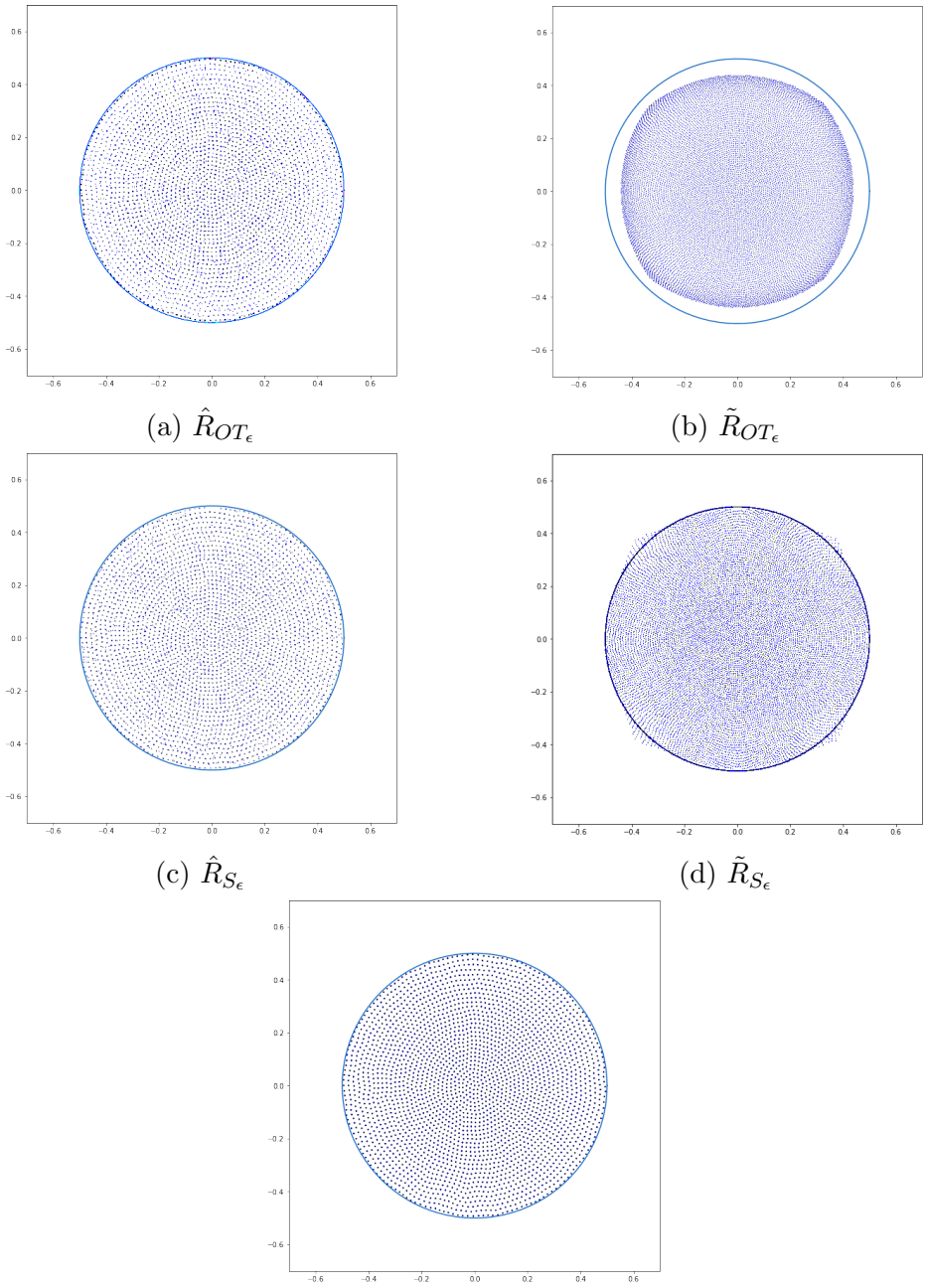


Figure 3.9: Test Case 5 : Ray-traced reflection $\bar{\nu}^M$ from $\tilde{R}_{S_{1/8\sqrt{N}}}$(point cloud) together with the exact desired distribution (color graph)



(a) $\hat{R}_{OT_\epsilon}$ (b) $\tilde{R}_{OT_\epsilon}$

(c) $\hat{R}_{S_\epsilon}$ (d) $\tilde{R}_{S_\epsilon}$

(e) ν_N

Figure 3.10: Test Case 1 : From left to right and from top to bottom the ray-traced images of μ^M (M=128*128) using the reflectors $\hat{R}_{OT_\epsilon}$, $\tilde{R}_{OT_\epsilon}$, $\hat{R}_{S_\epsilon}$, $\tilde{R}_{S_\epsilon}$ (check definition 9 for the explanation of the notations) and finally also the QMC discretization used for ν_N, $N = 64 * 64$ points. The regularization parameter was taken to be $\epsilon = \frac{1}{2*64}$ for all four solutions.

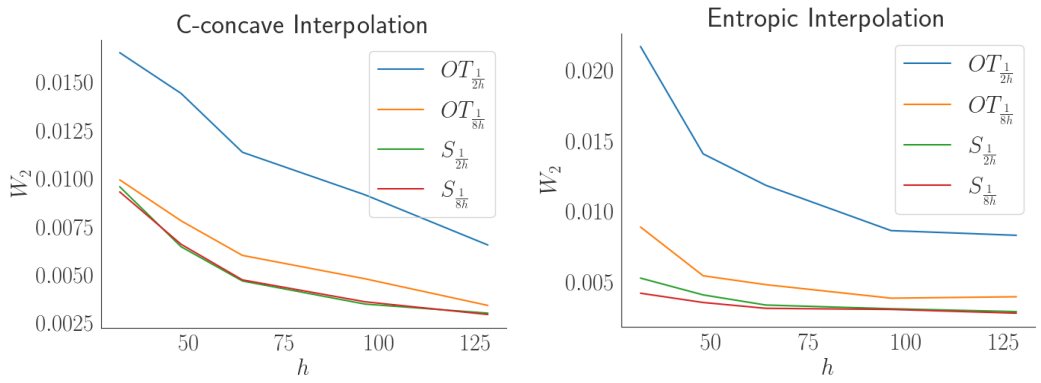


Figure 3.11: Test Case 1 ($h := \sqrt{N}$) : W_2 distance between ray-traced image with 128^2 points and exact target.

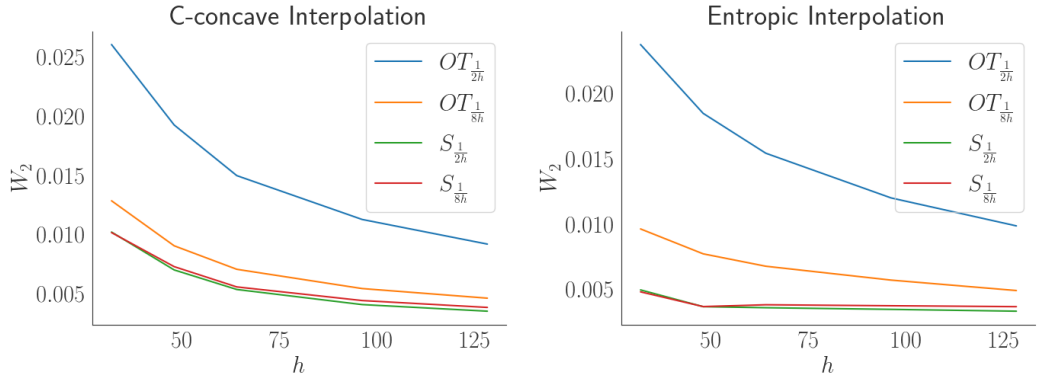


Figure 3.12: Test Case 2 ($h = \sqrt{N}$) : W_2 distance between ray-traced image with 128^2 points and exact target.

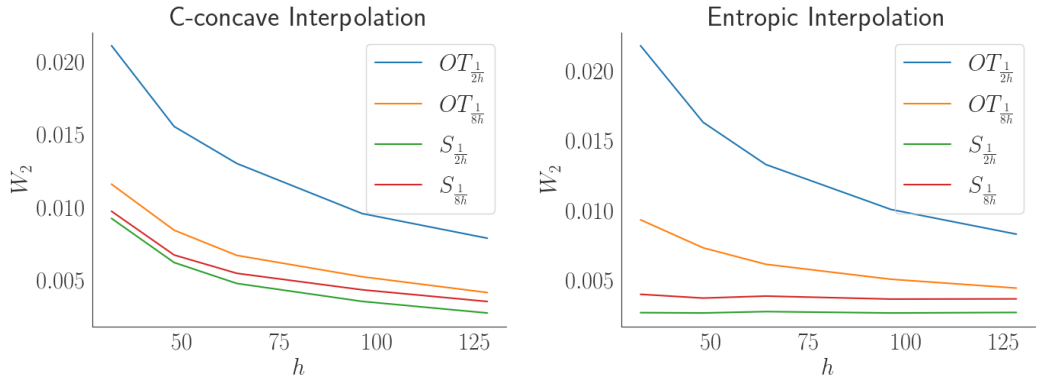


Figure 3.13: Test Case 3 ($h = \sqrt{N}$) : W_2 distance between ray-traced image with 128^2 points and exact target.

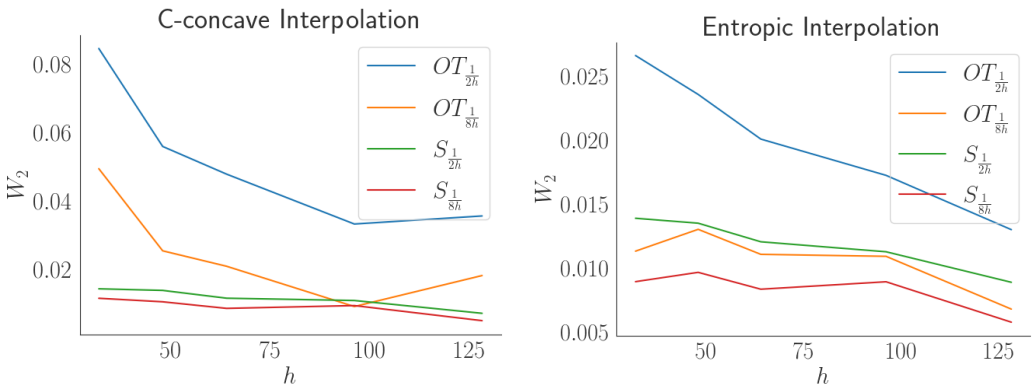


Figure 3.14: Test Case 4 ($h = \sqrt{N}$) : W_2 distance between ray-traced image with 128^2 points and exact target.

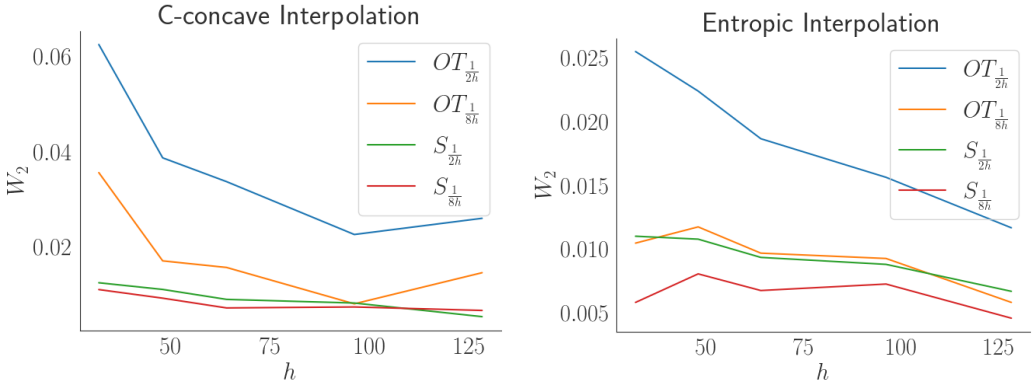


Figure 3.15: Test Case 5 ($h = \sqrt{N}$) : W_2 distance between ray-traced image with 128^2 points and exact target.

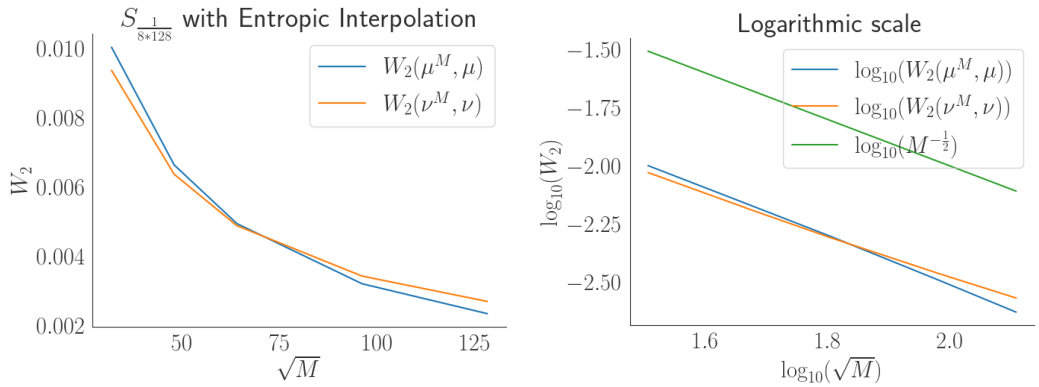


Figure 3.16: Test Case 1 : Convergence of error terms $W_2(\mu^M, \mu)$, $W_2(\bar{\nu}^M, \nu)$, original (left) and logarithmic (right) scales.

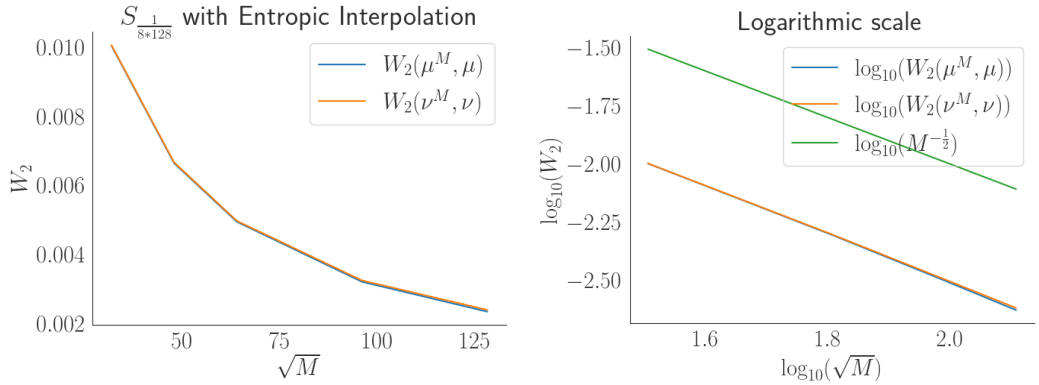


Figure 3.17: Test Case 3 : Convergence of error terms $W_2(\mu^M, \mu)$, $W_2(\bar{\nu}^M, \nu)$, original (left) and logarithmic (right) scales.

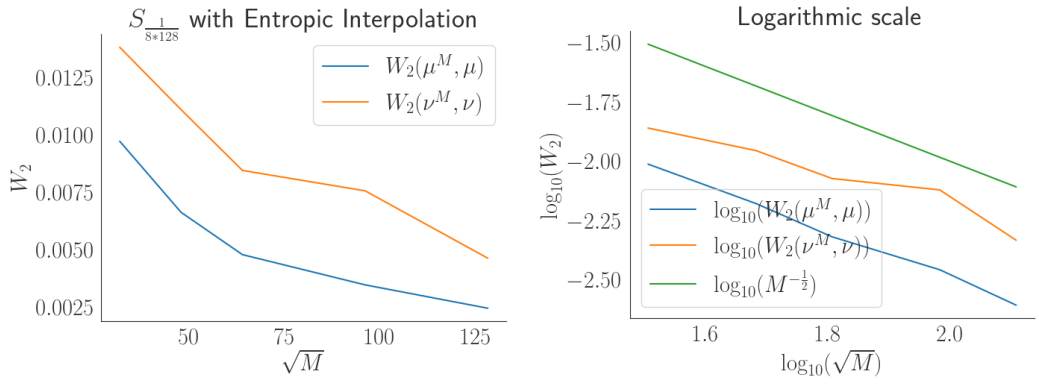


Figure 3.18: Test Case 5 : Convergence of error terms $W_2(\mu^M, \mu)$, $W_2(\bar{\nu}^M, \nu)$, original (left) and logarithmic (right) scales.

Chapter 4

Optimal transport regularization of the extended source problem

Introduction

In this chapter, we present the study of the extended source problem. We restrict our attention to the reflectors that are generated as the optimal transport solutions of a point source problem and analyze the relation between the extended source and point source problems: Let $\mathcal{R}_{f_0}$ be a reflector obtained by solving the point source problem with source $\mu_0 \in \mathcal{P}(\mathbb{S}^{d-1}_+)$ and target $\nu_0 \in \mathcal{P}(\mathbb{S}^{d-1}_-)$ and let $\mathcal{F}(\nu_0)$ be the reflection of extended source μ from this reflector. The reflector will inherit the regularity of the optimal transport point source problem.

We study the properties of this map $\mathcal{F} : \mathcal{P}(\mathbb{S}^{d-1}_-) :\to \mathcal{P}(\mathbb{S}^{d-1})$ and prove its continuity under some regularity assumption for the data. Note that if we can invert $\mathcal{F}$, then we can solve the extended source problem for the desired target $\nu \in \mathcal{P}(\mathbb{S}^{d-1}_-)$ by finding $\nu_0 = \mathcal{F}^{-1}(\nu)$ and building the corresponding point source reflector. Even though this parametrization fixes the over-determined feature of the initial extended source problem (see remark 1.2), $\mathcal{F}$ is formally a very non-linear convolution. We resort to the minimization of the residual $\nu_0 \mapsto \mathcal{L}(\mathcal{F}(\nu_0), \nu)$ with an ad-hoc misfit/loss function $\mathcal{L}$.

In the final section, we present our numerical simulations, analyzing the choice of different optimization strategies, choice of ray-tracing and dependence on the reflector height.

Through this chapter, we will work with the $d = 2$ case in order to emphasize the effect of the extended source problem and how to tackle it, clearly

seen in this simple case as well, while avoiding non-uniform or weighted grids, discussed in Chapter 3, necessary for the $d = 3$ case.

4.1 The extended source reflector parametrized by the point source problem

In this section, we analyze the relation between the extended source and point source problems. For a given extended source μ and a desired height $h_{\mathcal{R}}$ of the reflector, we take μ_0 as a point source distribution and consider the reflector generated by the solution of the point source problem with some target distribution ν_0. Then we compute the reflection ν of extended source μ from this reflector and study the relationship between ν_0 and ν.

The reflector $\nu_0 \to \mathcal{R}$ at a given height $h_{\mathcal{R}}$ is uniquely constructed from f_0 the Kantorovich potential of the point source optimal transport problem:

$$\mathcal{R}_{f_0} := \left\{ \vec{x}_0 e^{f_0(x_0)} | x_0 \in X_0,\ f_0(\pi/2) = h_{\mathcal{R}} \right\} \tag{4.1}$$

The reflector curve is a function over the point source angles $x_0 \in X_0 \subset \mathbb{S}^1_+$ from the origin $O_0 = O_{\mathbb{R}^2}$. For simplicity, we take the source S to be an interval on the horizontal axis, $[-0.5, 0.5]$.

Recall (reflection law (1.1)) that the reflection map T_0 for the light emitted from the origin O is given by the outward normal n_0 of the reflector:

$$T_0(x_0) = \vec{x}_0 - 2\langle \vec{x}_0, \vec{n}_0(x_0) \rangle \vec{n}_0(x_0) \tag{4.2}$$

In order to get an expression for the reflection map from other source points $O_s := (s, 0)$, $s \in S$, we need to express the normal in the corresponding angle parametrization denoted x_s.

The regularity of the reflectors obtained by solving the point source problem using optimal transport, summarized in proposition 3.4, allows to have the following re-parametrization using the angle parameters $x_s \in X_s$ from the other source points $O_s := (s, 0)$, $s \in S$:

Proposition 4.1 (Re-parametrization of $\mathcal{R}_{f_0}$)**.** *Let us assume that $X_0 = X_s = \mathbb{S}^1_+$ for all s and $h_{\mathcal{R}} > 1$. Then, for all $s \in S$ there exists a reparametrization A_s of X_0 into X_s and a function $f_s : X_s \to \mathbb{R}^+$ (see figure 4.1) such that:*

(i) The following re-parametrization of the reflector holds :

$$\mathcal{R}_{f_0} = \mathcal{R}_{f_s} := \{ \vec{x}_s\, e^{f_s(x_s)},\ x_s \in X_s \} \tag{4.3}$$

(ii) *The outward normal angle in the (4.3) parametrization, denoted* $n_s(x_s)$ *is given by*

$$\tan(n_s(x_s)) = \frac{\partial_{x_s} f_s(x_s)\,\cos(x_s) + \sin(x_s)}{\cos(x_s) - \partial_{x_s} f_s(x_s)\,\sin(x_s)} \tag{4.4}$$

(iii) *Assuming that the densities* μ_0 *and* ν_0 *are from* $H_{b,\alpha}$ *for some* $0 < \alpha < 1$ *and* $b > 0$*, the map* $f_0 \mapsto f_s$ *is continuous for the* $\mathcal{C}^1(\mathbb{S}^1_+)$ *topology.*

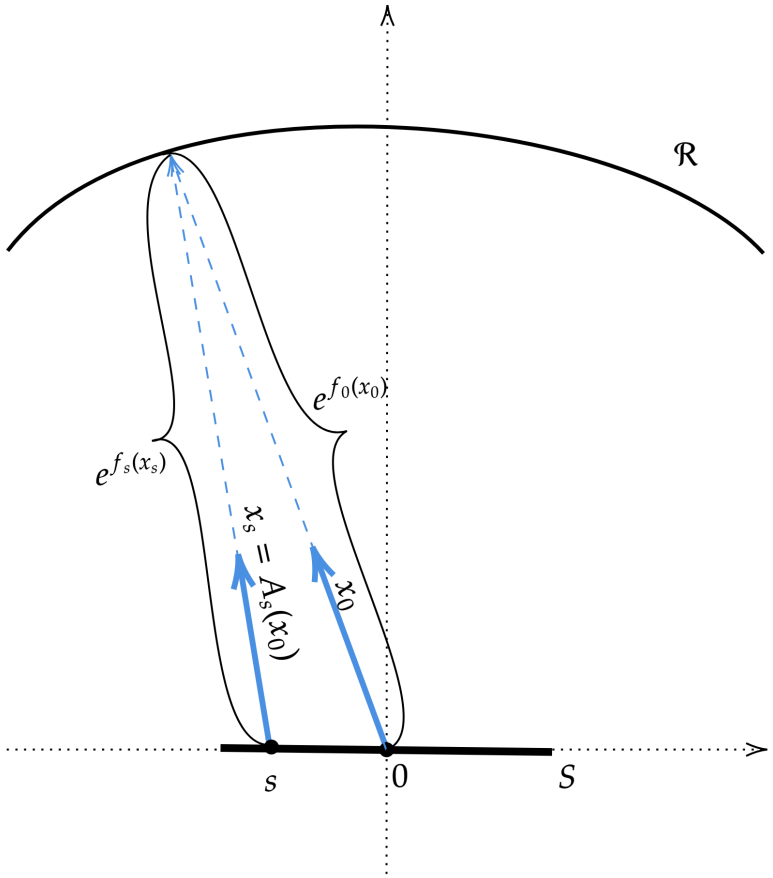


Figure 4.1: Reparametrization from $x_0 \in X_0$ to $x_s \in X_s$

Proof. A preliminary is to verify that the source patch S remains strictly inside the reflector convex envelope. Based on the envelope of parabolae property (see chapter 3.1), the abscissa of the intersection of the reflector with the axis supporting the patch (orthogonal to $\frac{\vec{\pi}}{2}$) is bounded below by the abscissa of the intersection of the axis with the parabola: $x_0 \in \mathbb{S}^1_- \mapsto \frac{C}{1-\cos(\pi - x_0)}$ and

symmetrically $x_0 \in \mathbb{S}^1_- \mapsto \frac{C}{1-\cos(x_0)}$ which is reached respectively for $x_0 = 0$ and $x_0 = \pi$. The constant is fixed by $\frac{C}{1-\cos(\frac{3\pi}{2})} = h_{\mathcal{R}}$ as in (1.5) and therefore the intersection is bounded below by $\frac{h_{\mathcal{R}}}{2}$, hence the condition $h_{\mathcal{R}} > 1$.

(i) All points $O_s = (0, s)$ on S are therefore in the convex envelope of $\mathcal{R}_{f_0}$. The strict convexity of the reflector therefore guarantees that any point $(s, 0)$ can be connected to any point on the reflector without intersecting the reflector anywhere else. This provides the re-parametrization (4.3) and also the uniqueness of $x_s \mapsto f_s(x_s)$.

(ii) is a direct consequence of the parametrization (4.3).

(iii) Define $A_s(x_0)$ as the angle of the vector connecting the shifted source O_s to the point $\mathcal{R}_{f_0}(x_0) := \vec{x}_0\, e^{f_0(x_0)}$. Using parametrization (4.3) one has

$$A_s(x_0) := \arccos\left(\frac{\cos(x_0)e^{f_0(x_0)} - s}{\sqrt{e^{2f_0(x_0)} - 2se^{f_0(x_0)}\cos(x_0) + s^2}}\right) \tag{4.5}$$

and

$$f_s(A_s(x_0)) := \log\left(\sqrt{e^{2\,f_0(x_0)} - 2se^{f_0(x_0)}\cos(x_0) + s^2}\right). \tag{4.6}$$

The map $x_0 \mapsto A_s(x_0)$ is bijective, smooth and differentiable like f_0 (a consequence of proposition 3.4). By construction at all points on the reflector $n_0(x_0) = n_s(A_s(x_0))$ where n_0 and n_s are respectively the normals of parametrizations $\mathcal{R}_{f_0}$ and $\mathcal{R}_{f_s}$. Taking the derivative in x_0 we get

$$\partial_{x_0} A_s \partial_{x_0} n_s(A_s) = \partial_{x_0} n_0 \tag{4.7}$$

and using the strict convexity of the reflector (proposition 3.4), we find that $\partial_{x_0} A_s$ cannot vanish. Applying the inverse function theorem A_s is therefore a diffeormorphism from S^1_+ onto itself. We can now re-write f_s using the new parametrization $x_s = A_s(x_0)$:

$$f_s(x_s) = \log\left(\sqrt{e^{2f_0(A_s^{-1}(x_s))} - 2s\cos(A_s^{-1}(x_s))e^{f_0(A_s^{-1}(x_s))} + s^2}\right) \tag{4.8}$$

Which is continuously differentiable with derivative:

$$\begin{aligned}\partial_{x_s} f_s(x_s) =& \frac{e^{f_0(A_s^{-1}(x_s))}(\partial_{x_s} A_s^{-1})(x_s)}{e^{2f_0(A_s^{-1}(x_s))} - 2s\cos(A_s^{-1}(x_s))e^{f_0(A_s^{-1}(x_s))} + s^2}\, Q \\ Q :=& e^{f_0(A_s^{-1}(x_s))}\partial_{x_0} f_0(A_s^{-1}(x_s)) + s\sin(A_s^{-1}(x_s)) \\ & - s\cos(A_s^{-1}(x_s))\partial_{x_0} f_0(A_s^{-1}(x_s))\end{aligned} \tag{4.9}$$

Note that f_s and $\partial_{x_s} f_s(x_s)$ are expressed analytically using f_0, $\partial_{x_0} f$, A_s^{-1} and $\partial_{x_s} A_s^{-1}$ (as fractions with non-vanishing denominators and bounded numerators). Therefore, in order to demonstrate C^1 convergence, all we need to do is to establish the continuous dependency of A_s^{-1} and $\partial_{x_s} A_s^{-1}$ on f_0.

This is just a consequence of the inverse function theorem: As $\partial_{x_0} A_s$ depends continuously on f_0 and $\partial_{x_s} A_s^{-1}$ can be expressed by $1/\partial_{x_0} A_s$.

The continuity of $f_0 \mapsto f_s$ for the C^1 topology follows.

□

Using this re-parametrization, we can express T_s, the reflection map from the reflector $\mathcal{R}$ for rays coming from the point $s \in S$, in terms of now re-parametrized normal n_s:

$$\vec{T_s}(x_s) = \vec{x}_s - 2\langle \vec{x}_s, \vec{n}_s(x_s) \rangle \vec{n}_s(x_s) \tag{4.10}$$

Where n_s can be given either using (4.4) or using n_0 and an inverse of the transform A_s:

$$n_s(x_s) = n_0(A_s^{-1}(x_s)) \tag{4.11}$$

As discussed in Chapter 1.3, the reflected light distribution ν from the given reflector $\mathcal{R}$ illuminated by the extended source of light S with a distribution μ can be given in terms of T_s. Following this construction, we define the "Forward Map" $\nu_0 \to \mathcal{F}(\nu_0)$ as the reflection of μ from $\mathcal{R}_{\nu_0}$:

$$\mathcal{F}(\nu_0) := \int_S \nu_s \, ds \text{ where } \nu_s = T_{s\#}\mu_s \tag{4.12}$$

Note that using proposition 4.1, we can construct T_s from f_s and n_s. Therefore, we can summarize the construction of the map $\mathcal{F}$ in the following steps:

A: $\nu_0 \mapsto f_0$ (or equivalently, T_0)
B: $f_0 \mapsto \{f_s\}_{s\in S}$ (or equivalently, $\{T_s\}_{s\in S}$)
C: $\{f_s\}_{s\in S} \mapsto \mathcal{F}(\nu_0) := \int_S T_{s\#}\mu_s \, ds$

Theorem 4.2 (Continuity of $\mathcal{F}$)**.** *Under the assumptions of proposition (3.4)-(4.1) and assuming that $\mu_s \ll \mu_0$ for all $s \in S$, $\mathcal{F}$ is continuous for the weak convergence in $H_{b,\alpha}(\mathbb{S}^1_-) \subset \mathcal{P}(\mathbb{S}^1_-)$*

Proof. Throughout this proof, for a target density ν_0 we will denote by superscript $^{\nu_0}$ the mathematical objects induced by solving the point source optimal transport problem between μ_0 and ν_0, e.g. f^{ν_0} will be a Kantorovich potential for this problem, T^{ν_0} will be a reflection map from $\mathcal{R}_{f_0^{\nu_0}}$ and n^{ν_0} will be a normal of $\mathcal{R}_{f_0^{\nu_0}}$.

Let $\{\nu_{0,k}\}_{k\in\mathbb{N}}$ be a weakly convergent sequence in $H_{b,\alpha}(\mathbb{S}^1_-)$ converging to ν_0. We need to verify that $\mathcal{F}(\nu_{0,k})$ also converges weakly to $\mathcal{F}(\nu_0)$. As $\mathcal{F}(\nu_0)$ can be expressed as $\int_S T^{\nu_0}_{s\#}\mu_s$, by the linearity of the integration, all we need is to verify that $T^{\nu_{0,k}}_{s\#}\mu_s$ converge weakly to $T^{\nu_0}_{s\#}\mu_s$ for all s.

The stability of $\nu_0 \mapsto T_0^{\nu_0}$ is a classical result (Theorem 2.6), where the convergence of sequence $(T_0^{\nu_{0,k}})$ (built from (2.11) using the sequence of $(f_{0,k})$) holds in probability:

$$\forall \epsilon > 0 \quad \mu_0\left[\left\{x \in \mathbb{S}_+ \,|\, d(T_0^{\nu_0}(x), T_0^{\nu_{0,k}}(x)) > \epsilon\right\}\right] \xrightarrow{k\to\infty} 0 \tag{4.13}$$

Note that under the assumptions, $T_0^{\nu_0}$ is induced by the reflection from the reflector with the continuous normal $n = n^{\nu_0}$ (see (4.2)). Hence the convergence of $T_0^{\nu_{0,k}}$ is equivalent to the convergence of $n^{\nu_{0,k}}$. Since we have the C^1 continuity of f_0 to f_s, and since for all s, $T_s^{\nu_{0,k}}$ is just another reflection using the reparametrized normal $n_s^{\nu_{0,k}}$, the above convergence in probability holds also for all $T_s^{\nu_{0.k}}$.

Finally we check that for the sequence of maps $T_s^{\nu_{0.k}} : \mathbb{S}_+ \to \mathbb{S}_-$, converging in probability to the map $T_s^{\nu_0}$ with respect to the measure μ_0, the pushforward measures ${T_s^{\nu_{0.k}}}_{\#}\mu_s$ converge weakly to ${T_s^{\nu_0}}_{\#}\mu_s$. For this we check the following convergence for all bounded Lipschitz functions ϕ (see remark 2.1):

$$\int_{\mathbb{S}_-} \phi(y) {T_s^{\nu_{0.k}}}_{\#}\mu_s(y) \xrightarrow{k\to\infty} \int_{\mathbb{S}_-} \phi(y) {T_s^{\nu_0}}_{\#}\mu_s(y) \tag{4.14}$$

Fix the function ϕ with a Lipschitz constant L_ϕ and an $\epsilon > 0$. Using the change of variable formula, Lipschitz property of ϕ and boundedness of $\mathbb{S}$ we get:

$$\left|\int_{\mathbb{S}_-}\phi(y){T_s^{\nu_{0.k}}}_{\#}\mu_s(y)-\int_{\mathbb{S}_-}\phi(y){T_s^{\nu_0}}_{\#}\mu_s(y)\right|\leq$$
$$\int_{\mathbb{S}_+}\left|\phi(T_s^{\nu_{0.k}}(x))-\phi(T_s^{\nu_0}(x))\right|\mu_s(x)\leq$$
$$\int_{\mathbb{S}_+}L_\phi d_{\mathbb{S}}(T_s^{\nu_{0.k}}(x),T_s^{\nu_0}(x))\mu_s(x)\leq$$
$$\int_{\mathbb{S}_+\setminus B_\epsilon}\epsilon L_\phi\mu_s(x)+\int_{B_\epsilon}diam(\mathbb{S}_+)L_\phi\mu_s(x)\leq$$
$$\epsilon L_\phi+diam(\mathbb{S}_+)L_\phi\mu_s[B_\epsilon] \tag{4.15}$$

Where B_ϵ is the set:

$$B_\epsilon:=\left\{x\in\mathbb{S}_+\,|\,d(T_s^{\nu_0}(x),T_s^{\nu_{0.k}}(x))>\epsilon\right\}.$$

Since $T_s^{\nu_{0.k}}$ converge in probability w.r.t. μ_0, the quantity $\mu_0[B_\epsilon]\xrightarrow{k\to\infty}0$. This, together with the absolute continuity $\mu_s<<\mu_0$ implies $\mu_s[B_\epsilon]\xrightarrow{k\to\infty}0$, which concludes the proof. □

The forward map $\mathcal{F}$ plays a crucial role in this work. It can be interpreted as a non-linear convolution, as detailed in the following remark.

Remark 4.3 ($\mathcal{F}$ is formally a non-linear convolution)**.** *We start back from (4.12) and assume (to simplify the exposition) that $\mu_s=\mu_0$ for all s. Meaning that the radiation pattern is identical for all points on the finite source. Using the change of variable formula with the maps T_s, we get, for all $y\in supp(\mathcal{F}(\nu_0))$:*

$$\begin{aligned}\left(\mathcal{F}(\nu_0)\right)(y)&=\int_S T_{s\#}\mu_0(y)ds\\&=\int_S\mu_0\left(T_s^{-1}(y)\right)\left(\partial_{x_s}T_{s|T_s^{-1}(y)}\right)^{-1}ds\\&=\int_S\nu_0\left(T_0\circ T_s^{-1}(y)\right)\partial_{x_0}T_{0|T_s^{-1}(y)}\left(\partial_{x_s}T_{s|T_s^{-1}(y)}\right)^{-1}ds\end{aligned}$$

We now assume that for a fixed y the mapping $s\mapsto\mathcal{Y}_y(s)=T_0\circ T_s^{-1}(y))$ (see figure 4.2) is injective and make the change of variable

$$\left(\mathcal{F}(\nu_0)\right)(y)=\int_{\mathcal{Y}_y(S)}\nu_0(y')\partial_{x_0}T_{0|T^{-1}_{\mathcal{Y}_{y'}^{-1}(y)}(y)}\left(\partial_{x_s}T_{\mathcal{Y}_{y'}^{-1}(y)|T^{-1}_{\mathcal{Y}_{y'}^{-1}(y)}(y)}\right)^{-1}dy' \tag{4.16}$$

The formula is complicated but shows that ν_0 is convolved with a Kernel involving the Jacobians of the maps T_s and T_0:

$$\mathcal{K}(y,y') := \partial_{x_0} T_{0|T^{-1}_{\mathcal{Y}^{-1}_{y'}(y)}(y)} \left(\partial_{x_s} T_{\mathcal{Y}^{-1}_y(y')|T^{-1}_{\mathcal{Y}^{-1}_y(y')}(y)} \right)^{-1} \tag{4.17}$$

and depending on the reflector, hence also on ν_0. The map $s \mapsto \mathcal{Y}_y(s)$ can be interpreted as follows (see also figure 4.2) : given a reflection direction y and a point s on the finite source, find the shooting angle x_s from that point. Then shoot a ray from the center source point O_0 with the same angle $x_0 = x_s$ and record the outgoing angle. Given two outgoing angles y and y', its inverse returns the coordinate on the finite source for which the shooting angle with y' reflection is the same as the shooting angle from O_0 yielding y. For instance $y = y'$ gives $\mathcal{Y}_y^{-1}(y) = 0$ and $\mathcal{K}(y,y) = 1$.

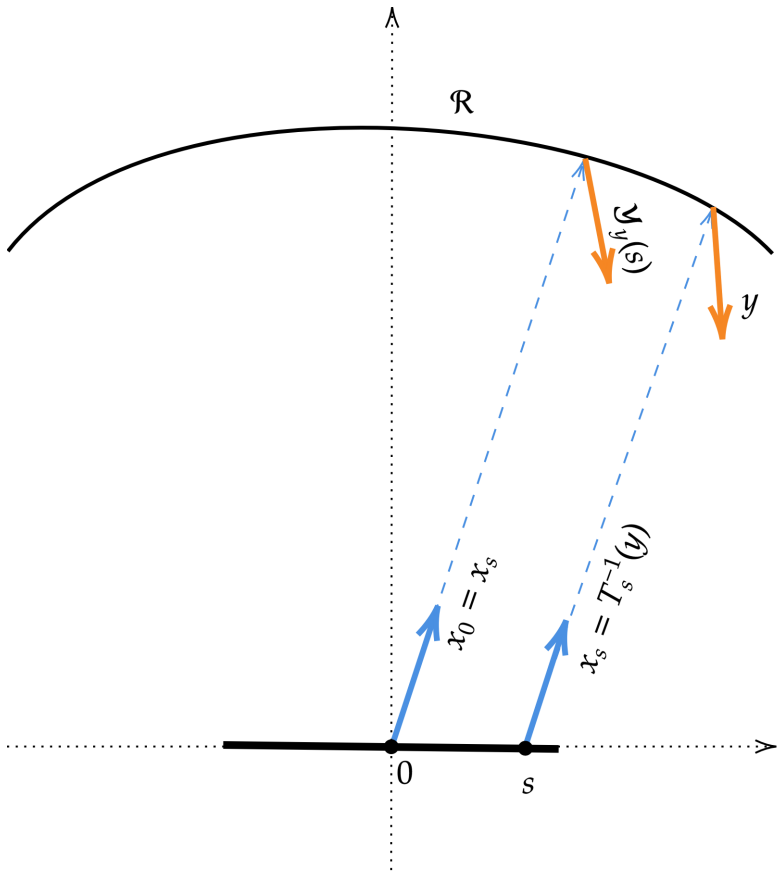


Figure 4.2: The map $s \to \mathcal{Y}_y(s)$ defined as $T_0 \circ T_s^{-1}(y)$

The convolutional nature of the forward map $\mathcal{F}$ is known in the optics community, although we are not aware of any formal description of the convolution kernel in terms of the reflection map. In some works (see [WZLM21] and references therein) the kernel is approximated to perform the deconvolution. This approximation either happens by disregarding the dependence on the variable y, or by restricting source and target measures to some simple shapes.

4.2 Regularization of the extended source problem

As already mentioned, if we can invert the map $\mathcal{F}$, then we can solve the extended source problem for the desired target $\nu \in \mathcal{P}(\mathbb{S}^{d-1}_{-})$ by finding $\nu_0 = \mathcal{F}^{-1}(\nu)$ and building the corresponding point source reflector. Even though this parametrization fixes the over-determined feature of the initial extended source problem (see remark 1.2), non-linear convolution nature of $\mathcal{F}$ makes it problematic to invert. We resort to the minimization of the residual $\nu_0 \mapsto \mathcal{L}(\mathcal{F}(\nu_0), \nu)$ with an ad-hoc misfit/loss function $\mathcal{L}$.

Overall, this approach fits in the framework of a regularisation approach as is customary for ill-posed non-linear inverse problems (see for instance [BB18] for a recent review). It is a concept of trying to find a "best approximation" of the solution within some class, with respect to some loss/misfit function. In terms of the reflector problem for the desired target ν, it is based on the following ingredients:

1. *A parametric set* of admissible reflectors. This is a regularization part.

2. *A forward map*, which for a given source distribution and an element of the parameter set (corresponding to the reflector), produces the reflected distribution.

3. *A misfit/loss function* that gives information about the "closeness" of the reflected and desired distributions. Ideally, this should be a distance, or at least convex, positive and reaching the minimum value of 0 only when the reflected distribution is equal to the desired one.

Then the regularized solution is the reflector within the parametric set, reflection from which minimizes the loss function, and the value of the loss function is a measure of its quality.

In the following sections, we discuss those ingredients and our choices, that provide the compact parameter set and the continuous loss, guaranteeing the existence of a minimizer.

4.2.1 Parameter set and the forward map

In general, parametric set of admissible reflectors could be any discrete parameterization of curves (Splines, Bezier ....) where parametrizing variable would be control points and tangents for example (see e.g. [BB19] [BKM$^+$20]). Although such parameterizations are associated with extra complications when trying to guarantee some desired properties of the reflector, e.g convexity/concavity.

The point source problem is sometimes used to approach the extended source problem. For example, this is the approach followed in [FCR10] (see also [WZLM21], [LFHL10] and the references therein). This is also the approach pursued in this work, in order to leverage the relation between the point source and extended source problem, discussed in the previous section:

We take parameter set to be the set of probability measures $\mathcal{P}(\mathbb{S}_-)$ and denote the parameter by ν_0. The reflector is constructed using the Kantorovich potential f_0 of the point source optimal transport problem between μ_0 and ν_0, as in 4.1. This choice guarantees the regularity properties of the reflector, summarized in proposition 3.4.

The forward map can be implemented in practice using the ray-tracing (chapter 1.4). Formally, in our case we take the forward map to be (4.12) from the previous section, summarized in the following steps:

A: $\nu_0 \mapsto f_0$ (or equivalently, T_0)
B: $f_0 \mapsto \{f_s\}_{s\in S}$ (or equivalently, $\{T_s\}_{s\in S}$)
C: $\{f_s\}_{s\in S} \mapsto \mathcal{F}(\nu_0) := \int_S T_{s\#}\mu_s \, ds$

4.2.2 The loss function

In this work, we use an optimal transport based loss. In Chapter 3.2.5, we discussed the use of W_2 distance on the space $\mathbb{S}^1_-$ for error estimation for the point source problem. Since the image of the forward map for the extended source problem is also a probability distribution on $\mathbb{S}^1_-$, it is possible to use $W_2(\mathcal{F}(\nu_0), \nu)$ as a loss. However, as we plan to use gradient-based optimization methods, we avoid taking the square root and instead work with the squared functional $W_2^2(\mathcal{F}(\nu_0), \nu)$. We approximate this value using Sinkhorn divergences (2.30), an accurate approximation for small ϵ, which is

also smooth (see Theorem 2.14). In order to get a dimensionless loss, we also use the normalization :

$$J(\nu_0) := \mathcal{L}(\mathcal{F}(\nu_0), \nu) := S_\epsilon(\mathcal{F}(\nu_0), \nu)/S_\epsilon(\mathcal{F}(\nu), \nu) \tag{4.18}$$

Recall that the denominator can only vanish if the desired prescribed target ν is equal to $\mathcal{F}(\nu)$, but due to the convolution-like nature of the map $\mathcal{F}$ (see remark 4.3), this can not happen as long as the reflector height $h_\mathcal{R} \neq \infty$ and source S contains more than one point.

Our choices of the parametric set, the forward map, and the loss guarantee the existence of a minimizer, as summarized in the following:

Theorem 4.4. *The global cost we minimize is the composition of the forward map $\mathcal{F}$ and the loss. If the loss is continuous for the weak topology on measures (like (4.18)) and the forward map is continuous (Theorem 4.2), the compactness of $H_{b,\alpha}(Y)$ for compact $Y \subset \mathbb{S}_-$ (remark 3.3) guarantees the existence of a minimizer.*

Also note that although Sinkhorn divergence is convex with respect to its input measures, we are not guaranteed that this will carry to (4.18), since we obtain it by composing Sinkhorn divergence with a forward map $\mathcal{F}$. In practice, we can initialize the optimizations using the point source solutions, which provide a good first approximation, which is crucial in avoiding local minimizers which are not global (if such exists).

Illustration

Figure 4.3 is a plot of the Loss value using a parametric family of point source target mixing two Gaussians with varying expectations:

$$(t, t') \in [0, 1]^2 \mapsto \mathcal{L}\left(\mathcal{F}(\nu_0^{(t,t')}), \nu\right) \tag{4.19}$$

where $\nu := \mathcal{F}(\nu_0^{(\frac{1}{3},\frac{1}{3})})$ and

$$\nu_0^{(t,t')} = \mathcal{N}_{u(t),\frac{\pi}{21}} + \mathcal{N}_{v(t'),\frac{\pi}{24}}, \quad u(t) = \frac{20\pi}{16} + t\frac{2\pi}{16}, \quad v(t') = \frac{26\pi}{16} + t'\frac{2\pi}{16}.$$

The graph is smooth and convex within the observed domain.

Remark 4.5 (On the differentiability of J)**.** *Computation of the loss J can be summarized into 3 steps:*

A: Computation of the solution of the point source problem and construction of the reflector.

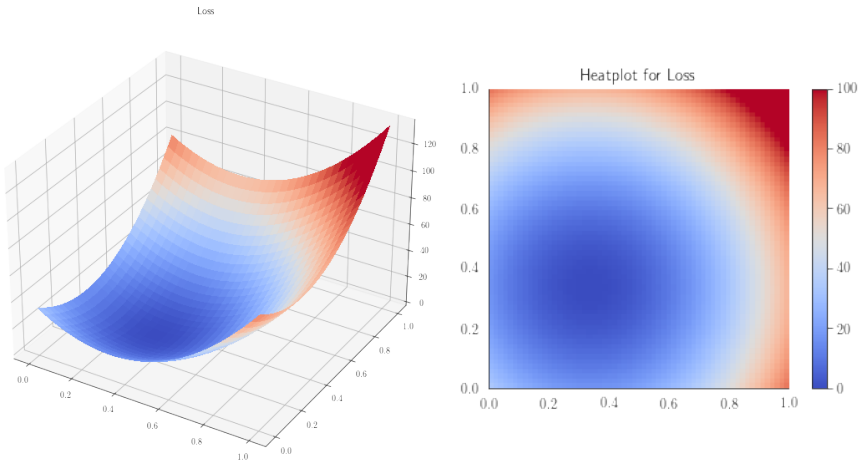


Figure 4.3: Graph plot of (4.19).

B: Computation of the forward map $\mathcal{F}(\nu_0)$.

C: Computation of the Sinkhorn divergence between the obtained measure $\mathcal{F}(\nu_0)$ *and the desired measure* ν.

For step C, the differentiability of the discrete entropic optimal transport with respect to the input data is given in 2.13, which by definition also apples to the Sinkhorn divergences. Also in step A we use the Sinkhorn divergence potential f_{S_ϵ}*, which is differentiable with respect to* ν_0.

But for step B, the definition of the forward map itself depends on proposition 4.1 and the convexity of the reflector. This is only established for the non-entropic $\epsilon = 0$ *reflector.*

In practice, f_{S_ϵ} *is a good approximation of* f_0 *and we used the autodifferentiation and gradient-based optimization methods of* `Pytorch` *without any difficulties.*

Remark 4.6 (Other possible choices of the loss function)**.** *In [BKM+20], the regularization method is applied to another optical system, extended source to far-field lens problem for* $d = 3$*. They use the* L_2 *norm between the densities as a loss, based on the point-wise comparison of desired and obtained target densities:*

$$J_{L_2}(\nu_0) := \sqrt{\frac{1}{D}\int_S \int_{\mathbb{S}^2_+} \left(\nu(y) - \mathcal{F}(\nu_0)(y)\right)^2 dsdy} \tag{4.20}$$

It is also worth noting that in [BKM+20] the parameter set used with this loss is not the set of probability measures, but the set of coefficients of the bi-cubic splines used for parametrizing the lens surface.

Also, the simple Gold method ([Gol64]) is used in the context of freeform optics in industry and also in [FCR10], but only based on heuristics. In remark 4.8, we discuss that this method is a simplified minimization scheme of the Kullback-Leibler divergence (2.22).

4.3 Optimization methods for minimizing J

Optimization problems arise in various fields, e.g. statistics, economics, physics etc, therefore there have been various methods devised for solving them through the history of mathematics. With the rise of machine learning and the need to train artificial neural networks, there has been a surge in the development of new efficient and easily implementable optimization methods (see e.g. [Rud17] for a recent review).

Here we present the methods that we used for minimizing the loss function (4.18) to solve the extended source problem. A comparison of their performance for this task is discussed in section 4.4.

4.3.1 Gradient descent

Gradient descent (also known as batched gradient descent or steepest descent) is now a standard method in optimization, that dates back to Cauchy ([Cau47]).

For the initial guess $\nu_0^{(0)}$ and a fixed step size or learning rate parameter $lr \in (0, \infty)$ gradient descent method iterates in k with an update:

$$\nu_0^{(k+1)} = \nu_0^{(k)} - lr \cdot \nabla_{\nu_0} J(\nu_0^{(k)}). \tag{4.21}$$

Convergence of this method (and its speed) depends on the differentiability and convexity properties of J, as well as the choice of the parameter lr. For the comprehensive analysis see e.g. [Rus06]. In practice, this form of gradient descent is rarely used, and instead is enhanced by some modifications. For the recent discussion of those modifications, see e.g. [Rud17].

4.3.2 Adam algorithm

Adam (Adaptive momentum) algorithm, proposed in [KB17] is a modification of the gradient descent method. It keeps the information about already computed gradient values and uses it to adapt the learning rate.

For the initial guess $\nu_0^{(0)}$, a fixed learning rate parameter $lr \in (0, \infty)$, initial momentum parameters m_0, v_0 (usually taken to be identically 0) and parameters $\beta_1, \beta_2 \in (0, 1)$, Adam algorithm iterates in k with an update:

$$
\begin{aligned}
g_{k+1} &= \nabla_{\nu_0} J(\nu_0^{(k)}) \\
m_{k+1} &= \beta_1 m_k + (1-\beta_1) g_{k+1} \\
v_{k+1} &= \beta_2 v_k + (1-\beta_2) g_{k+1}^2 \\
\bar{m}_{k+1} &= \frac{m_{k+1}}{1-\beta_1^{k+1}} \\
\bar{v}_{k+1} &= \frac{v_{k+1}}{1-\beta_2^{k+1}} \\
\nu_0^{(k+1)} &= \nu_0^{(k)} - lr \cdot \frac{\bar{m}_{k+1}}{\sqrt{\bar{v}_{k+1}} + 10^{-8}}
\end{aligned} \tag{4.22}
$$

where all arithmetic operations are understood point-wise.

Remark 4.7 (Convergence of Adam algorithm)**.** *In [KB17], the convergence proof of the Adam algorithm was provided for the convex loss and specific choices of the parameters. However, in [RKK19] it is demonstrated that the proof has a flaw and it is possible to construct examples where Adam algorithm will not converge.*

4.3.3 Gold's method

The analogy between the forward map and a non linear convolution $\nu = \mathcal{K}_{\nu_0} \star \nu_0$ is explained in remark 4.3. Gold method is a heuristic method of de-convolution (see e.g. [Gol64]). It is used in the context of freeform optics in [FCR10].

Assuming $\mathcal{K}$ is known, for a given ν, Gold algorithm iteratively corrects ν_0 pointwise using:

$$
\nu_0^{(k+1)} := \nu_0^{(k)} \left(\frac{\nu}{\mathcal{K} \star \nu_0^{(k)}} \right)^{\alpha}, \ \alpha > 0 \tag{4.23}
$$

This is easy to implement. Clearly when convergent $\nu_{\mathcal{F}}^{(\infty)} := \mathcal{K} \star \nu_0^{(\infty)} = \nu$. In the case of the reflector problem, $\mathcal{K}_{\nu_0}$ depends on ν_0 and is given in (4.17). We can replace the convolution by the forward map $\nu_{\mathcal{F}} := \mathcal{F}(\nu_0)$. Finally remark that $\nu_0^{(k+1)}$ has to remain a probability measure, thus a re-normalisation is necessary after every iteration.

Remark 4.8 (A variational formulation of the Gold method)**.** *We explain below that (4.23) is actually linked to the following optimization problem:*

$$
\nu_0^{(k+1)} := \operatorname*{arginf}_{\nu_0} \mathrm{KL}(\nu_0 \,|\, \nu_0^{(k)}) + \alpha \ \mathrm{KL}(\nu_{\mathcal{F}} \,|\, \nu) \tag{4.24}
$$

where α is a small positive relaxation parameter and

$$\mathrm{KL}(\nu_{\mathcal{F}} \,|\, \nu) := \left\langle \log(\frac{\nu_{\mathcal{F}}}{\nu}) - 1, \nu_{\mathcal{F}} \right\rangle + \langle 1, \nu \rangle \text{ if } \nu_{\mathcal{F}} \ll \nu, \qquad (4.25)$$
$$+\infty \text{ otherwise}$$

is the Kullback-Leibler divergence already introduced in (2.22). It is strictly convex, takes its minimum at ν, and has an infinite slope at 0. Its Gateaux derivative in $\nu_{\mathcal{F}}$ is formally given by $\langle \delta \,\mathrm{KL}(\nu_{\mathcal{F}} \,|\, \nu), \delta\nu \rangle = \left\langle \log(\frac{\nu_{\mathcal{F}}}{\nu}), \delta\nu \right\rangle$. It forces $\nu_{\mathcal{F}}$ to have the same support as ν, therefore it requires in practice to bin the rays (see section 1.4.3).

For a small α, (4.24) may be interpreted as a convex penalization of the direct minimisation of the Kullback-Leibler loss:

$$\mathcal{L}_{\mathcal{KL}}(\nu_{\mathcal{F}}, \nu) := \mathrm{KL}(\mathcal{F}(\nu_0) \,|\, \nu) \qquad (4.26)$$

If the resulting sequence $(\nu_0^{(k)})$ converges it reaches a minimiser. The variational formulation (4.24) has strong analogies with the theory of Wasserstein Gradient Flows (see e.g [San15]) and some of the techniques developed in this context are likely to be applicable (for instance $\sum_k \mathrm{KL}(\nu_0^{(k+1)} \,|\, \nu_0^{(k)})$ is a convergent series).

Getting back to Gold method, the optimality condition for (4.24) leads to:

$$\log\left(\frac{\nu_0^{(k+1)})}{\nu_0^{(k)}}\right) = -\alpha \frac{\partial \mathcal{F}}{\partial \nu_0}(\nu_0^{(k+1)}) \cdot \log\left(\frac{\mathcal{F}(\nu_0^{(k+1)})}{\nu}\right) \qquad (4.27)$$

This is a non-linear implicit system in $\nu_0^{(k+1)}$, and $\frac{\partial \mathcal{F}}{\partial \nu_0}(.)$ is a Jacobian operator or matrix. If we instead replace it with an identity, (4.23) follows directly by taking the exponential of this expression and can be seen as a cheap explicit proxy of (4.24).

4.4 Numerical Results

4.4.1 Experimental setting

Reflector Height. The parameter $h_{\mathcal{R}}$ first introduced in 1.5 "measures" how close the extended source problem is to the point source problem. In our study, it will vary between 1 and 9. When $h_{\mathcal{R}} \to \infty$, the extended source problem approaches the point source problem.

Source Distribution and discretization. The source patch interval $S = [-0.5, 0.5]$ will be fixed and the measure μ_s will always be uniform in s, that is, for all $s \in S$, $\mu_s = \mu_0$. Our approach is not limited to such measures but this assumption is the simplest and common for applications. For the source distribution μ_0, plotted in figure 4.6, we chose a distribution close to uniform within some angle opening and decays rapidly outside. To achieve these requirements, we take the sum of 16 Normal distributions, with means distributed uniformly within the interval $[9\pi/32, 23\pi/32]$ and deviation $\sigma = \pi/32$.

The number of points discretizing the angle spaces $X_{0,s}$ and the source interval S, denoted respectively N_A and N_S, are chosen such that $\frac{\pi \cdot h_{\mathcal{R}}}{N_A} \simeq \frac{1}{N_S}$: the grid steps on the reflector and the source patch are of the same order. The number of rays N shot is given for backward ray tracing as $N = N_A \times N_S$. Setting N and $h_{\mathcal{R}}$ therefore also fixes the discretization size. We use N_A for the angular discretization of the supports of the targets ν_0 and ν.

On our computer[1] taking $N = 5 \cdot 10^6$ and $h_{\mathcal{R}} = 5$, the 6GB GPU memory was working at full capacity (5.8 out of 6GB) and the computation of the Loss function with backward raytracing needs approximately 30 seconds. In comparison, it takes approximately 6 seconds for each iteration with $N = 10^5$ and the used memory is approximately 1GB. This is the setting for all presented computations below.

Parametrization of ν_0. Formally, the optimization variable is $\nu_0 \in \mathcal{P}(\mathbb{S}^1_+)$. In practice it is parametrized using a classic machine learning method given below that guarantees that the optimization variable keeps a fixed total sum of 1:

The actual optimization variable is a vector $\lambda \in \mathbb{R}^{N_A}$ defined as

$$\lambda_i := \log(\nu_{0,i}) + \log(\sum_i e^{\nu_{0,i}}) \tag{4.28}$$

where the $\{\nu_{0,i}\}$s discretize ν_0 and sum to 1. The point source target entering the loss function is recovered by the inverse transform

$$\nu_{0,i} := \frac{e^{\lambda_i}}{\sum_j e^{\lambda_j}} \tag{4.29}$$

Optimal transport computations of the reflector The implementation of the reflector computation and Sinkhorn divergence is based on

[1]We run the code on the laptop with a 64bit processor: Intel Core i7-8850H CPU @ 2.60GHz x 12 and GPU: Nvidia Quadro P3200 6GB with 1792 CUDA cores.

`Pytorch` and the optimal transport platform `Geomloss` (our implementation is available at [Cha20]). As observed in Chapter 3, the performance of the Sinkhorn algorithm and the bias induced by the entropic regularization depends on a blurring parameter ϵ. As seen in Chapter 3, as long as this parameter is of order $1/N_A$, the effect of different values on the solution obtained from the Sinkhorn divergence is not crucial. Moreover, our goal is not in obtaining a perfect point source solution but a good parametrization, therefore the accuracy of the point source solution is not crucial either. Therefore we use the value $\epsilon = 1/N_A$ in all our point source computations. (note that for the computation of the Sinkhorn divergence value for the loss, $S_\epsilon(\mathcal{F}(\nu_0), \nu)$, we use the different value of $\epsilon = 0.0001$).

Forward map.
We will be using the ray-tracing as discussed in Chapter 1.4. To compute the normals we will use the entropic canonical extension (3.28). Unless otherwise stated, we will use the backward ray-tracing (Chapter 1.4.2). We will use the binning only for plotting purposes. The loss will be computed on the point cloud without binning. Binning will play a role only for the simulations involving a Gold's method, as it requires point-wise operations which can not be performed on the point clouds.

Optimization methods. We will compare three approaches: Explicit *Gradient Descent* adjusting the gradient step/learning rate experimentally (the gradient is obtained using `Pytorch` autodifferentiation), *Adam* algorithm [KB17] as implemented in `Pytorch` and we also implemented *Gold* method (4.23). We use a learning rate $lr = 50$ for gradient descent, and $lr = 0.1$ for Adam algorithm. Also, for Gold's method, we use the power parameter $\alpha = 0.5$, which plays a similar role as the learning rate. Unless otherwise stated, we will always intitialize with $\nu_0 = \nu$ the prescribed target distribution (this is also the solution for $h_{\mathcal{R}} = +\infty$).

4.4.2 Dirac Targets and the convolution effect

We start with a test case that illustrates the convolution effect (remark 4.3) and helps to interpret more general solutions. As discussed in Chapter 3.1, the simplest point source reflector is the parabola mapping any point (the focal point) source distribution to the direction of the focal axis. We use a Dirac target distribution $\nu = \delta_{3\pi/2}$, $h_{\mathcal{R}} = 5$. We use *backward ray tracing* and *Adam* optimization.

Figure 4.4 compares the optimization with two initialization : $\nu_0 = \nu$ the Dirac mass itself and the normal distribution $\mathcal{N}_{\frac{3\pi}{2}, \frac{\pi}{41}}$. We do not represent

the reflector as it does not carry much information. Instead, we plot (left) the "optimal " point source target parametrization of the reflector ν_0 generated by the optimization. The Dirac initialization is stationary and the Gaussian converges to the Dirac solution (right). The convolution effect of the finite source onto the parabolic reflector is observed (center). In order to produce this plot, rays are binned as explained in section 1.4.3.

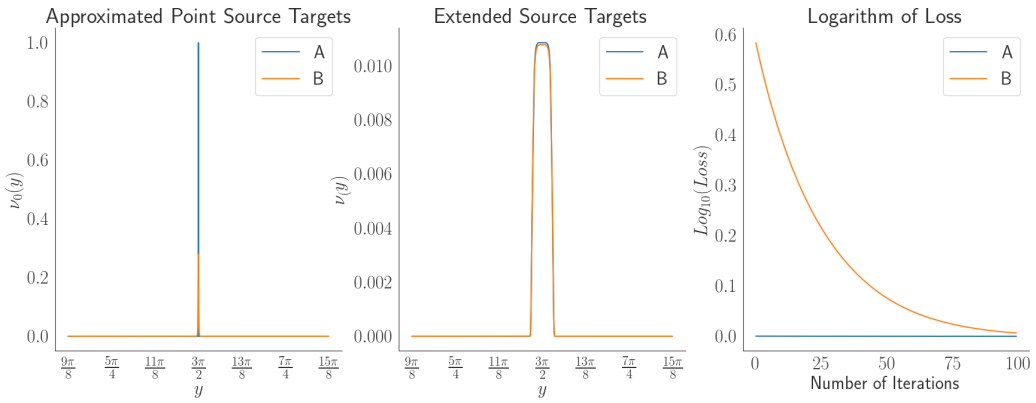


Figure 4.4: Dirac target distribution, two different initializations. (A): Initialized by the Dirac distribution. (B) Initialized by the Gaussian distribution. Left: final '"optimal " point source target parametrization of the reflector. Right: normalized loss function value along the optimization. Center: Target distribution simulated by ray tracing on the reflector generated by the optimization.

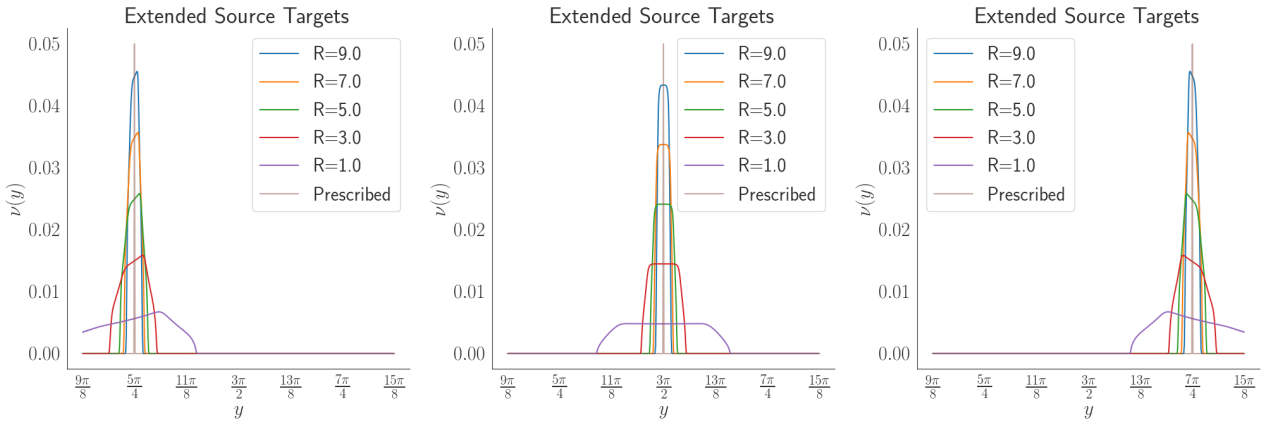
We can also illustrate the convolution effect by playing with the parameter $h_{\mathcal{R}}$. In figure 4.5 we show the target distribution generated from the reflection of the finite source for parabolic reflectors with axial direction angles $\frac{5\pi}{4}$, $\frac{3\pi}{2}$ and $\frac{7\pi}{4}$, and increasing heights $h_{\mathcal{R}}$ 1, 3, 5, 7 and 9. When $h_{\mathcal{R}}$ becomes larger, we approach the point source regime with a Dirac target distribution.

4.4.3 Comparison of optimization methods

Here we will present a comparison of the optimization methods (Adam, Gradient descent, Gold) for the following test cases (see figure 4.6) with a reflector height $h_{\mathcal{R}} = 5$ and backward ray tracing.

Test Case 1: "Uniform": $\nu = \frac{2}{\pi}\chi_{]\frac{5\pi}{4},\frac{7\pi}{4}[}$ the characteristic function of the intervall $]\frac{5\pi}{4},\frac{7\pi}{4}[$.

Test Case 2: Mixture of "Two Gaussians": $\nu = \mathcal{N}_{\frac{3\pi}{2}+\frac{\pi}{13},\frac{\pi}{21}} + \mathcal{N}_{\frac{3\pi}{2}-\frac{\pi}{7},\frac{\pi}{24}}$ (plus normalization).



(a) Axial direction $\frac{5\pi}{4}$ (b) Axial direction $\frac{3\pi}{2}$ (c) Axial direction $\frac{7\pi}{4}$

Figure 4.5: Finite source reflection on parabolae with different axial directions and heights $h_{\mathcal{R}}$.

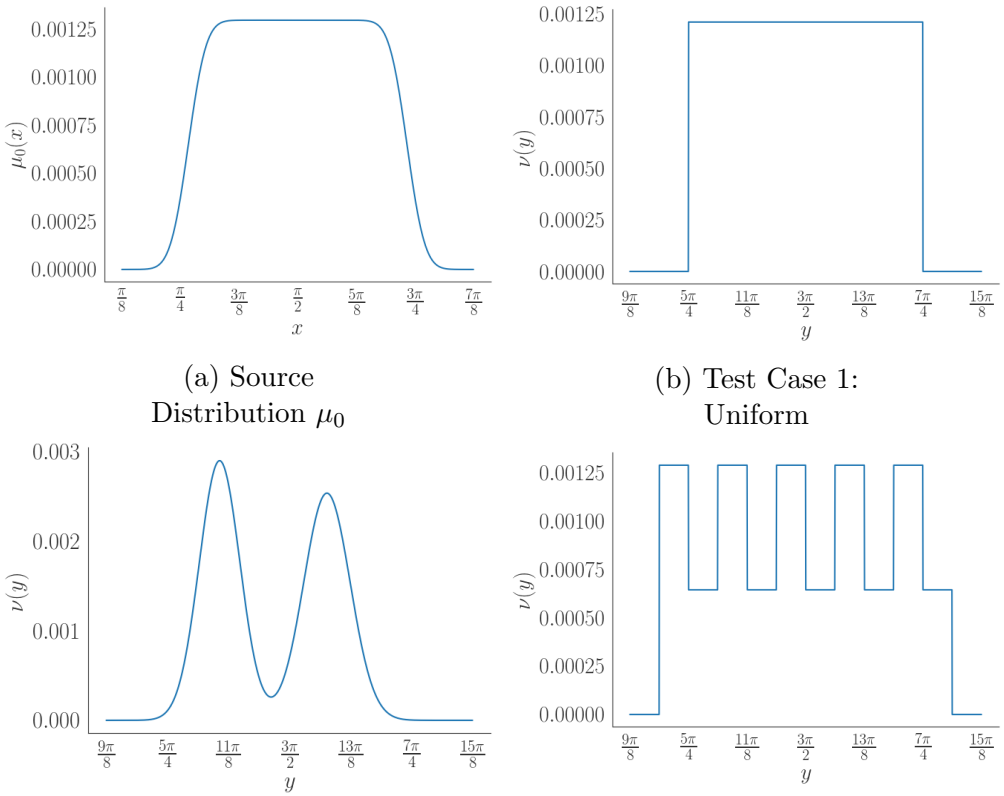
Test Case 3: "Binary" This test case was inspired by applications where the target distribution requires the values of the density to be "pixelized",. We alternate density values of 2 and 1 within the interval $[19\pi/16, 29\pi/16]$ with the step $\pi/16$, and a background noise $(1.e-10)$, then normalize.

Figure 4.7 compares the results obtained using the different optimization methods and backward ray tracing. The left column (approximated point source target) is the "optimal" ν_0 and the center column the resulting target obtained by ray tracing (binned) on the corresponding reflector. The discontinuous targets ν (test cases (A) and (C)) are clearly not in the range of the forward operator $\mathcal{F}$. The point source parametrization of the reflector performs a regularization through the already mentioned nonlinear convolution. The optimal solution still makes use of diracs/parabola near the discontinuities as it provides the strongest slopes. Gold's method fails except for the smooth case (b), and also it is very sensitive to small density values.

In figure 4.8, we explore the choice of the raytracing method (Chapter 1.4) with Adam optimization. Parameters have been tuned to use the same number of rays in both cases. It seems not to impact the optimization and justifies a preference for the more computationally efficient backward ray tracing.

In figure 4.9, we explore the dependence of the extended source problem on the parameter $h_{\mathcal{R}}$ (Chapter 1.4) with Adam optimization and backward ray tracing. We can see that as $h_{\mathcal{R}}$ decreases to 1, it becomes impossible to accurately approximate the desired targets, likely due to the ill-posed nature of the extended source problem (remark 1.2). The optimization method

still improves the initial approximation and minimizes the loss. Note that sometimes the loss values for the case $h_{\mathcal{R}} = 1$ are smaller than others, while the approximation is visibly worse. This is due to the fact that our loss function shows improvement from the initial guess. Since our initial guess is the solution of the point source problem, it is a worse approximation when the height is smallest, and hence the improvement is higher than for the other cases. Finally note that sometimes the loss values increase drastically, but then decrease again. This happens due to the fact that standard Adam algorithm implementation in `pytorch` does not check if the step size is appropriate for the descent direction and in rare cases it "overshoots". However, as Adam algorithm adjusts the step size after every iteration, it still manages to "recover" from such overshooting and decrease the loss.



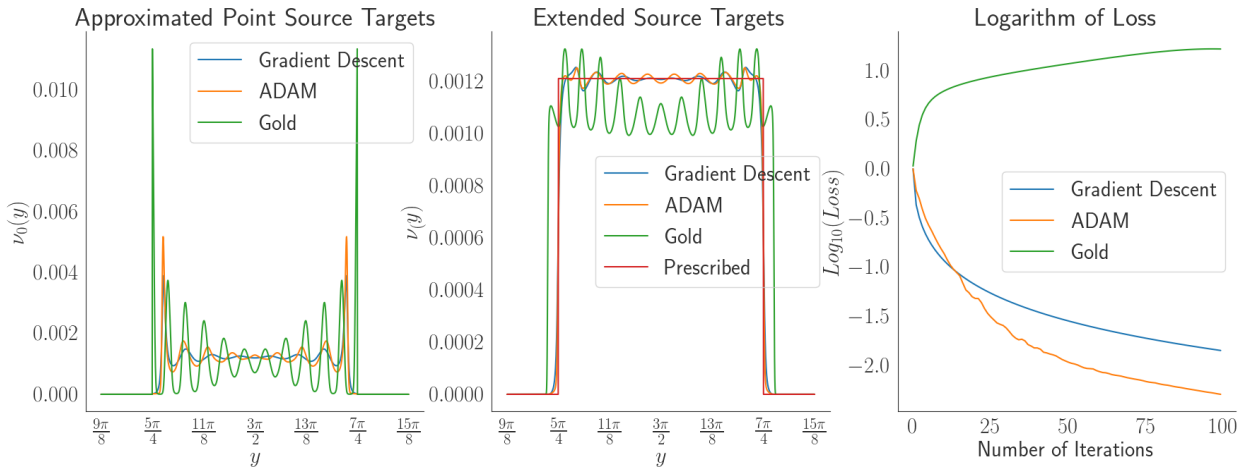
(a) Source Distribution μ_0

(b) Test Case 1: Uniform

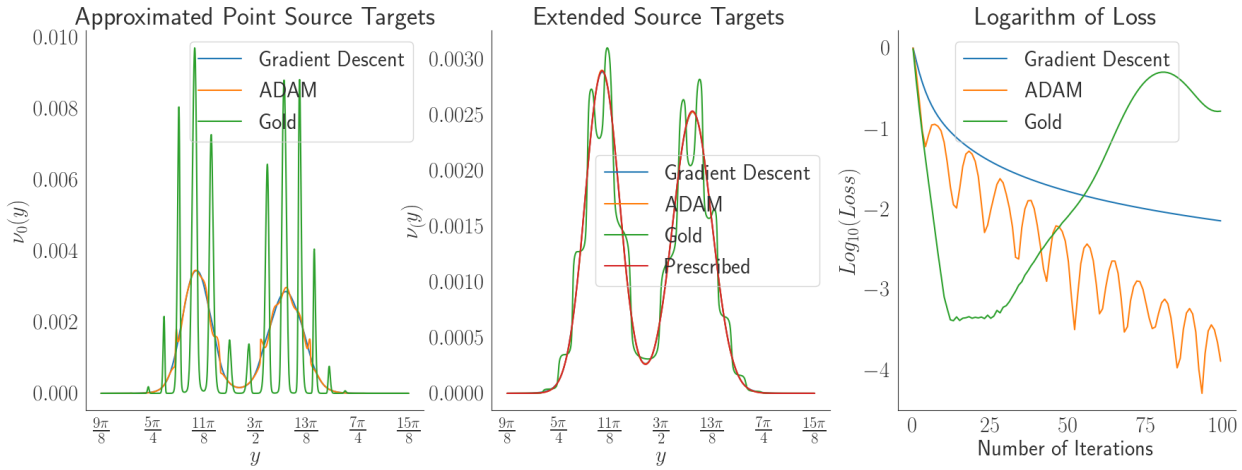
(c) Test Case 2: Two Gaussians

(d) Test Case 3: Binary

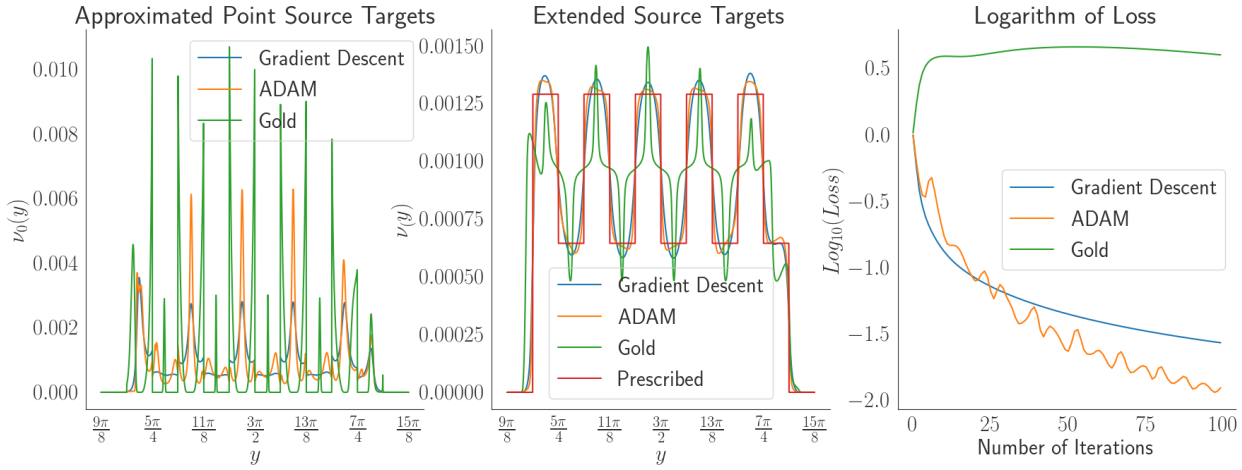
Figure 4.6: μ_0 and Different desired Target densities



(a) Test Case 1: Uniform

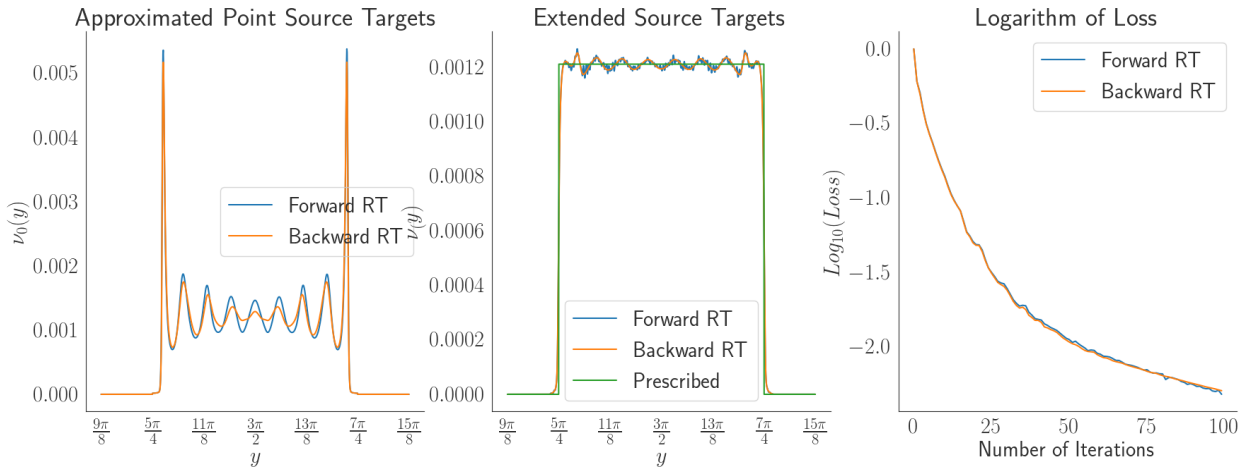


(b) Test Case 2: Two Gaussians

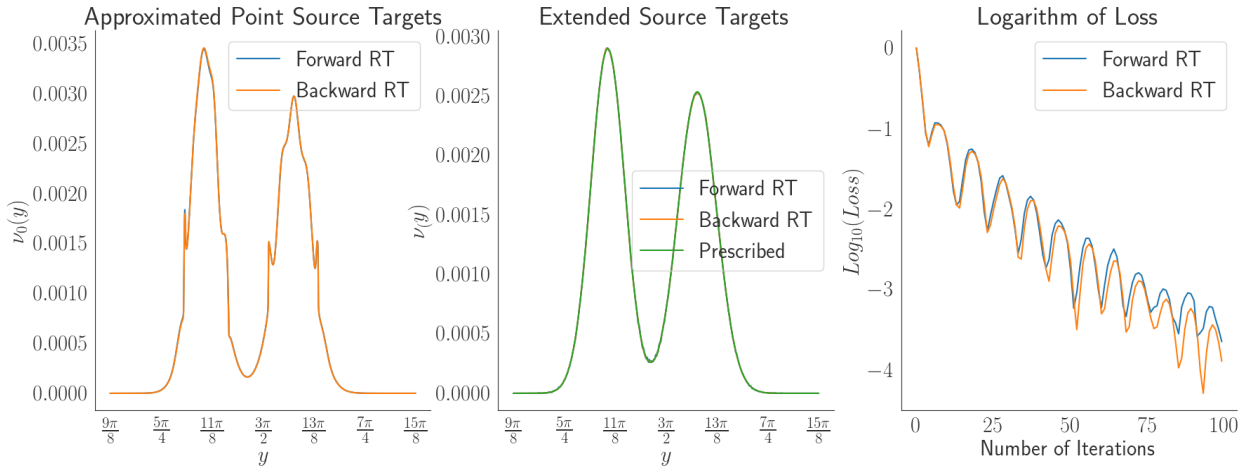


(c) Test Case 3: Binary

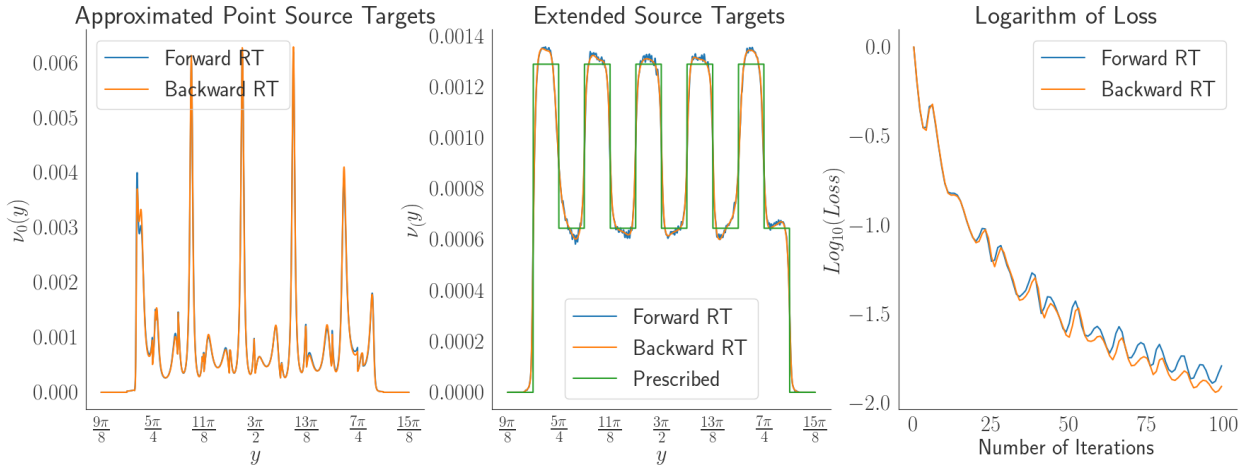
Figure 4.7: Comparison of different optimization methods



(a) Test Case 1: Uniform

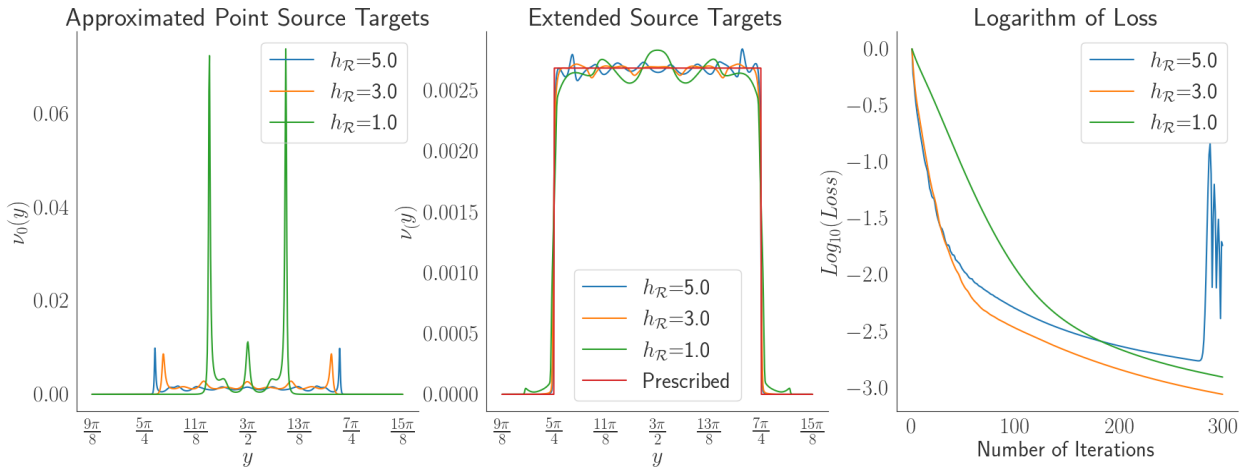


(b) Test Case 2: Two Gaussians

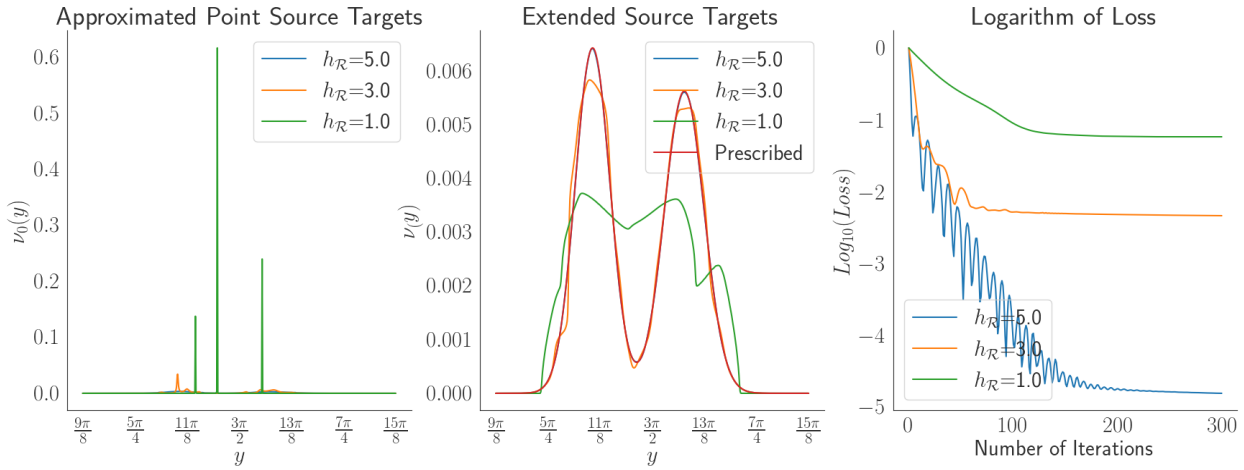


(c) Test Case 3: Binary

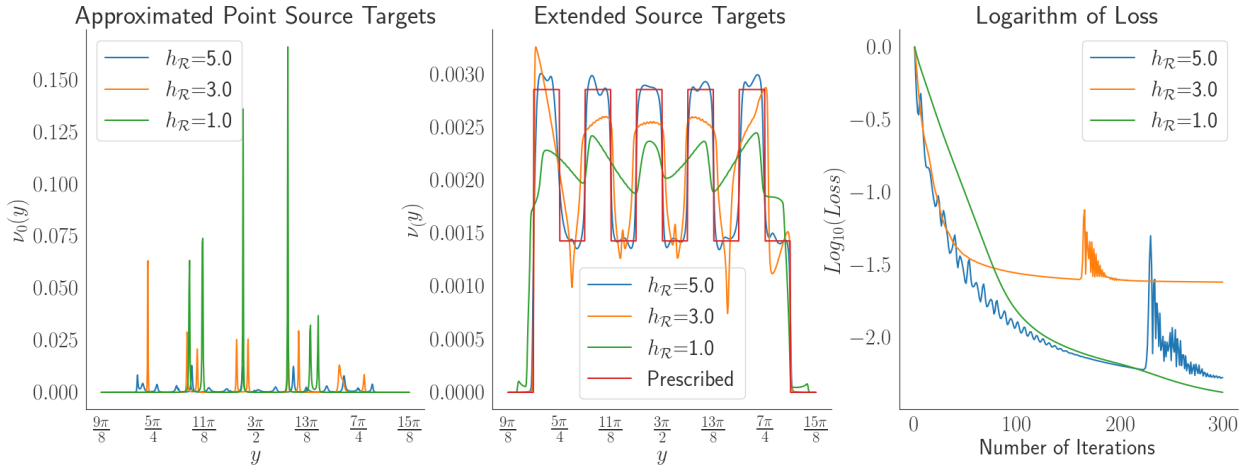
Figure 4.8: Comparison of forward/backward ray tracing



(a) Test Case 1: Uniform



(b) Test Case 2: Two Gaussians



(c) Test Case 3: Binary

Figure 4.9: Solving the extended source problem for different reflector heights $h_{\mathcal{R}}$

Conclusion and future work

The proposed optimal transport parameterization of the reflector offers theoretical guarantees for the optimization of a suitable loss function and provides a regularized solution. Our 1-D preliminary numerical study shows the approach is robust and converges at least to a local minimum. This research can be pursued in many directions, which we summarize in the following sections.

Heuristics and implementations

First of all, the extension of the code to accommodate $d = 3$ case (2D reflectors) will be of interest for the optics community and especially industry, since the practical applications are set in this case.

Testing Semi-Discrete optimal transport solvers instead of the entropic solvers would also be relevant.

It is possible to use the loss (4.18) for a machine learning approach, to parameterize the map $\mathcal{F}^{-1} : \nu \mapsto \nu_0$ with a convolutional neural network.

General formulation of the regularization approach allows the change of the optical setup, in the sense of changing the geometrical reflection law into different, possibly more realistic reflection models as long as they are computationally efficient and differentiable (at least in the computational "automatic differentiation" sense). We are thinking for example about analytical BRDF models used in computer graphics and industrial design to approximate the scattering effects of various materials.

Finally, it is also possible to apply a multi-scale optimization strategy based on restarting with initializations obtained from an increasing sequence of the source interval S. In other words, instead of using all N_S discretization points of the source interval S, start from a few discretization points on

the smaller interval $[-\eta, \eta]$ in the middle and gradually increase η and add discretization points to reach the desired amount. This approach could also save the computational speed for the iterations at the early stage, as the total number of rays will be less when using fewer points in the source interval.

Analytical properties of the various ingredients

Establishing if the loss (4.18) is convex is important for having a theoretical guarantee that the solution found using our approach is a global minimizer.

Note that, even for the smooth test case 2, on figure 4.9 we observe that the minimizer of $\mathcal{F}$ for the height $h_{\mathcal{R}} = 1$ is not in the class $H_{b,1}$, for which the weak continuity was established.

Imposing on the source and target densities of the point source problem to be in $C^{1,\alpha}$ (see notation 4) provides that the Kantorovich potential is in $C^{3,\alpha}$ (see proposition 3.4, and also [Loe13]). But we only use continuous differentiability of the reflector normal (given by the second derivative of Kantorovich potential) in (4.7), in order to establish the requirements for the inverse function theorem for the reparametrization map A_s. Except for that point, Kantorovich potential from $C^{1,\alpha}$ would be sufficient to guarantee weak continuity of the forward map $\mathcal{F}$.

In [Loe13] it is established that just boundedness away from 0 and infinity of the source and target densities can guarantee $C^{1,\alpha}$ Kantorovich potential. Therefore, providing an alternative proof of a non-vanishing derivative of A_s, using only Holder continuous normal, would allow proving the weak continuity of the forward map $\mathcal{F}$ on a more appropriate domain of measures with densities bounded away from 0 and infinity.

Another interesting open question, related to the reparametrization A_s is whether the reparametrized functions f_s are c-concave: Remember that f_0, as a Kantorovich potential of an optimal transport problem with the cost (3.5), is a c-transform of the second Kantorovich potential g_0. This translates into the property of the reflector $\mathcal{R}_{f_0}$ that it has tangential paraboloid with focus at $O = O_0$ at every point (see chapter 3.1). But we are not aware of any result, establishing if (under any regularity assumption) there also exists the family of tangential paraboloids with a focus at O_s. Such a result would mean that every T_s is an optimal map between the source measure μ_s and

the reflected measure $\nu_s = T_{s\#}\mu_s$.

Geometry of the extended source problem

Let $\mathcal{R}$ be a reflector with the height $h_{\mathcal{R}} > 1$, defined on the whole $\mathbb{S}_+$. This reflector induces a reflection map $T : S \times \mathbb{S}_+ \to Y$. This map induces a decomposition of the set $S \times \mathbb{S}_+ = \bigcup_{y \in Y} T^{-1}(y)$.

Understanding the relation between a given reflector and the corresponding decomposition $\bigcup_{y \in Y} T^{-1}(y)$ could provide useful insights on the solvability (or unsolvability) of the given extended source problem.

Hypothesis 1. *Under the assumptions of propositions 3.4 and 4.1, for all $y \in Y$, $T^{-1}(y)$ is a rectifiable curve in $S \times \mathbb{S}_+$.*

Note that in general, $T^{-1}(y)$ might not be curves: for the parabolic reflector with the focus at the origin and some $y_0 \in Y$ as an axial direction, $T^{-1}(y_0)$ contains at least one point from each $s \times \mathbb{S}_+$, and all of $0 \times \mathbb{S}_+$.

Under the above hypothesis, the reflector $\mathcal{R}$ reflecting source measure μ into the target measure ν (in the sense of $T_\#\mu = \nu$) induces a decomposition $T^{-1}(y)$, such that, the disintegration of the measure μ on this decomposition and integrating along the curves $T^{-1}(y)$ yields the measure ν. In other words, integrating the measure μ along the curve $T^{-1}(y)$ gives the value $\nu(y)$.

Remark 4.9. *Such decomposition of the source space is also observed in many-to-few dimensional optimal transport (see [CMP17] [MP18]), where the decomposition is selected amongst the ones induced by the level sets of $\nabla_y c(x, y)$ and is used to construct the optimal map T.*

Appendix A

Regularity assumptions on the cost $c(x, y)$

The regularity assumptions on the cost function $c(x, y)$, which are required for obtaining regularity results presented in this work, are satisfied for all costs that we use through this work. They are formulated for the spaces X and Y which are manifolds of the same dimension d. We present them here for completeness:

A1 ("Twist condition") The map $y \mapsto \nabla_x c(x, y)$ is injective for any $(x, y) \in X \times Y$.

A2 ("Non-degeneracy") $\det(\nabla_{x_i} \nabla_{y_j} c(x, y)) \neq 0$ for any $(x, y) \in X \times Y$.

A3 ("Ma-Trudinger-Wang (MTW) condition") $\forall x \in X, \forall p, \xi, \eta \in T_x X$(Tangent space of X at x) such that $\xi \perp \eta$ following holds:

$$\sum_{i,j,k,l} \nabla_{p_i,p_j} a_{k,l}(x, p)\xi_i\xi_k\eta_k\eta_l \leq 0$$

Where $a_{k,l}(x, p) := \nabla^2_{x_k,x_l} c(x, (\nabla_x c(x, y))^{-1}(p)$

Remark A.1. *The meaning of the MTW condition is hard to see from the bare definition. Intuitively it puts some geometric requirements for $\nabla_x c$. More precisely, the MTW condition presented above, is equivalent, for C^4 costs (see [LT20]), to the following condition, known as Loeper's condition:*
$\forall x, x_0 \in X$ *and* $y_1, y_2 \in Y$, $\forall \theta \in (0, 1)$ *such that*

$$\nabla_x c(x, y_\theta) = \theta \nabla_x c(x, y_1) + (1 - \theta)\nabla_x c(x, y_2)$$

following holds:

$$\max\{c(x_0, y_1) - c(x, y_1), c(x_0, y_2) - c(x, y_2)\} \geq c(x_0, y_\theta) - c(x, y_\theta) + o(|x - x_0|^2)$$

Where the second order term $o(|x - x_0|^2)$ *may also depend on* θ*.*

Appendix B

Wasserstein distance between push-forwards

Lemma B.1. *Given two Polish spaces (X, d_Y) and (Y, d_Y), with Borel probability measures μ_1, μ_2 on X and Lipschitz continuous map $T : X \to Y$ with Lipschitz constant Lip(T), Then*

$$W_p(T_{\#}\mu_1, T_{\#}\mu_2) \leq Lip(T) W_p(\mu_1, \mu_2)$$

Proof. Let (f, g) be optimal pair of Kantorovich potentials for $W_p(T_{\#}\mu_1, T_{\#}\mu_2)$. Then, for all $x, x' \in X$

$$f(T(x)) + g(T(x')) \leq d_Y^p(T(x), T(x')) \leq Lip(T)^p d_X^p(x, x')$$

Where first inequality holds due to fact that admissible pairs for maximization in Kantorovich duality approach cost function from below.

Inequality (B) implies that functions $\frac{f(T(\cdot))}{Lip(T)^p}$ and $\frac{g(T(\cdot))}{Lip(T)^p}$ are admissible pair in the dual form of $W_p(\mu_1, \mu_2)$. This leads to following:

$$\begin{aligned}
W_p^p(T_{\#}\mu_1, T_{\#}\mu_2) &= \min_{\gamma} \int_{Y \times Y} d_Y^p((T_{\#}\mu_1, T_{\#}\mu_2) d\gamma \\
&= \int_Y f(y)\, dT_{\#}\mu_1(y) + \int_Y g(y)\, dT_{\#}\mu_2(y') \\
&= \int_X f(T(x))\, d\mu_1(x) + \int_X g(T(x))\, d\mu_2(x) \\
&= Lip(T)^p \left(\int_X \frac{f(T(x))}{Lip(T)^p} d\mu_1(x) + \int_X \frac{g(T(x))}{Lip(T)^p} d\mu_2(x) \right) \\
&\leq Lip(T)^p W_2^p(\mu_1, \mu_2)
\end{aligned}$$

Taking p-th root on both sides leads to the desired inequality.

□

Appendix C

Local density property for curved spaces

The density property, discussed in Remark 2.16 requires stronger assumptions on the local scale when the underlying space X can have a non-zero curvature.

Definition 10 (Local density property)**.** *The sequence of measures $\mu^{(k)}$ on X satisfy the local density property (at a lengthscale $k^{-\frac{1}{2}}$), if there exists $s \in [2, \infty)$ and constants $C_1, C_2 \in \mathbb{R}$, such that:*

For any $x_0 \in X$ there exists a local coordinate system $\xi := (\xi_1, ..., \xi_d)$ centered at x_0 with the property that for any sequence of functions $h_k : 2D_k \to R$, satisfying $|\partial^{|\alpha|} h_k(x)| \leq C_1 e^{-|x|^2/C_1}$ for all multiindices $|\alpha| \leq s$, one has the following bound:

$$k^{-\frac{n}{2}} \int_{D_k} h_k (F_{x_0}^{(k)})_{\#} (\mu^k - \mu) \leq C_2 k^{-1}$$

Where D_k is a poly-disc centered at the origin with a radius of $\log(k)$ and $F_{x_0}^{(k)}$ is the scaled coordinate map from the neighbourhood of x_0 into $\mathbb{R}^d$, defined by $F_{x_0}^{(k)}(x) := k^{1/2}\xi(x)$.

Bibliography

[AG09] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *CIME summer school*, Italy, 2009.

[BB18] Martin Benning and Martin Burger. Modern regularization methods for inverse problems, 2018.

[BB19] Matt Brand and Daniel A. Birch. Freeform irradiance tailoring for light fields. *Opt. Express*, 27(12):A611–A619, Jun 2019.

[BCIR21] J.-D Benamou, G Chazareix, W L Ijzerman, and G Rukhaia. Point Source Regularization of the Finite Source Reflector Problem. working paper or preprint, September 2021.

[Ber17] R. J. Berman. The Sinkhorn algorithm, parabolic optimal transport and geometric Monge-Ampere equations. *ArXiv e-prints*, December 2017.

[BHP14] Kolja Brix, Yasemin Hafizogullari, and Andreas Platen. Solving the monge-ampère equations for the inverse reflector problem. *Mathematical Models and Methods in Applied Sciences*, 25, 04 2014.

[BHP15] Kolja Brix, Yasemin Hafizogullari, and Andreas Platen. Designing illumination lenses and mirrors by the numerical solution of monge–ampère equations. *Journal of the Optical Society of America A*, 32(11):2227, Oct 2015.

[Bil99] Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[BIR20] Jean-David Benamou, Wilbert L Ijzerman, and Giorgi Rukhaia. An Entropic Optimal Transport Numerical Approach to the Reflector Problem. working paper or preprint, April 2020.

[BKM+20] Egor V. Byzov, Sergey V. Kravchenko, Mikhail A. Moiseev, Evgeni A. Bezus, and Leonid L. Doskolovich. Optimization method for designing double-surface refractive optical elements for an extended light source. *Opt. Express*, 28(17):24431–24443, Aug 2020.

[BM21] Guillaume Bonnet and Jean-Marie Mirebeau. Monotone discretization of the Monge-Ampère equation of optimal transport. working paper or preprint, June 2021.

[Caf92] Luis Caffarelli. The regularity of mappings with a convex potential. *Journal of The American Mathematical Society - J AMER MATH SOC*, 5, 01 1992.

[Cau47] Augustin-Louis Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus de l'Académie des sciences*, page 536, 1847.

[Cha20] Guillaume Chazareix. *Extended Source Reflector Problem Code https://github.com/NightWinkle/ExtendedSourceReflectorProblem.* 2020.

[CM94] R. Cominetti and J. San Martin. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1):169–187, 1994.

[CMP17] Pierre-André Chiappori, Robert J. McCann, and Brendan Pass. Multi-to one-dimensional optimal transport. *Communications on Pure and Applied Mathematics*, 70(12):2405–2444, 2017.

[CPSV15] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced Optimal Transport: Geometry and Kantorovich Formulation. working paper or preprint, August 2015.

[CPSV18] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between

optimal transport and Fisher–Rao metrics. *Found. Comput. Math.*, 18, 2018.

[CRL+20] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence, 2020.

[CT19] Giovanni Conforti and Luca Tamanini. A formula for the time derivative of the entropic cost and applications, 2019.

[Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

[Dav92] P. A. Davies. Methods of choosing sample rays in ray-tracing computations. *Appl. Opt.*, 31(34):7277–7282, Dec 1992.

[FCR10] Florian R. Fournier, William J. Cassarly, and Jannick P. Rolland. Fast freeform reflector generation using source-target maps. *Opt. Express*, 18(5):5295–5304, Mar 2010.

[Fey19] Jean Feydy. *GEOMLOSS https://www.kernel-operations.io/geomloss/*. 2019.

[FG15] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, August 2015.

[FSV+18] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. working paper or preprint, October 2018.

[GBMB+04] Pablo Gimenez-Benitez, Juan Carlos Miñano, Jose Blen, Rubén Mohedano Arroyo, Julio Chaves, Oliver Dross, Maikel Hernandez, and Waqidi Falicoff. Simultaneous multiple surface optical design method in three dimensions. *Optical Engineering*, 43(7):1489 – 1502, 2004.

[Gla89] Andrew S. Glassner, editor. *An Introduction to Ray Tracing*. Academic Press Ltd., GBR, 1989.

[GO03] Tilmann Glimm and Vladimir Oliker. Optical design of single reflector systems and the monge–kantorovich mass transfer problem. *Journal of Mathematical Sciences*, 117:4096–4108, 09 2003.

[Gol64] R Gold. An iterative unfolding method for response matrices. 12 1964.

[Kan42] Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[KB17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[KMV16] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Adv. Diff. Equations*, 21:1117–1164, 2016.

[KO03] Sergey A. Kochengin and Vladimir I. Oliker. Computational algorithms for constructing reflectors. *Computing and Visualization in Science*, 6(1):15–21, 2003.

[Lec19] Hugo Leclerc. *Pysdot software - https://pypi.org/project/pysdot/*. 2019.

[Lév15] Bruno Lévy. A numerical algorithm for L2 semi-discrete optimal transport in 3D. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1693 – 1715, November 2015.

[LFHL10] Yi Luo, Zexin Feng, Yanjun Han, and Hongtao Li. Design of compact and smooth free-form optical system with uniform illuminance for led source. *Opt. Express*, 18(9):9055–9063, Apr 2010.

[LMS18] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211:969–1117, 2018.

[Loe09] Grégoire Loeper. On the regularity of solutions of optimal transportation problems. *Acta Mathematica*, 202(2):241 – 283, 2009.

[Loe13] Gregoire Loeper. Regularity of optimal maps on the sphere: the quadratic cost and the reflector antenna. *Arch. Ration. Mech. Anal.*, 2013.

[LRPC18] Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance, 2018.

[LS18] Bruno Lévy and Erica L. Schwindt. Notions of optimal transport theory and how to implement them on a computer. *Comput. Graph.*, 72:135–148, 2018.

[LT20] G. Loeper and N. S. Trudinger. Weak formulation of the mtw condition and convexity properties of potentials, 2020.

[Lé13] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport, 2013.

[Mér11] Quentin Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1584–1592, August 2011. 18 pages.

[Mey18] Jocelyn Meyron. *Semi-discrete optimal transport and applications in non-imaging optics*. Theses, Université Grenoble Alpes, October 2018.

[MMDCMT16] Pedro Machado Manhães De Castro, Quentin Mérigot, and Boris Thibert. Far-field reflector problem and intersection of paraboloids. *Numerische Mathematik*, 134(2):389–411, October 2016.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

[MP18] R. McCann and B. Pass. Optimal transportation between unequal dimensions. *arXiv: Analysis of PDEs*, 2018.

[OR15] Adam M. Oberman and Yuanlong Ruan. An efficient linear programming method for optimal transportation, 2015.

[Pal19] Soumik Pal. On the difference between entropic cost and the optimal transport cost, 2019.

[PC18] G. Peyré and M. Cuturi. Computational Optimal Transport. *ArXiv e-prints*, March 2018.

[PF14] Guido De Philippis and Alessio Figalli. The monge-ampère equation and its link to optimal transportation, 2014.

[Pra07] Aldo Pratelli. On the equality between monge's infimum and kantorovich's minimum in optimal mass transportation. *Annales de l'I.H.P. Probabilités et statistiques*, 43(1):1–13, 2007.

[RKK19] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *CoRR*, abs/1904.09237, 2019.

[RtI20] Lotte B. Romijn, Jan H.M. ten Thije Boonkkamp, and Wilbert L. IJzerman. Inverse reflector design for a point source and far-field target. *Journal of Computational Physics*, 408, 5 2020.

[RtTBI19] L.B. Romijn, J.H.M. ten Thije Boonkkamp, and W.L. IJzerman. Freeform lens design for a point source and far-field target. In *Optical Design and Fabrication 2019 (Freeform, OFT)*, page FT1B.2. Optical Society of America, 2019.

[Rud17] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2017.

[Rus06] Andrzej Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006.

[San15] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015.

[Sch16] Bernhard Schmitzer. Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems. *arXiv e-prints*, page arXiv:1610.06519, Oct 2016.

[Vil08] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[Wan96] Xu-Jia Wang. On the design of a reflector antenna. *Inverse Problems*, 12(3):351–375, jun 1996.

[Wan04] Xu-Jia Wang. On the design of a reflector antenna ii. *Calculus of Variations and Partial Differential Equations*, 20(3):329–341, Jul 2004.

[WFZ+18] Rengmao Wu, Zexin Feng, Zhenrong Zheng, Rongguang Liang, Pablo Benítez, Juan C. Miñano, and Fabian Duerr. Design of freeform illumination optics. *Laser & Photonics Reviews*, 12(7):1700310, 2018.

[Wom17] R. S. Womersley. Efficient Spherical Designs with Good Geometric Properties. *ArXiv e-prints*, September 2017.

[WXL+13] Rengmao Wu, Liang Xu, Peng Liu, Yaqin Zhang, Zhenrong Zheng, Haifeng Li, and Xu Liu. Freeform illumination design: a nonlinear boundary problem for the elliptic monge–ampére equation. *Opt. Lett.*, 38(2):229–231, Jan 2013.

[WZLM21] Shili Wei, Zhengbo Zhu, Wenyi Li, and Donglin Ma. Compact freeform illumination optics design by deblurring the response of extended sources. *Opt. Lett.*, 46(11):2770–2773, Jun 2021.

RÉSUMÉ

Dans ce travail, nous abordons un problème inverse en optique anidolique consistant à déterminer une surface capable de réflechir une distribution de lumière source à une distribution cible en champ lointain, toutes deux prescrites. La source lumineuse peut être ponctuelle ou étendue. Lorsque la source est une source ponctuelle , la distribution est supportée uniquement sur les directions des rayons optiques. Dans ce contexte, le problème inverse est bien posé pour des distributions de probabilité source et cible arbitraires. Il peut être reformulé comme un problème de transport optimal et constitue un exemple célébre de transport optimal sous un coût de déplacement non euclidien. Nous explorons l'utilisation du transport optimal entropique et de l'algorithme Sinkhorn associé pour le résoudre numériquement. La modélisation du réflecteur étant basée sur les potentiels de Kantorovich, plusieurs questions se posent. Premièrement, sur la convergence de l'approximation entropique discrète et nous suivons ici les travaux récents de Berman et en particulier les exigences de discrétisation qui y sont imposées. Deuxièmement, nous montrons que la correction du biais induit par le transport entropique Optimal peut être atteinte en utilisant la notion récente de divergences Sinkhorn. Pour le problème de source ponctuelle, nous discutons des outils mathématiques et numériques nécessaires pour produire et analyser les résultats numériques obtenus. Nous trouvons que l'algorithme Sinkhorn peut être adapté à la résolution du problème de la source ponctuelle au réflecteur en champ lointain. Nous ne connaissons pas de formulation mathématique similaire dans le cas de la source étendue : la distribution de lumière source a support sur l'espace produit: domaine physique-directions des rayons. Nous proposons de tirer parti de la formulation variationnelle bien posée du problème de source ponctuelle pour construire une paramétrisation lisse du réflecteur et de l'application modélisant la réflexion. Sous cette paramétrisation, nous pouvons construire une fonction de coût lisse à optimiser pour trouver la meilleure solution dans cette classe de réflecteurs. Les deux étapes, la paramétrisation et la fonction de coût, sont liées à des distances de transport entropiques optimales. Nous profitons également des progrès récents des techniques d'optimisation et des implémentations efficaces de l'algorithme Sinkhorn pour réaliser une étude numérique.

MOTS CLÉS

Transport Optimal entropique, Algorithme de Sinkhorn, Problème inverse du reflecteur, Source lumineuse étendue, Optimisation non linéaire.

ABSTRACT

In this work, we address the "freeform optics" inverse problem of designing a reflector surface mapping a prescribed source distribution of light to a prescribed target far-field distribution, for the point light source and the extended light source. When the source is a point source, the light distribution has support only on the optics ray directions. In this setting, the inverse problem is well-posed for arbitrary source and target probability distributions. It can be recast as an optimal transport problem and is a classic example of an optimal transport problem with a non-euclidean displacement cost. We explore the use of entropic Optimal Transport and the associated Sinkhorn algorithm to solve it numerically. As the reflector modeling is based on the Kantorovich potentials, several questions arise. First, on the convergence of the discrete entropic approximation and here we follow the recent work of Berman and in particular the imposed discretization requirements therein. Secondly, the correction of the bias induced by the entropic Optimal Transport using the recent notion of Sinkhorn divergences is shown to be necessary to achieve satisfactory results. For the point source problem, we discuss the necessary mathematical and numerical tools needed to produce and analyze the obtained numerical results. We find that Sinkhorn algorithm may be adapted to the resolution of the point source to far-field reflector problem. We are not aware of any similar mathematical formulation in the extended source case: i.e. the source has an "étendue" with support in the product space: physical domain-ray directions. We propose to leverage the well-posed variational formulation of the point source problem to build a smooth parameterization of the reflector and the map modeling the reflection. Under this parametrization, we can construct a smooth cost function to optimize for the best solution in this class of reflectors. Both steps, the parameterization and the cost function, are related to entropic optimal transport distances. We also take advantage of recent progress in the optimization techniques and the efficient implementations of Sinkhorn algorithm to perform a numerical study.

KEYWORDS

Entropic optimal transport, Sinkhorn algorithm, Inverse reflector problem, extended source, Non-linear optimization.