# Semi-discrete optimal transport and applications in non-imaging optics

Jocelyn Meyron

▶ **To cite this version:**

Jocelyn Meyron. Semi-discrete optimal transport and applications in non-imaging optics. Computer science. Université Grenoble Alpes, 2018. English. NNT : 2018GREAT104 . tel-02135220

**HAL Id: tel-02135220**
**https://theses.hal.science/tel-02135220**

Submitted on 21 May 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Communauté
UNIVERSITÉ Grenoble Alpes

# THÈSE

Pour obtenir le grade de

**DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES**

Spécialité : **SIGNAL IMAGE PAROLE TELECOMS**

Arrêté ministériel : 25 mai 2016

Présentée par


## Jocelyn MEYRON

Thèse dirigée par **Dominique ATTALI, GIPSA-lab** et
codirigée par **Boris THIBERT, Laboratoire Jean Kuntzmann** et
**Quentin MÉRIGOT, Laboratoire de Mathématiques d'Orsay**

préparée au sein du **Laboratoire Jean Kuntzmann**
dans **l'École Doctorale Électronique Électrotechnique, Automatique, Traitement du Signal (EEATS)**


# Transport optimal semi-discret et applications en optique anidolique

Thèse soutenue publiquement le **16 octobre 2018**,
devant le jury composé de :

**Monsieur Marco CUTURI**
Professeur, Université Paris-Saclay, Rapporteur
**Monsieur Bruno LÉVY**
Directeur de recherche, INRIA Nancy, Rapporteur
**Monsieur Quentin MÉRIGOT**
Professeur, Université Paris-Sud, Codirecteur de thèse
**Monsieur Boris THIBERT**
Maître de conférences, Université Grenoble Alpes, Codirecteur de thèse
**Madame Dominique ATTALI**
Directrice de recherche, CNRS, Directrice de thèse
**Madame Julie DIGNE**
Chargée de recherche, CNRS, Examinateur
**Monsieur André LIEUTIER**
Ingénieur, Dassault Systèmes, Examinateur
**Monsieur Édouard OUDET**
Professeur, Université Grenoble Alpes, Examinateur, Président du Jury

**UNIVERSITÉ DE GRENOBLE ALPES**

**EEATS**

**Électronique Électrotechnique Automatique & Traitement du signal**

# T H È S E

pour obtenir le titre de

## docteur en sciences

de l'Université de Grenoble Alpes

**Mention : Mathématiques appliquées**

Présentée et soutenue par

Jocelyn MEYRON

## Transport optimal semi-discret et applications en optique anidolique

Thèse dirigée par Quentin MÉRIGOT, Boris THIBERT et Dominique ATTALI

préparée au sein du Laboratoire Jean Kuntzmann (LJK)

soutenue le 16 octobre 2018

**Jury :**

| | | | |
|---|---|---|---|
| *Rapporteurs :* | Marco CUTURI | - | Université Paris-Saclay |
| | Bruno LÉVY | - | INRIA Nancy |
| *Directeurs :* | Quentin MÉRIGOT | - | Université Paris-Sud |
| | Boris THIBERT | - | Université Grenoble Alpes |
| | Dominique ATTALI | - | CNRS, Université Grenoble Alpes |
| *Examinateurs :* | Julie DIGNE | - | CNRS, Université Lyon 1 |
| | André LIEUTIER | - | Dassault Systèmes |
| | Édouard OUDET | - | Université Grenoble Alpes |

「何でもは知らない。知っていることだけ。」羽川翼。

# Remerciements

Je souhaite tout d'abord remercier les deux rapporteurs de cette thèse à savoir Marco Cuturi et Bruno Lévy pour leurs relectures attentives de ce manuscrit et leurs commentaires pertinents. Je souhaite également remercier l'ensemble des autres membres du jury : le président du jury Édouard Oudet ainsi que les examinateurs Julie Digne et André Lieutier pour avoir accepté de faire partie de mon jury et pour leur enthousiasme lors de la présentation de mes travaux.

Je remercie aussi mes encadrants de thèse Quentin et Boris pour le temps qu'ils m'ont consacré durant ces trois années. Je ne pense pas que j'aurais pu arriver au bout de ce travail sans leur soutien et leur bonne humeur. En particulier, une petite pensée pour toutes les affaires oubliées par Boris dans mon bureau. Je n'oublie pas aussi Dominique qui m'a encadré au début de ma thèse et qui a du prendre un peu de recul pour deux très bonnes raisons.

Je n'oublie pas aussi mes anciens professeurs (notamment M. Naili au lycée et M. Adad en classes préparatoires) pour m'avoir transmis leur amour pour les mathématiques m'ayant permis entre autres d'aboutir à ce travail.

Au laboratoire Jean Kuntzmann, je souhaite remercier l'ensemble du personnel en particulier le personnel administratif et informatique. Je remercie aussi mes collègues des bureaux 122-123-124. Tout particulièrement, je n'oublierai pas les moments passés dans le bureau 124 animé par (ils se reconnaîtront) *le papa et la maman du LJK* (merci notamment pour tous ces petits plats), *le futur retraité* (petit ange parti trop tôt) et *Monsieur le djeuns* (même s'il n'est pas là très souvent). Merci pour toutes ces découvertes musicales, mes connaissances en chanson française se sont grandement améliorées durant la période passée avec vous, si je ne devais retenir qu'un artiste ce serait sans doute *Monsieur K* i.e. Monsieur « Je suis dans l'avion de Pablo (wouh) ». J'ai également pu découvrir la discographie plutôt diversifiée de *Jul*. Je n'oublierai pas non plus toutes ces citations dignes des plus grands poètes : « La baguette n'attrape pas la soupe » ou encore « C'est dans les pâtes les plus sèches que l'on met le plus de beurre ». J'espère aussi que la collection de balles rebondissantes continuera de s'agrandir pour toujours plus de fun. Je remercie également les doctorants (mais pas que) avec qui j'ai pu échanger durant cette thèse (triés par ordre croissant en leur associant un entier aléatoire entre 1 et 1337 avec une graine égale à 42) : Florent, Alexis, Galaad, Alexandre, Kévin, Emmanuel, Zhu, Sophie, Bach, Arnaud, Charles, Cathy, David, Jean-Baptiste O, Simon, Émilie, Dimitri, Laurence, Abdelkader, Lionel, Victor et ceux que j'oublie.

Je remercie aussi ma mère et ma grand-mère pour m'avoir soutenu durant ce long travail et notamment durant la rédaction de ce manuscrit.

Je termine ces remerciements sur quelques tâches de café, boisson sans laquelle je n'aurais pas pu arriver au bout de ce travail.

**Résumé** — Dans cette thèse, nous nous intéressons à la résolution de nombreux problèmes d'optique anidolique. Plus précisément, il s'agit de construire des composants optiques qui satisfont des contraintes d'illumination à savoir que l'on veut que la lumière réfléchie (ou réfractée) par ce composant corresponde à une distribution fixée en avance. Comme applications, nous pouvons citer la conception de phares de voitures ou de caustiques. Nous montrons que ces problèmes de conception de composants optiques peuvent être vus comme des problèmes de transport optimal et nous expliquons en quoi cette formulation permet d'étudier l'existence et la régularité des solutions. Nous montrons aussi comment, en utilisant des outils de géométrie algorithmique, nous pouvons utiliser une méthode numérique efficace, la méthode de Newton amortie, pour résoudre tous ces problèmes. Nous obtenons un algorithme générique capable de construire efficacement un composant optique qui réfléchit (ou réfracte) une distribution de lumière prescrite. Nous montrons aussi la convergence de l'algorithme de Newton pour résoudre le problème de transport optimal dans le cas où le support de la mesure source est une union finie de simplexes. Nous décrivons également la relation commune qui existe entre huit différents problèmes de conception de composants optiques et montrons qu'ils peuvent tous être vus comme des équations de Monge-Ampère discrètes. Nous appliquons aussi la méthode de Newton à de nombreux problèmes de conception de composants optiques sur différents exemples simulés ainsi que sur des prototypes physiques. Enfin, nous nous intéressons à un problème apparaissant en transport optimal numérique à savoir le choix du point initial. Nous développons trois méthodes simples pour trouver de "bons" points initiaux qui peuvent être ensuite utilisés comme point de départ dans des algorithmes de résolution de transport optimal.
**Mots clés :** transport optimal, géométrie algorithmique, conception de réflecteurs.

**Abstract** — In this thesis, we are interested in solving many inverse problems arising in optics. More precisely, we are interested in designing optical components such as mirrors and lenses that satisfy some light conservation constraints meaning that we want to control the reflected (or refracted) light in order match a prescribed intensity. This has applications in car headlight design or caustic design for example. We show that optical component design problems can be recast as optimal transport ones for different cost functions and we explain how this allows to study the existence and the regularity of the solutions of such problems. We also show how, using computational geometry, we can use an efficient numerical method namely the damped Newton's algorithm to solve all these problems. We will end up with a single generic algorithm able to efficiently build an optical component with a prescribed reflected (or refracted) illumination. We show the convergence of the Newton's algorithm to solve the optimal transport problem when the source measure is supported on a finite union of simplices. We then describe the common relation between eight optical component design problems and show that they can all be seen as discrete Monge-Ampère equations. We also apply the Newton's method to optical component design and show numerous simulated and fabricated examples. Finally, we look at a problem arising in computational optimal transport namely the choice of the initial weights. We develop three simple procedures to find "good" initial weights which can be used as a starting point in computational optimal transport algorithms.
**Keywords:** optimal transport, computational geometry, reflector design.

# Contents

# General introduction

DESIGNING physical materials with desired reflection or refraction patterns is attracting more and more attention in many research fields. This includes the field of *non-imaging optics* where one is interested in designing optical components that transfer the radiation emitted by a light source onto a prescribed target illumination. The question of building such components naturally appears when one wants to optimize the use of light energy to decrease light loss or light pollution for example. Applications include the design of car headlights that have a specific shape to avoid directly lighting up incoming cars; street lamps to avoid wasting energy; solar ovens or hydroponic agriculture to optimize their production. Another interesting field of application is more a creative one where one wants to design components that produce aesthetically pleasing *caustics*, one can for instance think of architecture or computer graphics.

In traditional optics, one wants to transport one image from a source domain onto a target domain. More precisely, one is given a bijection between the incident source rays and the target positions and one wants to find the surface that reflects (for a mirror) or refracts (for a lens) the source rays onto the target positions respecting the given bijection. Roughly speaking, if we have a bijection $S : X \to Y$ from a source domain $X$ to a target domain $Y$ and if we model the surface by a height field $\varphi : X \to \mathbb{R}$, then we want to find such function $\varphi$ such that $T_\varphi(x) = S(x)$ for all $x \in X$. Here $T_\varphi : X \to Y$ is a function that models the behaviour of the component when it is hit by a ray coming from a point in $X$. In practice we can assume it is of the form $T_\varphi(x) = F(x, \nabla\varphi(x))$ where $F$ is known (it can for instance be Snell's law) and the problem can be recast as integrating the gradients $\nabla\varphi(x)$, meaning finding a surface with normals $\vec{n}_\varphi(x) = (\nabla\varphi(x), -1)/\left\|(\nabla\varphi(x), -1)\right\|$. Let us remark that this is a first order partial differential equation (PDE) on $\varphi$. In non-imaging optics, $S$ is not an input anymore and the problem amounts to finding both the bijection $S$ and the normal field $\vec{n}_\varphi$, making the problem more complex as it corresponds now to a second order PDE. Recently, a lot of research [WMB+05; PP05] has been dedicated to the development of efficient methods to solve such problems.

The classical approach to tackle this problem is to first estimate the bijection $S$ using a heuristic and then integrate the normal vector field $\vec{n}_\varphi$ [Kis+12; Yue+14]. Recent methods make use of a mathematical tool called *optimal transport* [Sch+14]. The optimal transport problem, as first posed by Gaspard Monge in the $18^{th}$ century, consists in finding a way of transporting a *source* measure onto a *target* measure while minimizing a cost function. In the last few years, this problem has received a lot of attention due to its relations with several areas of theoretical and applied mathematics such as optimization, partial differential equations, machine learning or geometry processing. Using optimal transport seems well suited in non-imaging optics since one is interested in *transporting* the source rays onto the target positions while minimizing some error. One category of methods uses optimal transport as a heuristic to estimate the bijection $S$ and then integrate the normal field as in traditional optics.

Another category of methods make use of recent striking results that show that optimal transport can be used to recover both the bijection $S$ and the parametrization $\varphi$ at the same time. We can, for instance, cite the work of Oliker and Caffarelli [CO08] or Wang [Wan96] who formulated the problem of finding the mirror that reflects an ideal point light source located at the origin onto a prescribed *far-field* (located at infinity) target density as an optimal transport problem. Other works showed that we can have a similar formulation for other optical component design problems: mirror and lens design for a collimated light source (a source that emits parallel rays) [GT13]; lens design for a point light source [GH09]. In all the cases, the Kantorovich potential gives access to a parametrization of the optical component. There is also a close relation between these results and the work of Alexandrov and Pogorelov [Pog64] or the Minkowski problem (reconstruct a convex surface given its Gaussian curvature). Let us also remark that, in most cases, the far-field setting is not realistic as in practice one wants to focus the target image at a finite distance, setting that is called the *near-field*. One can wonder if near-field problems as well as other optics problems can be formulated in a similar fashion, for instance when the light source is not ideal anymore. It was shown that near-field problems are not related to optimal transport but can be recast as solving so-called *generated Jacobian equations* [GK17] whose structure is very different from the equations arising in optimal transport. When the source is not ideal, we speak about *extended* sources and, in this case, for one surface point there can be multiple incident rays making the problem ill-posed. In this thesis, we focus on problems that can be formulated in terms of optimal transport.

Furthermore, we can distinguish optimal transport methods by the assumptions they make on the support of the source and target measures and on the cost function. In this thesis, we focus on the so-called *semi-discrete* setting where the source measure is continuous and the target measure is a sum of Dirac masses. Oliker [CO08] shows that this formulation is well suited to prove the existence and regularity of solutions to non-imaging optics problems. For instance, he shows that the solution to the mirror design for a point light source in the continuous case can be seen as the limit of solutions to semi-discrete problems which are easier to study. This setting has also multiple advantages:

- under some assumptions on the cost function and the support of the source density, one can show that the semi-discrete optimal transport problem can be recast as the maximization of a concave function involving the so-called *Laguerre* diagram;
- one can also show that the solutions satisfy some geometric properties namely being convex. This kind of geometric results is important in optics since convexity helps in milling the component and ensures that the normal field is well-behaved;
- efficient numerical methods exist and are well-studied from a theoretical viewpoint. We can for instance cite the *damped Newton's algorithm* whose convergence as well as convergence rates were proven in different settings: for the quadratic cost in the plane or for more general costs satisfying the so-called *Ma-Trudinger-Wang* condition.

In this thesis, we study in detail the relation between semi-discrete optimal transport and non-imaging optics. We also made the choice to also discretize the support of the source measure by triangulating it.

The semi-discrete setting is interesting to use in non-imaging optics because of many aspects but it also raises some difficulties. First, it involves computing the optimal transport between a set of codimension 1 (for example $\mathbb{S}^2$ for non-imaging problems involving a point light source) and a point cloud which is not a well studied topic. Then, the support of the measures are not structured i.e. are not Cartesian grids, contrary to recent methods in discrete optimal transport which take advantage of the grid structure to develop very efficient numerical methods. Another important point is the fact that the cost functions appearing in the optical component design problems are not "classical" which raises some theoretical issues about the existence of solutions. For instance, in the case of the mirror design for a point light source, the cost function $c$ is given by $c(x, y) = -\ln(1 - \langle x \mid y \rangle)$. Thus, if we want to be able to solve efficiently non-imaging optics problems using semi-discrete optimal transport, we need to make many refinements:

- study the convergence of the damped Newton's algorithm to solve optimal transport when the source measure is supported on lower-dimensional subsets (for instance a triangulated surface in $\mathbb{R}^3$);
- develop a robust way of computing the Laguerre diagram which, in the settings we consider, corresponds to computing the intersection between a triangulation and a *Power diagram*;
- choose a good initialization. As in traditional Newton's methods, the choice of the initial point heavily influences the convergence of the algorithm. It is even more important in the semi-discrete case since we will see that the convergence is only proven when the initial point satisfies some constraints.

In this thesis, we show that all the choices we made (optimal transport, semi-discrete setting, triangulating the support of the source measure) will allow us to develop a fully generic and efficient algorithm that is able to solve eight different non-imaging optics problems. We also show that we are able to handle large discretizations of the target, which was not possible before. To summarize, we leverage the formulation of these problems in terms of optimal transport to construct a coherent framework in which one can solve many optical component design problems in a suitably discretized fashion.

# Detailed outline

### Chapter 1

In the first chapter, we introduce the problem of optimal transport, its different formulations, the different cases we can consider and the associated numerical methods. We then look more in depth at the semi-discrete setting which is at the core of this thesis. We introduce the Laguerre diagram as well as the main numerical method that will be used throughout this thesis namely the damped Newton's algorithm. Our main contribution in this chapter is the study of semi-discrete optimal transport problem in an usual setting namely when the source measure is supported on a union of lower dimensional subsets (simplices) in $\mathbb{R}^d$ for the quadratic cost. A particular case of this setting is a triangulated surface in $\mathbb{R}^3$. We show the convergence of the damped Newton's method with linear speed under two assumptions:

- *regularity* of the source measure. More precisely, we enforce a condition that can be surprising at first that is a strong connectedness assumption which ensures that it is not possible to disconnect the support of the source measure by removing a finite number of points. We will see that this condition is necessary to prove the strong concavity of the Kantorovich functional and thus the convergence of the Newton's method;
- *genericity* of the target point cloud with respect to the support of the source measure.

These two hypotheses will be made clear in Section 1.4.1. We also explain why they are reasonable assumptions to obtain the convergence of the Newton's method to solve optimal transport in this setting. More precisely, we prove the following theorem (which is a simplified version of theorems 13 and 14):

**Theorem.** *Assume $\mu$ is a regular simplicial measure and $\nu$ a discrete probability measure whose support $Y = \{y_1, \dots, y_N\} \subset \mathbb{R}^d$ is in generic position with respect to the support of $\mu$. Then the damped Newton's method converges in a finite number of steps with a linear rate.*

After proving this theorem, we apply these results on some geometry processing algorithms on triangulations in $\mathbb{R}^3$. We can for instance cite the optimal quantization of a probability density over a triangulated surface (approximate a probability density with a point cloud) or remeshing (re-triangulate a mesh such that the new distribution of triangles respects a prescribed density), see Figure 1 for two examples. These results are published in [MMT18a].

### Chapter 2

In this chapter, we look at the connection between optimal transport and non-imaging optics problems. In particular, we explain, following [CGH08], how optimal transport helps in defining a notion of *Brenier* solution to optical component design problems. This formulation is interesting since it directly enforces some geometric properties of the solution namely being convex. We then place ourselves in a semi-discrete setting and show that the optical component

Figure 1 – **Examples of applications of optimal transport for source measures supported on triangulated surfaces.** Left: optimal quantization of a triangulated surface for a uniform density. Middle and right: remeshing using optimal transport (initial and new mesh).

design problems considered in this thesis have the same structure and can be recast as solving a non-linear system of equations of the form, see Equation (DMA)

$$\text{Find } \psi \in \mathbb{R}^N \text{ such that } \forall i \in \{1, \dots, N\}, \int_{\text{Lag}_i(\psi)} \mathrm{d}\mu(x) = \nu_i,$$

where $\mu$ is the source measure, $\nu = (\nu_i)_{1 \leqslant i \leqslant N}$ the discrete target probability measure and $\text{Lag}_i(\psi)$ is the *Laguerre cell* of $y_i$. This equation involves prescribing the amount of light reflected (or refracted) in a finite number of directions and can be seen as as a discretization of the so-called *Monge-Ampère* equation that appears in continuous optimal transport. Another of our contributions is that we study the cost function arising when building lenses for point light sources. In particular, we show that, under some assumptions, it satisfies the *Ma-Trudinger-Wang* condition. This condition appears when studying the regularity of solutions to optimal transport problems in the continuous setting. It is also important in the semi-discrete setting since it guarantees that the Laguerre cells are connected and the convergence of the damped Newton's method.

## Chapter 3

In this chapter, we see how we can leverage the formulation presented in the previous chapter to develop a generic framework to solve eight optical component design problems. This genericity will also be a consequence of the fact that, in the cases considered in this thesis, the Laguerre diagram can be expressed in the following fashion:

$$\text{Lag}_i(\psi) = \text{Pow}_i(P) \cap X$$

where $X$ is the support of the source measure which is a subset of $\mathbb{R}^2 \times \{0\}$ (for a collimated light source) or $\mathbb{S}^2$ (for a point light source) and $\text{Pow}_i(P)$ the Power cell of a point $p_i$. Since, in this chapter, we will suppose that $X$ is a triangulation, this will allow us to reuse the algorithm developed in the first chapter. We will end up with a generic algorithm that needs no parameter and that is able to solve all the optical component design problems considered

in this thesis. We also show how we can solve the *near-field* setting using a simple iterative procedure. We finish the chapter by showing numerous simulated and fabricated examples, see Figure 2 for two examples. These results are published in [MMT18b].

Figure 2 – **Examples of simulated and physical results obtained with our method.** Left: three lenses that refract the three channels of a color image for three collimated light sources. Right: fabricated mirror that reflects a collimated light source.

**Chapter 4**

In this last chapter, we look at an important aspect of computational optimal transport namely the choice of the initial point in the Newton's method. Indeed, to obtain the convergence of the damped Newton's algorithm that we use throughout this thesis, we need to ensure that all the Laguerre cells have positive mass at every stage of the algorithm and in particular at the beginning. We will see in Chapter 3 that this condition can be hard to satisfy in practice. To find such weights, we detail three simple procedures, prove their convergence and illustrate their behaviours on numerous numerical examples. We also explain that discrete optimal transport can also benefit from having a "better" initialization. Let us note that this is still an ongoing work.

## Publications

The publications associated with this thesis are the following:

- *Light in Power: A General and Parameter-free Algorithm for Caustic Design*, Quentin Mérigot, Jocelyn Meyron, Boris Thibert, ACM Transaction on Graphics (Transactions On Graphics, Proc. SIGGRAPH Asia), [MMT18b]
- *An algorithm for optimal transport between a simplex soup and a point cloud*, Quentin Mérigot, Jocelyn Meyron, Boris Thibert, SIAM Journal on Imaging Sciences, 11.2 (2018), pp. 1363–1389, [MMT18a]

# Introduction générale

Concevoir des matériaux qui réfléchissent ou réfractent des motifs choisis par l'utilisateur est un sujet de plus en plus étudié. Cela inclut *l'optique non imageante* qui s'intéresse à la conception de composants optiques transférant l'énergie lumineuse émise par une source de lumière vers une densité d'illumination cible prescrite à l'avance. Cette question apparaît naturellement lorsque l'on veut optimiser l'utilisation d'énergie lumineuse afin de diminuer les pertes énergétiques ou la pollution lumineuse. Les applications sont nombreuses et incluent par exemple la conception de phares de voitures ayant une forme spécifique pour éviter d'éblouir les autres voitures; les lampadaires pour éviter de gaspiller de l'énergie; les four solaires et l'agriculture hydroponique afin d'optimiser leurs productions. Un autre domaine d'application, plus créatif, est la conception de *caustiques* où l'on souhaite créer des composants optiques produisant des motifs agréables à l'œil.

Traditionnellement, en optique, l'objectif est de transporter une image d'un domaine source vers un domaine cible. Étant donnée une bijection entre les rayons sources et les positions cibles, l'objectif est de déterminer la surface qui réfléchit (pour un miroir) ou réfracte (pour une lentille) les rayons sources vers les positions cibles en respectant la bijection. Plus précisément, si on note $S : X \to Y$ une bijection d'un domaine source $X$ vers un domaine cible $Y$ et si nous modélisons la surface par un champ de hauteur $\varphi : X \to \mathbb{R}$, le but est de trouver une fonction $\varphi$ telle que $T_\varphi(x) = S(x)$ pour tout $x \in X$. Ici, $T_\varphi$ est une fonction modélisant le comportement du composant optique quand il est touché par un rayon venant d'un point $x \in X$. En pratique, on peut supposer qu'elle est de la forme $T_\varphi(x) = F(x, \nabla\varphi(x))$ où $F$ est connue (la loi de Descartes par exemple) et le problème revient à intégrer les gradients $\nabla\varphi(x)$ c'est-à-dire trouver une surface dont les normales sont données par $\vec{n}_\varphi(x) = (\nabla\varphi(x), -1)/\left\|(\nabla\varphi(x), -1)\right\|$. On peut remarquer que cela correspond à une équation aux dérivées partielles (EDP) du premier ordre en $\varphi$. En optique non imageante, $S$ n'est plus une donnée et le problème implique de trouver à la fois la bijection $S$ et le champ de normales $\vec{n}_\varphi$. Cela rend le problème plus complexe puisqu'il correspond maintenant à une EDP d'ordre deux. Récemment, beaucoup de travaux [WMB+05; PP05] se sont intéressés au développement de méthodes efficaces pour résoudre de tels problèmes.

L'approche standard pour aborder ce problème est, dans un premier temps, d'estimer la bijection $S$ en utilisant une heuristique et ensuite d'intégrer le champ de normales $\vec{n}_\varphi$ [Kis+12; Yue+14]. Des méthodes récentes utilisent un outil mathématique appelé le *transport optimal* [Sch+14]. Le problème du transport optimal, posé pour la première fois par Gaspard Monge au $18^{me}$ siècle, consiste à trouver la manière optimale de transporter une *mesure source* vers une *mesure cible* tout en minimisant une fonction coût. Récemment, ce problème a reçu beaucoup d'attention grâce à ces relations avec de nombreux domaines des mathématiques à la fois théoriques et appliquées comme l'optimisation, l'étude d'équations aux dérivées partielles, l'apprentissage automatique ou encore en *geometry processing*. Utiliser le transport optimal en optique non imageante semble être adapté car on souhaite *transporter* les rayons sources sur les positions cibles tout en minimisant une certaine erreur. Une catégorie de méthodes utilise

donc le transport optimal comme une heuristique pour estimer $S$ et ensuite intègre le champ de normales comme en optique traditionnelle.

Une autre catégorie de méthodes utilisent des résultats récents qui montrent que le transport optimal peut être directement utilisé pour déterminer à la fois la bijection $S$ et la paramétrisation $\varphi$. On peut par exemple citer le travail d'Oliker et Caffarelli [CO08] ou Wang [Wan96] qui ont formulé le problème consistant à trouver le miroir qui réfléchit la lumière émise par une source ponctuelle idéale située à l'origine vers une densité cible située à l'infinie (on parle de problème en *champ lointain*) en termes de transport optimal. D'autres travaux montrent que des formulations similaires existent pour d'autres problèmes de conception de composants optiques : conception de miroirs et de lentilles pour une source collimatée (émettant des rayons parallèles) [GT13]; conception de lentilles pour une source ponctuelle [GH09]. Dans tous les cas, il a été montré que le potentiel de Kantorovich donne accès à une paramétrisation du composant optique. Il existe aussi une forte relation entre ces résultats et les travaux d'Alexandrov et Pogorelov [Pog64] ou bien le problème de Minkowski (reconstruire une surface convexe à partir de la donnée de sa courbure de Gauss). On peut aussi remarquer que dans la plupart des cas, le problème en champ lointain n'est pas le plus réaliste car en pratique on souhaite plutôt focaliser l'image cible à une distance finie, ce problème est dit en *champ proche*. Nous pouvons nous demander si de tels problèmes, voire d'autres problèmes apparaissant en optique, peuvent aussi se formuler d'une manière semblable, comme par exemple lorsque la source de lumière n'est plus idéale. Il a été montré que les problèmes en champ proche ne sont pas des problème de transport optimal mais font partis d'une autre catégorie à savoir des *équations au Jacobien généré* [GK17] dont la structure est très différente des équations apparaissant en transport optimal. Quand la source n'est plus idéale, on parle de source *étendue* et dans ce cas, pour un point de la surface, il peut exister plusieurs rayons incidents, rendant le problème mal posé. Dans cette thèse, nous nous concentrons sur les problèmes qui peuvent être formulés en termes de transport optimal.

De plus, nous pouvons différencier les méthodes numériques utilisées pour résoudre le transport optimal en fonction des hypothèses qu'elles font sur le support des mesures source et cible et sur la fonction coût. Dans cette thèse, nous nous concentrons sur le cadre *semi-discret* où la mesure source est continue et la mesure cible est une somme de masses de Dirac. Oliker [CO08] a montré que cette formulation est adaptée pour montrer l'existence et la régularité des solutions de problèmes d'optique non imageante. Par exemple, il a montré que la solution au problème de conception d'un miroir qui réfléchit la lumière émise par une source ponctuelle dans le cas continu peut être vue comme la limite de solutions de problèmes semi-discrets, qui sont plus simples à étudier. Ce cadre a aussi d'autres avantages:

- sous certaines hypothèses sur la fonction coût et le support de la densité source, on peut montrer que le problème de transport optimal semi-discret équivaut à maximiser une fonction concave faisant intervenir le *diagramme de Laguerre*;

- nous pouvons aussi montrer que les solutions vérifient certaines propriétés géométriques, à savoir des propriétés de convexité. Ces résultats sont importants en optique car la convexité aide au fraisage des composants et assure que le champ de normales est bien

défini;

- des méthodes numériques efficaces existent et sont bien comprises d'un point de vue théorique. Nous pouvons par exemple citer la *méthode de Newton amortie* dont la convergence ainsi que la vitesse de convergence ont été obtenues dans différents cas : pour le coût quadratique dans le plan ou pour des fonctions coûts plus générales qui satisfont la condition de *Ma-Trudinger-Wang*.

Dans cette thèse, nous étudions en détails la relation entre le transport optimal semi-discret et l'optique non imageante. Nous faisons aussi le choix de discrétiser le support de la mesure source en le triangulant.

Il est intéressant d'utiliser le cadre semi-discret en optique non imageante par beaucoup d'aspects mais il soulève également certaines difficultés. Tout d'abord, il implique de calculer le transport optimal entre un ensemble de codimension 1 (par exemple $\mathbb{S}^2$ pour les problèmes où la source est ponctuelle) et un nuage de points, problème qui n'est pas très bien compris. De plus, le support des mesures n'est pas structuré dans le sens où ce ne sont pas des grilles cartésiennes, contrairement à certaines méthodes récentes en transport optimal discret qui tirent avantage de cette structure pour développer des méthodes numériques très efficaces. Un autre point important est le fait que les fonctions coûts apparaissant dans les problèmes de conception de composants optiques ne sont pas « classiques », ce qui soulève des problèmes sur l'existence de solutions pour de tels problèmes. Par exemple, dans le cas de la conception d'un miroir pour une source ponctuelle, la fonction coût est donnée par $c(x, y) = -\ln(1 - \langle x \mid y \rangle)$. Par conséquent, si nous voulons être capable de résoudre efficacement des problèmes d'optique non imageante, nous devons faire de nombreux raffinements :

- étudier la convergence de la méthode de Newton amortie pour résoudre le transport optimal quand la mesure source est supportée sur des ensembles de dimension non pleine (par exemple sur une surface triangulée dans $\mathbb{R}^3$);
- développer une méthode robuste pour calculer le diagramme de Laguerre qui, dans les cas que l'on considèrent, peut être vu comme l'intersection entre une triangulation et un *diagramme de puissance*;
- choisir une bon itéré initial Il est connu que dans les méthodes de Newton, le choix de lu point initial influence grandement la convergence de l'algorithme. Il est encore plus important dans le cadre semi-discret car nous verrons que la convergence est prouvée uniquement quand ce point satisfait certaines contraintes.

Dans cette thèse, nous montrons que les choix que nous avons faits (transport optimal, cadre semi-discret, trianguler le support de la mesure source) vont nous permettre de développer un algorithme générique et efficace capable de résoudre huit différents problèmes de conception de composants optiques. Il nous permettra aussi de considérer de très grandes discrétisations de la mesure cible. Pour résumer, nous tirons profit de la formulation de ces problèmes en termes de transport optimal pour construire un cadre cohérent dans lequel nous pouvons résoudre de nombreux problèmes de conception de composants optiques convenablement discrétisés.

# Plan détaillé

### Chapitre 1

Dans ce premier chapitre, nous introduisons le problème du transport optimal, ses différentes formulations, les différents cas que l'on peut considérer ainsi que les méthodes numériques associées. Nous étudions ensuite plus en détail le cadre semi-discret qui est au cœur de cette thèse. Nous introduisons le diagramme de Laguerre ainsi que la méthode numérique principale qui sera utilisé tout au long de ces travaux à savoir la méthode de Newton amortie. Notre principale contribution dans ce chapitre est l'étude du transport optimal semi-discret dans un cadre inhabituel à savoir quand la mesure source est supportée sur une union d'ensembles de dimension non pleine (des simplexes) dans $\mathbb{R}^d$ pour le coût quadratique. Un cas particulier de ce cadre est quand la mesure source est supportée sur une surface triangulée dans $\mathbb{R}^3$. Nous montrons la convergence linéaire de la méthode de Newton amortie sous deux hypothèses :

- *régularité* de la mesure source. Plus précisément, nous imposons une condition de forte connexité qui assure que le support de la mesure source ne peut pas être déconnecté en enlevant un nombre fini de points. Nous verrons que cette condition est nécessaire pour prouver la stricte concavité de la fonctionnelle de Kantorovich et donc la convergence de la méthode de Newton;
- *généricité* du nuage de points cible par rapport au support de la mesure source.

Nous clarifierons ces deux hypothèses dans la Section 1.4.1. Nous expliquons ensuite en quoi ces hypothèses sont raisonnables pour obtenir la convergence de la méthode de Newton pour résoudre le problème du transport optimal dans ce cadre. Plus précisément, nous démontrons le théorème suivant (qui est une version simplifiée des théorèmes 13 et 14) :

**Théorème.** *Si $\mu$ est une mesure simpliciale régulière et $\nu$ une mesure de probabilité discrète dont le support $Y = \{y_1, \dots, y_N\} \subset \mathbb{R}^d$ est en position générique par rapport au support de $\mu$, alors la méthode de Newton amortie converge linéairement en un nombre fini d'étapes.*

Après avoir démontré ce théorème, nous appliquons ces résultats sur différents problèmes géométriques faisant intervenir des surfaces triangulées dans $\mathbb{R}^3$. Cela inclut par exemple la quantification optimale d'une densité de probabilité sur une surface triangulée (approcher une densité de probabilité par un nuage de points) ou le remaillage (re-trianguler un maillage de telle sorte que la distribution des triangles respecte une certaine densité), voir Figure 3 pour deux exemples. Ces résultats sont publiés dans [MMT18a].

### Chapitre 2

Dans ce chapitre, nous nous intéressons à la relation entre le transport optimal et l'optique non imageante. En particulier, nous expliquons, en nous inspirant de [CGH08], comment le transport optimal est utile pour définir une notion de solution *à la Brenier* pour les problèmes

Figure 3 – **Exemples d'applications du transport optimal pour des mesures sources supportées sur des surfaces triangulées.** À gauche : quantification optimale d'une surface triangulée pour une densité uniforme. Au milieu et à droite : remaillage en utilisant le transport optimal (maillages initial et final).

de conception de composants optiques. Cette formulation est intéressante car elle encode directement certaines propriétés géométriques de la solution à savoir sa convexité. Nous nous plaçons ensuite dans le cadre semi-discret et montrons que les problèmes de conception de composants optiques considérés dans cette thèse ont la même structure et sont équivalents à résoudre un système non linéaire d'équations de la forme, voir Équation (DMA)

$$\text{Trouver } \psi \in \mathbb{R}^N \text{ tel que } \forall i \in \{1, \dots, N\}, \int_{\mathrm{Lag}_i(\psi)} \mathrm{d}\mu(x) = \nu_i,$$

où $\mu$ est la mesure source, $\nu = (\nu_i)_{1 \leqslant i \leqslant N}$ la mesure de probabilité cible et $\mathrm{Lag}_i(\psi)$ est la *cellule de Laguerre* de $y_i$. Cette équation impose de prescrire la quantité de lumière réfléchie (ou réfractée) dans un nombre fini de directions et peut être vue comme une discrétisation de l'équation de *Monge-Ampère* apparaissant en transport optimal. Une autre de nos contributions est l'étude de la fonction coût apparaissant lorsque l'on veut concevoir des lentilles pour des sources ponctuelles. En particulier, nous montrons, sous certaines hypothèses, qu'elle vérifie la condition de *Ma-Trudinger-Wang*. Cette condition apparaît lors de l'étude de la régularité des solutions de transport optimal dans le cadre continu. Elle est aussi importante dans le cadre semi-discret car elle garantit la connexité des cellules de Laguerre ainsi que la convergence de la méthode de Newton amortie.

## Chapitre 3

Dans ce chapitre, nous voyons comment on peut tirer parti de la formulation présentée dans le chapitre précédent pour développer un cadre générique pour résoudre huit problèmes de conception de composants optiques. Cette généricité sera aussi une conséquence du fait que dans les cas que l'on considère dans cette thèse, le diagramme de Laguerre a la forme suivante

$$\mathrm{Lag}_i(\psi) = \mathrm{Pow}_i(P) \cap X$$

où $X$ est le support de la mesure source qui est un sous-ensemble de $\mathbb{R}^2 \times \{0\}$ (pour une source

collimatée) ou $\mathbb{S}^2$ (pour une source ponctuelle) et $\mathrm{Pow}_i(P)$ la cellule de puissance du point $p_i$. Puisque dans ce chapitre, nous supposons que $X$ est une triangulation, cela nous permettra de réutiliser l'algorithme développé dans le premier chapitre. Nous obtiendrons un algorithme générique, sans paramètre et qui est capable de résoudre tous les problèmes de conception de composants optiques considérés dans cette thèse. Nous montrons aussi comment nous pouvons utiliser une procédure itérative simple pour résoudre le problème en champ proche. Nous terminons ce chapitre en présentant différents exemples simulés et des prototypes physiques, voir Figure 4 pour deux exemples. Ces résultats sont publiés dans [MMT18b].

Figure 4 – **Exemple simulé et prototype physique obtenus avec notre méthode.** À gauche : trois lentilles qui réfractent les trois canaux d'une image couleur avec trois sources collimatées. À droite : miroir qui réfléchit une source collimatée.

**Chapitre 4**

Dans ce dernier chapitre, nous nous intéressons à un autre aspect important du transport optimal numérique à savoir le choix de l'itéré initial dans la méthode de Newton. En effet, afin d'obtenir la convergence de la méthode de Newton amortie que nous utilisons dans toute cette thèse, nous avons besoin d'assurer que toutes les cellules de Laguerre ont une masse strictement positive à chaque étape de l'algorithme et en particulier au début. Nous verrons dans le Chapitre 3 que cette condition est difficile à satisfaire en pratique. Nous détaillons trois procédures permettant de trouver un tel itéré, nous montrons leur convergence et nous les illustrons sur différents exemples. Nous expliquons également en quoi les méthodes numériques pour le transport optimal discret pourrait être améliorées en choisissant un meilleur itéré initial. Ce chapitre correspond à un travail en cours.

## Publications

Les publications associées avec cette thèse sont les suivantes :

- *Light in Power: A General and Parameter-free Algorithm for Caustic Design*, Quentin Mérigot, Jocelyn Meyron, Boris Thibert, ACM Transaction on Graphics (Transactions On Graphics, Proc. SIGGRAPH Asia), [MMT18b]
- *An algorithm for optimal transport between a simplex soup and a point cloud*, Quentin Mérigot, Jocelyn Meyron, Boris Thibert, SIAM Journal on Imaging Sciences, 11.2 (2018), pp. 1363–1389, [MMT18a]

Chapter 1

# Transporting a simplex soup on a point cloud

## Contents



In mathematics, the *optimal transport* problem, first introduced by Monge [Mon81], consists in finding an optimal way of transporting one probability measure onto another one. In the last few years, it has received a lot of attention in mathematics (see e.g. [Vil09; San15]), in mathematical physics, in machine learning but also in computational geometry and in geometry processing because of the intimate connection between optimal transport maps for the quadratic cost and Power diagrams [OP89; AHA98; Mér11; Goe+12; CMT15; Lév15]. Since it allows to measure distances and interpolate between functions (and even more general objects) by taking into account both the intensity of the function and its graph, it has been used in numerous applications such as image processing [TPG16] or machine learning [Cut13; Fro+15]. Furthermore, it also defines a geometry on the space of probability measures which can be for instance used to solve partial differential equations, see for instance [GM17] that uses optimal transport to enforce the incompressibility constraint for the Euler equation. In

this first chapter, we look at a degenerate setting of semi-discrete optimal transport where the source measure is supported on a collection of lower-dimensional subsets of $\mathbb{R}^d$.

In Section 1.1, we introduce the optimal transport problem, recall its different formulations and the main results on the existence of solutions. In Section 1.2, we review the existing numerical methods in different settings. In Section 1.3, we look more precisely at the so-called *semi-discrete* setting. We introduce the concept of *Laguerre diagram* and study the main conditions necessary for the transport maps to be well-defined as well as the main algorithm that we will use throughout this thesis, namely the *damped Newton's method*. We also show why this setting can be interesting through different applications. In Section 1.4, we introduce the setting we will look in this chapter namely the optimal transport between a simplicial measure and a finitely supported measure. We give the main theorem on the convergence with linear speed of the damped Newton's method to solve the optimal transport in this particular setting. The proof relies on two conditions: (i) a genericity condition on the point cloud with respect to the support of the source measure and (ii) a (strong) connectedness condition on the support of the source measure. The convergence will then be a direct consequence of the regularity and the strict motonicity of the Kantorovich functional. Finally, in Section 1.5, we present various numerical illustrations of the effectiveness of this algorithm through different examples such as optimal quantization of a probability density over a surface, remeshing or rigid point set registration on a mesh.

## 1.1 Generalities on optimal transport

In this section, we introduce the main formulations of optimal transport and state the main results of existence and uniqueness of optimal transport maps. We start by introducing the concept of the pushforward measure.

**Definition 1.1** (Pushforward measure)
*Let $X$ and $Y$ be two measurable spaces, $\mu$ a measure on $X$ and a function $T : X \to Y$. We define the* pushforward *of the measure $\mu$ by $T$ as the measure $T_{\#}\mu$ such that*

$$\forall A \subset Y \text{ Borel set, } T_{\#}\mu(A) = \mu(T^{-1}(A)).$$

*If $\mu$ and $\nu$ are two measures on $X$ and $Y$, and if $T_{\#}\mu = \nu$, then $T$ preserves the mass between $X$ and $Y$.*

We can now state the *Monge* formulation of optimal transport.

**Definition 1.2** (Monge formulation of optimal transport)
*For two probability measures $\mu$ and $\nu$ supported respectively on $X$ and $Y$ subsets of $\mathbb{R}^d$, and a cost function $c : X \times Y \to \mathbb{R}$, the optimal transport problem between $\mu$ and $\nu$ consists in finding a map called the* transport map $T : X \to Y$ *that minimizes the* global transportation cost. *It*

*can be summarized as*

$$\inf_{T:X\to Y} \int_X c(x,T(x))\mathrm{d}\mu(x) \text{ under the constraint } T_{\#}\mu = \nu. \tag{M}$$

An illustration of this problem can be found in Figure 1.

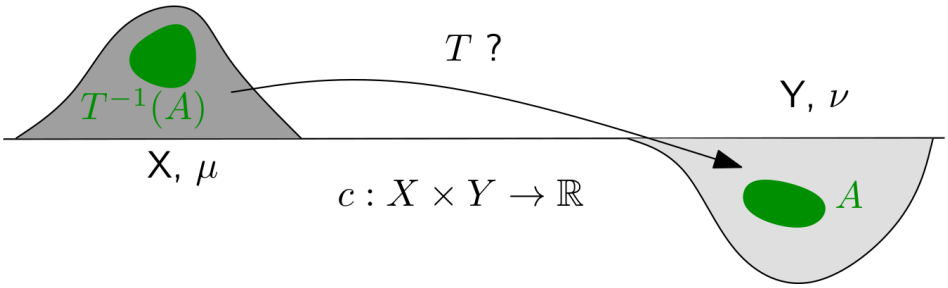


Figure 1 – **Illustration of the optimal transport problem.** The goal of optimal transport is to find a mass-preserving map $T : X \to Y$ that transports mass between $(X, \mu)$ and $(Y, \nu)$ while minimizing the transport cost $c$.

With this formulation, $T$ must be a valid map making impossible to split mass which could be necessary for some applications. This prevents for instance the case where $\mu$ is a Dirac mass and $\nu$ is not. To circumvent this issue Kantorovich proposed in [Kan58] to relax the constraint $T_{\#}\mu = \nu$ as follows:

**Definition 1.3** (Kantorovich formulation of optimal transport)
*For two probability measures $\mu$ and $\nu$ supported respectively on $X$ and $Y$ subsets of $\mathbb{R}^d$, and a cost function $c : X \times Y \to \mathbb{R}$, we are looking for a probability measure called the* transport plan *$\gamma$ on the product space $X \times Y$ with marginals $\mu$ and $\nu$ i.e.*

$$\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) \text{ where } \Gamma(\mu,\nu) = \{\gamma \mid (P_X)_{\#}\gamma = \mu \text{ and } (P_Y)_{\#}\gamma = \nu\} \tag{K}$$

*where $P_X : (x, y) \in X \times Y \mapsto x$ and $P_Y : (x, y) \in X \times Y \mapsto y$ are the projections on $X$ and $Y$.*

Remark 1. *There is a relation between the Monge problem and the relaxed Kantorovich problem: if $(I \times T)_{\#}\mu$ (where $I$ denotes the identity map) is a transport plan then $T_{\#}\mu = \nu$, i.e. $T$ is a transport map, so that the Kantorovich cost is smaller than the Monge cost. See Chapter 1 of [San15] for more details on the relation between the two formulations.*

Remark 2. *When $c(x, y) = \|x - y\|^p$ for $p \geqslant 1$, then the optimal cost to transport $\mu$ onto $\nu$ is called the* Wasserstein *$p$ distance and is thus defined as*

$$\mathrm{W}_p(\mu,\nu) = \min_{\gamma} \left\{ \int_{X\times Y} \|x-y\|^p \,\mathrm{d}\gamma(x,y) \mid \gamma \in \Gamma(\mu,\nu) \right\}^{1/p}$$

*The function $W_p$ defines a distance over the space of probability measures and can be used to define a notion of geodesics, see [San15].*

With this formulation, the problem becomes a linear program with convex constraints. One can then take the dual formulation that amounts to solving

$$\sup\left\{\int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu \mid \varphi \in L^1(\mu), \psi \in L^1(\nu), \forall x, y,\ \varphi(x) + \psi(y) \leqslant c(x,y)\right\}. \qquad (K^*)$$

The functions $\varphi$ and $\psi$ that solve this problem are called *Kantorovich potentials*.

This dual formulation will be at the heart of many numerical methods for solving optimal transport notably in the semi-discrete case. We now state the main existence results for optimal transport plans for the primal and dual Kantorovich problems. The proofs of these theorems can be found for instance in [Vil03; San15].

**Theorem 3.** *Let $\mu$ and $\nu$ be two probability measures supported on $X$ and $Y$ compact metric spaces and $c : X \times Y \to \mathbb{R}$ be lower semi-continuous and bounded from below. Then* (K) *admits a solution.*

**Theorem 4.** *Let $\mu$ and $\nu$ be two probability measures supported on $X$ and $Y$ compact and $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ be a continuous cost function. Then $(K^*)$ admits a solution.*

The next theorem is a duality result that ensures that, under some assumptions, the Kantorovich problem (K) and its dual formulation $(K^*)$ have the same solution.

**Theorem 5.** *If $X$ and $Y$ are two compact manifolds in $\mathbb{R}^d$ and $c : X \times Y \to \mathbb{R}$ is uniformly continuous and bounded then $(K^*)$ admits a solution that coincides with the solution of* (K).

Remark 6. *The previous theorem can be stated for very general spaces $X$ and $Y$ (Polish spaces). Here, to avoid introducing other notions, we restrict ourselves to the setting in which we will use this result.*

As a previous theorem states, the existence of optimal transport maps is obtained through fairly general assumptions. However, uniqueness results are obtained under stricter conditions. For instance, when the cost function is of the form $c(x, y) = h(x - y)$ where $h$ is a strictly convex function, then Brenier's theorem [Bre91] allows to get an expression for the optimal transport map.

**Theorem 7.** *Let $\mu$ and $\nu$ be two probability measures supported on a compact domain $\Omega \subset \mathbb{R}^d$, then there exists a unique transport plan $\gamma$ for the cost $c(x, y) = h(x - y)$ for $h$ strictly convex. It is of the form $(I, T)_{\#}\mu$ provided that $\mu$ is absolutely continuous and $\partial\Omega$ negligible. Moreover, there is the following relation between $T$ and a Kantorovich potential $\varphi$:*

$$T(x) = x - (\nabla h)^{-1}(\nabla\varphi(x)).$$

*In particular, when $c$ is the quadratic cost $c(x, y) = \|x - y\|^2$ then*

$$T(x) = \nabla_x\left(\frac{x^2 - \varphi(x)}{2}\right) = \nabla F(x) \text{ where } F \text{ is convex.}$$

We finish this section by recalling the link between optimal transport and the Monge-Ampère equation, see [BFO12] for applications and numerics.

REMARK 8 (Monge-Ampère equation). *When $\mu$ and $\nu$ are two probability densities on $\mathbb{R}^d$, and $c$ is the quadratic cost function, then the previous theorem asserts that the optimal transport map $T = \nabla F$ is the gradient of a convex function $F$. Injecting this expression into $T_{\#}\mu = \nu$, we get that $F$ solves (in a suitable weak sense) the non-linear differential equation called the* Monge-Ampère *equation*

$$\nu(\nabla F(x)) \det(\mathrm{D}^2 F(x)) = \mu(x) \text{ such that } \nabla F(X) = Y \text{ and } F \text{ convex.} \tag{MA}$$

*When the cost $c$ is not the quadratic cost but satisfies the so-called* (Twist) *condition (see next section), the equation becomes trickier to study and is of the general form*

$$\det(\mathrm{D}^2\varphi(x) - \mathrm{D}^2_{xx} c(x, T(x))) = \det(D^2_{xy} c(x, T(x)) \frac{\mu(x)}{\nu(T(x))}$$

*where $\varphi$ is a Kantorovich potential. See [Vil09] for more details.*

## 1.2 Computational optimal transport

In this section, we look at the main numerical methods to solve optimal transport problems. We do not detail here the semi-discrete setting since it will be studied in depth in Section 1.3. We distinguish the methods by the assumptions they make on the support of the source and target measures and/or the cost function. Indeed, we will see that one has to make a compromise between the complexity of the algorithm and how general the cost function can be.

In the following of this chapter, we will replace $\psi$ by its opposite in the dual formulation of the Kantorovich problem to make the results clearer, $(K^*)$ becomes

$$\sup \left\{ \int_X \varphi \mathrm{d}\mu - \int_Y \psi \mathrm{d}\nu \mid \varphi \in L^1(\mu),\ \psi \in L^1(\nu),\ \forall x, y,\ \varphi(x) - \psi(y) \leqslant c(x, y) \right\}. \tag{$K^*$}$$

### 1.2.1 Discrete setting

**Uniform measures on sets with the same cardinal.** When both probability measures are uniform and supported on point clouds with the same cardinal $N$, finding optimal transport maps is equivalent to the assignment problem appearing in combinatorial optimisation. One of the first methods is the *auction* algorithm developed by Bertsekas [Ber81; Ber88]. It can be found in Algorithm 1. This algorithm is based on a coordinate-wise ascent on the dual cost. Since it is known that coordinate-wise ascent methods can be stuck at some points which are not maximizers, the auction algorithm enforces another condition known as the *$\epsilon$-complementary slackness* condition to avoid these points.

Roughly speaking, the auction algorithm is an iterative method which builds a bijection $T$

**Algorithm 1:** Bertsekas' auction algorithm

**Input:** cost function $c$, increment $\epsilon > 0$
**Output:** bijection $T : X \to Y$, Kantorovich potential $\psi \in \mathbb{R}^Y$
$\psi \leftarrow 0$
$S \leftarrow \emptyset$
**while** $\exists x \in X \setminus S$ **do**
  $y_0 \leftarrow \operatorname{argmin}_{y \in Y}(c(x, y) + \psi(y))$
  $\psi(y_0) \leftarrow \psi(y_0) + \epsilon$
  **if** $\exists x' \in X$ *s.t.* $T(x') = y_0$ **then**
    $S \leftarrow S \setminus \{x'\}$
  **end**
  $S \leftarrow S \cup \{x\}, \quad T(x) \leftarrow y_0$
**end**
**return** $T, \psi$

(corresponding to the transport map) between $X$ and $Y$ while maintaining a set of assigned points of $X$ and a weight vector $\psi$. At each iteration, if $x$ is an unassigned point, we look for the target point $y_0$ such that $y_0 \in \operatorname{argmin}_{y \in Y}(c(x, y) + \psi(y))$. We then try to increase the corresponding weight $\psi(y_0)$ by a fixed increment $\epsilon > 0$ and checks if $y_0$ has already a correspondence $x'$. If it is the case, we swap $x$ and $x'$, assign $x$ to $y_0$ and starts a new iteration. The original version of this algorithm has a complexity of $\mathrm{O}(N^3 C/\eta)$ where $C = \max_{x,y \in X \times Y} c(x, y)$ and $\eta$ the numerical error. There exists a scaled version that has a worst-case complexity of $\mathrm{O}(N^3 \log(C/\eta))$, More details on the convergence analysis of this algorithm can be found in [Mér13].

Remark 9. *The expression* $\min_{y \in Y}(c(x, y) + \psi(y))$ *corresponds to the* $c$-transform *of the function* $\psi$ *evaluated at the point* $x$ *which is a notion appearing in optimal transport. It is also used in the semi-discrete setting, see Section 1.3 and [LS17].*

**General discrete case.** The other most popular methods are based on the entropic regularization of optimal transport introduced in [Cut13]. Let us note the work of Schmitzer [SS13] that establishes a connection between these methods and Bertsekas' auction algorithm. The core of these methods is the *Sinkhorn* or *Iterative Proportional Fitting Procedure* (IPFP) procedure, which we now briefly explain. The algorithm maintains two Kantorovich potentials $\varphi$ and $\psi$ and update them iteratively. We denote by $\epsilon$ the regularization parameter. The basis of this method is that when $\epsilon$ tends to 0 then the regularized problem should converge to the non-regularized one, see [Car+17] for a study of the quadratic cost. Furthermore, the update step on $\varphi$ can be written as

$$\varphi^{k+1}(x) = \epsilon \log(\mu_x) - \epsilon \log \left( \sum_{y \in Y} \exp(-\frac{1}{\epsilon}(c(x, y) + \psi^k(y))) \right). \qquad \text{(Sinkhorn/Update)}$$

One can see that when $\epsilon$ tends to 0, this expression becomes

$$\varphi^{k+1}(x) = \operatorname*{argmin}_{y \in Y}(c(x,y) + \psi^k(y)).$$

One can also remark that this step corresponds exactly to the update step in the auction algorithm (see the previous paragraph).

**Algorithm 2:** Sinkhorn-Knopp algorithm

**Input:** measures $\mu, \nu$, cost matrix $C$, regularization parameter $\epsilon > 0$, $k_{max} > 0$
**Output:** Kantorovich potentials $\varphi, \psi$
$u_0 \leftarrow \mathbf{1}_X, v_0 \leftarrow \mathbf{1}_Y$
$G_\epsilon = \exp(-\frac{C}{\epsilon})$
**for** $0 \leqslant k \leqslant k_{max}$ **do**
  $u_{k+1} \leftarrow \mu/(G_\epsilon v_k)$
  $v_{k+1} \leftarrow \nu/(G_\epsilon^T u_{k+1})$
**end**
**return** $\epsilon \ln(u_{k_{max}}), -\epsilon \ln(v_{k_{max}})$

Making the change of variable $(u, v) = (\exp(\varphi/\epsilon), \exp(-\psi/\epsilon))$, it is easy to see that each iteration practically consists only in two matrix-vector products, the matrix $G_\epsilon$ being the so-called *Gibbs* kernel, see Algorithm 2. It is therefore crucial to make this algorithm efficient that these matrix-vector products can be computed in subquadratic time. This is known to be possible when the point sets are distributed on regular grids and when the transport cost is $\ell^p$, using a simple tensorization trick. It is also possible when both point sets lie on the same surface and when the cost is the squared geodesic distance [Sol+15]. Due to its easy implementation and effectiveness, this formulation has been used with success in numerous applications such as barycenters of measures [Sol+15] or surface matching [Fey+17]. The main weakness of the approach is the choice of the regularization parameter $\epsilon$. In general, the relation between objects of the regularized and non-regularized problems is still not well understood.

### 1.2.2 Continuous setting

When both measures are continuous, one of the first numerical methods was the so-called *Benamou-Brenier* algorithm proposed in [BB00]. It is based on a computational fluid mechanics formulation of optimal transport between two probability densities $\rho_0$ and $\rho_1$ defined on $\mathbb{R}^d$. The square of the Wasserstein distance $\mathrm{W}_2$ can be expressed as follows:

$$\mathrm{W}_2^2(\rho_0, \rho_1) = \inf_{\rho, v} \left\{ \int_{\mathbb{R}^d} \int_{t=0}^{1} \rho(t,x) \left\| v(t,x) \right\|^2 \mathrm{d}t\mathrm{d}x \mid \rho(0,\cdot) = \rho_1, \rho(1,\cdot) = \rho_1, \partial_t \rho + \nabla \cdot (\rho v) = 0 \right\}.$$

This dynamic formulation allows to use efficient finite element or finite difference discretizations or augmented Lagrangian algorithms for solving the optimization problem.

The authors of [LR05] proposed to use a Newton's algorithm by linearizing the Monge-

Ampère operator. Other methods are based on efficient discretizations of the *Monge-Ampère* equation (MA) such as monotone schemes [BFO12; BCM16]. The main difficulties that appear when solving Monge-Ampère type equations are: first, imposing the boundary condition $\nabla F(X) = Y$ for $F$ convex which is known as *second boundary value problem* (BV2) is hard to impose. In practice, this condition can be handled using for instance a Hamilton-Jacobi equation on the boundary, see [BFO14]. Secondly, the fact that it is a degenerate elliptic equation imposes to design specific numerical schemes satisfying some kind of motonicity property on the space of convex functions. This scheme should also be designed to enforce the convexity of the discrete solution. More details on the study of such equations and the regularity of its solutions can be found in [Gut12].

The next section will be dedicated to the study of the so-called *semi-discrete* setting i.e. when the source measure is continuous and the target measure is supported on a point cloud. We will show that in this setting, one can develop efficient and robust numerical methods for the quadratic cost $c(x, y) = \|x - y\|^2$.

## 1.3 Optimal transport in the semi-discrete setting

We now suppose that the target measure $\nu$ is finitely supported on a point cloud $Y = \{y_1, \dots, y_N\} \subseteq \mathbb{R}^d$, i.e. $\nu = \sum_{1 \leqslant i \leqslant N} \nu_i \delta_{y_i}$ where $\sum_{i=1}^N \nu_i = 1$. This setting will be referred to as the *semi-discrete* setting.

### 1.3.1 Kantorovich functional

The dual Kantorovich problem presented earlier becomes the following

$$\sup \left\{ \int_X \varphi \mathrm{d}\mu - \sum_{i=1}^N \psi_i \nu_i \mid \varphi \in L^1(\mu),\ \psi \in \mathbb{R}^N,\ \forall x, i,\ \varphi(x) - \psi_i \leqslant c(x, y_i) \right\}. \tag{$K^*$}$$

We now take a closer look at the constraint $\varphi(x) - \psi_i \leqslant c(x, y_i)$. We can rewrite it as $\varphi(x) \leqslant c(x, y_i) + \psi_i$. An easy computation show that if $(\varphi, \psi)$ satisfies this constraint then $(\psi^c, \psi)$, where $\psi^c(x) = \inf_{1 \leqslant i \leqslant N} (c(x, y_i) - \psi_i)$, also satisfies it. Another calculation (see [LS17] for instance) shows that maximizers of ($K^*$) are of the form $(\psi^c, \psi)$. Following [AHA98; GM96], we then denote by $\Phi$ the following function that we call the *Kantorovich functional*

$$\Phi(\psi) = \int_X \inf_{1 \leqslant i \leqslant N} (c(x, y_i) + \psi_i) \, \mathrm{d}\mu(x) - \sum_{i=1}^N \nu_i \psi_i \tag{Kantorovich}$$

The goal of this section is to show that under some assumptions on the cost function $c$, one can relate the semi-discrete optimal transport problem to the maximization of the functional $\Phi$.

Looking at the expression of $\Phi$, it is natural to define the following so-called *Laguerre*

*diagram* of a point cloud $Y$.

**Definition 1.4** (Laguerre diagram)
*For a point cloud $Y = \{y_1, \dots, y_N\} \subset \mathbb{R}^d$, a cost function $c : X \times Y \to \mathbb{R}$ and weights $\psi \in \mathbb{R}^N$, we define the Laguerre cell of $y_i \in Y$ by*

$$\mathrm{Lag}_i(\psi) = \{x \in X \mid \forall j,\ c(x, y_i) + \psi_i \leqslant c(x, y_j) + \psi_j\}. \tag{Laguerre}$$

*The* Laguerre diagram *is the collection of all the Laguerre cells $(\mathrm{Lag}_i(\psi))_{1 \leqslant i \leqslant N}$.*

Remark 10. *The Laguerre diagram is a very important object in semi-discrete optimal transport due to its relation with computational geometry. Indeed, one can remark that in the case of the quadratic cost $c(x, y) = \|x - y\|^2$, one has the relation $\mathrm{Lag}_i(\psi) = \mathrm{Pow}_i(\psi) \cap X$ where $\mathrm{Pow}_i(\psi)$ denotes the usual Power cell of $y_i$ defined by*

$$\mathrm{Pow}_i(\psi) = \{x \in \mathbb{R}^d \mid \forall j,\ \|x - y_i\|^2 + \psi_i \leqslant \|x - y_j\|^2 + \psi_j\}.$$

*This implies that the Laguerre diagram is the restriction of a Power diagram to $X$. This idea will be studied in more depth in the following section.*

We now look at a necessary condition to ensure that transport maps are well-defined and that the Laguerre diagram forms a partition of $X$.

**Definition 1.5** (Negligibility condition)
*Given $X \subset \mathbb{R}^d$ and a point cloud $Y \subset \mathbb{R}^d$, a measure $\mu$ supported on $X$ is said to satisfy the* Negligibility *condition if*

$$\forall \psi \in \mathbb{R}^N,\ \forall i \neq j,\ \mu(L_{ij}(\psi)) = 0, \tag{Neg}$$

*where $L_{ij}(\psi)$ is the hyperplane separating $\mathrm{Lag}_i(\psi)$ and $\mathrm{Lag}_j(\psi)$ defined by, for $i \neq j$*

$$L_{ij}(\psi) = \{x \in X \mid c(x, y_i) + \psi_i = c(x, y_j) + \psi_j\}.$$

We can show that this condition is verified in some standard settings appearing in optimal transport theory:

1. When $c$ is the quadratic cost i.e. $c(x, y) = \|x - y\|^2$. An easy calculation shows that, for $i \neq j$, $L_{ij}(\psi) = X \cap \{x \in \mathbb{R}^d \mid 2\langle x \mid y_i - y_j\rangle = \psi_j - \psi_i\}$ meaning that $L_{ij}(\psi)$ is an hyperplane and thus its $d$-dimensional measure vanishes.
2. When $c$ satisfies the so-called *Twist* condition [Vil09] which can be stated as the following:

$$\forall x \in X, y \in Y \mapsto \mathrm{D}c(x, y) \text{ is injective.} \tag{Twist}$$

   The fact that the (Twist) condition is a sufficient condition for having the (Neg) condition is a consequence of the implicit function theorem, see [KMT16] for a detailed proof.

The link between the Kantorovich functional and solutions of semi-discrete optimal transport lies in the fact that, under some conditions on the geometry of the cost function namely the

(Neg) condition, the function $\Phi$ is differentiable as stated by the next theorem.

**Theorem 11** (Regularity of $\Phi$)**.** *If $\mu$ is an absolutely continuous measure that satisfies the* (Neg) *condition then we can rewrite $\Phi$ as*

$$\Phi(\psi) = \sum_{i=1}^{N} \int_{\mathrm{Lag}_i(\psi)} (c(x, y_i) + \psi_i) \mathrm{d}\mu(x) - \sum_{i=1}^{N} \nu_i \nu_i. \tag{1.3.1}$$

*It is also of class $\mathcal{C}^1$ and its gradient is given by*

$$\frac{\partial \Phi}{\partial \psi_i}(\psi) = G_i(\psi) - \nu_i \text{ where } G_i(\psi) = \int_{\mathrm{Lag}_i(\psi)} \mathrm{d}\mu(x).$$

*Proof.* Under the (Neg) condition, maps of the form $T_\psi : x \in X \mapsto \mathrm{argmin}_{1 \leqslant i \leqslant N}(c(x, y_i) + \psi_i)$ are well defined $\mu$-almost everywhere and we can deduce another expression for $\Phi$:

$$\Phi(\psi) = \sum_{i=1}^{N} \int_{\mathrm{Lag}_i(\psi)} (c(x, y_i) + \psi_i) \mathrm{d}\mu(x) - \sum_{i=1}^{N} \nu_i \nu_i.$$

Then, for any $\gamma \in \mathbb{R}^N$, let us remark that: $\min_{1 \leqslant i \leqslant N}(c(x, y_i) + \gamma_i) \leqslant c(x, T_\psi(x)) + \gamma_{T_\psi(x)}$. We also have:

$$\Phi(\psi) = \int_X (c(x, T_\psi(x)) + \psi_{T_\psi(x)}) \mathrm{d}\mu(x) - \sum_{i=1}^{N} \nu_i \psi_i.$$

Thus

$$\begin{aligned} \Phi(\psi) - \Phi(\gamma) &\leqslant \int_X (\psi_{T_\psi(x)} - \gamma_{T_\psi(x)}) \mathrm{d}\mu(x) - \sum_{i=1}^{N} (\psi_i - \gamma_i) \\ &\leqslant \sum_{i=1}^{N} \left( \int_{\mathrm{Lag}_i(\psi)} \mathrm{d}\mu(x) - \nu_i \right) (\psi_i - \gamma_i) = \langle G(\psi) - \nu \mid \psi - \gamma \rangle. \end{aligned}$$

We deduce that the supdifferential $\partial^+ \Phi(\psi)$ is not empty meaning that $\Phi$ is concave. Moreover, under the (Neg) condition, we can show that $G$ is continuous, see for instance Proposition 2.3 in [KMT16]. Moreover, the vector $G(\psi) - \nu$ depends continously on $\psi$, thus $\partial^+ \Phi(\psi)$ is reduced to one point and $\Phi$ is $\mathcal{C}^1$ smooth with its gradient given by $\nabla \Phi(\psi) = G(\psi) - \nu$. □

This result implies that finding $\psi \in \mathbb{R}^N$ such that $\nabla \Phi(\psi) = 0$ directly induces an optimal transport map $T_\psi$ between $\mu$ and $\nu$, see Figure 2.

If we now set $G = (G_1, \ldots, G_N)$, then the optimal transport problem between $\mu$ and $\nu = \sum_i \nu_i \delta_{y_i}$ amounts to the resolution of the following finite-dimensional non-linear system of equations:

$$\text{Find } \psi \in \mathbb{R}^N \text{ such that } \quad G(\psi_1, \ldots, \psi_N) = (\nu_1, \ldots, \nu_N). \tag{DMA}$$

Remark 12. *Equation* (DMA) *can be regarded as a discretization of the Monge-Ampère equation* (MA)*, hence the abbreviation.*

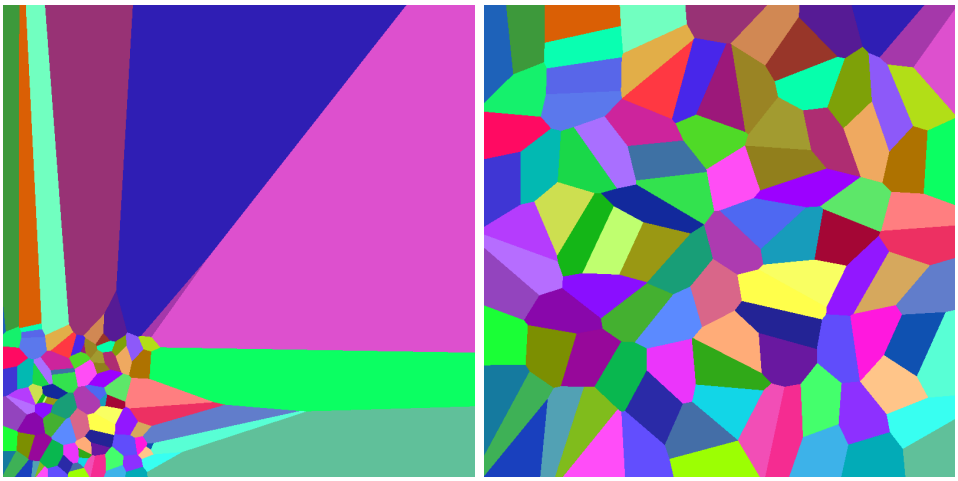

Figure 2 – **Illustration of Laguerre diagrams in semi-discrete optimal transport.** The target point cloud is randomly sampled inside $[0, \frac{1}{3}]^2$ while the source measure is uniform on the square $[0, 1]^2$. The target measure is uniform. Left: initial diagram; Right: final diagram where for all $i \in \{1, \dots, N\}$ , $\mu(\mathrm{Lag}_i(\psi)) = \frac{1}{N}$.

### 1.3.2 Numerical methods

We now look at the main methods to solve semi-discrete optimal transport which we showed is equivalent to solving Equation (DMA). We can separate the methods into two types: the ones based on variations of the so-called *Oliker-Prussner* algorithm and the ones based on Newton methods.

#### Oliker-Prussner based algorithms

The problem of optimal transport between a probability density on $\mathbb{R}^d$ and a finitely supported measure has been considered in many works, and can be traced back to Alexandrov and Pogorelov. One of the first methods was proposed by the authors of [OP89] who analysed a coordinatewise-decrement algorithm for a problem similar but not quite equivalent to optimal transport – namely, a Monge-Ampère equation with Dirichlet boundary conditions. In the following, we will call this method the *Oliker-Prussner* algorithm, it can be found in Algorithm 3. More precisely, this algorithm consists in an iterative procedure that updates a weight vector $\psi \in \mathbb{R}^N$ so as to satisfy at the end Equation (DMA). At each step, we look at the points $y_i$ such that $G_i(\psi) \leqslant \nu_i - \epsilon$ for a fixed $\epsilon > 0$. We then look at the minimal decrement $t \geqslant 0$ we can choose to have that $G(\psi - t\mathbf{1}_i) \geqslant \nu_i$ where $\mathbf{1}_i$ denotes the vector of $\mathbb{R}^N$ with a zero everywhere except at $i$ where there is a 1. We then update $\psi$ by decreasing $\psi_i$ by $t$ and start a new iteration. After convergence, one has $\|G(\psi) - \nu\|_\infty \leqslant N\epsilon$.

This coordinatewise-decrement approach was extended to an optimal transport setting in [CKO99b], leading to a $\mathrm{O}(N^3/\eta)$ algorithm where $N$ is the number of Dirac masses and $\eta$ is the desired numerical error. The complexity of the algorithm makes it difficult to use

**Algorithm 3:** Oliker-Prussner algorithm to solve semi-discrete optimal transport

**Input:** source measure $\mu$, target measure $\nu = \sum_{i=1}^{N} \nu_i \delta_{y_i}$, cost $c$, decrement $\epsilon > 0$
**Output:** weights $\psi$ that solves the optimal transport between $\mu$ and $\nu$ for the cost $c$

- Initialization: Define

$$\psi_i^0 = \begin{cases} 0 & \text{if } i = 0 \\ C := \max_{X\times Y} c(x,y) - \min_{X\times Y} c(x,y) & \text{for } i \neq 0 \end{cases}$$

  Then it is easy to see that $G_0(\psi^0) = 1$ and $G_i(\psi^0) = 0$ for $i \neq 0$
- While there exists $i \neq 0$ such that $G_i(\psi^k) \leqslant \nu_i - \epsilon$, define

$$t_i = \min\{t \geqslant 0 \mid G_i(\psi^k - t\mathbf{1}_i) \geqslant \nu_i\}$$

  and set $\psi^{k+1} = \psi^k - t_i \mathbf{1}_i$.

**return** $\psi^k$

it in practical situation where the number of Diracs masses can be of order $10^5 \sim 10^6$ for a reasonable numerical error $\eta \approx 10^{-8}$. We also note that, contrary to the auction algorithm detailed in the previous section, there exists no scaling method.

### Newton methods

Aurenhammer, Hoffmann and Aronov [AHA98] proposed a variational formulation for semi-discrete optimal transport, but do not analyse its algorithmic consequences further. This variational formulation was combined with *quasi*-Newton [Mér11; Lév15; CMT15] or Newton [Goe+12; Su+13] methods with good experimental results but without analyzing the convergence. The convergence of a *damped* Newton's algorithm was established first in [Mir15] for the Monge-Ampère equation with Dirichlet condition and was extended to optimal transport for cost functions satisfying the so-called *Ma-Trudinger-Wang* condition in [KMT16]. In the next paragraph, we introduce the damped Newton's method.

This method solves Equation (DMA) using Algorithm 4. In this algorithm, we denote by $A^+$ the *pseudo-inverse* of a matrix $A$. As usual for Newton's methods, the convergence will be a natural consequence of the $\mathcal{C}^1$ regularity of $G$ and of a strict monotonicity property for $\mathrm{D}G$ (see Theorem 13 in the next section). The strict monotonicity of $G$ only holds near points $\psi \in \mathbb{R}^N$ such that every Laguerre cell contains a positive fraction of the mass, i.e. $\psi \in \mathcal{K}^+$ where

$$\mathcal{K}^+ = \{\psi \in \mathbb{R}^N \mid \forall i \in \{1,\dots,N\},\ G_i(\psi) > 0\}. \tag{1.3.2}$$

The role of the damping step in Algorithm 4 (i.e. the choice of $\ell$ in the loop) is to ensure that $\psi^k$ always remain in $\mathcal{K}^+$. Also, since $G$ is invariant under the addition of a constant to all weights, we cannot expect *strict* monotonicity of $G$ in all directions. We denote by $\{\mathrm{cst}\}^\perp$ the orthogonal complement of the space of constant functions on $Y$ for the canonical scalar

product on $\mathbb{R}^N$, i.e.

$$\{\mathrm{cst}\}^\perp = \{v \in \mathbb{R}^N \mid \sum_{1 \leqslant i \leqslant N} v_i = 0\}.$$

**Algorithm 4:** Damped Newton's algorithm

**Input** A measure $\mu$, a finitely supported measure $\nu = \sum_{1 \leqslant i \leqslant N} \nu_i \delta_{y_i}$,
A numerical error $\eta > 0$,
A family of weights $\psi^0 \in \mathbb{R}^N$ such that $\varepsilon_0 := \min\left[\min_i G_i(\psi^0),\ \min_i \nu_i\right] > 0$

**While** $\left\|G(\psi^k) - \nu\right\| \geqslant \eta$

- Compute $v^k = -\mathrm{D}G(\psi^k)^+(G(\psi^k) - \nu)$
- Determine the minimum $\ell \in \mathbb{N}$ such that $\psi^{k,\ell} := \psi^k + 2^{-\ell} v^k$ satisfies

$$\begin{cases} \min\limits_{1 \leqslant i \leqslant N} G_i(\psi^{k,\ell}) \geqslant \varepsilon_0 \\ \left\|G(\psi^{k,\ell}) - \nu\right\| \leqslant (1 - 2^{-(\ell+1)}) \left\|G(\psi^k) - \nu\right\| \end{cases}$$

- Set $\psi^{k+1} = \psi^k + 2^{-\ell} v^k$ and $k \leftarrow k + 1$.

**Output** A family of weights $\psi^k$ solving (DMA) up to $\eta$, i.e. $\left\|G(\psi^k) - \nu\right\| \leqslant \eta$.

### 1.3.3 Applications

We finish this section by mentioning some applications of semi-discrete optimal transport:

- **Blue noise sampling**: the authors of [Goe+12] used semi-discrete optimal transport to develop an algorithm to generate blue-noise sampling of densities supported on 2D domains. At the end of this chapter, we extend this algorithm to also work on densities supported on triangulated surfaces, see Section 1.5.
- **Non-imaging optics**: we will see in Chapter 3 that numerous inverse problems appearing in optics can be recast as optimal transport problems. In these problems, one wants to design optical component design that optimally transfer a source illumination into a prescribed target radiation. Optimal transport can be used to formulate these problems in a common framework.
- **Euler equation for incompressible fluids**: the authors of [Goe+15] used optimal transport to develop a particle-based approach to solve PDEs appearing in computational fluid mechanics. In particular, optimal transport is an efficient way of imposing the incompressibility constraint.

Illustrations of these applications can be found in Figure 3.

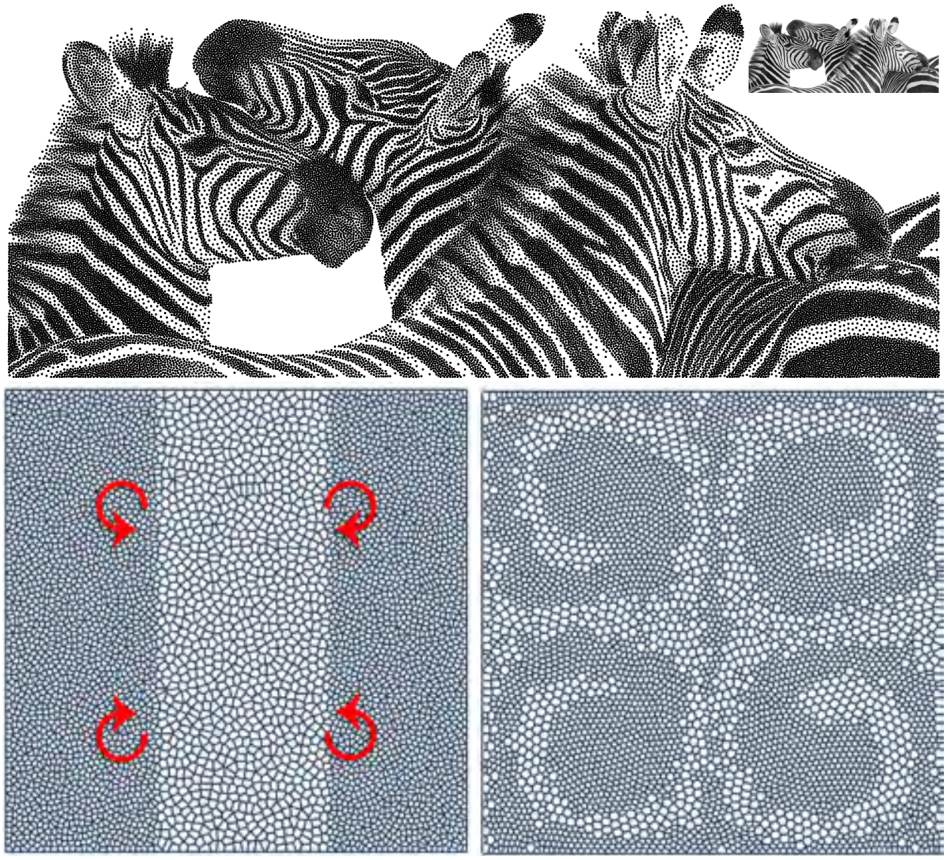

Figure 3 – **Examples of applications of semi-discrete optimal transport.** Top: Blue noise sampling of a greyscale image (image extracted from [Goe+12]); Bottom: Euler equation for incompressible fluids using a formulation based on Power diagrams (images extracted from [Goe+15]).

## 1.4 Optimal transport between a simplex soup and a point cloud

In this section, we look at a more singular setting where the source measure $\mu$ is not a probability density anymore, but is instead supported on what we call a *simplex soup*, i.e. a finite union of simplices in $\mathbb{R}^d$. We will allow the dimension of the simplices to range from 2 to $d$. We call such a measure a *simplicial measure*. The situation where one or more simplices in the collection have dimension strictly less than $d$ is difficult both in theory (as Brenier's theorem does not apply, and the optimal transport might not exist or not be unique) and in practice. In Section 1.4.4, we will show the convergence of Algorithm 4 to solve the optimal transport problem even in this degenerate setting. This section corresponds to the publication [MMT18a].

This result can be applied to different settings where the source and target measures are concentrated on lower-dimensional objects. We investigate at the end of this chapter applications

such as *optimal quantization* of a probability density over a surface, *remeshing* or *point set registration* on a mesh. Another interesting application is the optimal transport problem between measures concentrated on graphs of functions [MTW05], which are lower-dimensional subsets of $\mathbb{R}^d$. Such problems occur for instance in signal analysis and machine learning [Tho+16]. The cost involved in this setting is of the form $c(x, y) = ||x - y||^2 + |f(x) - g(y)|^2$. When the functions $f$ and $g$ are strictly convex and their gradients are less than one, the cost $c$ satisfies the Ma-Trudinger-Wang condition [MTW05] and we can apply the results of [KMT16]. When $f$ and $g$ do not satisfy these assumptions, our result shows that the damped Newton's algorithm still converges. Our work can be used in other applications involving optimal transport map between point sets and surfaces. The authors of [Dig+14] use optimal transport to reconstruct a simplicial complex from a point set: the simplicial complex is initially chosen as the Delaunay triangulation of the input point set and optimal transport is used to get an error metric to iteratively simplify the complex. Optimal transport has also been used in surface mapping. For instance, in [Man+17], the authors use optimal transport to find a low distortion map between two surfaces, without user interaction.

### 1.4.1 Formulation as a non-linear system of equations

From now on, we assume that the cost function is the quadratic cost $c(x, y) = \|x - y\|^2$ and that the source measure $\mu$ is a simplicial probability measure, as defined below.

**Definition 1.6** (Simplex soup)
*A* simplex soup *is a finite family $\Sigma$ of simplices of $\mathbb{R}^d$. The dimension of a simplex $\sigma$ is denoted $d_\sigma$. The support of the simplex soup $\Sigma$ is the set $K = \cup_{\sigma\in\Sigma}\sigma$.*

**Definition 1.7** (Simplicial measure)
*We call* simplicial measure *a measure $\mu = \sum_{\sigma\in\Sigma}\mu_\sigma$, where $\Sigma$ is a simplex soup, and where the measure $\mu_\sigma$ has density $\rho_\sigma$ with respect to the $d_\sigma$-dimensional Hausdorff measure on $\sigma$, i.e.*

$$\forall B \subseteq \mathbb{R}^d \text{ Borel, } \mu(B) = \sum_{\sigma\in\Sigma}\int_{B\cap\sigma}\rho_\sigma(x)d\mathcal{H}^{d_\sigma}(x).$$

Before summarizing the main properties of $G$, we will need the following additional definition.

**Definition 1.8** (Regular simplicial measure)
*A simplicial measure $\mu$ supported on $\bigcup_{\sigma\in\Sigma}\sigma$ is called* regular *if*

- *the dimension of every simplex $\sigma$ is $\geqslant 2$,*
- *for every $\sigma \in \Sigma$, $\rho_\sigma : \sigma \to \mathbb{R}$ is continuous and $\min_\sigma \rho_\sigma > 0$,*
- *it is not possible to disconnect the support $K = \bigcup_{\sigma\in\Sigma}\sigma$ by removing a finite number of points, i.e. $\forall S \subseteq K$ finite, $K \setminus S$ is connected.*

The main result on the properties of $G$ is the following.

**Theorem 13.** *Assume $\mu$ is a regular simplicial measure and that the points $y_1, \dots, y_n$ are in generic positions (according to Definition 1.9). Then,*

- *$G$ has class $\mathcal{C}^1$ on $\mathbb{R}^N$.*
- *$G$ is strictly monotone in the following sense*

$$\forall \psi \in \mathcal{K}^+, \forall v \in \{\mathrm{cst}\}^\perp \setminus \{0\}, \quad \langle \mathrm{D}G(\psi)v \mid v \rangle < 0.$$

The statement of this theorem is similar to Theorems 1.3 and 1.4 in [KMT16]. However, the results of [KMT16] were established under the assumption that the Laguerre cells induced by the cost function are convex in some "$c$-exponential chart", which is the discrete version of the so-called Ma-Trudinger-Wang property [MTW05; Loe09]. In the setting considered here, the Laguerre cells can be disconnected, so that we cannot expect them to be convex in any chart. Consequently, the strategy used in [KMT16] cannot be applied here, and we need to find an alternative way to establish the regularity of $G$. What we show here is that a mild genericity assumption on the points $y_1, \dots, y_N$ ensures that $G$ is $\mathcal{C}^1$ even when the source measure is singular, i.e. supported over a lower-dimensional subset of $\mathbb{R}^d$. The price to pay for this, however, is that we do not (and cannot expect to) get quantitative estimates on the speed of convergence of the algorithm as in [KMT16]. In particular, the existence of $\tau^*$ in the following theorem is obtained through a compactness argument.

**Theorem 14.** *Under the hypotheses of the previous theorem, the proposed Damped Newton's algorithm (see Algorithm 4) converges in a finite number of steps. Moreover, the iterates satisfy*

$$\left\| G(\psi^{k+1}) - \nu \right\| \leqslant \left(1 - \frac{\tau^\star}{2}\right) \left\| G(\psi^k) - \nu \right\|,$$

*where $\tau^* \in ]0, 1]$ depends on $\mu, \nu$ and $\epsilon_0$.*

As we will see in Section 1.5, the behaviour of Algorithm 4 seems better in practice meaning that the number of Newton's iterations is small even for large point sets.

We now show that the optimal transport problem we consider amounts to solving the system (DMA). The results mentioned here are very classical when the source measure is supported on a full dimensional subset of $\mathbb{R}^d$. Here, in order to handle lower-dimensional simplex soups, we need to introduce a notion of *genericity*. In the following, we denote by $[x_0, \dots, x_k]$ the convex hull of the points $x_0, \dots, x_k$.

**Definition 1.9** (Generic point set)
*A point set $\{y_1, \dots, y_N\} \subset \mathbb{R}^d$ is in* generic position *with respect to a $k$-dimensional simplex $\sigma = [x_0, \dots, x_k]$ if the following condition holds for every integer $p \in \{1, \dots, k\}$, every $\ell \in \{1, \dots, \min(d, N-1)\}$, every distinct $i_0, \dots, i_\ell \in \{1, \dots, N\}$ and every distinct $j_0, \dots, j_p \in \{0, \dots, k\}$:*

$$\dim(\{y_{i_1} - y_{i_0}, \dots, y_{i_\ell} - y_{i_0}\}^\perp \cap \mathrm{vect}(x_{j_1} - x_{j_0}, \dots, x_{j_p} - x_{j_0})) = \max(p - \ell, 0) \quad \text{(Generic)}$$

*The point set is in generic position with respect to a simplex soup $K = \cup_{\sigma \in \Sigma} \sigma$ if it is in generic*

*position with respect to all the simplices $\sigma \in \Sigma$. See Figure 4 for an illustration of this condition in $\mathbb{R}^3$.*

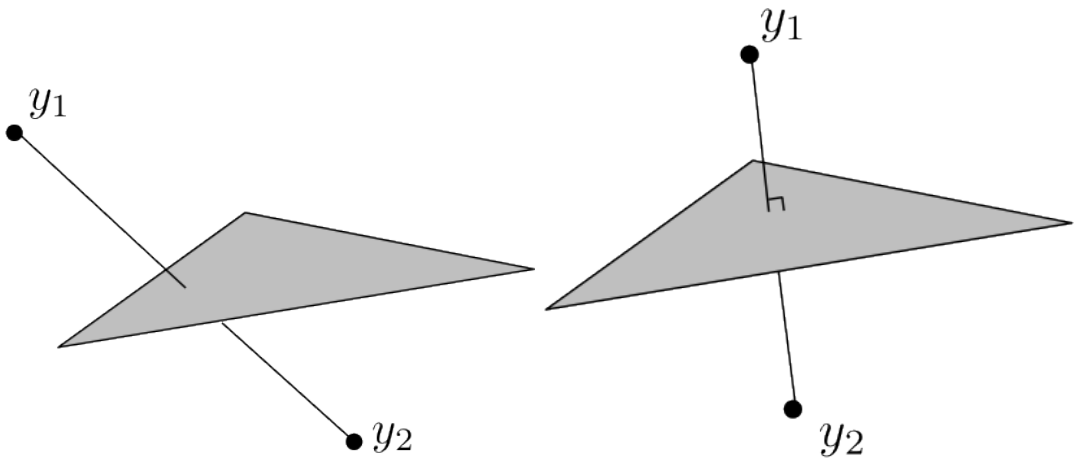


Figure 4 – **Illustration of the genericity condition for a triangle $\sigma$ in $\mathbb{R}^3$.** Left: $\{y_1, y_2\}$ is in generic position with respect to $\sigma$. Right: $\{y_1, y_2\}$ is not in generic position since the line $(y_1, y_2)$ is orthogonal to $\sigma$ .

**Definition 1.10** (Power diagram)
*The ith power cell induced by weights $\psi \in \mathbb{R}^N$ on a point set $\{y_1, \dots, y_N\}$ is defined by*

$$\mathrm{Pow}_i(\psi) := \{x \in \mathbb{R}^d \mid \forall j \in \{1, \dots, n\}, \|x - y_i\|^2 + \psi_i \leqslant \|x - y_j\|^2 + \psi_j\}.$$

Remark 15. *Note that the Laguerre cells for the quadratic cost are intersections of Power cells with the simplex soup, namely*

$$\mathrm{Lag}_i(\psi) = \mathrm{Pow}_i(\psi) \cap K. \tag{1.4.3}$$

*Condition* (Generic) *ensures in particular that for any choice of weights $(\psi_i)_{1 \leqslant i \leqslant N}$ the $(d-\ell)$-dimensional facets of the Power diagram induced by $(y_i)_{1 \leqslant i \leqslant N}, (\psi_i)_{1 \leqslant i \leqslant N}$ intersect the $p$-dimensional facets of $\sigma$ in a trivial way, when $(d-\ell) + p \leqslant d$.*

We also need the following technical lemma that states that, under genericity, the Laguerre cells form a partition of a simplex soup almost everywhere. This is a variation of the (Neg) condition for this setting.

Lemma 16. *Assume that $\mu$ is a simplicial measure and that $y_1, \dots, y_N$ is in generic position (Def 1.9). Let $\psi \in \mathbb{R}^N$ and define $\mathrm{Lag}_{i,j}(\psi) = \mathrm{Lag}_i(\psi) \cap \mathrm{Lag}_j(\psi)$ Then,*

$$\forall i \neq j, \quad \mu(\mathrm{Lag}_{i,j}(\psi)) = 0 \text{ and } \forall i, \quad \mu(\partial \mathrm{Lag}_i(\psi)) = 0.$$

*Proof.* Let $\sigma = [x_0, \dots, x_k]$ be a $k$-dimensional simplex in the support of $\mu$. Then, from the genericity assumption, one has $\dim(\mathrm{vect}(x_1 - x_0, \dots, x_k - x_0) \cap \{y_i - y_j\}^\perp) = k - 1$, so that

in particular $\dim(\sigma \cap \mathrm{Lag}_{i,j}(\psi)) \leqslant k-1$. This gives

$$\mu_\sigma(\mathrm{Lag}_{i,j}(\psi)) = \int_{\sigma \cap \mathrm{Lag}_{i,j}(\psi)} \rho_\sigma(x) \mathrm{d}\mathcal{H}^k(x)\mathrm{d}x = 0.$$

Summing these equalities over $\sigma \in \Sigma$, we get $\mu(\mathrm{Lag}_{i,j}(\psi)) = 0$. The second equality then follows from $\partial \mathrm{Lag}_i(\psi) \subseteq \bigcup_{j \neq i} \mathrm{Lag}_{i,j}(\psi)$. □

The relation between solutions of (DMA) and optimal transport maps is explained in the following proposition.

Proposition 17. *Let $\mu$ be a simplicial measure supported on $K$, and let $y_1, \dots, y_N$ be in generic position (Def 1.9). If $\psi \in \mathbb{R}^N$ satisfies* (DMA)*, then, the map*

$$T_\psi : x \in K \mapsto \operatorname*{argmin}_i \|x - y_i\|^2 + \psi_i.$$

*is well-defined $\mu$-a.e. and is an optimal transport map between $\mu$ and $\nu$.*

*Proof.* The fact that $T_\psi$ is well-defined almost everywhere follows from Lemma 16. Remark that $T_\psi$ is a transport map between $\mu$ and $T_{\#}\mu$ . Denote $\psi(y_i) := \psi_i$. Then, by definition of $T_\psi$, one has for any transport map $T$ $\|x - T_\psi(x)\|^2 + \psi(T_\psi(x)) \leqslant \|x - T(x)\|^2 + \psi(T(x))$. Integrating this inequality gives

$$\int_K (\|x - T_\psi(x)\|^2 + \psi(T_\psi(x)))\mathrm{d}\mu(x) \leqslant \int_K (\|x - T(x)\|^2 + \psi(T(x)))\mathrm{d}\mu(x).$$

Since $T$ and $T_\psi$ are both transport maps between $\mu$ and $\nu$, a change of variable gives

$$\int_K \psi(T_\psi(x))\mathrm{d}\mu(x) = \sum_{1 \leqslant i \leqslant N} \psi_i \nu_i = \int_K \psi(T(x))\mathrm{d}\mu(x).$$

Subtracting this equality from the inequality above directly gives the result. □

The next sections are dedicated to the proof of Theorem 13. It is split into three parts: first, we prove the $\mathcal{C}^1$ regularity of the function $G$; then its strict motonicity and finally we analyze the convergence of Algorithm 4.

### 1.4.2 $\mathcal{C}^1$ regularity

The main result of this section is the following theorem that states that under genericity conditions, the function $G : \mathbb{R}^N \to \mathbb{R}^N$ appearing in (DMA) is of class $\mathcal{C}^1$.

**Theorem 18.** *Let $\mu$ be a regular simplicial measure supported on a simplex soup $\Sigma$ (as in Definition 1.8) and let $Y = \{y_1, \dots, y_N\}$ be a generic point set. Then,*

- *the function $G$ appearing in* (DMA) *has class $\mathcal{C}^1$ on $\mathbb{R}^N$;*

- *denoting* $\mathrm{Lag}_{i,j}(\psi) := \mathrm{Lag}_i(\psi) \cap \mathrm{Lag}_j(\psi)$, *the derivatives of* $G$ *are given by*

$$\begin{cases} \frac{\partial G_i}{\partial \psi_j}(\psi) = \sum_{\sigma \in \Sigma} \frac{1}{2\left\|\Pi_{\sigma^0}(y_i - y_j)\right\|} \int_{\mathrm{Lag}_{i,j}(\psi) \cap \sigma} \rho_\sigma(x) \mathrm{d}\mathcal{H}^{d_\sigma - 1}(x) & \forall i \neq j \\ \frac{\partial G_i}{\partial \psi_i}(\psi) = -\sum_{j \neq i} \frac{\partial G_i}{\partial \psi_j}(\psi) & \forall i. \end{cases} \tag{1.4.4}$$

  *where* $\Pi_{\sigma^0} : \mathbb{R}^d \to \sigma^0$ *denotes the orthogonal projection on the linear subspace* $\sigma^0$ *tangent to* $\sigma$.

Remark 19. *Note that in contrast with Theorem 4.1 in [KMT16], the map* $G$ *is continuous on the whole space* $\mathbb{R}^N$ *and not only on the set* $\mathcal{K}^+$ *defined in* (1.3.2). *Without the genericity hypothesis, one cannot hope a global regularity result of this kind.*

- *Let* $\mu$ *be the uniform probability measure on* $K = [0,1]^2 \subseteq \mathbb{R}^2$ *(union of two triangles), and let* $y_1 = (\frac{1}{2}, 0)$, $y_2 = (-\frac{1}{2}, 0)$ *and* $y_3 = (1, 0)$. *Set* $\psi^t = (0, t, 0)$. *Then,*

$$\frac{\partial G_1}{\partial \psi_3}(\psi^t) = \mathcal{H}^1(K \cap \mathrm{Lag}_1(\psi^t) \cap \mathrm{Lag}_3(\psi^t)) = \begin{cases} 0 & \text{when } t > \frac{-6}{4} \\ 1 & \text{when } t < \frac{-6}{4}, \end{cases}$$

  *thus showing that* $G$ *is not globally* $\mathcal{C}^1$.
- *The regularity hypothesis would never be satisfied when one of the simplex is one-dimensional, thus explaining the first hypothesis in our definition of regular simplicial measure (Def. 1.8). Note also that this lack of genericity translates into a lack of regularity for* $G$. *Indeed, take* $\mu$ *the uniform measure over a segment* $[a, b]$. *Then, the partial derivative*

$$\frac{\partial G_i}{\partial \psi_j}(\psi) = \mathcal{H}^0(\mathrm{Lag}_i(\psi) \cap \mathrm{Lag}_j(\psi) \cap [a, b]) = \mathrm{Card}(\mathrm{Lag}_i(\psi) \cap \mathrm{Lag}_j(\psi) \cap [a, b])),$$

  *can only take values in* $\{0, 1\}$ *and must be discontinuous or constant.*

The end of this section if devoted to the proof of Theorem 18. We first remark that by linearity of the integrals in the definition of $G$ with respect to $\mu$, the theorem will hold for a simplicial measure if it holds for any measure with density supported on a simplex. We therefore let $\sigma$ be a $k$-dimensional simplex of $\mathbb{R}^d$ and $\mu = \mu_\sigma$ be a measure on $\sigma$ with continuous density $\rho_\sigma : \sigma \to \mathbb{R}$ with respect to the $k$-dimensional Hausdorff measure on $\sigma$. We also introduce

$$G_{\sigma,i}(\psi) := \int_{\mathrm{Lag}_i(\psi) \cap \sigma} \rho_\sigma(x) \mathrm{d}\mathcal{H}^k(x) \mathrm{d}x. \tag{1.4.5}$$

The following lemma will be used to compute the first derivatives of the function $G_{\sigma,i}$.

Lemma 20. *Let* $\rho : \mathbb{R}^k \to \mathbb{R}$ *be a continuous function on* $\mathbb{R}^k$ *and let* $z_1, \cdots, z_N \in \mathbb{R}^k$ *be vectors whose conic hull is* $\mathbb{R}^k$ *(i.e.* $\forall x \in \mathbb{R}^k, \exists \lambda_1, \ldots, \lambda_N \geqslant 0$ *s.t.* $x = \sum_i \lambda_i z_i$*). Given* $\lambda \in \mathbb{R}^k$, *define*

$$\hat{K}(\lambda) := \{x \in \mathbb{R}^k \mid \forall i \in \{1, \ldots, N\},\ \langle x \mid z_i \rangle \leqslant \lambda_i\}, \tag{1.4.6}$$

$$\hat{G}(\lambda) := \int_{\hat{K}(\lambda)} \rho(x) \mathrm{d}\mathcal{H}^k(x). \tag{1.4.7}$$

*Then,*

- *Assume that the $z_i$ are non-zero. Then, the function $\hat{G}$ is continuous.*
- *Assume that all the vectors $z_i$ are pairwise independent (i.e. not collinear, implying in particular that they are non-zero). Then $\hat{G}$ has class $\mathcal{C}^1$ and its partial derivatives are*

$$\frac{\partial \hat{G}}{\partial \lambda_i}(\lambda) = \frac{1}{\|z_i\|} \int_{\hat{K}(\lambda) \cap \{x \mid \langle x | z_i \rangle = \lambda_i\}} \rho(x) \mathrm{d}\mathcal{H}^{k-1}(x) \tag{1.4.8}$$

*Proof.* Let $e_1, \dots, e_N$ be the canonical basis of $\mathbb{R}^N$. We will proceed in two steps: we first show that under the first assumption $\hat{G}$ is continuous. Then, in a second step, we use this property on an other function to prove the $\mathcal{C}^1$ regularity of the function $\hat{G}$.

**Step 0.** Note that, because the conic hull of the $z_i$ equals $\mathbb{R}^k$, the polytope $\hat{K}(\lambda)$ is always compact. Moreover, one easily sees that if $\lambda \leqslant \lambda'$ (coordinate-wise), one has $\hat{K}(\lambda) \subseteq \hat{K}(\lambda')$. This implies that

$$\forall R \geqslant 0, \quad \exists C_R \subseteq \mathbb{R}^d \text{ compact s.t. } \forall \lambda' \in \mathbb{R}^N \ \max_i |\lambda'_i - \lambda_i| \leqslant R \Rightarrow \hat{K}(\lambda') \subseteq C_R. \tag{1.4.9}$$

We now sketch how to prove the continuity of the function $\hat{G}$ near any $\lambda \in \mathbb{R}^N$. Let $t \in [-R, R]$. We can assume that $t \geqslant 0$. First, note that the symmetric difference $\hat{K}(\lambda) \Delta \hat{K}(\lambda + te_i)$ is contained in a slab, or more precisely

$$\hat{K}(\lambda) \Delta \hat{K}(\lambda + te_i) \subseteq C_R \cap \{x \in \mathbb{R}^d \mid \langle x \mid z_i \rangle \in [\lambda_i, \lambda_i + t]\},$$

and that the width of the slab is $t / \|z_i\|$. This gives

$$\left| \hat{G}(\lambda) - \hat{G}(\lambda + te_i) \right| \leqslant \int_{\hat{K}(\lambda) \Delta \hat{K}(\lambda + te_i)} \rho(x) \mathrm{d}\mathcal{H}^{k-1}(x) \leqslant \left[ \frac{\mathrm{diam}(C_R)^{d-1} \max_{C_R} |\rho|}{\|z_i\|} \right] t$$

A similar bound obviously exists for $t \leqslant 0$. Using this estimate on each coordinate axis, one obtains the continuity of $\hat{G}$ (and in fact, this proof even shows that $\hat{G}$ is locally Lipschitz). This proves the first statement.

**Step 1.** We now prove the second statement, and assume that $\rho$ is continuous and the $z_i$ are pairwise independent. Fix some index $i_0 \in \{1, \dots, N\}$ and take $\lambda \in \mathbb{R}^N$. We consider the convex set $L := \{x \in \mathbb{R}^k \mid \forall i \neq i_0 \ \langle x \mid z_i \rangle \leqslant \lambda_i\}$. For any $t \geqslant 0$, using the function $u : x \in \mathbb{R}^k \mapsto \langle x \mid z_{i_0} \rangle - \lambda_{i_0}$, one has $\hat{K}(\lambda + te_{i_0}) \setminus \hat{K}(\lambda) = L \cap u^{-1}([0, t])$. Applying the co-area formula with the function $u$ whose gradient is $\nabla u = z_{i_0}$, we can evaluate the slope

$$\begin{aligned}
\frac{1}{t}(\hat{G}(\lambda + te_{i_0}) - \hat{G}(\lambda)) &= \frac{1}{t} \int_{L \cap u^{-1}([0,t])} \rho(x) d\mathcal{H}^k(x) \\
&= \frac{1}{t} \int_0^t \int_{L \cap u^{-1}(s)} \frac{\rho(x)}{\|z_{i_0}\|} \mathrm{d}\mathcal{H}^{k-1}(x) \mathrm{d}s \\
&= \frac{1}{t} \int_0^t g_{i_0}(\lambda + se_{i_0}) \mathrm{d}s
\end{aligned} \tag{1.4.10}$$

where we have set

$$g_{i_0}(\overline{\lambda}) := \int_{\hat{K}_{i_0}(\overline{\lambda})} \frac{\rho(x)}{\|z_{i_0}\|} \mathrm{d}\mathcal{H}^{k-1}(x) \quad \text{with } \hat{K}_{i_0}(\overline{\lambda}) = \{x \in \hat{K}(\overline{\lambda}) \mid \langle x \mid z_{i_0}\rangle = \overline{\lambda}_{i_0}\}$$

Note that by construction, $\hat{K}_{i_0}(\overline{\lambda})$ is the facet of $\hat{K}(\overline{\lambda})$ with exterior normal $z_{i_0}/\|z_{i_0}\|$. Assume for now that we are able to prove that the functions $g_{i_0}$ are continuous. Then, by the fundamental theorem of calculus and by Equation (1.4.10) one has $\frac{\partial \hat{G}}{\partial \lambda_{i_0}}(\lambda) = g_{i_0}(\lambda)$. Since we have assumed that $g_{i_0}$ is continuous, this shows that the function $\hat{G}$ has continuous partial derivatives and is therefore $\mathcal{C}^1$, and gives the desired expression for its partial derivatives.

**Step 2.** Our goal is now to establish the continuity of the function $g_{i_0}$. In order to do that, we will parameterize the facet $\hat{K}_{i_0}(\lambda)$ using the hyperplane $V = \{z_{i_0}\}^\perp$ and $\Pi_V$ the orthogonal projection on this hyperplane. Then, since $x \in \hat{K}_{i_0}(\lambda)$ satisfies $\langle x \mid z_{i_0}\rangle = \lambda_{i_0}$ one has $x = \Pi_V(x) + \lambda_{i_0}\frac{z_{i_0}}{\|z_{i_0}\|^2}$ and thus

$$g_{i_0}(\lambda) = \frac{1}{\|z_{i_0}\|} \int_{\Pi_V(\hat{K}_{i_0}(\lambda))} \rho\left(y + \lambda_{i_0}\frac{z_{i_0}}{\|z_{i_0}\|^2}\right) \mathrm{d}\mathcal{H}^{k-1}(y)$$

By compactness, $\rho$ is uniformly continuous on $C_R$, where $C_R$ is defined in Equation (1.4.9): there exists a function $\omega_R : \mathbb{R}^+ \to \mathbb{R}^+$ satisfying $\lim_{r\to 0}\omega_R(r) = 0$ and such that for all $x, y \in C_R$, $|\rho(x) - \rho(y)| \leqslant \omega_R(\|x - y\|)$. Using the function $\rho_\lambda(y) := \rho(y + \lambda_{i_0} z_{i_0}/\|z_{i_0}\|^2)$ and the notation $\tilde{\hat{K}}_{i_0}(\lambda) = \Pi_V(\hat{K}_{i_0}(\lambda))$, one has for every $\lambda'$

$$\begin{aligned}
&\|z_{i_0}\| \left|g_{i_0}(\lambda) - g_{i_0}(\lambda')\right| \\
&= \left| \int_{\tilde{\hat{K}}_{i_0}(\lambda)} \rho_\lambda(y) \mathrm{d}\mathcal{H}^{k-1}(y) - \int_{\tilde{\hat{K}}_{i_0}(\lambda')} \rho_{\lambda'}(y) \mathrm{d}\mathcal{H}^{k-1}(y) \right| \\
&\leqslant \left| \int_{\tilde{\hat{K}}_{i_0}(\lambda)} (\rho_\lambda(y) - \rho_{\lambda'}(y)) \mathrm{d}\mathcal{H}^{k-1}(y) \right| \\
&\quad + \left| \int_{\tilde{\hat{K}}_{i_0}(\lambda)} \rho_{\lambda'}(y) \mathrm{d}\mathcal{H}^{k-1}(y) - \int_{\tilde{\hat{K}}_{i_0}(\lambda')} \rho_{\lambda'}(y) \mathrm{d}\mathcal{H}^{k-1}(y) \right|
\end{aligned} \tag{1.4.11}$$

Suppose now that $\max_i |\lambda_i - \lambda'_i| \leqslant R$. Then the first term of the right hand side term is bounded by $\mathcal{H}^{k-1}(\Pi_V(C_R))\omega_R(|\lambda_{i_0} - \lambda'_{i_0}|/\|z_{i_0}\|)$ which tends to zero when $\lambda'$ tends to $\lambda$. For the second term, we note that

$$\begin{aligned}
\tilde{\hat{K}}_{i_0}(\lambda) &= \{y \in V \mid \forall i \neq i_0,\ \langle y + \lambda_{i_0} z_{i_0}/\|z_{i_0}\|^2 \mid z_i\rangle \leqslant \lambda_i\} \\
&= \{y \in V \mid \forall i \neq i_0,\ \langle y \mid \tilde{z}_i\rangle \leqslant \lambda_i - \lambda_{i_0}\langle z_i \mid z_{i_0}\rangle/\|z_{i_0}\|^2\},
\end{aligned}$$

where we have set $\tilde{z}_i = \Pi_V(z_i) = \Pi_{\{z_{i_0}\}^\perp}(z_i)$. The assumption that $z_i$ and $z_{i_0}$ are independent implies that the vectors $\tilde{z}_i$ are non-zero. We conclude using the first part of the Lemma that

the function

$$\lambda \mapsto \int_{\tilde{K}_{i_0}(\lambda)} \rho_{\lambda'}(y) \mathrm{d}\mathcal{H}^{k-1}(y)$$

is continuous. Using the inequality (1.4.11), we see that $\lim_{\lambda' \to \lambda} g_{i_0}(\lambda') = \lambda$. This shows that $g_{i_0}$ is continuous and concludes the proof of the lemma. □

Note that when applying this lemma to the $\mathcal{C}^1$ regularity of $G_{\sigma,i}$, the vectors $z_i$ will be used both to define the Laguerre cell and the simplex $\sigma$. We will also use the following easy consequence of the genericity hypothesis.

LEMMA 21. *Assume $\{y_1, \dots, y_N\} \subset \mathbb{R}^d$ is in generic position with respect to a $k$-dimensional simplex $\sigma = [x_0, \dots, x_k]$ and let $H = \mathrm{vect}(x_1 - x_0, \dots, x_k - x_0)$. Then,*

- *For every pairwise distinct $i, j, l \in \{1, \dots, n\}$, the vectors $z_1 = \pi_H(y_j - y_i)$ and $z_2 = \pi_H(y_l - y_i)$, where $\pi_H$ is the orthogonal projection on $H$, are not collinear.*
- *For every distinct $i, j \in \{1, \dots, n\}$, the vector $\pi_H(y_j - y_i)$ is not perpendicular to any of the $(k-1)$-dimensional facets of $\sigma$.*

*Proof.* By the genericity condition of Definition 1.9, $\{y_j - y_i\}^\perp \cap H$ is of dimension $k-1$. Furthermore, for a vector $u \in \{y_j - y_i\}^\perp \cap H$, one has $\langle u \mid y_j - y_i \rangle = 0$ and $\langle u \mid z_1 \rangle = 0$ which implies that $\{y_j - y_i\}^\perp \cap H = \{z_1\}^\perp \cap H$. Similarly, one has $\{y_l - y_i\}^\perp \cap H = \{z_2\}^\perp \cap H$. If $z_1$ and $z_2$ are collinear, then $\{y_j - y_i, y_l - y_i\}^\perp \cap H = (\{y_j - y_i\}^\perp \cap H) \cap (\{y_l - y_i\}^\perp \cap H)$ is of dimension $k-1$ which contradicts the genericity condition. The proof of the second item is straightforward. □

We conclude this section by proving the $\mathcal{C}^1$ regularity of $G_{\sigma,i}$. To do that, we will use Lemma 20 with a set of vectors $(z_i)$ that describe the boundary of the Power cell $\mathrm{Pow}_i$ and the simplex $\sigma$.

*Proof of Theorem 18.* Our goal is to show that $G_{\sigma,i}$ (defined in (1.4.5)) is $\mathcal{C}^1$–regular and to compute its partial derivatives. From now on, we fix some index $i_0 \in \{1, \dots, N\}$. Reordering indices if necessary, we assume that $i_0 = N$. We want to apply Lemma 20, and for that purpose we are first going to rewrite $\mathrm{Lag}_i(\psi) \cap \sigma$ under the form (1.4.6). Denote $H$ the $k$-dimensional affine space spanned by $\sigma$; translating everything if necessary, we can assume that $H$ is a linear subspace of $\mathbb{R}^d$. A simple calculation shows that the intersection of the $N$th power cell with $H$ is given by

$$\mathrm{Pow}_N(\psi) \cap H = \{x \in H \mid \forall i \in \{1, \dots, N-1\}, \langle x \mid z_i \rangle \leqslant \lambda_i\},$$

where $\lambda_i = \frac{1}{2}(\|y_i\|^2 + \psi_i - (\|y_N\|^2 + \psi_N))$ and $z_i$ is the orthogonal projection of $y_i - y_N$ on $H$. Since $\sigma$ is a $k$-dimensional simplex, it can be written as the intersection of $k+1$ half-spaces of $H$, i.e. $\sigma = \{x \in H \mid \forall j \in \{N, \dots, N+k\}, \langle x \mid z_j \rangle \leqslant 1\}$ for some non-zero vectors $z_i$ of $H$. Combining these two expressions, one gets

$$\mathrm{Lag}_N(\psi) \cap \sigma = \{x \in H \mid \forall i \in \{1, \dots, N+k\},\ \langle x \mid z_i \rangle \leqslant \lambda_i\}.$$

where $\lambda_i = 1$ for $i \in \{N, \dots, N+k\}$.

We will now show that the assumptions of Lemma 20 are satisfied. Since $\sigma$ is a nondegenerate simplex, $z_i \neq 0$ for every $i \geqslant N$ and the vectors $z_i, z_j$ for $i \neq j$ and $i, j \geqslant N$ are pairwise independent. From the first genericity property of Lemma 21, we know that $z_i = \Pi_H(y_i - y_N)$ and $z_j = \Pi_H(y_j - y_N)$ are independent ($i \neq j$ and $i, j < N$). From the second genericity condition, we also know that $z_i, z_j$ are independent when $i \neq j$ and $i < N$ and $j \geqslant N$. In order to apply Lemma 20 we need to extend the continuous density $\rho_\sigma : \sigma \subseteq H \to \mathbb{R}$ into a continuous density $\rho : H \to \mathbb{R}$. Since $\sigma$ is convex, this can be easily done using the projection map $\Pi_\sigma : H \to \sigma$, and by setting $\rho(x) = \rho_\sigma(\Pi_\sigma(x))$. Then, $\rho$ is continuous as the composition of two continuous maps (recall that since $\sigma$ is convex, the projection $\Pi_\sigma$ is 1-Lipschitz). With these constructions one has

$$G_{\sigma,N}(\psi) = \hat{G}(A(\psi)),$$

where $A : \mathbb{R}^N \to \mathbb{R}^{N+k}$ is the affine map

$$A(\psi) := \left(\frac{1}{2}(\|y_1\|^2 + \psi_1 - (\|y_N\|^2 + \psi_N)), \dots, \frac{1}{2}(\|y_{N-1}\|^2 + \psi_{N-1} - (\|y_N\|^2 + \psi_N), 1, \dots, 1\right)$$

with $k+1$ trailing ones. By Lemma 20, $\hat{G}$ has class $\mathcal{C}^1$, and the expression above shows that $G_{\sigma,N}$ is also $\mathcal{C}^1$. Moreover, denoting $A = (A_1, \dots, A_{N+k})$, one gets

$$\begin{aligned}
\forall i \neq N, \quad \frac{\partial G_{\sigma,N}}{\partial \psi_i}(\psi) &= \sum_{1 \leqslant j \leqslant N+k} \frac{\partial A_j}{\partial \psi_i}(\psi) \frac{\partial \hat{G}}{\partial \lambda_j}(A(\psi)) \\
&= \frac{1}{2} \frac{\partial \hat{G}}{\partial \lambda_i}(A(\psi)) \\
&= \frac{1}{2\|z_i\|} \int_{\hat{K}(A(\psi)) \cap \{x \in H \mid \langle x | z_i \rangle = \lambda_i\}} \rho_\sigma(x) \mathrm{d}\mathcal{H}^{k-1}(x) \\
&= \frac{1}{2\|z_i\|} \int_{\mathrm{Lag}_{i,N}(\psi) \cap \sigma} \rho_\sigma(x) \mathrm{d}\mathcal{H}^{k-1}(x),
\end{aligned}$$

thus establishing the first formula in (1.4.4). The second formula in this equation deals with the case $i = N$, and follows from a similar computation and from the expression

$$\left(\frac{\partial A_j}{\partial \psi_N}(\psi)\right)_{1 \leqslant j \leqslant N+k} = \left(-\frac{1}{2}, \dots, -\frac{1}{2}, 0, \dots, 0\right),$$

with $k+1$ trailing zeros. We have therefore established the theorem when $\mu = \mu_\sigma$. The case where $\mu = \sum_{\sigma \in \Sigma} \mu_\sigma$ is a simplicial measure follows by linearity. □

### 1.4.3 Strict monotonicity

As mentioned in Section 1.4.1, the second ingredient needed for the proof of the convergence of the damped Newton's algorithm is a motonicity property of $G$, or equivalently the strong concavity of the function $\Phi$ defined in Equation (1.3.1). This property relies heavily on the

"strong connectedness" of the support of $\mu$ assumed in the third item of Definition 1.8. We recall that we denote by $\{\mathrm{cst}\}^\perp = \{v \in \mathbb{R}^Y \mid \sum_{1 \leqslant i \leqslant N} v_i = 0\}$ the orthogonal of the constant functions on $Y$.

**Theorem 22.** *Let $\mu$ be a regular simplicial measure and assume that $y_1, \dots, y_N$ is generic with respect to the support of $\mu$ (Definition. 1.9). Then $G$ is strictly monotone in the sense that*

$$\forall \psi \in K^+,\ \forall v \in \{\mathrm{cst}\}^\perp \setminus \{0\},\ \langle \mathrm{D}G(\psi)v \mid v \rangle < 0.$$

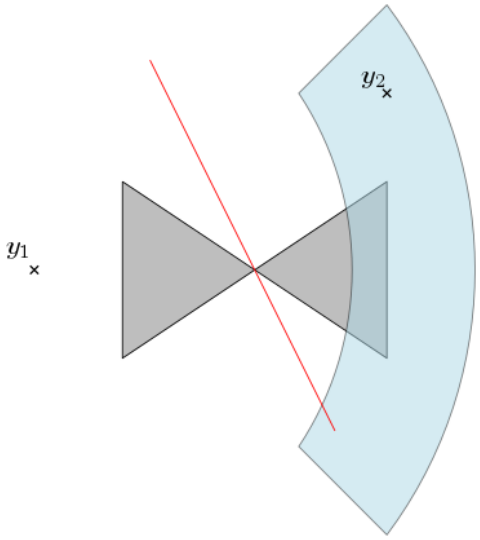


Figure 5 – **Example illustrating the connectivity property in the definition of a regular simplicial measure.** Here, $K$ is a simplex soup made of the two triangles in gray and the set of points $y_1, y_2$ such that $\mu(\mathrm{Lag}_{1,2}(\psi)) = 0$ has not a zero measure.

Remark 23. *Let us illustrate the fact that the connectedness of $K$ is not sufficient (i.e. why we require that it is impossible to disconnect the support $K$ of $\mu$ by removing a finite number of points). Consider the case where $K$ is made of the two 2-dimensional simplices embedded in $\mathbb{R}^2$, and displayed in grey in Figure 5. We assume that $\mu$ is the restriction of the Lebesgue measure to $K$ and that $Y = \{y_1, y_2\}$. Then, the Jacobian matrix of $G$ at $\psi$ is the 2-by-2 matrix given by*

$$\mathrm{D}G(\psi) = \begin{pmatrix} a & -a \\ -a & a \end{pmatrix} \quad \text{where } a = \frac{1}{2\,\|y_1 - y_2\|}\mathcal{H}^1(\mathrm{Lag}_{1,2}(\psi) \cap K).$$

*If we fix $y_1 \in \mathbb{R}^2$, it is easy to see that for any $y_2$ in the blue domain, there exists weights $\psi_1$ and $\psi_2$ such that the interface $\mathrm{Lag}_{1,2}(\psi)$ (in red) passes through the common vertex between the two simplices, thus implying that $a = 0$, hence $\mathrm{D}G(\psi) = 0$. In such setting, $G$ is not strictly monotone, the conclusion of Theorem 22 does not hold.*

The end of this section is devoted to the proof of Theorem 22.

### Preliminary lemmas

With a slight abuse, we call tangent space to a convex set $K$ the linear space $\mathrm{vect}(K-x)$ for some $x$ in $K$ (this space is independent of the choice of $x$). We denote $\mathrm{relint}(K)$ the *relative interior* of a convex set $K \subseteq \mathbb{R}^d$ and we call dimension of $K$ the dimension of the affine space spanned by $K$.

LEMMA 24. *Let $e, f$ be convex sets and $E$ and $F$ their tangent spaces. Assume that* $\mathrm{relint}(f) \cap \mathrm{relint}(e) \neq \emptyset$. *Then,*

$$\dim(e \cap f) = \dim(E \cap F).$$

*Proof.* Let $G$ be the tangent space to $e \cap f$, so that $\dim(e \cap f) = \dim(G)$. It suffices to show that $G = E \cap F$ to prove that $\dim(e \cap f) = \dim(E \cap F)$. The inclusion $G \subseteq E \cap F$ holds without hypothesis (a tangent vector to $e \cap f$ is always both a tangent vector to $e$ and to $f$). For the reciprocal inclusion, consider $x \in \mathrm{relint}(e) \cap \mathrm{relint}(f)$ and $v \in E \cap F$. Then, by definition of the relative interior, for $t$ small enough one has $x + tv \in e$ and $x + tv \in f$, i.e. $x + tv \in e \cap f$, so that $tv$ belongs to $G$. This shows $G \subseteq E \cap F$ and concludes the proof. □

LEMMA 25. *Let $f \subseteq f'$ and $e$ be three convex sets of $\mathbb{R}^d$, and $F \subseteq F'$ and $E$ be their tangent spaces. Assume that*

- $\mathrm{relint}(f) \cap \mathrm{relint}(e) \neq \emptyset$ ;
- $\dim(F') = \dim(F) + 1$ *and* $\dim(E \cap F') = \dim(E \cap F) + 1$.

*Then* $\dim(e \cap f') = \dim(e \cap f) + 1$.

*Proof.* Let us first show that $\mathrm{relint}(e) \cap \mathrm{relint}(f') \neq \emptyset$ .We consider a basis $e_1, \ldots, e_n$ of $F$ and a vector $e_{n+1} \in E \cap F'$ such that $E \cap F' = (E \cap F) \oplus \mathbb{R}e_{n+1}$ and $F' = F \oplus \mathbb{R}e_{n+1}$. Let $x_0$ be a point in the intersection $\mathrm{relint}(f) \cap \mathrm{relint}(e)$, which we assumed non-empty. There exists $\varepsilon > 0$ such that $\Delta := \mathrm{conv}(\{x_0 \pm \epsilon e_i \mid 1 \leqslant i \leqslant n\}) \subseteq f$. Using the assumption that $F'$ is the tangent space to $f'$, we know that there exists a point $y \in f'$ such that $v = y - x_0 \in F' \setminus F$. Consider the convex sets $\Delta_\pm$ spanned by $\Delta$ and one of the points $x_0 \pm v$, $\Delta_\pm = \mathrm{conv}(\Delta \cup \{x_0 \pm v\})$ . The convex set $\Delta_+ \cup \Delta_-$ is a neighborhood of $x_0$, meaning that there exists $t \neq 0$ such that $x_\pm := x_0 \pm te_{n+1} \in \mathrm{relint}(\Delta_\pm)$. Assume for instance $x_+ \in \mathrm{relint}(\Delta_+) \subseteq f'$. Since $\Delta_+$ has the same dimension as $f'$, one has $x_+ \in \mathrm{relint}(\Delta_+) \subseteq \mathrm{relint}(f')$ and by a standard property of the relative interior one has $(x_0, x_+] = (x_0, x_0 + te_{n+1}] \subseteq \mathrm{relint}(f')$. Finally, since $x_0$ belongs to the relative interior of $e$ and $e_{n+1} \in E$, the segment $(x_0, x_0 + te_{n+1}]$ must intersect the relative interior of $e$, proving that $\mathrm{relint}(e) \cap \mathrm{relint}(f') \neq \emptyset$.

Then using Lemma 24, we have $\dim(e \cap f) = \dim(E \cap F)$ and $\dim(e \cap f') = \dim(E \cap F') = \dim(e \cap f) + 1$. □

**Proof of the strict motonicity**

This theorem will follow using standard arguments, once one has established the connectedness of the graph induced by the Jacobian matrix. Let $\psi \in K^+$, $H := \mathrm{D}G(\psi)$ and consider the graph $\mathcal{G}$ supported on the set of vertices $V = \{1, \dots, N\}$ and with edges

$$E(\mathcal{G}) := \{(i,j) \in V^2 \mid i \neq j \text{ and } H_{i,j}(\psi) > 0\}.$$

Remark that thanks to the formula established in Theorem 18, the matrix $H$ is symmetric. This fact could also be deduced from the variational formulation: $H$ is the Hessian of the functional $\Phi$ defined in Equation (1.3.1). This implies in particular that $\mathcal{G}$ is not oriented.

LEMMA 26. *If* $\mathrm{Lag}_{i,j}(\psi)$ *intersects some* $k$*-dimensional simplex* $\sigma \in \Sigma$*, then the intersection is either a singleton or has dimension* $k-1$*.*

*Proof.* Denote $\sigma = [x_0, \dots, x_k]$ and assume that $m = \dim(\mathrm{Lag}_{i,j}(\psi) \cap \sigma) \geqslant 1$. Consider a $p$-dimensional facet $f = [x_{j_0}, \dots, x_{j_p}]$ of $\sigma$ and a facet $\mathrm{Lag}_{i_0,\dots,i_\ell}(\psi) = \bigcap_{k=0}^{\ell} \mathrm{Lag}_{i_k}(\psi)$ of $\mathrm{Lag}_{i,j}(\psi)$ (we take $i_0 = i$ and $i_1 = j$) such that $\dim(\mathrm{Lag}_{i_0,\dots,i_\ell}(\psi) \cap f) = m$ and assume that both facets are minimal for the inclusion. It is easy to see that this minimality property implies that the relative interiors of $f$ and $\mathrm{Lag}_{i_0,\dots,i_\ell}(\psi)$ must intersect each other. With Lemma 24, this ensures that

$$\begin{aligned} m &= \dim(\mathrm{Lag}_{i_0,\dots,i_\ell}(\psi) \cap f) && (1.4.12)\\ &= \dim(\{y_{i_1} - y_{i_0}, \dots, y_{i_\ell} - y_{i_0}\}^\perp \cap \mathrm{vect}(x_{j_1} - x_{j_0}, \dots, x_{j_p} - x_{j_0})) = p - \ell, && (1.4.13) \end{aligned}$$

where we used the genericity property (Def 1.9) to get the last equality. We now prove that $p = k$ and $\ell = 1$ by contraction. If we assume that $p < k$, there exists $j_{p+1} \in \{1, \dots, k\}$ distinct from $\{j_0, \dots, j_p\}$. Set $e = \mathrm{Lag}_{i_0\dots,i_\ell}(\psi)$, $f = [x_{j_0}, \dots, x_{j_p}]$ and $f' = [x_{j_0}, \dots, x_{j_{p+1}}]$. The genericity hypothesis allows us to apply Lemma 25. The conclusion of the lemma is that $\dim(\mathrm{Lag}_{i_0,\dots,i_\ell}(\psi) \cap f') = p + 1 - \ell > m$, which violates the definition of $m$. By contradiction one must have $p = k$. With the same arguments (removing a point $y_{i_n}$ for some $n \in \{0, \dots, \ell\}$ different from $y_i$ and $y_j$ from the list if $i_\ell \geqslant 1$) we can see that necessarily $\ell = 1$. With (1.4.12) we get $m = k - 1$, thus concluding the proof of the lemma. □

LEMMA 27. *The graph* $\mathcal{G}$ *is connected.*

*Proof.* Consider the finite set

$$S := \{x \in \mathbb{R}^d \mid \exists \sigma \in \Sigma, \exists i \neq j \in \{1, \dots, N\},\ \mathrm{Lag}_{i,j}(\psi) \cap \sigma = \{x\}\}.$$

For any simplex $\sigma \in \Sigma$, denote $\sigma^* = \sigma \setminus S$, and let $K^* = K \setminus S$. By definition of a regular simplicial measure (Def. 1.8), we know that $K^*$ is connected. Let $C = \{i_1, \dots, i_c\}$ be a connected component of the graph $\mathcal{G}$, and define $L = \bigcup_{i \in C} \mathrm{Lag}_i(\psi)$ and $L' = \bigcup_{i \notin C} \mathrm{Lag}_i(\psi)$.

**Step 1** We first show that for any simplex $\sigma \in \Sigma$, one must have either $\sigma^* \subset \mathrm{int}(L)$ *or* $\sigma^* \subset \mathrm{int}(\mathbb{R}^d \setminus L)$. For this, it suffices to prove that for any $\sigma \in \Sigma$, $\sigma^* \cap \partial L = \emptyset$. We argue by

contradiction, assuming the existence of a point $x \in \partial L \cap \sigma^*$. Then, by definition of $\partial L$, there exists $i \in C$ and $j \notin C$ such that $x \in \mathrm{Lag}_{i,j}(\psi)$. Since $x \in \sigma^*$, we know that $x$ does not belong to $S$. This implies that $\mathrm{Lag}_{i,j}(\psi) \cap \sigma$ cannot be a singleton. By the previous Lemma, this gives $\dim(\sigma \cap \mathrm{Lag}_{i,j}(\psi)) = d_\sigma - 1$ so that

$$H_{i,j}(\psi) = \mathrm{const}(y_i, y_j) \int_{\sigma \cap \mathrm{Lag}_{i,j}(\psi)} \rho_\sigma(x) \mathrm{d}\mathcal{H}^{d_\sigma - 1}(x) > 0.$$

This shows that $i$ and $j$ are in fact adjacent in the graph $\mathcal{G}$ and contradicts $j \notin C$.

**Step 2** We now prove that $C$ is equal to $\{1, \dots, N\}$ by contradiction. We group the simplices $\sigma \in \Sigma$ according to whether $\sigma^*$ belongs to $\mathrm{int}(L)$ or to $\mathrm{int}(\mathbb{R}^d \setminus L)$. The sets $K_i^*$ are open for the topology induced on $K^*$ because $K_1^* = \mathrm{int}(L) \cap K^*$ and $K_2^* = \mathrm{int}(L') \cap K^*$. Since they are also non empty, this violates the connectedness of $K^*$. We can conclude that $C = \{1, \dots, N\}$, i.e. $\mathcal{G}$ is connected. □

*Proof of Theorem 22.* First note that the matrix $H$ is symmetric and therefore diagonalizable in an orthonormal basis. Gershgorin's circle theorem immediately implies that the eigenvalues of the matrix are negative. The theorem will be established if we are able to show that the nullspace of $H$ (i.e. the eigenspace corresponding to the eigenvalue zero) is the 1-dimensional space generated by constant functions. The computations presented here are similar to the ones in [CGS10, Lemma 3.3]. Consider $v$ in the nullspace and let $i_0$ be an index where $v$ attains its maximum, i.e. $i_0 \in \mathrm{argmax}_{1 \leqslant i \leqslant n} v_i$. Then using $Hv = 0$, hence $(Hv)_{i_0} = 0$, one has

$$0 = \sum_{i \neq i_0} H_{i,i_0} v_i + H_{i_0,i_0} v_{i_0} = \sum_{i \neq i_0} H_{i,i_0} v_i - \sum_{i \neq i_0} H_{i,i_0} v_{i_0} = \sum_{i \neq i_0} H_{i,i_0}(v_i - v_{i_0}).$$

This follows from $H_{i_0,i_0} = -\sum_{i \neq i_0} H_{i,i_0}$. Since for every $i \neq i_0$, one has $H_{i,i_0} \geqslant 0$ and $v_{i_0} - v_i \geqslant 0$, this implies that $v_i = v_{i_0}$ for every $i$ such that $H_{i,i_0} \neq 0$. By induction and using the connectedness of the graph $\mathcal{G}$, this shows that $v$ has to be constant, i.e. $\mathrm{Ker}(H) = \mathrm{vect}(\{\mathrm{cst}\})$. □

### 1.4.4 Convergence analysis

In this section, we show the convergence of a damped Newton's algorithm for a general function $G : \mathbb{R}^N \to \mathbb{R}^N$ that satisfies some regularity and strict monotonicity conditions. As a direct consequence, using the results of Sections 1.4.2 and 1.4.3, we show the convergence with a linear speed of the damped Newton's algorithm to solve the non-linear equation (DMA). We denote by $\mathcal{P}_N$ the set of $\nu = (\nu_1, \cdots, \nu_N) \in \mathbb{R}^N$ that satisfies $\nu_i \geqslant 0$ and $\sum_i \nu_i = 1$. For a given function $G : \mathbb{R}^N \to \mathcal{P}_N$ and $\varepsilon > 0$, we define the set

$$\mathcal{K}^\epsilon := \left\{ \psi \in \mathbb{R}^N \mid \forall i,\ G_i(\psi) \geqslant \epsilon \right\},$$

where $G(\psi) = (G_i(\psi))_{1 \leqslant i \leqslant N}$. We then have the following proposition, which is an adaptation to our setting of Theorem 1.5 in [KMT16] and Proposition 2.10 in [Mir15].

PROPOSITION 28. *Let $G : \mathbb{R}^N \to \mathcal{P}_N$ be a function which is invariant under the addition of a constant, i.e. a multiple of $(1, \dots, 1) \in \mathbb{R}^N$, and $\varepsilon > 0$. We assume the following properties:*

1. *(Compactness) For every $a \in \mathbb{R}$, the following set is compact:*

$$\mathcal{K}_a^\epsilon := \mathcal{K}^\epsilon \cap \left\{ \psi \in \mathbb{R}^N \mid \sum_{i=1}^N \psi_i = a \right\} = \left\{ \psi \in \mathbb{R}^N \mid \forall i,\ G_i(\psi) \geqslant \epsilon \text{ and } \sum_{i=1}^N \psi_i = a \right\}.$$

2. *($\mathcal{C}^1$ regularity) The function $G$ is of class $\mathcal{C}^1$ on $\mathcal{K}^\epsilon$.*

3. *(Strict monotonicity) We have:*

$$\forall \psi \in \mathcal{K}^\epsilon,\ \forall v \in \{\mathrm{cst}\}^\perp \setminus \{0\},\ \langle \mathrm{D}G(\psi)v \mid v \rangle < 0$$

*Then Algorithm 4 converges with linear speed. More precisely, if $\nu \in \mathcal{P}_N$ and $\psi^0 \in \mathbb{R}^N$ are such that $\epsilon_0 = \frac{1}{2} \min\left(\min_i G_i(\psi^0), \min_i \nu_i\right) > 0$, then the iterates $(\psi^k)$ of Algorithm 4 satisfy the following inequality, where $\tau^* \in (0, 1]$ depends on $\epsilon_0$:*

$$\left\| G(\psi^{k+1}) - \nu \right\| \leqslant \left(1 - \frac{\tau^\star}{2}\right) \left\| G(\psi^k) - \nu \right\|.$$

*Proof.* Let $\nu \in \mathcal{P}_N$ and $\psi^0 \in \mathbb{R}^N$ such that $\epsilon_0 = \frac{1}{2} \min\left(\min_i G_i(\psi^0), \min_i \nu_i\right)$ is positive. For convenience, we denote in this proof $\epsilon := \epsilon_0$. We put $a = \sum_{i=1}^N \psi_i^0$. We are going to show that there exists $\tau' \in ]0, 1]$ such that for every $\psi \in \mathcal{K}_a^\epsilon$ and every $\tau \in (0, \tau')$, one has

$$\psi_\tau \in \mathcal{K}_a^\epsilon \quad \text{and} \quad \|G(\psi_\tau) - \nu\| \leqslant \left(1 - \frac{\tau}{2}\right) \|G(\psi) - \nu\|,$$

where $\psi_\tau = \psi - \tau v$ and $v = \mathrm{D}G^+(\psi)(G(\psi) - \nu)$. This directly implies the convergence of Algorithm 4 by putting $\tau^* = \frac{\tau'}{2}$.

Since $\mathcal{K}_a^\epsilon$ is a compact set, the continuous map $\mathrm{D}G$ is uniformly continuous on $\mathcal{K}_a^\epsilon$, i.e. there exists a function $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ that satisfies $\lim_{x \to 0} \omega(x) = \omega(0) = 0$ and such that

$$\forall \psi, \tilde{\psi} \in \mathcal{K}_a^\epsilon,\ \left\| \mathrm{D}G(\psi) - \mathrm{D}G(\tilde{\psi}) \right\| \leqslant \omega(\left\| \psi - \tilde{\psi} \right\|).$$

Note also that the modulus of continuity $\omega$ can be assumed to be an increasing function. For any $\psi \in \mathcal{K}_a^\epsilon$, we let $v = \mathrm{D}G^+(\psi)(G(\psi) - \nu)$ and $\psi_\tau = \psi - \tau v$ for any $\tau \geqslant 0$. Since $G$ is of class $\mathcal{C}^1$, a Taylor expansion in $\tau$ gives

$$G(\psi_\tau) = G(\psi - \tau \mathrm{D}G^+(\psi)(G(\psi) - \nu)) = (1 - \tau)G(\psi) + \tau\nu + R(\tau) \tag{1.4.14}$$

where $R(\tau) = \int_0^\tau (\mathrm{D}G(\psi_t) - \mathrm{D}G(\psi))v \mathrm{d}t$ is the integral remainder. Then, we can bound the

norm of $R(\tau)$

$$\begin{aligned}\|R(\tau)\| &= \left\| \int_0^\tau (\mathrm{D}G(\psi_t) - \mathrm{D}G(\psi))v \mathrm{d}t \right\| \\ &\leqslant \|v\| \int_0^\tau \omega(\|\psi_t - \psi\|)\mathrm{d}t = \|v\| \int_0^\tau \omega(t\,\|v\|)\mathrm{d}t \\ &\leqslant \tau\,\|v\|\,\omega(\tau\,\|v\|)\end{aligned}$$

where we have used the fact that $\omega$ is an increasing function.

**Step 1** We first want to show that for every $\psi \in \mathcal{K}_a^\epsilon$ there exists $\tau(\psi) > 0$ such that

$$\forall \tau \in (0, \tau(\psi)) \qquad \psi_\tau \in \mathcal{K}_a^\epsilon \quad \text{and} \quad \|G(\psi_\tau) - \nu\| \leqslant \left(1 - \frac{\tau}{2}\right) \|G(\psi) - \nu\| . \tag{1.4.15}$$

Recall that for every $i \in \{1, \dots, N\}$ one has $\nu_i \geqslant 2\epsilon$ and $G_i(\psi) \geqslant \epsilon$. Thus one gets

$$G_i(\psi_\tau) \geqslant (1-\tau)G_i(\psi) + \tau\nu_i + R_i(\tau) \geqslant (1+\tau)\epsilon - \|R(\tau)\| .$$

So if we choose $\tau$ such that $\|R(\tau)\| \leqslant \tau\epsilon$ then $G_i(\psi_\tau) \geqslant \epsilon$ and $\psi_\tau \in \mathcal{K}^\epsilon$. Now, since $\lim_{x\to 0} \omega(x) = 0$, there exists $\alpha_1 > 0$ such that for every $0 \leqslant t \leqslant \alpha_1$, one has $\omega(t) \leqslant \epsilon/\|v\|$. This implies that if $\tau \leqslant \alpha_1/\|v\|$, then $\|R(\tau)\| \leqslant \tau\epsilon$ and consequently $\psi_\tau \in \mathcal{K}^\epsilon$. Note that $G(\psi) - \nu$ belongs to $\{\mathrm{cst}\}^\perp$ and that $\mathrm{D}G(\psi)$ is an isomorphism from $\{\mathrm{cst}\}^\perp$ to $\{\mathrm{cst}\}^\perp$. We deduce that $\psi_\tau - \psi = \tau v$ belongs to $\{\mathrm{cst}\}^\perp$, hence $\psi_\tau \in \mathcal{K}_a^\epsilon$.

From Equation (1.4.14), we have $G(\psi_\tau) - \nu = (1-\tau)(G(\psi) - \nu) + R(\tau)$. So, to get the second condition of Equation (1.4.15), it is sufficient to show that $\|R(\tau)\| \leqslant (\tau/2)\,\|G(\psi) - \nu\|$. The estimation on $\|R(\tau)\|$ and the definition of $v$ gives us

$$\|R(\tau)\| \leqslant \tau \left\|\mathrm{D}G^+(\psi)\right\| \|G(\psi) - \nu\|\,\omega(\tau\,\|v\|).$$

Still from the continuity of $\omega$ at 0, we can find $\alpha_2 > 0$ such that for every $\tau \leqslant \alpha_2/\|v\|$ one has $\omega(\tau\,\|v\|) \leqslant \epsilon/2\,\|\mathrm{D}G^+(\psi)\|$, thus $\|R(\tau)\| \leqslant (\tau/2)\,\|G(\psi) - \nu\|$. Therefore, by putting $\tau(\psi) := \min(\alpha_1/\|v(\psi)\|, \alpha_2/\|v(\psi)\|, 1)$, Equation (1.4.15) is proved. Note that we impose $\tau(\psi)$ to be less than 1.

**Step 2** The function $G$ is of class $\mathcal{C}^1$ on $\mathcal{K}_a^\epsilon$. For every $\psi$ in $\mathcal{K}_a^\epsilon$, $\mathrm{D}G(\psi)$ is an isomorphism from $\{\mathrm{cst}\}^\perp$ to $\{\mathrm{cst}\}^\perp$ and its inverse $\mathrm{D}G^+(\psi)$ depends continuously on $\psi$. Since $\sum_i G_i(\psi) = \sum_i \nu_i$, $G(\psi) - \nu$ belongs to $\{\mathrm{cst}\}^\perp$, so the function $v(\psi) = \mathrm{D}G^+(\psi)(G(\psi) - \nu)$ is also continuous by composition. If $G(\psi) \neq \nu$, the strict monotonicity of $G$ ensures that $v(\psi) \neq 0$ and so $\tau(\psi) = \min(\alpha_1/\|v(\psi)\|, \alpha_2/\|v(\psi)\|, 1)$ is also continuous in $\psi$. If $G(\psi) = \nu$, then $v(\psi) = 0$. However, by continuity of $v$, the function $\tilde{\psi} \mapsto \tau(\tilde{\psi})$ is constant equal to 1 in a neighborhood of $\psi$. Hence the function $\psi \mapsto \tau(\psi)$ is globally continuous. Therefore, the infimum of $\tau(\psi)$ over the compact set $\mathcal{K}_a^\epsilon$ is attained at a point of $\mathcal{K}_a^\epsilon$, thus is strictly positive. We deduce that

we can take a uniform bound $\tau(\psi) =: \tau' > 0$ in Equation (1.4.15) that does not depend on $\psi$. This directly implies the convergence of the damped Newton algorithm with linear speed. □

*Proof of Theorem 14.* The function $G$ appearing in (DMA) satisfies the regularity condition (Theorem 18) and the monotonicity condition (Theorem 22) needed in Proposition 28. It remains to show the compactness condition. Let us take $a \in \mathbb{R}$ and let us show that $\mathcal{K}_a^\epsilon$ is compact. It is easy to see that $\mathcal{K}_a^\epsilon$ is closed since $G$ is continuous. Let $\psi \in \mathcal{K}_a^\epsilon$, $i \neq j$ and $x \in \mathrm{Lag}_i(\psi)$. Then one has

$$\psi_i \leqslant \psi_j + \|x - y_j\|^2 - \|x - y_i\|^2 \leqslant \psi_j + \mathrm{diam}(K \cup Y)^2,$$

where $\mathrm{diam}(K \cup Y)$ is the diameter of $K \cup Y$. So the differences $|\psi_i - \psi_j|$ are bounded by $\mathrm{diam}(K \cup Y)^2$. Combined with the fact that $\sum_i \psi_i$ is constant, one has that $\psi$ is bounded by a constant independent on $\psi$. Thus, $K_a^\epsilon$ is compact. □

## 1.5 Numerical results

In this section, we solve the optimal transport problem in $\mathbb{R}^3$ between triangulated surfaces (possibly with holes, with or without a boundary) and point clouds, for the quadratic cost and show it can be used in different settings: *optimal quantization* of a probability density over a surface, *remeshing* and *point set registration* on a mesh. The source density is assumed to be affine on each triangle of the triangulated surface. One crucial aspect of the algorithm is the exact computation of the combinatorics of the Laguerre cells, *i.e.* the intersection between a triangulated surface and a 3D power diagram, see Equation (1.4.3). Another important aspect is the initialization step in Algorithm 4, *i.e.* finding a set of weights $\psi^0$ which guarantees that all the initial Laguerre cells have a positive mass. This aspect is described in detail in Chapter 4. We first explain the algorithm we use to compute the Laguerre cells and the function $G$ along with its Jacobian matrix D$G$ before presenting some results and applications.

### 1.5.1 Implementation details

We describe here an algorithm to compute the combinatorics of the intersection of a Power diagram $\mathrm{Pow}(P) := (\mathrm{Pow}_i(P))_i$ of a weighted point cloud $P = \{(p_i, \omega_i)\}$ with a triangulated surface $K = \cup_{\sigma \in \Sigma} \sigma$ where $\sigma$ is a triangle. Note that in general the intersection of a Power cell with $K$ is not convex and can even have several connected components (as illustrated for instance in Figure 6 in the second and third rows). We encode here the triangulated surface $K$ with a connected graph $G_1$ where $G_1$ is the 1-skeleton of $K$ (i.e. the collection of its vertices and edges) seen as a subset of $\mathbb{R}^3$. Similarly, the intersection of the 2D faces of the Power diagram with the triangulated surface $K$, namely $G_2 = \cup_i (K \cap \partial\, \mathrm{Pow}_i)$, is encoded by a graph.

Remark 29. *Remark that for every $i$, the intersection between the triangulated surface $K$ and the boundary of the three dimensional convex set* $\mathrm{Pow}_i(P)$ *is generically a union of closed*

*polygonal lines. Let us also remark that $G_2$ can be disconnected. For instance, if $K$ is a triangulation of the sphere, then the intersection between a Power cell $\mathrm{Pow}_i(\psi)$ that "traverses" the sphere will have two connected components.*

The core of the algorithm to compute $G_1 \cup G_2$ is then composed of the following steps:

1. We first split the edges in the graph $G_1$ at points in $G_1 \cap G_2$. Since $G_1$ is connected, this can be done by a simple traversal, in which we need to intersect the edges of the triangulation with the 2-dimensional Power cells.
2. We then traverse $G_2$ starting from vertices in $G_1 \cap G_2$ by intersecting the 2-dimensional Power cells with triangles. $G_2$ might be disconnected, but we can discover the connected components using the non-visited vertices in $G_1 \cap G_2$. This step provides us with both the geometry and connectivity of $G_1 \cup G_2$, and also an orientation coming from the underlying triangulated surface $K$.
3. The graph $G_1 \cup G_2$ is embedded on the triangulated surface $K$, and the connected components of $K \setminus (G_1 \cup G_2)$ are (open) convex polygons. Each of these polygons represents an intersection of the form $\mathrm{Pow}_i \cap \sigma$. The boundary of these polygons can easily be reconstructed from $G_1 \cup G_2$ and the orientation (obtained in the second step).

This algorithm is encapsulated into a function `power_diagram_surface_intersection(tri, pow, f)` where `tri` represents the graph $G_1$, `pow` is the 3D power diagram (which is actually represented by its dual graph called a *regular triangulation*) and `f` is a functor which will be called on each polygon $P$ that is the intersection between a triangle $\sigma$ and a Power cell $\mathrm{Pow}_i(\psi)$.

As always in computational geometry algorithms, we need *predicates* and *constructions*. Here, the main predicates we need are the intersection tests between a 2D face and a segment (for the first step) and between a Power edge (1D face) and a triangle (used in the second step). These predicates are implemented in an exact manner using the filtered predicates mechanism provided by the `CGAL` library [CGA16]. The same library allows to efficiently compute 3D Power diagrams using a randomized algorithm which is quadratic in the worst-case but close to linear most of the time.

Remark 30. *Other algorithms exist for intersecting a Power diagram to a triangulated surface. For instance, the authors of [SNA17] use an elementary property of Voronoi diagrams to develop an efficient and parallelizable algorithm to intersect a Voronoi diagram in $\mathbb{R}^d$ with a triangulated surface. The `GEOGRAM` library [Lév15] also provides functions to compute the intersection of a 2D or 3D Voronoi diagram with a triangulation or a tetrahedrization in $\mathbb{R}^3$.*

The core of the algorithm, meaning the computation of the intersection between a triangulated surface and a 3D Power diagram is implemented in `C++` amounting to 1.6k lines of code (counting the predicates). Once the combinatorics are computed, one needs to integrate the source density over the Laguerre cells, this has to be done in an exact manner (see the next

paragraph for more details). In our setting, we assumed the source density to be piecewise affine. Such density $\mu$ is represented by a class able to answer the following question: what is the value of $\mu(p)$ knowing that $p$ belongs to a triangle $t$? This is implemented using a simple barycentric interpolation. On top of it, we made `Python` bindings using the `pybind11` library to leverage the speed and ease-of-use of Python libraries like `NumPy`, `SciPy` an `Matplotlib`. This allowed us to quickly obtain a working damped Newton's algorithm and then focus on the applications (see the next section for details). To give an idea of how the computation time is divided: if we take a triangulation of the sphere $\mathbb{S}^2$ for $K$ with a uniform measure and a point cloud $Y$ of size $10^5$ uniformly sampled in $[-1/2, 1/2]^3$, then the running time to compute $G(\psi)$ and $\mathrm{D}G(\psi)$ is divided into two parts: 73 % for computing the Laguerre diagram and 27 % for computing the integrals.

**Numerical integration.** The computation of $G_i(\psi)$ and $\frac{\partial G_i}{\partial \psi_j}(\psi)$ requires the evaluation of integrals of the form $\int_{\mathrm{Lag}_i(\psi)\cap\sigma} \rho_\sigma(x)\mathrm{d}\mathcal{H}^2(x)$ and $\int_{\mathrm{Lag}_{i,j}(\psi)\cap\sigma} \rho_\sigma(x)\mathrm{d}\mathcal{H}^1(x)$ where $\rho_\sigma : \mathbb{R}^3 \to \mathbb{R}^+$ is an affine density and $\sigma$ a triangle with vertices in $\mathbb{R}^3$. In order to evaluate these integrals exactly, we use the classical Gaussian quadrature formulas. In our setting, we have that if $t = [a, b, c]$ is a triangle, $s = [a, b]$ a line segment and $\rho : t \to \mathbb{R}$ an affine density, then

$$\int_t \rho(x)\mathrm{d}\mathcal{H}^2(x) = \mathrm{Area}(t) \cdot \rho\left(\frac{a+b+c}{3}\right) \text{ and } \int_s \rho(x)\mathrm{d}\mathcal{H}^1(x) = \|b-a\| \cdot \rho\left(\frac{a+b}{2}\right).$$

**Computation of the descent direction $v^k$.** We can solve efficiently the linear system $\mathrm{D}G(\psi^k)v^k = -(G(\psi^k) - \nu)$ since $\mathrm{D}G(\psi^k)$ is sparse. In practice, we use a sparse Cholesky solver, but a conjugate gradient method can also be used.

### 1.5.2 Numerical results and applications

We compute the optimal transport map between a piecewise linear measure defined on a triangulated surface $K$ and a discrete measure defined on a 3D point cloud. Even if we can handle non uniform measures, in the examples presented here, the source density is uniform over the triangulation: $\rho_\sigma = 1/\mathrm{Area}(K)$ for every $\sigma \in \Sigma$, where $\mathrm{Area}(K)$ is the area of $K$. The point cloud is chosen to be a sampling of point on the mesh with some noise added. In the examples, the solutions are computed up to an error of $\eta = 10^{-6}$. In practice, we do not check the genericity assumption of Definition 1.9 and the solver worked in all our experiments.

The first two rows of Figure 6 displays results for a uniform target measure and the last two for a non-uniform one. Remark that in this case the non uniformity creates smaller Laguerre cells on the right side. Note that the centroids of the Laguerre cells provide naturally a correspondence between the point cloud and the triangulated surface: we associate to each $y_i$ the centroid of the Laguerre $\mathrm{Lag}_i(\psi^k)$, where $\psi^k$ is the output of Algorithm 4. In practice, the number of iterations remains small even for large point sets. For instance, if we choose $10,000$ noisy samples on the torus, the algorithm takes 16 iterations to solve the problem.

The transport map can be seen on two examples in Figure 7. For a uniform target density (first figure), the transport map can roughly be seen as the orthogonal projection on the mesh. However it is more complex for a non-uniform one (second figure). Corresponding error plots (in logarithmic scale) for a numerical error of $\eta = 10^{-14}$ can be found in Figure 8. We observe in Figure 8 and Figure 9 that the error rate decreases superlinearly. More precisely, we observe in Figure 9 that the convergence becomes quadratic when the Newton step parameter $\tau := 1/2^{\ell}$ (where $\ell$ is given in Algorithm 4) is equal to 1. We also observe in Figure 10 that the damped Newton's method converges in much less iterations than the BFGS algorithm [LN89].

REMARK 31. *We also underline that the Laguerre cells can be non geodesically convex and even disconnected (as illustrated in the second and third columns of Figure 6) which shows that our method handles more general settings than [KMT16], i.e. cost functions whose Laguerre cells cannot be convex in any chart (violating the hypothesis of [KMT16], Definition 1.1).*

We now show how to use this algorithm as a building block for higher level operations such as optimal quantization of surfaces, remeshing and point set registration.

### Optimal quantization of a surface

*Optimal quantization* is a sampling technique used to approximate a density function with a point cloud, or more accurately a finitely supported measure. It has many applications in image dithering or in computer graphics (see [Goe+12] for more details). Here, we show how to perform this kind of sampling on triangulated surfaces. Given a triangulated surface $K \subset \mathbb{R}^3$ and a density $\mu$ on $K$, the problem can be stated as follows

$$\min_{y_1,\dots,y_N \in \mathbb{R}^3} \mathrm{W}_2\left(\mu, \frac{1}{N}\sum_{i=1}^{N} \delta_{y_i}\right),$$

where $\mathrm{W}_2(\mu,\nu)$ is the Wasserstein distance between $\mu$ and $\nu$ for the quadratic cost. A procedure to find a solution to this minimization problem is the following: we first define $Y^0$ as the set of vertices of $K$ and consider the constant probability measure $\nu^0$ on $Y^0$. For each $k \geq 0$, we solve the optimal transport between $\mu$ on $K$ and $\nu^k$ on $Y_k$ and pick one point, for instance the centroid, per Laguerre cell. We iterate this procedure by choosing for the new point cloud $Y^{k+1}$ the set of the previously computed centroids and for $\nu^{k+1}$ the uniform measure over $Y^{k+1}$. After a few iterations, this gives us a (locally) optimal quantization of $K$. Figure 11 shows examples of sampling on different surfaces with different densities.

### Remeshing

We now consider the following problem: given a triangulated surface $K$, a density $\mu$ supported on this mesh, we want to build a new mesh such that the distribution of triangles respect this density, meaning that we want more triangles where the density is bigger. This has applications for instance in finite element methods for solving partial differential equations

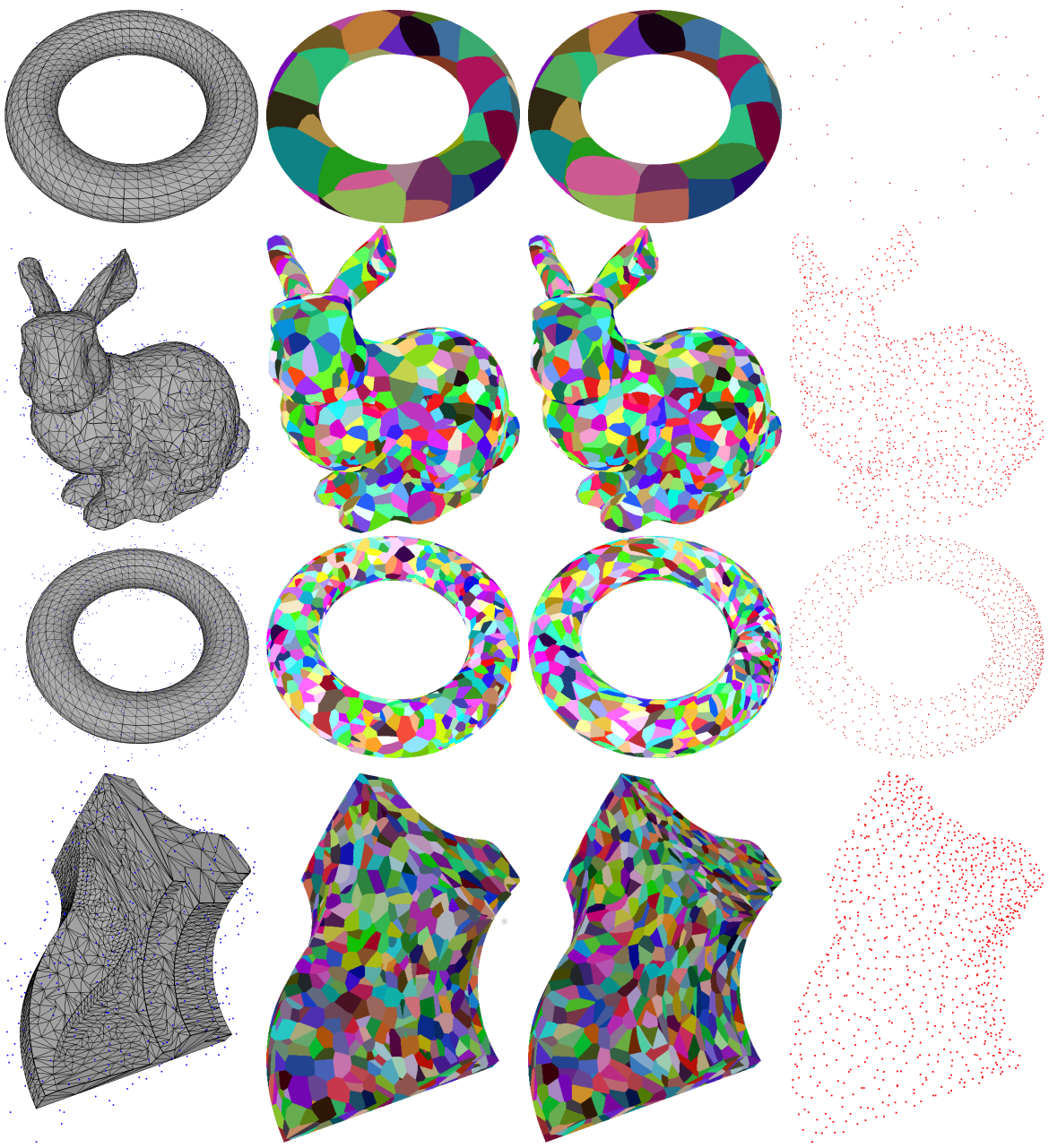

Figure 6 – **Optimal transport between a connected triangulated surface in $\mathbb{R}^3$ and a target measure supported on a 3D point cloud.** From left to right: Mesh and initial point cloud (in blue), Initial Laguerre cells, Final Laguerre cells, Centroids of the final Laguerre cells. The source measure is uniform. In the first two rows, the target density is uniform while in the last two, it linearly decreases from left to right. In the first row, $N = 50$ while in the other rows, $N = 1000$. Computation time (number of iterations): 3s (4) / 41s (6) / 74s (13) / 58s (22) for a numerical error $\eta = 10^{-6}$.

where the quality of the mesh used for discretization matters. To do this, we can use the following simple procedure: we consider the uniform discrete measure $\nu$ supported on the

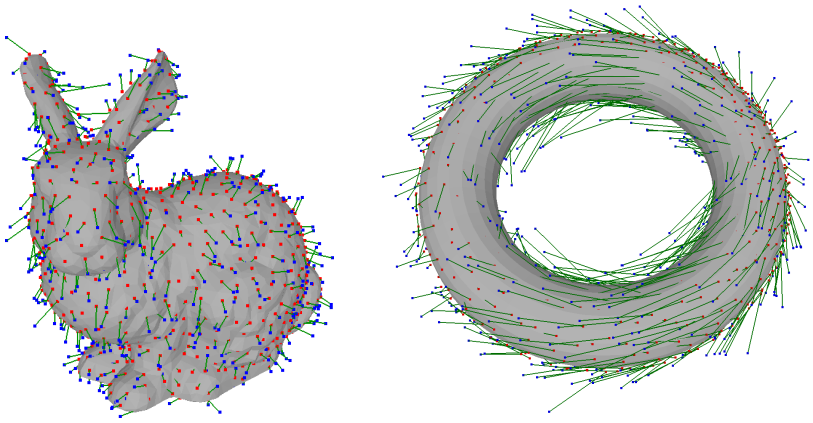

Figure 7 – **Visualization of a transport map.** Here, the transport map is visualized as a collection of segments (in green) connecting an initial point $y_i$ (in blue) to the centroid (in red) of its Laguerre cell $\mathrm{Lag}_i(\psi)$. The first figure corresponds to a uniform target density on the Stanford bunny model (second row of Figure 6), while the second one corresponds to a non-uniform target density on the torus (third row of Figure 6). On both examples, $N = 1000$.

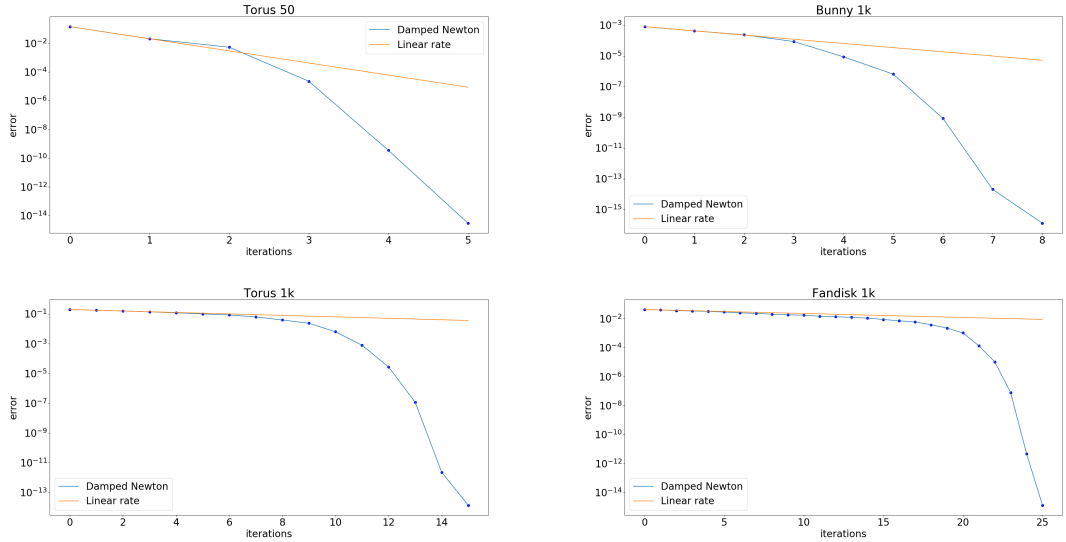


Figure 8 – **Evolution of the error rate** $(\|G(\psi^k) - \nu\|)_k$ for the four examples of Figure 6 for a numerical error $\eta = 10^{-14}$ (the y-axis is in logarithmic scale).

vertices of $K$; we solve the optimal transport between $\mu$ on $K$ and $\nu$; the new mesh will be taken as the dual (in the graph sense) of the final Laguerre diagram. See Figure 12 for two examples for different source densities.

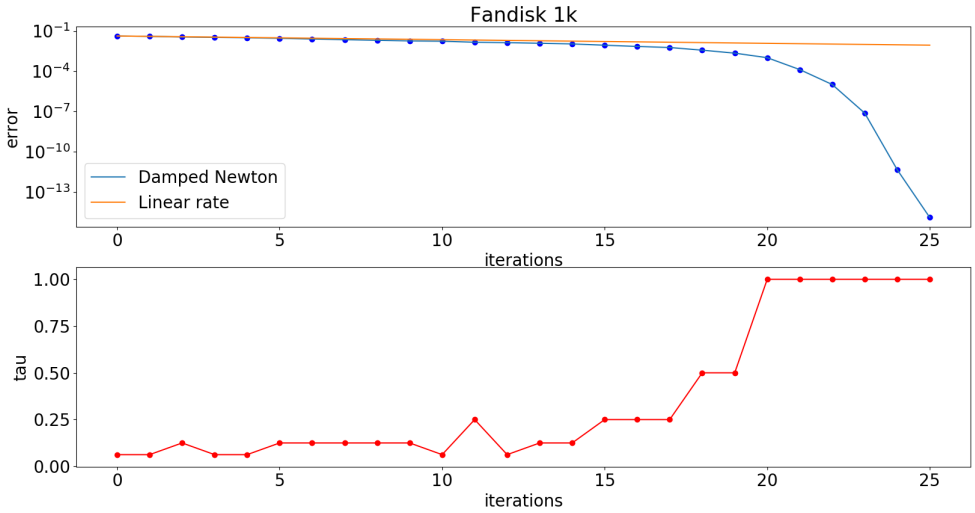


Figure 9 – **Evolution of the error rate and the Newton step parameter.** Top: evolution of the error rate; Bottom: evolution of $\tau := 1/2^{\ell}$ of Algorithm 4 for the Fandisk model. In the top row, the y-axis is in logarithmic scale. The numerical error is set to $\eta = 10^{-14}$.


Figure 10 – **Comparison between the damped Newton's method (Algorithm 4) and the BFGS algorithm** on the third example of Figure 6. We stopped the BFGS algorithm after 1000 iterations (the y-axis is in logarithmic scale).

### Point set registration

We finally consider the rigid point set registration on a mesh. Given a triangulated surface $K$ and a point cloud $Y$, we want to find a rigid transformation $T$ such that the $L^2$ distance

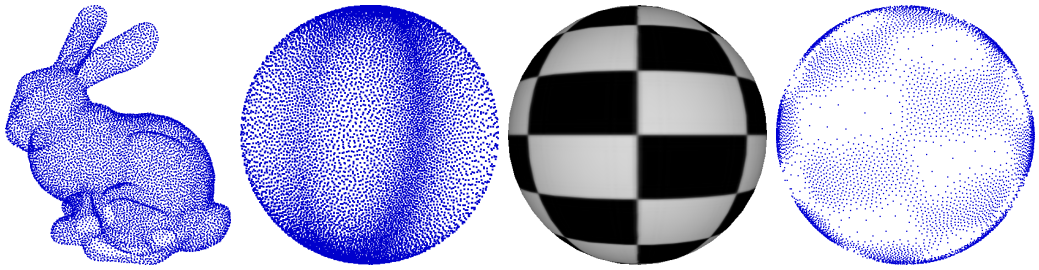

Figure 11 – **Optimal quantization of triangulated surfaces for different densities and surfaces.** From left to right: uniform density $\mu = 1$ on the Stanford Bunny (10k points); non-linear density $\mu(x, y, z) = e^{-3|y|}$ on the sphere (10k points); checkerboard texture and sampling for the density corresponding to the UV-mapping of the texture on the hemisphere (5k points).

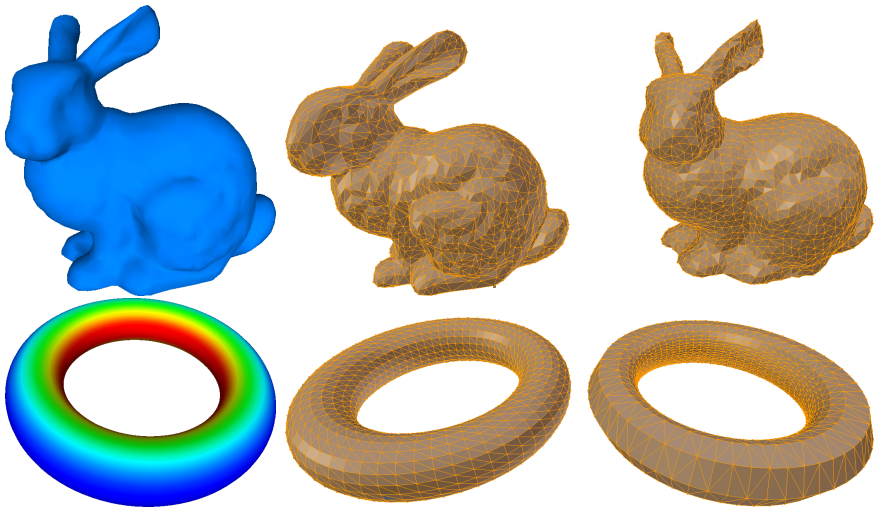

Figure 12 – **Remeshing using optimal transport.** From left to right: source density; initial mesh and remeshed surface. First row: Uniform density: $\mu = 1$; Second row: $\mu$ is proportional to a mean curvature estimator of the source mesh. Number of vertices for each model: Bunny: 2.2k; Torus: 5.6k.

between $K$ and $T(Y)$ is minimal that is to say it solves the following minimization problem

$$\min_{T \text{ rigid}} \sum_{y \in Y} \min_{x \in K} \|T(y) - x\|^2 .$$

The most popular method to do this is the Iterative Closest Point (ICP) algorithm developed in [BM92]. For this algorithm, we need to be able to compute for each point $y_i$ from the point cloud $Y$ its closest point on the mesh $K$. We can replace the traditional nearest neighbor query with the following routine: we solve the optimal transport between the constant probability measure $\mu$ on $K$ and the constant probability measure $\nu$ on $Y$, then associate each point $y_i$ to a point (for instance the centroid) of the Laguerre cell $\mathrm{Lag}_i(\psi)$ where $\psi \in \mathbb{R}^N$ are the final weights. It amounts to considering the following problem

$$\min_{T \text{ rigid}} \mathrm{W}_2 \left( \mu, \frac{1}{N} \sum_{i=1}^{N} \delta_{T(y_i)} \right)$$

where $\mu$ is the uniform measure over $K$. The resulting algorithm is called Optimal Transport ICP (OT-ICP). See Figure 13 for one example. In our results, OT-ICP converges in much less iterations than standard ICP, namely 3 iterations versus 20 iterations for the same stopping criterion in our two test cases. The quality of the final point clouds is approximately the same for the two algorithms. The main disadvantage of OT-ICP is its running time which remains higher than the traditional method.

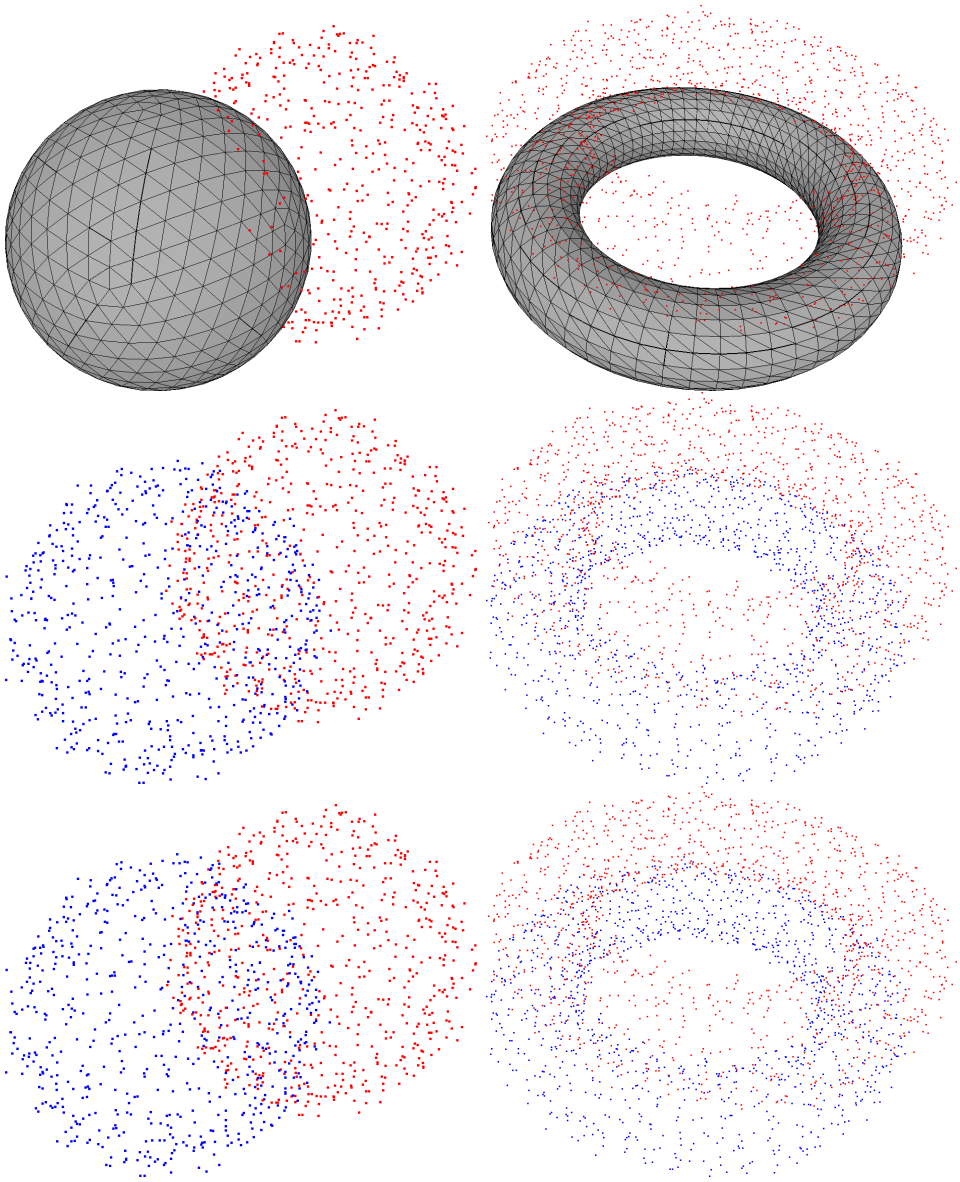

Figure 13 – **Comparison between ICP and OT-ICP.** From top to bottom: initial mesh (in grey) and initial point cloud (in red); initial (red) and final (blue) point clouds using traditional ICP; initial (red) and final (blue) using OT-ICP.

Chapter 2

# Optimal transport formulation of non-imaging optics problems

## Contents



In this chapter, we show how one can study many different inverse problems arising in optics in a unified framework using optimal transport. More precisely, we describe the intimate relation between optimal transport for different cost functions and optical component design problems.

We introduce in Section 2.1 the field of non-imaging optics, explain how it relates to the design of optical components (such as mirrors or lenses) and give the main settings that we will look into namely the *far-field* and *near-field* settings. We also describe the numerical methods that exist to solve such problems. In Section 2.2, we explain in a particular setting namely the mirror design for a collimated source and a target at infinity, how one can formalize the problem using optimal transport and the notion of weak solution *à la Brenier*. We show existence of solutions in the semi-discrete and continuous settings. In Section 2.3, we present a method inspired by the *supporting paraboloids* algorithm and show how, using optimal transport, we can develop a common framework to solve different optical component design problems for an ideal light source in the *far-field* setting. Finally, in Section 2.4, we explain in more details the *Ma-Trudinger-Wang* (MTW) condition [MTW05]. We show that the cost function appearing in the design of lenses for point light sources satisfies this MTW condition using similar arguments

as in [GH09]. The main interest for us is that it guarantees: i) that the Laguerre cells are connected; ii) the convergence of the damped Newton's method presented in Chapter 1.

## 2.1 Non-imaging optics

In this section, we present the field of non-imaging optics as well as the existing numerical methods developed to solve the problems arising in this domain.

### 2.1.1 Introduction

The field of non-imaging optics deals with the design of optical components whose goal is to transfer the radiation emitted by a light source onto a prescribed target. This question is at the heart of many applications where one wants to optimize the use of light energy by decreasing light loss or light pollution. An illustration can be found in Figure 1 with a point light source $\mu$ and a target at infinity $\nu$.

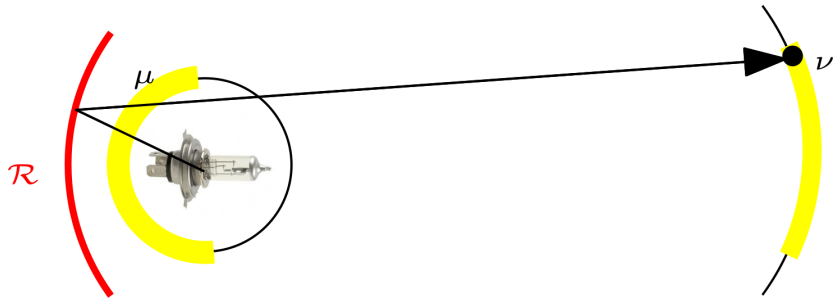


Figure 1 – **An example of problem arising in non-imaging optics.** We want to design the mirror $\mathcal{R}$ that reflects the light emitted from a point source $\mu$ towards the prescribed target illumination $\nu$ located at infinity.

Such problems appear in the design of car beams [And+15], public lighting, solar ovens and hydroponic agriculture. This problem has also been considered under the name of *caustic design*, with applications in architecture, interior decoration [FDL10; DH15]. A caustic is the envelope of rays reflected or refracted by an object. In the following, we will call a *setting* a pair (light source, target illumination), and more precisely (type of light source, position of the target). When the target illumination is located at infinity, we can model it as a collection of directions and we say that it is a **far-field** problem. On the contrary, when the target is located at a finite distance, we say that it is a **near-field** problem. We can also consider different types of light sources, the most common types are the following:

- **ideal sources**: such sources are light sources where there is only one incident ray for a given surface point. For instance, the sun is not an ideal source. This category includes collimated light sources that emits rays parallel to each other; or point light sources that emits rays in every direction.

- **extended sources**: unlike ideal ones, extended light sources possess a radius and thus are capable of producing shadows with fuzzy edges. In practice, they are modelled as a collection of point light sources.

In this chapter, we consider the problem of designing a wide variety of mirrors and lenses that satisfy different kinds of light energy constraints. To be a little bit more specific, in each problem that we consider, one is given a light source and a desired illumination after reflection or refraction which is called the *target*. The goal is to design the geometry of a mirror or lens which transports exactly the light emitted by the source onto the target. The design of such optical components can be thought of as an *inverse problem*, where the *direct problem* would be the simulation of the target illumination from the description of the light source and the geometry of the mirror or lens, using for instance ray tracing techniques.

In practice, the mirror or lens needs to satisfy aesthetic and pragmatic design constraints. In many situations, such as for the construction of car lights, physical moulds are built by milling and the mirror or lens is built on this mould. Sometimes the optical component itself is directly milled. This imposes some constraints that can be achieved by imposing convexity or smoothness conditions. The convexity constraint is classical since it allows in particular to mill the component with a tool of arbitrary large radius. Conversely, concavity allows to mill the mould of the component. Also, convex mirrors are easier to chrome-plate, because convex surfaces have no bumps in which the chrome would spuriously concentrate [CBC77].

In the next section, we summarize a few of the existing numerical methods that exist to solve such problems.

### 2.1.2 Existing numerical methods

The field of non-imaging optics has been extensively studied in the last thirty years. We give below an overview of the main approaches to tackle several optical component design problems. These methods can be mainly separated into two categories: (i) ad-hoc (or parametric) methods whose goal is to develop procedures specific to a particular setting; (ii) freeform optics which tries to create tools that can be used more freely to design optical components. In the following, we will indistinctly use "reflector" to refer to a mirror and "refractor" to refer to a lens.

**Ad hoc methods.** Ad-hoc methods are methods specific to each setting. For instance, in [WMB+05], many different methods are presented for different optical component design problems. More precisely, they study the underlying optical phenomena to find a configuration in which they are able to design "image-forming concentrators" that is to say optical components that transform a source radiation into a prescribed target one. They also study different combinations of basic components such as spherical mirrors, parabolic concentrators, hyperbolic concentrators, lenses... They can also reduce the dimensionality of the problem by assuming for instance that the optical component has some symmetry. Most of these methods only work in specific settings as they are restrained to only use some types of surface and thus lack degrees of freedom. On the contrary, in this chapter, we are interested in being able to design

components for a variety of different settings with a common formulation.

This is why we now look at the second category of methods namely the freeform-based ones. A freeform surface, contrary to traditional optics, has no translational or rotational symmetry. We will differentiate the methods depending on the setting namely, the type of source and location of the target. We start by looking at the setting where the source is ideal and the target located at infinity.

### Ideal source and far-field target

**Non convex energy minimization methods.** Many different methods to solve inverse problems arising in non-imaging optics rely on variational approaches. When the energies to be minimized are not convex, they can be handled by different kind of iterative methods. A survey on inverse surface design from light transport behaviour can be found in [PP05].

One class of methods uses stochastic optimization. In [FDL10], the optical component (mirror or lens) is represented as a $\mathcal{C}^2$ B-spline triangle mesh and a stochastic optimization is used to adjust the heights of the vertices so as to minimize a light energy constraint. Note that this approach is very costly, since a forward simulation (i.e. simulating the behaviour of the light through the component) needs to be performed at every step and the number of steps is very high in practice. Furthermore, using this method, lots of artifacts in the final caustic images are present.

Stochastic optimization has also been used in [Pap+11] to design reflective or refractive caustics for collimated light sources. At the center of the method is the Expectation Minimization algorithm initialized with a Capacity Constrained Voronoi Tessellation (CCVT) using a variant of the Lloyd's algorithm [Llo82]. The source is a uniform directional light and is modeled using an array of curved microfacets. Microfacets are tiny facets that approximate a surface, they are used in computer graphics for instance to approximate reflections. The target is represented by a mixture of Gaussian kernel functions. This method cannot accurately handle low intensity regions and artifacts due to the discretization are present. Microfacets were also used in [Wey+09] to represent the mirror. Due to the sampling procedure, this method cannot correctly handle smooth regions and does not scale well with the size of the target.

The method proposed in [Yue+12] uses transparent sticks made of acrylate resin to represent the refractive surface. This allows to reduce production cost, to be more entertaining for the user since a single set of sticks can produce different caustic patterns. The main problem with this approach is the computational complexity since they need to solve a NP-hard mixed integer programming problem.

**Ray-mapping and normal integration.** The approaches of [Kis+12; Yue+14; Sch+14] have in common that they first compute some bijection between the incident rays and their position on the target screen and then use an iterative method to compute the shape of the

refractive surface. The method of [Yue+14] uses a continuous parametrization and thus cannot correctly handle totally black and high-contrast regions (boundaries between very dark and very bright areas).

In [Sch+14], the authors propose a method to build lenses that can refract complicated and highly contrasted targets. They first use optimal transport on the target space to compute a mapping between the refracted rays of an initial lens and the desired normals, then perform a post-processing step to build a surface whose normals are close to the desired ones. The authors of [FFL16] also considers the design of lenses for point light sources by first computing a ray map between the source and target rays. They then build the surface with prescribed normals using a least-squares design approach. The main difficulty here is this normal integration step. Indeed, the problem may not have a solution since it is non-convex.

**Monge-Ampère equations.** When the source and target lights are modeled by continuous functions, the problem amounts to solving a generalized Monge-Ampère equation, either in the plane for collimated light sources, or on the sphere for point light sources. These partial differential equations are highly non-linear. The existence and regularity of the solutions, namely of the mirror or lens surfaces, have been extensively studied. When the light source is a point, the regularity of the solutions has been studied for mirrors [CO08; CGH08] and lenses [GH09] and when the light source is collimated, one recovers the usual Monge-Ampère equation in $\mathbb{R}^2$ [GT13].

**Optimal transport based methods in non-imaging optics.** In fact, the Monge-Ampère equations corresponding to the non-imaging problems considered in this chapter can be recast as optimal transport problems. This was first observed by [Wan04] and [GO03] for the mirror problem with a point light source. Many algorithms related to optimal transport have been developed to address non-imaging problems. For collimated sources, one can use wide-stencils finite difference schemes [Pri+13], or numerical solvers for quadratic semi-discrete optimal transport, such as [Mér11] or [Goe+12]. For point sources, there exist variants of the Oliker-Prussner algorithm for the mirror problem [CKO99a] or the lens problem [GH09]. These variants are both based on the idea of constructing the mirror or lens as the intersection of simple objects such as paraboloids (for mirrors) or ellipsoids (for lenses). This is why these methods are also called *supporting paraboloids* or *supporting ellipsoids*. We will refine this idea, see Section 2.3 and Chapter 3. The main weakness of these methods is that they have a $\mathrm{O}(N^4)$ complexity, restricting their use to small discretizations. A quasi-Newton method based on the supporting paraboloids method has been proposed in [CMT15] for uniform point-source reflector design, and can handle around $10^5$ Dirac masses. Finally we note that the approach of [Sch+14] to build lenses also relies on optimal transport. However, the optimal transport step is used as a heuristic to estimate the normals of the surface, and not to directly construct a solution to the non-imaging problem. A post-processing step is then performed by minimizing a non-convex energy composed of five weighted terms.

#### Ideal source and near-field target

When the target is located at a finite distance, the problem does not correspond to an optimal transport problem anymore. In particular, the cost function is not "separable" anymore into a sum of a function of the weight $\psi$ and a function of $(x, y)$ where $x$ is an incident ray and $y$ one of the prescribed target directions. The cost function satisfies a variant of the so-called Ma-Trudinger-Wang condition [MTW05], for which Trudinger established a regularity theory [Tru12]. The corresponding partial differential equation is part of a more general category called *Generated Jacobian equations*. When the source is a point, regularity of the solutions have been studied in [KO97] for reflectors and in [GH14] for refractors. When the source is collimated, the refractor problem has been studied in [GT13]. Let us note that the easiest way to prove existence of weak solutions is through a semi-discrete approach i.e. by approximating the target illumination by a finitely supported probability measure. For more general results (and in particular for reflectors for collimated light sources) on the regularity of the solutions of such equations, see [GK17]. The results from the same article encapsulate a lot of different settings: collimated or point light source with a target light at a finite distance. A numerical method based on the supporting ellipsoids method has been proposed to solve the near-field reflector problem for a point light source [KO97] but can only handle a discretization of a few dozens of Dirac masses.

#### Extended light sources

When the light source is not ideal, we say that the light source is *extended*. The optical component design problem becomes ill-posed since more than one ray can touch the component surface for one given point. The authors of [FCR09] proposed a method based on the supporting ellipsoids algorithm developed for the *far-field* setting but is not able to efficiently solve problems of relatively high resolution.

## 2.2 Mirror design for a collimated source and target at infinity

In this section, the goal is to show that continuous inverse problems arising in optics can be approached by semi-discrete ones. To illustrate this, we look at the following optical component design problem: find the mirror that reflects a collimated light source onto a prescribed illumination at infinity. It is known [GT13] that this problem amounts to solving a Monge-Ampère equation for the quadratic cost. We will study the existence of weak (Brenier) solutions for this problem. The strategy will be the following:

1. Prove the existence of solutions when $\nu$ is a finitely supported probability measure, which can be understood as the *semi-discrete* version of the original problem;
2. Prove that a solution of the continuous problem can be seen as the limit of a sequence of solutions of semi-discrete problems.

Proving the existence of solutions using intermediary semi-discrete problems it not a novel idea, as it has been used in [CO08] to prove the existence of solutions of mirror design for a point light source. It is a generalization of the work by Alexandrov and Pogorelov [Pog64] and has relations with the semi-discrete Minkowski problem [Gu+13] (reconstruct a convex surface with prescribed Gaussian curvature).

Let us model the collimated light source by a probability measure $\mu$ supported on a domain $\Omega \subset \mathbb{R}^2$. The target light illumination will be represented by a probability measure $\nu$ on $\mathbb{S}^2$. The mirror surface will be denoted by $\mathcal{R}$. According to Snell's law, for an incident ray $i \in \mathbb{S}^2$, a normal vector $n \in \mathbb{S}^2$, the reflected ray $R(i,n)$ is given by:

$$R : \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{S}^2, (i,n) \mapsto i - 2\langle i \mid n \rangle n.$$

In our setting, since the light is collimated, all the incident rays $i$ are parallel to $e_z = (0,0,1)$. The mirror will be represented by a graph of a function $\varphi$ over $\Omega$ meaning that $\mathcal{R} = \{(x, \varphi(x)) \mid x \in \Omega\}$. For a point $x \in \Omega$, the normal of the surface $\mathcal{R}$ at $x$ is given by $\vec{n}(x) = (\nabla\varphi(x), -1)/\left\|(\nabla\varphi(x), -1)\right\|$. If we define

$$F : v \in \mathbb{R}^2 \mapsto R\left(e_z, \frac{(v,-1)}{\|(v,-1)\|}\right) \in \mathbb{S}^2 \setminus \{e_z\},$$

Let us remark that $v \in \mathbb{R}^2 \mapsto \frac{(v,-1)}{\|(v,-1)\|}$ is a bijection from $\mathbb{R}^2$ to the open lower hemisphere $\mathbb{S}^2_- = \{y \in \mathbb{S}^2 \mid \langle y \mid e_z \rangle < 0\}$ and that $R(e_z, \mathbb{S}^2_-) = \mathbb{S}^2 \setminus \{e_z\}$. Thus $F$ is a diffeomorphism between $\mathbb{R}^2$ and $\mathbb{S}^2 \setminus \{e_z\}$. Then the reflection of an incident ray with origin $x \in \mathbb{R}^2$ with direction $e_z$ is given by $F(\nabla\varphi(x))$. We deduce that for a function $\varphi$, the reflected measure is $(F \circ \nabla\varphi)_{\#}\mu$. Since we want to prescribe this measure, the problem can then be stated as

$$\text{Find } \varphi : \Omega \to \mathbb{R} \text{ differentiable such that } (F \circ \nabla\varphi)_{\#}\mu = \nu.$$

With this formulation, the problem is posed on the target domain $\mathbb{S}^2$. Since $F$ is a bijection from $\mathbb{R}^2$ to $\mathbb{S}^2 \setminus \{e_z\}$, applying the inverse transformation $F^{-1}$, we can look at the problem on the source domain $\Omega$:

$$\text{Find } \varphi : \Omega \to \mathbb{R} \text{ differentiable such that } \nabla\varphi_{\#}\mu = (F^{-1})_{\#}\nu. \tag{2.2.1}$$

We now introduce the notion of Brenier solution that appears in optimal transport, see [San15].

**Definition 2.1** (Brenier solution)
*We say that a differentiable function $\varphi : \Omega \to \mathbb{R}$ is a solution of the mirror design problem in the* Brenier *sense if it is convex and satisfies Equation* (2.2.1).

Remark 32. *The notion of Brenier solution can also be defined in the semi-discrete setting. In this case, the function $\varphi$ will be differentiable $\mu$-almost everywhere.*

The rest of the section is dedicated to the proof of the following theorem that states the

existence of solutions to this problem. We will denote by $\mathrm{supp}(\mu)$ the support of a measure $\mu$ and $M_\mu(\Omega)$ the set of functions whose mean value is zero with respect to a measure $\mu$ on $\Omega$:

$$M_\mu(\Omega) = \left\{ \varphi : \Omega \to \mathbb{R} \mid \int_\Omega \varphi(x) \mathrm{d}\mu(x) = 0 \right\}.$$

**Theorem 33.** *Let $\mu$ be a compactly supported, absolutely continuous measure with a bounded density $\rho$ and let $\nu$ be a probability measure such that $\mathrm{supp}(\nu) \subset \mathbb{S}^2 \setminus \{e_z\}$ is compact. We also suppose that $\mu$ satisfies a Poincaré inequality*

$$\forall \varphi \in M_\mu(\Omega),\ \int_\Omega |\varphi(x)|^2 \mathrm{d}\mu(x) \leqslant \int_\Omega |\nabla \varphi(x)|^2 \mathrm{d}\mu(x). \tag{Poincaré}$$

*Then there exists a Brenier solution to the mirror design problem which is unique up to the addition of a constant.*

This theorem is already known, see [Pri+13]. We include here a proof to show how semi-discrete optimal transport can be used as a tool to prove the existence of weak solutions of such equations. As said in the introduction, we will start by looking at the case where $\nu$ is finitely supported. The result is the following proposition.

Proposition 34. *If $\nu$ is a finitely supported probability measure on $Y = \{y_1, \dots, y_N\}$, then Equation (2.2.1) admits a Brenier solution $\varphi$ differentiable almost everywhere which can be written as*

$$\forall x \in \Omega,\ \varphi(x) = \max_{1 \leqslant i \leqslant N} (\langle x \mid p_i \rangle - \psi_i)$$

*where $p_i$ is defined by the relation $F(p_i) = y_i$ and $\psi$ is the Kantorovich potential solution of the semi-discrete optimal transport problem between $\mu$ and $\sum_{i=1}^N \nu_i \delta_{p_i}$ for the quadratic cost. Moreover this solution is unique up to the addition of a constant.*

*Proof.* We first take $p_i$ such that $F(p_i) = y_i$ and define for $\psi \in \mathbb{R}^N$ the function $\varphi_\psi(x) = \max_{1 \leqslant i \leqslant N} (\langle x \mid p_i \rangle - \psi_i)$, then it is easy to see that $\varphi_\psi$ is convex. We now show that there exists a vector $\psi$ such that $\varphi_\psi$ is a solution of Equation (2.2.1).

We consider the semi-discrete optimal transport problem between $\mu$ and $\nu = \sum_{i=1}^N \nu_i \delta_{p_i}$ for the quadratic cost. According to Theorem 5, there exist Kantorovich potentials $\varphi^*$ and $\psi^*$ solution to this problem. We put $\varphi = \varphi_{\psi^*}$ and show that $\varphi$ is a solution of Equation (2.2.1). First, it is easy to see that $\nabla \varphi(x) = p_i$ if and only if $x \in \mathrm{Lag}_i(\psi)$ where $\mathrm{Lag}_i(\psi)$ is the Laguerre

cell of $p_i$, see Definition 1.4. Then, taking $A = \{y_{i_0}, \dots, y_{i_k}\} \subset Y$ we have:

$$\begin{aligned}(F \circ \nabla\varphi)_{\#}\mu(A) &= \mu(\nabla\varphi^{-1}(F^{-1}(A))) \\ &= \mu(\nabla\varphi^{-1}(\{p_{i_0}, \dots, p_{i_k}\})) \\ &= \mu\left(\bigcup_{j=1}^{k} \mathrm{Lag}_{i_j}(\psi)\right) \\ &= \sum_{j=1}^{k} \nu_{i_j} = \nu(A).\end{aligned}$$

Thus $(F \circ \nabla\varphi)_{\#}\mu = \nu$ and $\varphi$ is a solution of Equation (2.2.1). An illustration of the graph of this solution be found in Figure 2.

We now prove the uniqueness (up to the addition of a constant) of the solution. We take two solutions $\varphi_1$ and $\varphi_2$. They are both solutions of Equation (2.2.1) meaning that $\nabla\varphi_1(\Omega) \subset \{p_1, \dots, p_N\}$ as well as $\nabla\varphi_2(\Omega)$. Since they are also convex and differentiable $\mu$-almost everywhere, they are both of the form $\varphi_j : x \mapsto \max_{1 \leqslant i \leqslant N}(\langle x \mid p_i\rangle - \psi_i^j)$, where $j \in \{1, 2\}$ and $\psi^j \in \mathbb{R}^N$. Then the application $\nabla\varphi_j : x \in \Omega \mapsto p_i$ such that $x \in \mathrm{Lag}_i(\psi^j)$ is an optimal transport map meaning that $\nabla\varphi_1$ and $\nabla\varphi_2$ are both *optimal* transport maps for the quadratic cost. Brenier's theorem affirms that, for the quadratic cost, the transport plan is unique thus $\nabla\varphi_1 = \nabla\varphi_2$ $\mu$-almost everywhere. Translating $\varphi_1$ and $\varphi_2$, we can suppose that $\varphi_1, \varphi_2 \in M_\mu(\Omega)$. Thus, since $\varphi_1 - \varphi_2 \in M_\mu(\Omega)$, the Poincaré inequality gives:

$$\int_\Omega \|\varphi_1(x) - \varphi_2(x)\|^2 \,\mathrm{d}\mu(x) \leqslant \int_\Omega \|\nabla\varphi_1(x) - \nabla\varphi_2(x)\|^2 \,\mathrm{d}\mu(x) = 0.$$

We deduce that $\varphi_1 - \varphi_2$ is a constant function and the solution is unique up to the addition of a constant.

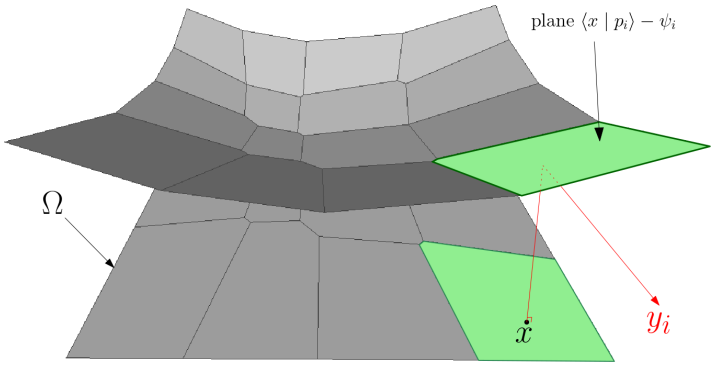


Figure 2 – **Illustration for the construction of a solution to the** (2.2.1) **problem in the semi-discrete setting.** The solution is a convex function constructed as the lower envelope of affine functions.

□

We will now show the following convergence result.

PROPOSITION 35. *We let $\mu$ be an absolutely continuous probability measure on a bounded domain $\Omega \subset \mathbb{R}^2$ with a bounded density $\rho$. We also suppose that $\mu$ satisfies the Poincaré inequality* (Poincaré).

*Let $(\nu^k)$ be a sequence of finitely supported probability measures weakly converging towards a probability measure $\nu$. We suppose there exists a compact set $K \subset \mathbb{S}^2 \setminus \{e_z\}$ such that* $\mathrm{supp}(\nu^k) \subset K$ *and* $\mathrm{supp}(\nu) \subset K$.

*Let $\varphi^k \in M_\mu(\Omega)$ denote the solution of* (2.2.1) *for the target measure $\nu^k$. Then the sequence $(\varphi^k)$ converges uniformly towards a function $\varphi \in M_\mu(\Omega)$ which is the unique solution of* (2.2.1) *for the target measure $\nu$.*

*Proof.* For any $k \geqslant 0$, by Proposition 34, there exists a unique solution $\varphi^k \in M_\mu(\Omega)$ a solution of (2.2.1) for the discrete target measure $\nu^k$.

The assumption on the support of the target measures $\nu^k$ implies that $\mathrm{supp}((F^{-1})_\# \nu^k) \subset F^{-1}(K)$ is compact since $F^{-1}$ is continuous. Thus the image of $\nabla\varphi^k$ is included in the compact set $F^{-1}(K)$. This means that $\nabla\varphi^k$ is uniformly bounded so that $\varphi^k$ is $L$-Lipschtiz (with $L$ independent of $k$). Furthermore, combined with the fact that $\Omega$ is bounded and that the mean value of $\varphi^k$ is 0 (since $\varphi^k \in M_\mu(\Omega)$), we get that $\varphi^k$ is uniformly bounded on $\Omega$. The Arzelà-Ascoli theorem implies that there exists a subsequence $(\varphi^{\sigma_k})_k$ that uniformly converges towards a function $\overline{\varphi} \in M_\mu(\Omega)$.

Then, since $\rho$ is bounded on $\Omega$, we have

$$\|\nabla\varphi^{\sigma_k} - \nabla\overline{\varphi}\|_{L^1(\mu)} = \int_\Omega \rho(x)(\nabla\varphi^{\sigma_k}(x) - \nabla\overline{\varphi}(x))\mathrm{d}x \leqslant \|\rho\|_\infty \|\nabla\varphi^{\sigma_k} - \nabla\overline{\varphi}\|_{L^1(\Omega)}.$$

Using Theorem 3.5 of [CCSM10], we have that $(\nabla\varphi^{\sigma_k})_k$ converges towards $\nabla\overline{\varphi}$ for the $L^1(\Omega)$ norm. Thus, using the previous inequality, we have that $(\nabla\varphi^{\sigma_k})_k$ also converges towards $\nabla\overline{\varphi}$ for the $L^1(\mu)$ norm.

Now, if we denote $\gamma^k = (\nabla\varphi^{\sigma_k}, \nabla\overline{\varphi})_\#\mu$, then a simple calculation shows that it is a transport plan between $\nu^{\sigma_k}$ and $\nu$, thus

$$\begin{aligned}\mathrm{W}_1(\nabla\varphi^{\sigma_k}_\#\mu, \nabla\overline{\varphi}_\#\mu) &\leqslant \int_{\Omega\times\Omega} \|x-y\|\,\mathrm{d}\gamma^k(x,y) = \int_\Omega \|\nabla\varphi^{\sigma_k}(x) - \nabla\overline{\varphi}(x)\|\,\mathrm{d}\mu(x)\\ &\leqslant \|\nabla\varphi^{\sigma_k} - \nabla\overline{\varphi}\|_{L^1(\mu)}\end{aligned}$$

We deduce that the sequence $(\nabla\varphi^{\sigma_k}_\#\mu)_k$ converges towards $\nabla\overline{\varphi}_\#\mu$ for the $\mathrm{W}_1$ norm. Since $(\nabla\varphi^{\sigma_k})_\#\mu = (F^{-1})_\#\nu^{\sigma_k}$, then $((F^{-1})_\#\nu^{\sigma_k})_k$ converges towards $\nabla\overline{\varphi}_\#\mu$ for the $\mathrm{W}_1$ norm. Then, Theorem 5.9 of [San15] ensures that $((F^{-1})_\#\nu^{\sigma_k})_k$ also weakly converges towards $\nabla\overline{\varphi}_\#\mu$. Using the continuity of $F^{-1}$, we also have that $((F^{-1})_\#\nu^{\sigma_k})_k$ weakly converges towards $F^{-1}_\#\nu$. Since the limit of a sequence is unique, we get $\nabla\overline{\varphi}_\#\mu = F^{-1}_\#\nu$ and $\overline{\varphi}$ is a solution of (2.2.1).

We now show that all the converging subsequences of $(\varphi^k)$ are converging towards the same limit. Let us take two subsequences converging towards two functions $\varphi_1$ and $\varphi_2$ in $M_\mu(\Omega)$. $\varphi_1$ and $\varphi_2$ are convex since they are the limits of convex functions. The first part of the proof shows that they both solve the same optimal transport problem between $\mu$ and $(F^{-1})_\#\nu$ for the quadratic cost. Using Brenier's theorem, see Theorem 7, we get that $\nabla\varphi_1 = \nabla\varphi_2$ $\mu$-almost everywhere. Furthermore $\varphi_1 - \varphi_2 \in M_\mu(\Omega)$, thus the Poincaré inequality gives that $\varphi_1 = \varphi_2$ almost everywhere.

We deduce that any converging subsequence of $(\varphi^k)$ converges towards the same limit. This implies that $(\varphi_k)$ converges towards the same limit $\overline{\varphi}$, the unique solution of the limit problem. □

*Proof of Theorem 33.* The last thing we need to do is to describe how to construct a sequence $(\nu^k)$ of finitely supported probability measures that weakly converges towards a continuous probability measure $\nu$ supported on a compact set $K \subset \mathbb{S}^2 \setminus \{e_z\}$. The construction is heavily inspired by the one done in [CO08].

We let $R > 0$ and take $k \in \mathbb{N}$ such that $k \geqslant 2$. Let $V_i^k$ for $i = 1, \dots, k$ be a partition of $K$ into $k$ subsets such that $\operatorname{diam}(V_i^k) < \frac{1}{R}$ and $\nu(V_i^k) > 0$. One can for instance choose a Voronoi diagram on $K$ for a sufficiently large number of well chosen sites. Take a point $y_i^k$ per subset $V_i^k$ and pose $\nu_i^k = \int_{V_i^k} d\nu(y)$. We now define the measure $\nu^k$ on $K$ by

$$\nu^k = \sum_{i=1}^{k} \nu_i^k \delta_{y_i^k}.$$

The sequence $(\nu^k)_k$ weakly converges towards $\nu$ as $k$ goes to infinity and for every $k$, $\operatorname{supp}(\nu^k) \subset K$ by construction. We can then apply Proposition 35 to the sequence $(\nu^k)_k$ and get the existence of a function $\varphi \in M_\mu(\Omega)$ solution of (2.2.1) for the target measure $\nu$. □

Similar (but more complex) proofs can be done to show the existence of weak solutions for other non-imaging problems in optics.

## 2.3 Light Energy Conservation equation

We present in this section several mirror and lens design problems arising in non-imaging optics. Let us note that we do not take into account multiple reflections or refractions as well as the Fresnel coefficient. The setting we will place ourselves in is the following: we are given an *ideal* light source as well as a desired illumination "at infinity" after reflection or refraction, called the target. The goal is to design the geometry of a mirror or lens which transports the energy emitted by the source onto the target. Even though the problems we consider are quite different from one another, they share a common structure in that they all correspond to a so-called generalized Monge-Ampère equation, whose discrete version is given by Equation

(DMA) that was studied in Chapter 1. Let us note that our method is heavily inspired by the supporting paraboloids method developed by [CKO99a].

In the following, the source illumination is denoted by $\rho$ with support $\Omega$ where $\Omega$ is a subset of $\mathbb{R}^2 \times \{0\}$ or $\mathbb{S}^2$ and the desired target illumination is described by a set of intensity values $\sigma = (\sigma_i)_{1 \leqslant i \leqslant N}$ supported on a finite set of directions $Y = \{y_1, \cdots, y_N\} \subset \mathbb{S}^2$. We also recall the following notions explained in Chapter 1:

- *Laguerre cell* of a point $y_i \in \mathbb{R}^3$ for a vector of weights $\psi \in \mathbb{R}^N$:

$$\mathrm{Lag}_i(\psi) = \{x \in \Omega \mid \forall j,\ c(x, y_i) + \psi_i \leqslant c(x, y_j) + \psi_j\}.$$

- Function $G$: $G(\psi) = (G_i(\psi))_{1 \leqslant i \leqslant N}$ where $G_i(\psi) = \rho(\mathrm{Lag}_i(\psi))$.
- *Discrete Monge-Ampère equation (DMA)*:

$$\text{Find } \psi \in \mathbb{R}^N \text{ such that } \forall i \in \{1, \ldots, N\},\ G_i(\psi) = \nu_i.$$

### 2.3.1 Mirror design

#### Convex mirror for a collimated light source

In this first problem, the light source is collimated meaning that it emits parallel vertical rays, and the source can be encoded by a light intensity function $\rho$ over a 2D domain. For simplicity, we assume that the domain is included in $\mathbb{R}^2 \times \{0\} \subset \mathbb{R}^3$ and that all the rays are parallel to the $z$ direction and directed upwards. The problem is to find the surface $\mathcal{R}$ of a mirror that sends the source intensity $\rho$ to the target intensity $\sigma$, see figures 3 (top left) and 4. This problem corresponds to a Monge-Ampère equation in the 2D plane, which corresponds to the quadratic optimal transport problem [Pri+13]. The following proposition describes how we can find such mirror $\mathcal{R}$.

Proposition 36. *For a collimated light source $\rho$ supported on $\mathbb{R}^2 \times \{0\}$, a target illumination at infinity $\sigma = \sum_{i=1}^N \sigma_i \delta_{y_i}$, a* convex *mirror $\mathcal{R}$ reflecting $\rho$ into $\nu$ can be parametrized by*

$$\mathcal{R}_\psi : x \in \mathbb{R}^2 \times \{0\} \ \mapsto (x, \max_{1 \leqslant i \leqslant N} (\langle x \mid p_i \rangle - \psi_i))$$

*where $p_i \in \mathbb{R}^2$, $\psi \in \mathbb{R}^N$ is a vector of elevations solving a discrete Monge-Ampère equation* (DMA) *for the cost $c(x, y) = -\langle x \mid y \rangle$.*

*Proof.* Since the number of reflected directions $Y$ is finite, the mirror surface $\mathcal{R}$ is composed of a finite number of planar facets, as illustrated in Figure 4. We define $\mathcal{R}_\psi$ as the graph of a convex function of the form $x \mapsto (\max_i \langle x \mid p_i \rangle - \psi_i)$; for every $i \in \{1, \cdots, N\}$, $p_i$ is the orthogonal projection of a unit normal of the plane (called *slope* in the following) that reflects according to Snell's law the vertical ray $(0, 0, 1)$ towards the direction $y_i$ (see Section 3.2 for the full expression) and $\psi_i$ is a real number that encodes the elevation of the supporting plane with slope $p_i$. We denote by $\psi := (\psi_i)_{1 \leqslant i \leqslant N}$ the set of elevations. The Visibility cell $\mathrm{V}_i(\psi)$ of

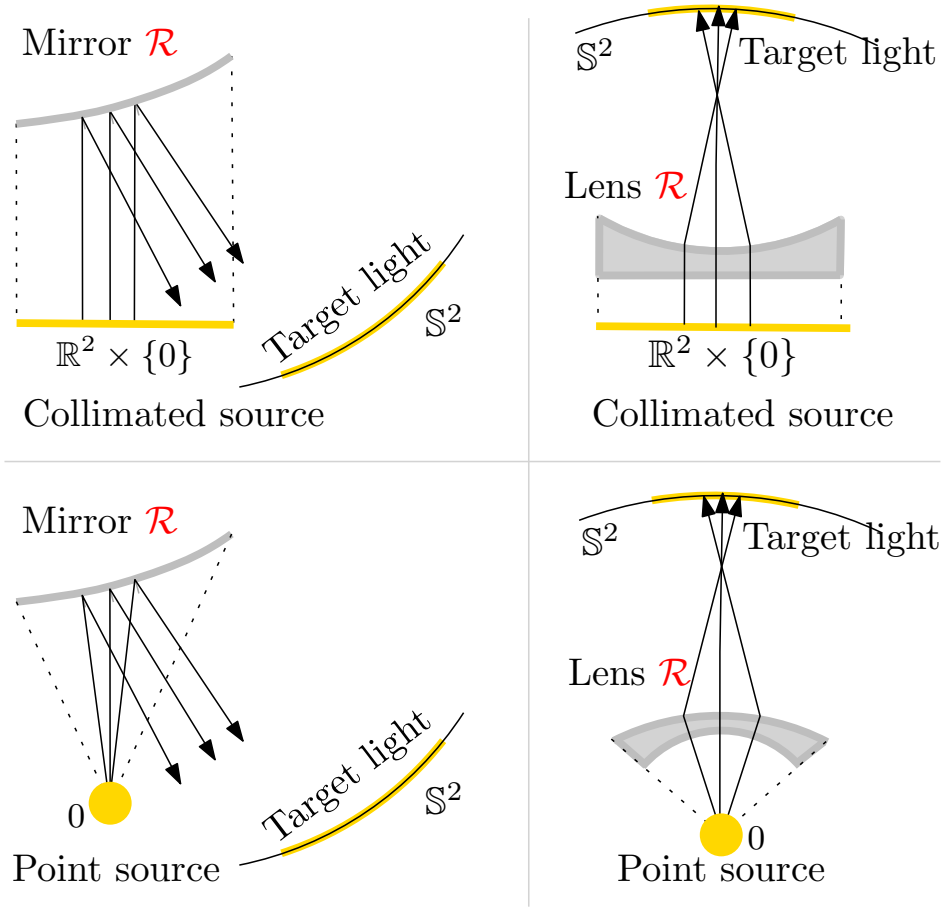


Figure 3 – **Four inverse problems arising in non-imaging optics.** In each case, the goal is to build the surface $\mathcal{R}$ of a mirror or a lens. Remark that for each problem, we provide two solutions (for instance, we can have convex and concave surfaces when the light source is collimated). *Top/Bottom:* Collimated light sources/Point light sources. *Left/Right:* Mirror/Lens design.

$y_i$ is defined as the set of locations $x \in \mathbb{R}^2 \times \{0\}$ whose rays are reflected towards the direction $y_i$, meaning that the vertical rays hit the $i$th facet of $\mathcal{R}_\psi$. Given the definition of $\mathcal{R}_\psi$, we have

$$\mathrm{V}_i(\psi) = \{x \in \mathbb{R}^2 \times \{0\} \mid \ \forall j, -\langle x \mid p_i \rangle + \psi_i \leqslant -\langle x \mid p_j \rangle + \psi_j\}.$$

Remark 37. *One can see that the Visibility cell corresponds to a Laguerre cell for the cost function* $c(x, y) = -\langle x \mid y \rangle$ *defined on* $(\mathbb{R}^2 \times \{0\}) \times \mathbb{S}^2$.

By construction, the vertical ray emanating from the point $x \in \mathrm{V}_i(\psi)$ touches the mirror surface $\mathcal{R}$ at an altitude $\langle x \mid p_i \rangle - \psi_i$ for a given $i$ and is reflected to the direction $y_i$, and therefore the amount of light reflected towards the direction $y_i$ equals the integral of $\rho$ over $\mathrm{V}_i(\psi)$. Note that one also has $\nabla \mathcal{R}_\psi(x) = p_i$ if $x \in \mathrm{V}_i(\psi)$. The *Collimated Source Mirror*

*problem* (CS/Mirror) then amounts to finding $\psi \in \mathbb{R}^N$ such that

$$\forall i \in \{1, \cdots, n\} \quad \int_{\mathrm{V}_i(\psi)} \rho(x)\mathrm{d}x = \sigma_i. \tag{LEC}$$

By construction, a solution to Equation (LEC) provides a parameterization $\mathcal{R}_\psi$ of a convex mirror that sends the collimated light source $\rho$ to the discrete target $\sigma$:

$$\mathcal{R}_\psi : \ x \in \mathbb{R}^2 \mapsto \ (x, \max_{1 \leqslant i \leqslant N} \langle x \mid p_i \rangle - \psi_i),$$

where $\mathbb{R}^2 \times \{0\}$ and $\mathbb{R}^2$ are identified. Notice that since the mirror is a graph over $\mathbb{R}^2 \times \{0\}$, the vectors $y_i$ cannot be upward vertical. □

Remark 38. *In practice we assume that* $y_i \in \mathbb{S}^2_- := \{y \in \mathbb{S}^2, \ \langle y \mid e_z \rangle \leqslant 0\}$ *and localize the position of the mirror by considering it only above the support* $X_\rho := \Omega = \{x \in \mathbb{R}^2 \times \{0\}, \ \rho(x) \neq 0\}$ *of* $\rho$.

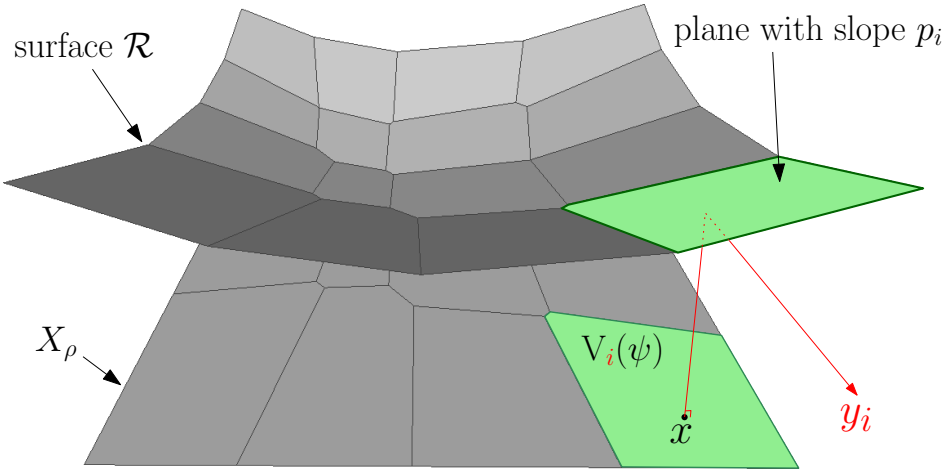


Figure 4 – **Convex Mirror for a collimated light source** (when $N = 16$). The mirror surface $\mathcal{R}$ is the graph of a convex piecewise affine functions. The support $\Omega$ of $\rho$ is decomposed into Visibility cells $(\mathrm{V}_i(\psi))_{1 \leqslant i \leqslant N}$. Every vertical ray above a point $x \in X_\rho$ belongs to a cell $\mathrm{V}_i(\psi)$, touches a plane with slope $p_i$ and is reflected to the direction $y_i$.

**Concave mirror.** The same approach also allows the construction of concave mirrors, using a concave function of the form $x \mapsto \min_i(\langle x \mid p_i \rangle + \psi_i)$. This amounts to replacing the Visibility cells by

$$\mathrm{V}_i(\psi) = \{x \in \mathbb{R}^2 \times \{0\} \mid \ \forall j, \ \langle x \mid p_i \rangle + \psi_i \leqslant \langle x \mid p_j \rangle + \psi_j\}$$

which is a Laguerre diagram for the cost $c(x, y) = \langle x \mid y \rangle$. In that case, a solution to Equation (LEC) provides a parametrization of a concave mirror $\mathcal{R}_\psi(x) = (x, \min_i \langle x \mid p_i \rangle + \psi_i)$ that sends the collimated light source $\rho$ to the discrete target $\sigma$.

### Concave mirror for a point source

In the second mirror design problem, all the rays are emitted from a single point in space, located at the origin, and the light source is described by an intensity function $\rho$ supported on the unit sphere $\mathbb{S}^2$. The problem we consider is to find the surface $\mathcal{R}$ of a mirror that sends the light intensity $\rho$ to the light intensity $\sigma$ (Fig. 3, bottom left). The following proposition gives a parametrization of such mirror $\mathcal{R}$.

PROPOSITION 39. *For a point light source $\rho$ supported on $\mathbb{S}^2$, a target illumination at infinity $\sigma = \sum_{i=1}^N \sigma_i \delta_{y_i}$, a* convex *mirror $\mathcal{R}$ reflecting $\rho$ into $\nu$ can be parametrized by*

$$\mathcal{R}_\psi : \ x \in \mathbb{S}^2 \ \mapsto \ \min_{1 \leqslant i \leqslant N} \frac{\psi_i}{1 - \langle x \mid y_i \rangle} \ x$$

*where $\psi \in \mathbb{R}_+^N$ is a vector of focal distances solving a discrete Monge-Ampère equation* (DMA) *for the cost $c(x, y) = -\ln(1 - \langle x \mid y \rangle)$.*

*Proof.* Following [CO08], we build a concave surface $\mathcal{R}$ that is composed of pieces of confocal paraboloids. More precisely, we denote by $P(y_i, \psi_i)$ the solid (i.e filled) paraboloid whose focal point is at the origin with focal distance $\psi_i$ and with direction $y_i$. We define the surface $\mathcal{R}_\psi$ as the boundary of the intersection of the solid paraboloids, namely $\mathcal{R}_\psi = \partial \left( \cap_i P(y_i, \psi_i) \right)$. The Visibility cell $\mathrm{V}_i(\psi)$ is the set of ray directions $x \in \mathbb{S}^2$ emanating from the light source that are reflected in the direction $y_i$. Since each paraboloid $\partial P(y_i, \psi_i)$ is parameterized over the sphere by $x \mapsto \psi_i x / (1 - \langle x \mid y_i \rangle)$ for $\psi_i > 0$, one has

$$\mathrm{V}_i(\psi) = \left\{ x \in \mathbb{S}^2 \mid \ \forall j, \frac{\psi_i}{1 - \langle x \mid y_i \rangle} \leqslant \frac{\psi_j}{1 - \langle x \mid y_j \rangle} \right\}.$$

The *Point Source Mirror problem* (PS/Mirror) then amounts to finding the vector $\psi$ that satisfies the Light Energy Conservation Equation (LEC). The mirror surface is then parameterized by

$$\mathcal{R}_\psi : \ x \in \mathbb{S}^2 \ \mapsto \ \min_i \frac{\psi_i}{1 - \langle x \mid y_i \rangle} \ x.$$

In practice, we assume that the target $Y$ is included in $\mathbb{S}^2_-$, that the support $X_\rho$ of $\rho$ is included $\mathbb{S}^2_+ := \{ y \in \mathbb{S}^2, \ \langle y \mid e_z \rangle \geqslant 0 \}$, and that the mirror is parameterized over $X_\rho$. □

REMARK 40. *We can rewrite $\mathrm{V}_i(\psi)$ as*

$$\mathrm{V}_i(\psi) = \left\{ x \in \mathbb{S}^2 \mid \ \forall j, \ \ln(\psi_i) - \ln(1 - \langle x \mid y_i \rangle) \leqslant \ln(\psi_j) - \ln(1 - \langle x \mid y_j \rangle) \right\}.$$

*This implies that the Visibility cell $\mathrm{V}_i(\psi)$ can be seen as the Laguerre cell $\mathrm{Lag}_i(\widetilde{\psi})$ for $\widetilde{\psi} = (\ln(\psi_i))_{1 \leqslant i \leqslant N}$ and the cost function $c(x, y) = -\ln(1 - \langle x \mid y \rangle)$ defined on $\mathbb{S}^2 \times \mathbb{S}^2$.*

REMARK 41. *One can also define the mirror surface as the boundary of the union (instead of*

*the intersection) of a family of solid paraboloids. Then, the Visibility cells become*

$$\mathrm{V}_i(\psi) = \left\{ x \in \mathbb{S}^2 \mid \ \forall j, \frac{\psi_i}{1 - \langle x \mid y_i \rangle} \geqslant \frac{\psi_j}{1 - \langle x \mid y_j \rangle} \right\}$$

*and a solution to Equation (LEC) provides a parameterization* $\mathcal{R}_\psi(x) = x \max_i \frac{\psi_i}{1-\langle x|y_i\rangle}$ *of the mirror surface. It is a Laguerre cell for* $\widetilde{\psi} = (\ln(\psi_i))_{1 \leqslant i \leqslant N}$ *and* $c(x, y) = \ln(1 - \langle x \mid y \rangle)$.

### 2.3.2 Lens design

In this section, the goal is to design lenses that refracts a given light source intensity to a desired one. Similarly to mirror design, we consider collimated and point light sources. We denote by $n_1$ the refractive index of the lens, by $n_2$ the ambient space refractive index, see Figure 5 and by $\kappa = \frac{n_1}{n_2}$ the ratio of the two indices.

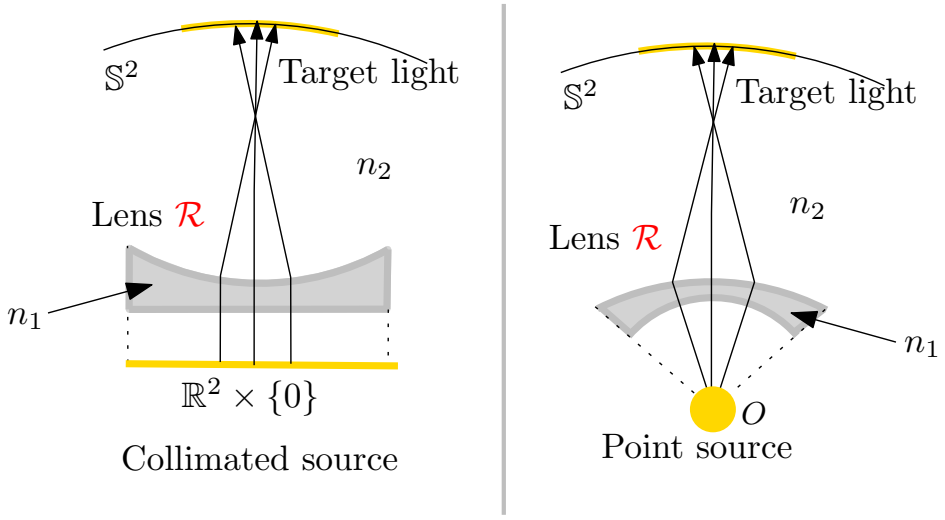


Figure 5 – **Lens design.** $n_1$ is the refractive index of the lens, $n_2$ the one of the ambient medium and $\kappa = \frac{n_1}{n_2}$. Left: collimated source; Right: point light source.

#### Concave lens for a collimated light source

We consider here a collimated light source encoded by a function $\rho$ supported on a 2D domain. The goal is to find the surface of a lens that sends $\rho$ to $\sigma$, see the top right diagram in Figure 3. We assume that the rays emitted by the source are vertical and that the bottom of the lens is flat and orthogonal to the vertical axis. There is no refraction angle when the rays enter the lens, and we therefore only need to build the top part of the lens.

By a simple change of variable, we show that this problem is equivalent to (CS/Mirror). More precisely, for every $y_i \in Y$, we now define $p_i$ to be the slope of a plane that *refracts* the vertical ray $(0, 0, 1)$ to the direction $y_i$ (see Section 3.2 for a detailed expression). We define $\mathcal{R}$

as the graph of a convex function of the form $x \mapsto (\max_i \langle x \mid p_i \rangle - \psi_i)$, where $\psi = (\psi_i)_{1 \leqslant i \leqslant N}$ is the set of elevations. We define the Visibility cell $\mathrm{V}_i(\psi)$ to be the set of points $x \in \mathbb{S}^2$ that are *refracted* to the direction $y_i$

$$\mathrm{V}_i(\psi) = \{x \in \mathbb{S}^2 \mid \forall j,\ -\langle x \mid p_i \rangle + \psi_i \leqslant -\langle x \mid p_j \rangle + \psi_j\}.$$

The *Collimated Source Lens problem* (CS/Lens) then amounts to finding weights $(\psi_i)_{1 \leqslant i \leqslant N}$ that satisfy (LEC). In that case the lens surface is parameterized by

$$\mathcal{R}_\psi :\ x \in \mathbb{S}^2 \ \mapsto\ (x, \max_{1 \leqslant i \leqslant N} \langle x \mid p_i \rangle - \psi_i).$$

In practice, we choose the directions $y_i$ in $\mathbb{S}^2_+$ and the mirror to be parameterized over the support $X_\rho$ of $\rho$.

**Convex lens.** Remark that we can also build convex lenses by considering parameterizations with concave functions of the form $x \mapsto \min_{1 \leqslant i \leqslant N}(\langle x \mid p_i \rangle + \psi_i)$. Figure 6 illustrates a concave and a convex solution to the same non-imaging optics problem.

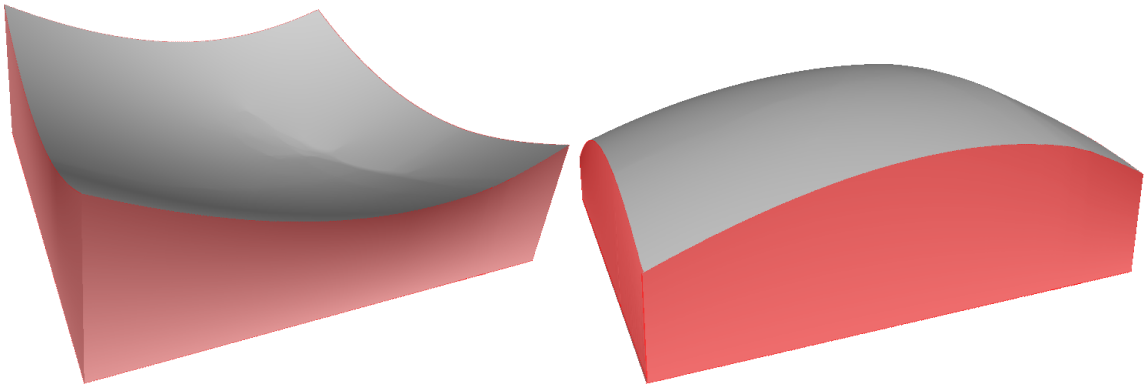

Figure 6 – **Concave (left) and convex (right) lens** for a uniform collimated light source and the same target.

### Convex lens for a point source

We now consider the same problem, except that we replace the collimated light source by a point source one. As in the collimated setting, we fix the bottom part of the lens. We choose a piece of sphere centered at the source, so that the rays are not deviated. The next proposition gives a parametrization of the outer part of the lens.

Proposition 42. *For a point light source $\rho$ supported on $\mathbb{S}^2$, a target illumination at infinity $\sigma = \sum_{i=1}^N \sigma_i \delta_{y_i}$, a* convex *lens $\mathcal{R}$ reflecting $\rho$ into $\nu$ can be parametrized by*

$$\mathcal{R}_\psi :\ x \in \mathbb{S}^2 \ \mapsto\ \min_{1 \leqslant i \leqslant N} \frac{\psi_i}{1 - \kappa' \langle x \mid y_i \rangle}\ x$$

*where $\psi \in \mathbb{R}_+^N$ is a vector solving an optimal transport problem for the cost $c(x,y) = -\ln(1 - \kappa'\langle x \mid y\rangle)$.*

*Proof.* As in [GH09], the lens is composed of pieces of ellipsoids of constant eccentricities $\kappa' := \frac{1}{\kappa} < 1$, where $\kappa$ is the ratio of the indices of refraction. Each ellipsoid $\partial E(y_i, \psi_i)$ can be parameterized over the sphere by $x \mapsto x\frac{\psi_i}{1-\kappa'\langle x|y_i\rangle}$ for $\psi_i > 0$. The Visibility cell of $y_i$ is then

$$\mathrm{V}_i(\psi) = \left\{ x \in \mathbb{S}^2 \mid \ \forall j, \frac{\psi_i}{1 - \kappa'\langle x \mid y_i\rangle} \leqslant \frac{\psi_j}{1 - \kappa'\langle x \mid y_j\rangle} \right\}.$$

which corresponds to a Laguerre cell for the cost function $c(x,y) = -\ln(1 - \kappa'\langle x \mid y\rangle)$ on $\mathbb{S}^2 \times \mathbb{S}^2$. The *Point Source Lens problem* (PS/Lens) then amounts to finding weights $(\psi_i)_{1\leqslant i\leqslant N}$ that satisfy (LEC). Remark that the top surface of the lens is then parameterized by

$$\mathcal{R}_\psi : \ x \in \mathbb{S}^2 \ \mapsto \ \min_{1\leqslant i\leqslant N} \frac{\psi_i}{1 - \kappa'\langle x \mid y_i\rangle}\, x.$$

In practice, we choose the set of directions $y_i$ to belong to $\mathbb{S}^2_+$ and the lens to be parameterized over the support $X_\rho \subset \mathbb{S}^2_+$ of $\rho$. □

Remark 43. *One can also choose to define the lens surface as the boundary of the union (instead of the intersection) of a family of solid ellipsoids. In that case, the Visibility cells are given by*

$$\mathrm{V}_i(\psi) = \left\{ x \in \mathbb{S}^2 \mid \ \forall j, \frac{\psi_i}{1 - \kappa'\langle x \mid y_i\rangle} \geqslant \frac{\psi_j}{1 - \kappa'\langle x \mid y_j\rangle} \right\}$$

*and a solution to Equation (LEC) provides a parameterization $\mathcal{R}_\psi(x) = x \max_i \frac{\psi_i}{1-\kappa'\langle x|y_i\rangle}$ of the lens surface. This corresponds to a Laguerre diagram on the sphere for $c(x,y) = \ln(1-\kappa'\langle x \mid y\rangle)$.*

### 2.3.3 Generic formulation

Let $X$ be a domain of either the plane $\mathbb{R}^2 \times \{0\}$ or the unit sphere $\mathbb{S}^2$, $\rho : X \to \mathbb{R}$ a probability density and $Y = \{y_1, \cdots, y_N\} \subset \mathbb{S}^2$ be a set of $N$ points. We define the function $G : \mathbb{R}^N \to \mathbb{R}^N$ by

$$G_i(\psi) = \int_{\mathrm{V}_i(\psi)} \rho(x)\mathrm{d}x$$

where $G(\psi) = (G_i(\psi))_{1\leqslant i\leqslant N}$ and $\mathrm{V}_i(\psi) \subset X$ is the Visibility cell of $y_i$, whose definition depends on the non-imaging problem. Using this notation, Equation (LEC) can be rephrased as finding weights $\psi = (\psi_i)_{1\leqslant i\leqslant N}$ such that

$$\forall i \in \{1, \cdots, N\}, \quad G_i(\psi) = \sigma_i. \tag{2.3.2}$$

Remark 44. *Many other problems arising in non-imaging optics amount to solving Equation (2.3.2). For example, the design of a lens that refracts a point light source to a desired near-field target can also be modeled by a Monge-Ampère equation that has the same struc-*

*ture [GH09]. In this case, the Visibility diagram correspond to the radial projection onto the sphere of pieces of confocal ellipsoids with non constant eccentricities and is not associated to an optimal transport problem.*

Let us also remark that this equation corresponds to a discrete Monge-Ampère equation (DMA) as presented in Chapter 1 meaning that we can use the damped Newton's algorithm, see Algorithm 4 to solve this equation. In the next chapter, we will discuss the specifics of this algorithm for non-imaging problems. Before doing that, we will take a closer look to the Ma-Trudinger-Wang condition and its relation to the regularity of the solutions of Monge-Ampère equations.

## 2.4 Ma-Trudinger-Wang condition for the refractor problem

This section deals with the so-called *Ma-Trudinger-Wang* (MTW) condition [MTW05] that appears in the regularity theory of solutions of optimal transport problems. This condition appears to be important when studying the convergence of the damped Newton's method presented in Chapter 1. Indeed, it has been shown in [Loe09] that the MTW condition allows to have a bound on the complexity of the Laguerre diagram and to show that the Laguerre cells are connected. This is used in [KMT16] to show that the algorithm has a superlinear convergence rate.

Here, we study this condition for cost functions related to the so-called *refractor problem* or *lens design problem* presented in the previous section. It has been shown in [GH09] that one can solve this problem by parametrizing the lens as an intersection (or union) of *ellipsoids*. As in the previous section, the ratio of the indices of refraction is denoted by $\kappa = \frac{n_1}{n_2}$ where $n_1$ is the refraction index of the interior medium and $n_2$ the ambient medium. We put $\kappa' = \frac{1}{\kappa}$. When $\kappa' < 1$, it is known that the problem is equivalent to solving the optimal transport problem on the sphere between the light source distribution $\rho$ and the target illumination $\nu$ for the cost function $c(x, y) = -\ln(1 - \kappa'\langle x \mid y\rangle)$ if we choose to build the lens as an intersection of ellipsoids or with the cost function $c(x, y) = \ln(1 - \kappa'\langle x \mid y\rangle)$ if we choose an union of ellipsoids. The authors of [GH09] proved that for an intersection of ellipsoids the MTW condition is *not* satisfied when $\kappa' < 1$. The case where $\kappa' < 1$ corresponds to what we call the *practical* setting. Indeed, in practical applications, the lens has a greater index of refraction than the one of the ambient medium. The goal of this section is to show that when the component is instead parametrized by a union of ellipsoids, the MTW condition is satisfied in the practical setting, when $\kappa' < 1$.

We use the following notations for the partial derivatives of a differentiable function $f : X \to Y$ and $x \in X$:

- $\mathrm{D}_{x_k} f(x) = f_{x_k}(x) := \frac{\partial f}{\partial x_k}(x)$,
- $\mathrm{D}_{x_i,x_j} f(x) = f_{x_i,x_j}(x) := \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$.

### 2.4.1 Ma-Trudinger-Wang condition

We start by introducing the Ma-Trudinger-Wang (MTW) condition. In the semi-discrete case, we mention a geometric version of this condition called the *Loeper's condition* [Loe09]. We suppose we are given a manifold $M \subset \mathbb{R}^d$, a set $Y \subset \mathbb{R}^d$ and a cost function $c : M \times Y \to \mathbb{R}$. We also suppose that $c$ is regular namely that for all $y \in Y$, the function $x \mapsto c(x, y)$ is of class $\mathcal{C}^{1,1}$ on $M$ and that it satisfies the (Twist) condition introduced in the previous chapter. We start by defining the $c$-exponential.

**Definition 2.2** ($c$-exponential)
*For a given $y_0 \in Y$, the c-exponential at $y_0$ for the cost function $c$ is denoted by* $\exp^c_{y_0}$ *and is defined by*

$$\exp^c_{y_0} : v \in T_{y_0}M = [\nabla_y c(x, y_0)]^{-1}(v) \in M.$$

*We can also define the c-exponential with respect to the first variable meaning that for $x_0 \in X$, we define* $\exp^c_{x_0}$

$$\exp^c_{x_0} : v \in T_{x_0}M = [\nabla_x c(x_0, y)]^{-1}(v) \in M.$$

*The choice between the two will be clear in context.*

Using this notion, we can state the weak version of the Ma-Trudinger-Wang condition, [Vil09].

**Definition 2.3** (Weak Ma-Trudinger-Wang condition)
*A function $c : M \times Y \to \mathbb{R}$ satisfies the* weak Ma-Trudinger-Wang *condition if*

$$\forall x \in M,\ \forall p, \xi, \eta \in T_xM,\ \xi \perp \eta \text{ and } \sum_{i,j,k,l} \mathrm{D}_{p_i,p_j} a_{k,l}(x, p)\xi_i\xi_j\eta_k\eta_l \leqslant 0 \qquad \text{(w-MTW)}$$

*where* D *is the* gradient *operator and $a_{k,l}(x, p) = \nabla^2_{x_k,x_l} c(x, \exp^c_x(p))$.*

When $Y = \{y_1, \dots, y_N\}$ is finite, this condition can be stated more clearly using the notion of Laguerre cells introduced in Section 1.3. Roughly, it amounts to saying that the Laguerre cells are convex in some *c-exponential* charts, see [Loe09; KMT16] for more details. The $c$-convexity of a set $\Omega$ is defined as below.

**Definition 2.4** ($c$-convexity)
*A set $\Omega \subset M$ is said to be c-convex if it is convex in every c-exponential chart meaning that*

$$\forall y_0 \in Y,\ (\exp^c_{y_0})^{-1}(\Omega) \text{ is convex.}$$

Loeper's condition in the semi-discrete case can then be stated as follows:

**Definition 2.5** (Loeper's condition)

*A function* $c : M \times Y \to \mathbb{R}$ *satisfies* Loeper's condition *if*

$$\forall y_0 \in Y,\ \forall \psi \in \mathbb{R}^N,\ \mathrm{Lag}_{y_0}(\psi) \text{ is } c\text{-convex.}$$

*This is equivalent to saying that:*

$$\forall y_0 \in Y,\ \forall y \neq y_0,\ v \in T_{y_0}M \mapsto c(\exp^c_{y_0} v, y_0) - c(\exp^c_{y_0} v, y) \text{ is quasiconvex}$$

*A function is said to be quasiconvex if all its sublevelsets are convex.*

Remark 45. *Under certain assumptions on the cost function c, one can show that this semi-discrete version of Loeper's condition is implied by classical conditions introduced in the smooth setting such as the MTW condition. See Remark 4.3 of [KMT16] for more details.*

### 2.4.2 (PS/Lens) for a union of ellipsoids

We show here that the cost function arising in the (PS/Lens) problem for a union of ellipsoids when $\kappa' < 1$ satisfies the (w-MTW) condition. The main result is the following theorem:

**Theorem 46.** *The function* $c(x, y) = \ln(1 - \kappa' \langle x \mid y \rangle)$ *for* $0 < \kappa' < 1$ *defined on* $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ *satisfies the* (w-MTW) *condition.*

Table 2.1 summarizes the cases where the MTW condition is satisfied or not for mirror and lens design for a point light source. For mirror design, a calculation shows that the (w-MTW) condition is verified when parametrizing the mirror as an intersection of paraboloids [Wan96; Loe11]. For a union of paraboloids, the authors of [CMT15] shows that the corresponding Laguerre cell can be composed of multiple connected components and thus does not satisfy the MTW condition. For lens design, it has been shown in [GH09] that the condition is verified for a union of hyperboloids for $\kappa' > 1$ and is not verified for an intersection of ellipsoids for $\kappa' < 1$. In this section, we show the last two remaining cases meaning the intersection of hyperboloids and the union of ellipsoids.

| MTW | NON MTW |
|---|---|
| (PS/Mirror) intersection paraboloids | (PS/Mirror) union paraboloids |
| (PS/Lens) union ellipsoids, $\kappa' < 1$ | (PS/Lens) intersection ellipsoids, $\kappa' < 1$ |
| (PS/Lens) union hyperboloids, $\kappa' > 1$ | (PS/Lens) intersection hyperboloids, $\kappa' > 1$ |

Table 2.1 – Summary of the settings that satisfy the (w-MTW) condition or not.

The calculations presented in this section are very similar to the ones done in [GH09], we also follow the same outline and use the same notations for clarity reasons. We start by seeing how the notion of tangential gradient and tangential Hessian matrix is expressed on the unit sphere $\mathbb{S}^{d-1}$.

Proposition 47. *Let* $f : \mathbb{S}^{d-1} \to \mathbb{R}$ *be a differentiable function, its tangential gradient* $\nabla f$

*and its tangential Hessian matrix* $\nabla^2 f$ *at a point* $x \in \mathbb{S}^{d-1}$ *are given by*

$$\nabla f(x) = \mathrm{D}f(x) - \langle \mathrm{D}f(x) \mid x \rangle x$$
$$\nabla^2 f(x) = \mathrm{D}^2 f(x) - \langle \mathrm{D}f(x) \mid x \rangle I$$

*where* $\mathrm{D}f(x)$ *and* $\mathrm{D}^2 f(x)$ *are the standard gradient and Hessian of* $f$ *in* $\mathbb{R}^d$ *and* $I$ *is the identity matrix.*

To prove this result, we first need to compute the $c$-exponential $\exp_x^c(p)$ that we will denote as $Y(x,p)$ in the following for clarity.

LEMMA 48. *For the cost function* $c(x,y) = \ln(1 - \kappa' \langle x \mid y \rangle)$ *with* $0 < \kappa' < 1$*, the point* $Y(x,p)$ *that satisfies* $\nabla_x c(x, Y(x,p)) = p$ *is given by the expression*

$$\forall x \in \mathbb{S}^{d-1},\ \forall p \in T_x \mathbb{S}^{d-1},\ Y(x,p) = \lambda(p) x + \left( \lambda(p) - \frac{1}{\kappa'} \right) p,$$

*where* $\lambda(p) = \frac{\|p\|^2 + \sqrt{h(p)}}{\kappa'(1+\|p\|^2)}$ *and* $h(p) = \kappa'^2 - (1-\kappa'^2)\|p\|^2$ *defined for* $\|p\|^2 \leqslant \frac{\kappa'^2}{1-\kappa'^2}$.

*Proof.* The first observation we can make is that, in the case where $\kappa' < 1$, Snell's law implies that there is a maximum incidence angle $0 \leqslant \theta \leqslant \theta_{max} = \arcsin(\kappa')$. An easy computation shows that this constraint is equivalent to having $\langle x \mid y \rangle > \kappa'$ for an incident ray $x$ and a refracted ray $y$.

We then let $x \in \mathbb{S}^{d-1}$ and $p \in T_x \mathbb{S}^{d-1}$, we want to find $Y(x,p)$ such that $\nabla_x c(x, Y(x,p)) = p$. A simple calculation gives the following expressions for the gradient and tangential gradient with respect to $x$

$$\mathrm{D}_x c(x,y) = \frac{-\kappa' y}{1 - \kappa' \langle x \mid y \rangle} \text{ and } \nabla_x c(x,y) = \frac{-\kappa' y + \kappa' \langle x \mid y \rangle x}{1 - \kappa' \langle x \mid y \rangle}.$$

We now search for $Y(x,p)$ in $\mathrm{vect}(x,p)$ meaning that we look for two scalars $\lambda, \mu \in \mathbb{R}$ such that $Y(x,p) = \lambda x + \mu p$. If we inject this expression in the tangential gradient, we find that $\mu = \lambda - \frac{1}{\kappa'}$. Then, using the fact that $\|Y(x,p)\|^2 = 1$, we get the following second order equation

$$(1 + \|p\|^2)\lambda^2 - \frac{2}{\kappa'} \|p\|^2 \lambda + \left( \frac{\|p\|^2}{\kappa'^2} - 1 \right) = 0.$$

An easy calculation shows that if $\|p\|^2 \leqslant \frac{\kappa'^2}{1-\kappa'^2}$ then the solutions are given by $\lambda_\pm(p) = \frac{\|p\|^2 \pm \sqrt{h(p)}}{\kappa'(1+\|p\|^2)}$ where $h(p) = \kappa'^2 - (1-\kappa'^2)\|p\|^2$. Then, because we need to have $\langle x \mid Y(x,p) \rangle = \lambda_\pm(p) > \kappa'$, we deduce that we have to choose the root with the plus sign, thus

$$Y(x,p) = \lambda_+(p) x + \left( \lambda_+(p) - \frac{1}{\kappa'} \right) p.$$

□

*Proof of Theorem 46.* In the following, we let $x \in \mathbb{S}^{d-1}$ and $p \in T_x\mathbb{S}^{d-1}$. The proof is divided into three parts. First, we compute the first order coefficients $c_{x_i}(x, Y(x,p)) = \mathrm{D}_{x_i} c(x, Y(x,p))$; then, we compute the second order coefficients $a_{k,l}(x,p)$; and finally, we show that the condition (w-MTW) is verified when $\kappa' < 1$.

**Step 1: Computation of $c_{x_i}(x, Y(x,p))$.** Using Lemma 48, we get

$$\mathrm{D}_x c(x, Y(x,p)) = p - \frac{\kappa'\lambda(p)}{1-\kappa'\lambda(p)} x.$$

We deduce that

$$c_{x_i}(x, Y(x,p)) = \mathrm{D}_{x_i} c(x, Y(x,p)) = p_i - \frac{\kappa'\lambda(p)}{1-\kappa'\lambda(p)} x_i.$$

Let us also remark that we have following relation that we be useful in the next part of the proof:

$$\langle \mathrm{D}_x c(x, Y(x,p)) \mid x \rangle = \frac{-\kappa'\lambda(p)}{1-\kappa'\lambda(p)}.$$

**Step 2: Computation of $a_{k,l}(x,p)$.** First, an easy computation shows that

$$\forall x, y \in \mathbb{S}^{d-1},\ \forall k, l,\ c_{x_k,x_l}(x,y) = -c_{x_k}(x,y) c_{x_l}(x,y). \tag{$\star\star$}$$

Let us remark that, contrary to [GH09], here we have a minus sign in front. Then:

$$c_{x_k,x_l}(x, Y(x,p)) = -\left(p_k - \frac{\kappa'\lambda(p)}{1-\kappa'\lambda(p)} x_k\right)\left(p_l - \frac{\kappa'\lambda(p)}{1-\kappa'\lambda(p)} x_l\right).$$

With that, we are ready to compute the coefficients $a_{k,l}(x,p)$. These coefficients depend on the tangential Hessian matrix, see Proposition 47.

$$\begin{aligned} a_{k,l}(x,p) &= \nabla^2_{x_k,x_l} c(x, Y(x,p)) = c_{x_k,x_l}(x, Y(x,p)) - \langle \mathrm{D}_x c(x, Y(x,p)) \mid x \rangle \delta_{k,l} \\ &= -(p_k - g(p)x_k)(p_l - g(p)x_l) + \delta_{k,l} g(p) \end{aligned}$$

where $g(p) = \frac{\kappa'\lambda(p)}{1-\kappa'\lambda(p)}$.

**Step 3: Checking the** (w-MTW) **condition.** Finally, we show that the condition (w-MTW) is verified when $\kappa' < 1$.

Then, the second derivatives of $a_{k,l}$ with respect to $p_i$ and $p_j$ are given by the following expressions where $\delta_{i,j}$ is the Kronecker symbol. To simplify the notations, we dropped the evaluation at $(x,p)$.

$$\mathrm{D}_{p_i} a_{k,l} = -\left[(\delta_{k,i} - \partial_{p_i} g x_k)(p_l - g x_l) + (p_k - g x_k)(\delta_{j,l} - \partial_{p_j} g x_l)\right] + \delta_{k,l} \partial_{p_i} g$$

Expanding this equation, we get:

$$\mathrm{D}_{p_i,p_j} a_{k,l} = -[\underbrace{-(\partial_{p_i,p_j} g x_k)(p_l - g x_l)}_{(1)} + \underbrace{(\delta_{k,i} - \partial_{p_i} g x_k)(\delta_{j,l} - \partial_{p_j} g x_k)}_{(2)} + \underbrace{(\delta_{j,k} - \partial_{p_j} g x_l)(\delta_{l,i} - \partial p_i g x_l)}_{(3)} + \underbrace{(p_k - g x_k)(-\partial_{p_i,p_j} g x_l)}_{(4)}] + \underbrace{\delta_{k,l}\partial_{p_i,p_j} g}_{(5)}$$

Now, we take $\xi, \eta \in T_x \mathbb{S}^{d-1}$ such that $\xi \perp \eta$ and we study the quantity $\mathrm{D}_{p_i,p_j} a_{k,l} \xi_i \xi_j \eta_k \eta_l$. To do that, we will look at each term separately. The notation $\overset{\sum}{=}$ means that we sum over $i, j, k, l$. We start by looking at the last term (5)

$$(5)\ \xi_i \xi_j \eta_k \eta_l = \partial_{p_i,p_j} g \xi_i \xi_j \delta_{k,l} \eta_k \eta_l = \partial_{p_i,p_j} g \xi_i \xi_j \eta_k^2 \overset{\sum}{=} \langle \mathrm{D}^2_{pp} g(p) \xi \mid \xi \rangle \left\| \eta \right\|^2 .$$

Then, we look at the first one (1):

$$\begin{aligned}(1)\ \xi_i \xi_j \eta_k \eta_l &= -\partial_{p_i,p_j} g x_k (p_l - g x_l) \xi_i \xi_j \eta_k \eta_l = -\partial_{p_i,p_j} g \xi_i \xi_j \eta_k x_k (p_l - g x_l) \eta_l \\ &\overset{\sum}{=} -\langle \mathrm{D}^2_{pp} g \xi \mid \xi \rangle \langle \eta \mid x \rangle (\langle p \mid \eta \rangle - g \langle x \mid \eta \rangle) = 0 \text{ since } \eta \perp x.\end{aligned}$$

The same computation holds for (4). Furthermore, for the second term (2), we have:

$$\begin{aligned}(2)\ \xi_i \xi_j \eta_k \eta_l &= (\delta_{k,i} - \partial_{p_i} g x_k)(\delta_{j,l} - \partial_{p_j} g x_l) \xi_i \xi_j \eta_k \eta_l \\ &= (\delta_{k,i}\delta_{j,l} - \delta_{k,i}\partial_{p_j} g x_l - \delta_{j,l}\partial_{p_i} g x_k + \partial_{p_i} g \partial_{p_j} g x_k x_l) \xi_i \xi_j \eta_k \eta_l \\ &= \xi_i \eta_i \xi_j \eta_j - \xi_i \eta_i \partial_{p_j} g \xi_j x_l \eta_l - \xi_i \partial_{p_i} g \xi_j \eta_j x_k \eta_k + \partial_{p_i} g \xi_i \partial_{p_j} g \xi_j \eta_k \eta_l \\ &\overset{\sum}{=} \langle \xi \mid \eta \rangle^2 - \langle \xi \mid \eta \rangle \langle \mathrm{D}_p g(p) \mid \xi \rangle \langle x \mid \eta \rangle - \langle \mathrm{D}_p g(p) \mid \xi \rangle \langle \xi \mid \eta \rangle \langle x \mid \eta \rangle + \langle \mathrm{D}_g(p) \mid \xi \rangle^2 \langle x \mid \eta \rangle^2 = 0.\end{aligned}$$

since $x \perp \eta$ and $\xi \perp \eta$. And the same computation works for (3).

Combining the terms, we get:

$$\sum_{i,j,k,l} \mathrm{D}_{p_i,p_j} a_{k,l} \xi_i \xi_j \eta_k \eta_l = \sum_{i,j,k,l} (5)\ \xi_i \xi_j \eta_k \eta_l = \langle \mathrm{D}^2_{pp} g(p) \xi \mid \xi \rangle \left\| \eta \right\|^2 .$$

We then have $g(p) = \frac{\kappa'^2}{1-\kappa'^2} + \frac{\sqrt{h(p)}}{1-\kappa'^2}$ where we recall that $h(p) = \kappa'^2 - (1-\kappa'^2)\left\|p\right\|^2$. This implies that $\nabla g(p) = -\frac{p}{\sqrt{h(p)}}$ and that

$$\langle \mathrm{D}^2_{pp} g(p) \xi \mid \xi \rangle = -\frac{h(p)\left\|\xi\right\|^2 + (1-\kappa'^2)\langle p \mid \xi \rangle^2}{h(p)^{3/2}}.$$

Finally, we get:

$$\sum_{i,j,k,l} \mathrm{D}_{p_i,p_j} a_{k,l} \xi_i \xi_j \eta_k \eta_l = -\frac{h(p)\,\|\xi\|^2 + (1-\kappa'^2)\langle p \mid \xi\rangle^2}{h(p)^{3/2}}\,\|\eta\|^2\,. \qquad (\star)$$

We conclude that if $\kappa' < 1$ then the right hand side is negative (since $h(p)$ is positive) and (w-MTW) is verified. □

REMARK 49. *It is easy to show using very similar arguments that the cost function* $c(x,y) = -\ln(\kappa'\langle x \mid y\rangle - 1)$ *(corresponding to an intersection of hyperboloids) does not satisfy the MTW condition when* $\kappa' > 1$.

REMARK 50. *We obtain very similar results as the one presented in [GH09] except that we have a minus sign in front of the expression* $(\star)$ *and that the function* $h$ *is different. This sign comes from the expression* $(\star\star)$ *and it is preserved under differentiation.*

Chapter 3

# A Parameter-free algorithm for mirror and lens design

## Contents



In this chapter, we will show how we can leverage the formulation of non-imaging optics problems in terms of optimal transport described in the previous chapter to develop an efficient method to solve them: we propose a generic algorithm to solve eight optical component design problems.

We start in Section 3.1 by showing that all the Visibility diagrams we consider have the same structure and can be obtained by intersecting a 3D Power diagram with a planar or spherical domain. Then, in Section 3.2, we describe an algorithm based on the damped Newton's method explained in Chapter 1 to solve many optical component design problems when the target illumination is located at infinity i.e. in the *far-field* setting. We also show that we can use a simple iterative procedure to also handle the case when the target is located at a finite distance i.e. the *near-field* setting. This allows us to consider more complex problems such as when the target is a color image and or when one wants builds multiple components that target the same illumination to account for occlusion. Finally, in Section 3.3, we illustrate our method with numerous simulated and fabricated examples. In particular, we show that we can handle high resolution target lights with more than 4 million Dirac masses. The results presented in this chapter come from the article [MMT18b].

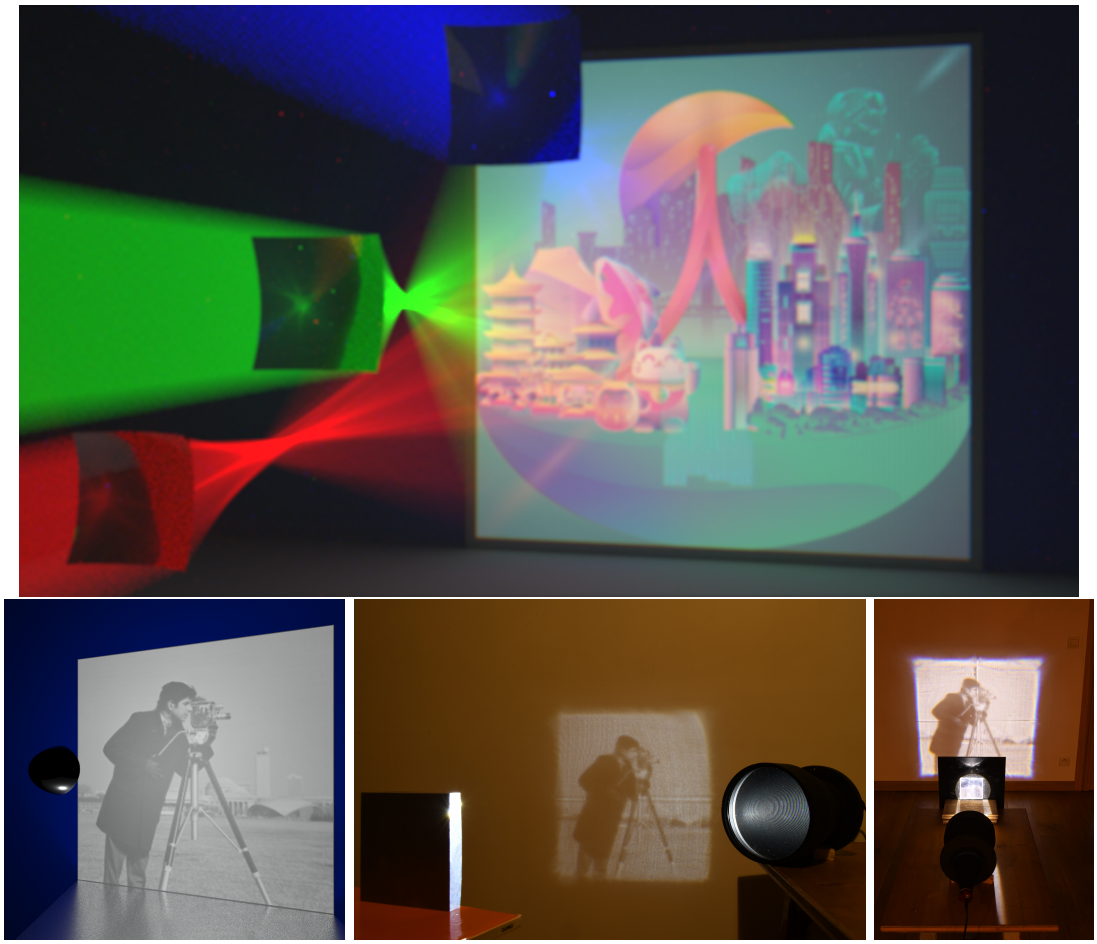

Figure 1 – **Examples of simulated and physical results.** Our algorithm can be used to design mirrors and lenses that reflect or refract collimated or point light sources onto a prescribed distribution of light. *From top to bottom and left to right:* Three lenses that refract the three channels of a color image; Mirror that reflects a point light source (located inside the mirror); Fabricated mirror that reflects a collimated light source; Fabricated lens that refracts a collimated light source.

## 3.1 Visibility diagram as a restricted Power diagram

The main difficulty in evaluating the function $G$ appearing in Equation (LEC) is to compute the Visibility cells $\mathrm{V}_i(\psi)$ associated to each optical modeling problem. We show in this section that the Visibility cells have always the same structure, allowing us to build a generic algorithm in Section 3.2. More precisely, in all the non-imaging problems presented in Section 2.3, the Visibility cells are of the form

$$\mathrm{V}_i(\psi) = \mathrm{Pow}_i(P) \cap X. \tag{3.1.1}$$

For a collimated source, $X$ denotes the plane $\mathbb{R}^2 \times \{0\}$ and for a point source, $X$ is the unit sphere $\mathbb{S}^2$. We recall that the sets $\mathrm{Pow}_i(P)$ are the usual Power cells of a weighted point cloud

$P = \{(p_i, \omega_i)\} \subset \mathbb{R}^3 \times \mathbb{R}$ defined as

$$\mathrm{Pow}_i(P) = \{x \in \mathbb{R}^3 \mid \forall j,\ \|x - p_i\|^2 + \omega_i \leqslant \|x - p_j\|^2 + \omega_j\}.$$

This will allow us to efficiently compute the Visibility cells as it is easier and more generic to intersect a Power diagram with a domain than directly computing the cells. The expression of the weighted point cloud $P = \{(p_i, \omega_i)\}$ depends on the problem as we will explain in the following.

### 3.1.1 Collimated source

In the (CS/Mirror) case, the following proposition explains the structure of the Visibility cells.

PROPOSITION 51. *Letting $X = \mathbb{R}^2 \times \{0\}$ and $\psi \in \mathbb{R}^N$, we have $\mathrm{V}_i(\psi) = \mathrm{Pow}_i(P) \cap X$ where the weighted point cloud $P = \{(p_i, \omega_i)\}$ is given by the following expressions*

- ***Convex** mirror:*
$$(p_i, \omega_i) = \left( \frac{\mathrm{proj}_{\mathbb{R}^2}(y_i - e_z)}{\langle y_i - e_z \mid e_z \rangle}, 2\psi_i - \|p_i\|^2 \right),$$
- ***Concave** mirror:*
$$(p_i, \omega_i) = \left( -\frac{\mathrm{proj}_{\mathbb{R}^2}(y_i - e_z)}{\langle y_i - e_z \mid e_z \rangle}, 2\psi_i - \|p_i\|^2 \right).$$

*where $\mathrm{proj}_{\mathbb{R}^2} : \mathbb{R}^3 \to \mathbb{R}^2 \times \{0\}$ denotes the orthogonal projection onto $\mathbb{R}^2 \times \{0\}$.*

*Proof.* We explain the formulas in the convex case. In the (CS/Mirror) case, the light source is collimated and $p_i \in \mathbb{R}^2 \times \{0\}$ is the slope of the plane which reflects (according to Snell's law) the upward vertical ray $e_z = (0, 0, 1)$ to the direction $y_i$. We can express Snell's law in vector form by $\vec{r} = \vec{i} - 2\langle \vec{i} \mid \vec{n} \rangle \vec{n}$ where $\vec{i}$ is the unit incident ray, $\vec{n}$ the unit normal and $\vec{r}$ the reflected ray. Then, a straightforward calculation shows that

$$p_i = \frac{\mathrm{proj}_{\mathbb{R}^2}(y_i - e_z)}{\langle y_i - e_z \mid e_z \rangle}.$$

The Visibility cell of $y_i$ is then given by

$$\begin{aligned}
\mathrm{V}_i(\psi) &= \{x \in X \mid\ \forall j, -\langle x \mid p_i \rangle + \psi_i \leqslant -\langle x \mid p_j \rangle + \psi_j\} \\
&= \{x \in X \mid\ \forall j, \|x - p_i\|^2 + 2\psi_i - \|p_i\|^2 - \|x\|^2 \leqslant \|x - p_j\|^2 + 2\psi_j - \|p_j\|^2 - \|x\|^2\} \\
&= \{x \in X \mid\ \forall j, \|x - p_i\|^2 + \omega_i \leqslant \|x - p_j\|^2 + \omega_j\}, \\
&= \mathrm{Pow}_i(P) \cap X,
\end{aligned}$$

where $\omega_i = 2\psi_i - \|p_i\|^2$. In the concave case, one just has to replace $p_i$ by its opposite. □

REMARK 52. *When looking at the (CS/Lens) problem, one just has to replace the expression of $p_i$ with the slope of the plane that* refracts *the upward vertical ray $e_z$ to the direction $y_i$. A straightforward calculation shows that in the* concave *case:*

$$p_i = -\frac{\mathrm{proj}_{\mathbb{R}^2}(y_i - \kappa e_z)}{\langle y_i - \kappa e_z \mid e_z \rangle}.$$

*In the* convex *case, one just replaces $p_i$ by its opposite.*

### 3.1.2 Point light source

We now look at the (PS/Lens) setting. The calculations are similar to the one presented in [CMT15] but are included here for completeness. The main result is the following proposition:

PROPOSITION 53. *Letting $X = \mathbb{S}^2$ and $\psi \in \mathbb{R}^N$, we have $\mathrm{V}_i(\psi) = \mathrm{Pow}_i(P) \cap X$ where the weighted point cloud $P = \{(p_i, \omega_i)\}$ is given by*

- ***Convex*** *lens:*
$$(p_i, \omega_i) = \left( -\kappa' \frac{y_i}{2\widetilde{\psi}_i}, -\frac{1}{\widetilde{\psi}_i} - \frac{\kappa'^2}{4\widetilde{\psi}_i^{\,2}} \right),$$
- $\widetilde{\textbf{\textit{Concave}}}$ *lens:*
$$(p_i, \omega_i) = \left( \kappa' \frac{y_i}{2\widetilde{\psi}_i}, \frac{1}{\widetilde{\psi}_i} - \frac{\kappa'^2}{4\widetilde{\psi}_i^{\,2}} \right)$$

*where $\kappa' = \frac{1}{\kappa}$, $\widetilde{\psi} = (\ln(\psi_i))_{1 \leqslant i \leqslant n}$.*

The notation $\widetilde{\textbf{Concave}}$ means that the component is not concave but converges towards a concave one when the discretization tends to infinity.

*Proof.* First, in order to transform the problem into an optimal transport one (see Remark 40 in the previous section), we let $\widetilde{\psi} = (\ln(\psi_i)))_{1 \leqslant i \leqslant N}$. We then take $x \in \mathrm{V}_i(\psi)$ which is equivalent to $x \in \mathrm{V}_i(\widetilde{\psi})$ (due to the fact that ln is non-decreasing), we have

$$i = \operatorname*{argmin}_{1 \leqslant j \leqslant N} \frac{\widetilde{\psi}_j}{1 - \kappa' \langle x \mid y_j \rangle} \iff i = \operatorname*{argmax}_{1 \leqslant j \leqslant N} \left( \widetilde{\psi}_j^{-1} - \langle x \mid \kappa' \widetilde{\psi}_j^{-1} y_j \rangle \right)$$

Then:

$$\begin{aligned}\max_{1\leqslant j\leqslant N}\left(\widetilde{\psi}_j^{-1}-\langle x\mid \kappa'\widetilde{\psi}_j^{-1}y_j\rangle\right) &= \max_{1\leqslant j\leqslant N}\left(\widetilde{\psi}_j^{-1}-\left\|x+\frac{\kappa'}{2}\widetilde{\psi}_j^{-1}y_j\right\|^2+\|x\|^2+\frac{1}{4}\left\|\kappa'\widetilde{\psi}_j^{-1}y_j\right\|^2\right)\\ &= \|x\|^2-\min_{1\leqslant j\leqslant N}\left(\left\|x+\frac{\kappa'}{2}\widetilde{\psi}_j^{-1}y_j\right\|^2-\widetilde{\psi}_j^{-1}-\frac{\kappa'^2}{4}\widetilde{\psi}_j^{-2}\right)\\ &= 1-\min_{1\leqslant j\leqslant N}\left(\|x-p_i\|^2+\omega_i\right)\end{aligned}$$

where we have denoted $p_i=-\kappa'\frac{y_i}{2\widetilde{\psi}_i}$ and $\omega_i=-\widetilde{\psi}_i^{-1}-\frac{\kappa'^2}{4}\widetilde{\psi}_i^{-2}$. We conclude that

$$x\in \mathrm{V}_i(\psi)\iff x\in \mathrm{Pow}_i(P)\cap X.$$

The formulas for the $\widetilde{\textbf{Concave}}$ can be recovered by doing very similar computations. □

REMARK 54. *Let us remark that setting $\kappa'=1$ in the previous equation allows us to recover the formulas for the (PS/Mirror) setting.*

All the formulas for the eight different settings are summarized in Table 3.1.

Table 3.1 – Formulas for the weighted points used to define the Power cells in Equation (3.1.1) for the various problems. In the lens design problem, $\kappa>0$ is the ratio of the indices of refraction, $\kappa>1$ in the (PS/Lens) setting, $\kappa'=\frac{1}{\kappa}$. Ccv means concave and Cvx convex. $\widetilde{\text{Ccv}}$ means that the optical component converges to a concave one when the discretization tends to infinity.

| **Setting** | Points | Weights |
|---|---|---|
| Cvx (CS/Mirror) | $p_i=\frac{\mathrm{proj}_{\mathbb{R}^2}(y_i-e_z)}{\langle y_i-e_z\mid e_z\rangle}$ | $\omega_i=2\psi_i-\|p_i\|^2$ |
| Ccv (CS/Mirror) | $p_i=-\frac{\mathrm{proj}_{\mathbb{R}^2}(y_i-e_z)}{\langle y_i-e_z\mid e_z\rangle}$ | $\omega_i=2\psi_i-\|p_i\|^2$ |
| Cvx (PS/Mirror) | $p_i=-\frac{y_i}{2\ln(\psi_i)}$ | $\omega_i=-\frac{1}{\ln(\psi_i)}-\frac{1}{4\ln(\psi_i)^2}$ |
| $\widetilde{\text{Ccv}}$ (PS/Mirror) | $p_i=\frac{y_i}{2\ln(\psi_i)}$ | $\omega_i=\frac{1}{\ln(\psi_i)}-\frac{1}{4\ln(\psi_i)^2}$ |
| Cvx (CS/Lens) | $p_i=-\frac{\mathrm{proj}_{\mathbb{R}^2}(y_i-\kappa e_z)}{\langle y_i-\kappa e_z\mid e_z\rangle}$ | $\omega_i=2\psi_i-\|p_i\|^2$ |
| Ccv (CS/Lens) | $p_i=\frac{\mathrm{proj}_{\mathbb{R}^2}(y_i-\kappa e_z)}{\langle y_i-\kappa e_z\mid e_z\rangle}$ | $\omega_i=2\psi_i-\|p_i\|^2$ |
| Cvx (PS/Lens) | $p_i=-\kappa'\frac{y_i}{2\ln(\psi_i)}$ | $\omega_i=-\frac{1}{\ln(\psi_i)}-\frac{\kappa'^2}{4\ln(\psi_i)^2}$ |
| $\widetilde{\text{Ccv}}$ (PS/Lens) | $p_i=\kappa'\frac{y_i}{2\ln(\psi_i)}$ | $\omega_i=\frac{1}{\ln(\psi_i)}-\frac{\kappa'^2}{4\ln(\psi_i)^2}$ |

In the next section, we will show how we can leverage the relation between a Visibility diagram and a Power diagram to use an algorithm based on the damped Newton's method, detailed in Algorithm 4, to solve efficiently and generically all the optical component design problems.

## 3.2 A generic algorithm

For each optical design problem, given a light source intensity function, a target light intensity function and an error parameter, we will detail an algorithm, namely Algorithm 5 outputs a triangulation of a mirror or a lens that satisfies the Light Energy Conservation Equation (LEC).

The goal is to find weights $\psi$ such that $G(\psi) = \sigma$ (see Equation (2.3.2)). In the previous section, we saw how we can compute the Visibility diagram as the restriction of a Power diagram with a planar or spherical domain. This allows us to reuse the damped Newton's method we presented in Chapter 1 to solve this equation. A key point of this algorithm is to enforce the Jacobian matrix $\mathrm{D}G(\psi)$ to always be of rank $N-1$. To this purpose, we need to enforce all along the process that

$$\forall i \in \{1, \dots, N\},\ G_i(\psi) > 0. \tag{3.2.2}$$

Indeed, first remark that since $G$ is invariant under the addition of a constant, the kernel of $\mathrm{D}G(\psi)$ always contains the vector $(1, \dots, 1)$. Now remark that if we have $G_i(\psi) = 0$, then the corresponding Visibility cell $\mathrm{V}_i(\psi)$ is empty, which implies that $\nabla G_i(\psi) = 0$ (the gradient being taken with respect to $\psi$). This is because the gradient of $G_i$ involves integral on the boundary $\partial \mathrm{V}_i(\psi)$, as shown for instance in Theorem 18 and Theorem 1.3 of [KMT16]. Hence, if $G_i(\psi) = 0$, then the rank of $\mathrm{D}G(\psi)$ is at most $N-2$ which prevents from using the Damped Newton method. Our method consists of three steps:

- **Initialization** (Section 3.2.1): We first discretize the source density into a piecewise affine density supported on a triangulation and the target one into a finitely supported measure. Then, we find initial weights $\psi^0$ satisfying the condition $\forall i,\ G_i(\psi^0) > 0$.
- **Damped Newton** (Section 3.2.2): We construct a sequence $\psi^k$ following Algorithm 6 until $\left\|G(\psi^k) - \sigma\right\| \leqslant \eta$. The main difficulty here is to evaluate $G(\psi^k)$ and $\mathrm{D}G(\psi^k)$.
- **Surface construction** (Section 3.2.3): Finally, we convert the solution $\psi^k \in \mathbb{R}^N$ into a triangulation. Depending on the non-imaging problem, this amounts to approximating an intersection (or union) of half-spaces (or solid paraboloids, or ellipsoids) by a triangulation.

### 3.2.1 Initialization

**Discretization of light intensity functions.** Our framework allows to handle any kind of collimated or point light source or target light intensity functions. The light source can be for example any positive function supported on the plane or the sphere (depending on the problem). We first approach the support of the source density $\rho$ by a triangulation $T$ and assume that the density $\rho : T \to \mathbb{R}^+$ is affine on each triangle with a value at each vertex. We then normalize $\rho$ by dividing it by the total mass $\int_T \rho(x)\mathrm{d}x$.

Similarly, the target light intensity function can also be any discrete probability measure supported on a finite set $Y \subset \mathbb{S}^2$. If the user provides a greyscale image, one can transform it into a discrete measure of the form $\sigma = \sum_i \sigma_i \delta_{y_i}$ using Lloyd's algorithm or more simply

**Algorithm 5:** Mirror / lens construction

**Input** A light source intensity function $\rho_{in}$.
A target light intensity function $\sigma_{in}$.
A tolerance $\eta > 0$.

**Output** A triangulation $\mathcal{R}_T$ of a mirror or lens.

**Step 1** Initialization (Section 3.2.1)
$T, \rho \leftarrow$ `DISCRETIZATION_SOURCE`$(\rho_{in})$
$Y, \sigma \leftarrow$ `DISCRETIZATION_TARGET`$(\sigma_{in})$
$\psi^0 \leftarrow$ `INITIAL_WEIGHTS`$(Y)$

**Step 2** Solve Equation (LEC): $G(\psi) = \sigma$ using Algorithm 6 (Section 3.2.2)
$\psi \leftarrow$ `DAMPED_NEWTON`$(T, \rho, Y, \sigma, \psi^0, \eta)$

**Step 3** Construct a triangulation $\mathcal{R}_T$ of $\mathcal{R}$ (Section 3.2.3)
$\mathcal{R}_T \leftarrow$ `SURFACE_CONSTRUCTION`$(\psi, \mathcal{R}_\psi)$

by taking one Dirac mass per pixel. We do the latter in all experiments. The target measure is also normalized by dividing with the discrete integral $\sum_i \sigma_i$. We need $\min_i \sigma_i > 0$ for the damped Newton's algorithm to converge, so if $\sigma_i = 0$, we simply remove the corresponding Dirac mass $\delta_{y_i}$, thus ensuring that no light is sent towards the direction $y_i$.

**Choice of the initial family of weights $\psi^0$.** As mentioned at the beginning of this section, we need to ensure that at each iteration all the Visibility cells have non-empty interiors. In particular, we need to choose a set of initial weights $\psi^0 = (\psi_i^0)_{1 \leqslant i \leqslant N}$ such that the initial Visibility cells are not empty. In our case, we can use simple heuristics:

- For the collimated light sources cases (CS/Mirror) and (CS/Lens), it is easy to see that if we choose $\psi_i^0 = \frac{\|p_i\|^2}{2}$ then $\omega_i = 0$, where $p_i$ is obtained using the formulas of Section 3.1, then the Visibility diagram becomes a Voronoi diagram, hence $p_i \in \mathrm{V}_i(\psi^0)$.
- For the mirror design for a point light source (PS/Mirror) case, an easy calculation shows that if we choose $\psi_i^0 = 1$, then $-y_i \in \mathrm{V}_i(\psi^0)$.
- For the lens design for a point light source (PS/Lens) case, we can show that if we also choose $\psi_i^0 = 1$, then $y_i \in \mathrm{V}_i(\psi^0)$.

Remark that the previous expressions for $\psi^0$ ensure that $G_i(\psi^0) = \rho(\mathrm{V}_i(\psi^0)) > 0$ only when the support $X_\rho$ of the light source is large enough. As an example in the (PS/Mirror) case, if for some $i$, $-y_i \notin X_\rho$, then we may have $G_i(\psi^0) = 0$. To handle this difficulty, we use a linear interpolation between $\rho$ and a constant density supported on a set that contains the $(-y_i)$'s. This strategy also works for the (CS/Mirror), (PS/Lens) and (CS/Lens) cases.

REMARK 55. *More details on the initialization methods for optimal transport and in particular for the linear interpolation procedure can be found in Chapter 4 of this thesis.*

### 3.2.2 Damped Newton's algorithm

When the light source is collimated (*i.e.* $X = \mathbb{R}^2 \times \{0\}$), the problem is known to be an optimal transport problem in the plane for the quadratic cost, the function $G$ is the gradient of a concave function, its Jacobian matrix $\mathrm{D}G$ is symmetric and $\mathrm{D}G \leqslant 0$. Moreover, if $G_i(\psi) > 0$ for all $i$ and if $X_\rho$ is connected, then the kernel of $\mathrm{D}G$ is spanned by $\psi = \mathrm{cst}$. This ensures the convergence of the damped Newton's algorithm, see Algorithm 4 and [KMT16], presented as Algorithm 6.

In the case of a point source, we make the change of variable $\widetilde{\psi} = \ln(\psi)$ and $\widetilde{G} = G \circ \exp$, so that $G(\psi) = \sigma$ if and only if $\widetilde{G}(\tilde{\psi}) = \sigma$. This change of variable turns the optical component design problem into an optimal transport problem, ensuring that $\widetilde{G}$ is the gradient of a concave function and that $\mathrm{D}\widetilde{G}$ is symmetric negative [CMT15], thus easily invertible. In the (PS/Mirror) problem with convex mirrors, the damped Newton algorithm is also provably converging [KMT16].

**Computation of $G$ and $\mathrm{D}G$.** By Section 3.1, the Visibility cells $\mathrm{V}_i(\psi)$ can be computed by intersecting a certain 3D Power diagram with a triangulation $T$ of the support $X_\rho$ of $\rho$. Such intersection can be computed for instance using the algorithm detailed in Section 1.5 or the ones described in [Lév15; SNA17]. Then,

$$G_i(\psi) = \int_{\mathrm{V}_i(\psi)} \rho(x)\mathrm{d}x$$

can be computed using first order quadrature formulas. The computation of $\mathrm{D}G$ is done using forward-mode automatic differentiation [Ral81], where we store the gradient of $G_i(\psi)$ as a sparse vector. Note that this works quite efficiently since all numbers that occur in the computation of $G_i(\psi)$ depend only on the values $\psi_j$ where $j$ is such that $(i, j)$ are neighbors in the Visibility diagram, i.e. $\mathrm{V}_i(\psi) \cap \mathrm{V}_j(\psi) \neq \emptyset$.

**Linear system.** Since $\mathrm{D}\widetilde{G}$ is sparse and symmetric negative, we can efficiently solve the linear systems using preconditioned conjugate gradient or Cholesky solver.

### 3.2.3 Surface construction

In the last step of Algorithm 5, we build a triangulation of the mirror or lens surface. The input is a family of weights $\psi$ solving Equation (LEC) and the parameterization function $\mathcal{R}_\psi$ whose expression is given in Section 2.3 and depends on the eight different cases. We triangulate each Visibility cell by taking the convex hull of the vertices of its boundary. A

**Algorithm 6:** Damped Newton's method for an optical component design problem

**Input** The source $\rho : T \to \mathbb{R}^+$ and target $\sigma = \sum_i \sigma_i \delta_{y_i}$; an initial vector $\psi^0$ and a tolerance $\eta > 0$.

**Step 1** Transformation into an optimal transport problem

**If** $X = \mathbb{R}^2 \times \{0\}$, then $\widetilde{\psi}^0 = \psi^0$ (and $\widetilde{G} = G$).
**If** $X = \mathbb{S}^2$, then $\widetilde{\psi}^0 = (\ln(\psi_i^0))_{1 \leqslant i \leqslant N}$ (and $\widetilde{G} = (G_i \circ \exp)_{1 \leqslant i \leqslant N}$).

**Step 2** Solve the equation: $\widetilde{G}(\widetilde{\psi}) = \sigma$ using Algorithm 4 on Page 25

**Return** $\psi := (\widetilde{\psi}_i^k)_{1 \leqslant i \leqslant N}$ if $X = \mathbb{R}^2 \times \{0\}$ or
$\psi := (\exp(\widetilde{\psi}_i^k))_{1 \leqslant i \leqslant N}$ if $X = \mathbb{S}^2$.

vertex of the triangulation will belong to at least one Visibility cell. For each vertex, we can compute exactly the normal to the (continuous, non-discretized surface) using Snell's law since we know the incident ray and the corresponding reflected/refracted direction $y_i$.

## 3.3 Results and discussion

In this section, we present several numerical examples for the different problems previously described as well as some other applications. In the experiments, we take $\kappa = 1.5$. Unless stated otherwise, the light source is chosen to be uniform and the discretization of the target (number of Diracs $N$) is chosen to be equal to the size of the image. The stopping criterion of the Newton's algorithm (Algorithm 6) is set to $\eta = 10^{-8}$.

The output of our algorithm is a triangulation equipped with a normal at each vertex. In all the simulations, we use the LuxRender rendering engine, with the Bidirectional Path Tracing rendering method combined with a Sobol sampler and Fresnel losses are not taken into account.

### 3.3.1 Simulated results

**Genericity.** Our algorithm is able to solve the eight different optical component design problems presented in Section 2.3. We present for instance in Figure 2 four examples for which we display the Visibility diagram of $X_\rho$ as well as the optical component (lens or mirror) above it, a mesh of the optical component and a forward simulation using LuxRender. Then, for the examples of Figures 3, 4, 6, and 7, we display the target distribution as an image; a mean curvature plot (blue represents low mean curvature and red high mean curvature) of the constructed mesh $\mathcal{R}_T$ and a forward simulation using LuxRender.

**High-contrast and complex target lights.** We can handle any kind of target distribution. Figures 3 and 4 shows several examples of mirror design for respectively a collimated and a point light source. Note that we are able to construct mirrors for smooth images such as the CAMERAMAN (first row) or LENA (second row) images as well as images with totally black areas (third and fourth rows). We are also able to handle target supported on *non-convex* sets such as the HIKARI and SIGGRAPH images. One can notice that since the area of the Visibility cells are equal to the greyscale values of the image then the triangles have roughly the same size, implying that one can recognize the target image in the mesh of the surface, see Figure 2 for zooms on different meshes. The mean curvature plot shows the discontinuities in the surface which come from the black areas in the image. Figures 6 and 7 show the same kind of results for the lens design problems (CS/Lens) and (PS/Lens).

**Non-uniform light sources.** Algorithm 5 can also be used with non-uniform light sources. In Figure 5, we compare the meshes that are generated in the (CS/Lens) case when the source is either uniform or a Gaussian (left). Because of the higher concentration of light, the details of the triangulation are more concentrated in the middle in the Gaussian case (middle) than in the uniform case (right).

**Convex / concave optical components.** As shown in Section 2.3, for each problem, one can choose between two different parameterizations. For instance, for the (CS/Lens) problem, one can build a lens which is either concave or convex, see Figure 6 for an illustration of these differences. Note that in the (PS/Lens) setting (which corresponds to the last row of Figure 2 and Figure 7), the light source is not supported on the full hemisphere $\mathbb{S}^2_+$ but instead on a smaller part of it. Indeed, choosing a smaller support for $\mu$ enforces that $\mathcal{R}_T$ is a graph above the plane instead of the hemisphere and thus avoids potential inter-refractions. As for all figures, we have performed no post-processing on $\mathcal{R}_T$ in order to emphasize the benefit of designing convex or concave optical components (convexity is a form of regularity). One also observes that when the lens is rotated with respect to the light source (Figure 9 and first row of Figure 14), or when the target screen is not at the right distance (Figure 13), the image is deformed in a monotonic and regular way. We believe this is due to the monotonicity properties of optimal transport and to the concavity/properties of the optical components.

**Comparison with previous work.** Figure 8 compares the state of the art results presented in [Sch+14] (second column) and the LuxRender renderings obtained by our method (third column) on two target distributions for the (CS/Lens) case with a collimated uniform light source in the near-field setting. Although the results are comparable, one can notice, in the second column, the presence of small artifacts between the black and white regions, for instance around the rings (notably in the center).

**Application to pillows** This problem consists in decomposing the optical component (mirror or lens) into several smaller optical components that are called *pillows*, as illustrated in the

first image of Figure 1, and are widely used in car headlight design. Each pillow independently satisfies a non-imaging problem with the same target light, but with a different source (since it receives only a portion of the light). Hence, the optical component made with all the pillows glued together is more reliable and allows for example to reduce the artifacts due to small occluders. Indeed if one object is in front of one or more lenses, the quality of the refracted image decreases but the image can still be recognized. Using pillows also gives some flexibility to the designer to improve the appearance and the volume occupied by the component. An example with 9 pillows can be found in Figure 10, and the effect of a small occluder.

**Application to color images** Using pillows, we can also target color images. Indeed, we can build one component for each of the Red, Green and Blue channels of an image. If we then place three lights (red, green and blue) in front of each component the 3 images will be perfectly aligned and thus produce the original color image, see the first image of Figure 1.

### 3.3.2 Physical prototypes

We built three lenses (see Figure 11) and two mirrors (see figures 12 and 1) corresponding to collimated light sources. The lenses are fabricated in PMMA (whose index of refraction is 1.49) and the mirrors in aluminum. All the lenses and mirrors have size $100mm \times 100mm$ and were milled in one pass on 3-axis CNC machines after milling the blank. For the CAMERAMAN and HIKARI lenses, we choose to focus the target image on a wall at 2 meters and the target is a square of size $600mm \times 600mm$. For the EINSTEIN'S SIGNATURE lens and the two mirrors, we choose to focus the target image on a wall at 1 meter and the target is a square of size $300mm \times 300mm$. The five components were milled in one pass on the *DMG-DC100V* machine with a $10mm$ radius end-mill.

The milling process is very sensitive. For the lenses, the end-mill is a super finishing ball mill D10, 3 teeth and is following a concentric spiral trajectory. For the mirrors the end-mill is a PCD ball mill hooped D10, 2 teeth and is following a parallel scanning trajectory. We observe that the precision of the milling is not accurate enough: when the collimated light source is traversing a lens or reflected by a mirror with no sandpapering and polishing, the light is dispersed and we do not recognize the target (see Figure 14, second row). We had to sandpaper them by hand before polishing them with a polishing paste. This clearly damages the lens surface: there is a trade-off between removing the artifacts due to the milling and smoothing too much the surface (see Figure 14, rows 3-5), thus damaging its refractive properties. Remark that thanks to the convexity property (see Figure 14, fourth row), the lens surface is quite regular and is more robust to sandpapering.

We can also observe some artifacts in the milling process. For instance, some corrugations are present in the CAMERAMAN lens (Figure 14 first row) and induce some artifacts in the rendered image (Figure 11, first row, at the top of the projected image). We observe that although the image are very contrasted, the projected image are very accurate. The boundary of the target is often slightly blurred and this is due to the boundary of the lens or mirror

where the milling was less good. Our model do not take into account the different wavelengths of the white color and we observe on the boundary of the projected images a small chromatic aberration (the boundary is slightly blue).

### 3.3.3 Limitations

The main limitation of the approach is the fact that we only deal with ideal light sources (a light bulb is for instance neither collimated nor a point). A second remark is that we do not account for self shadowing and internal reflections (although, this is not a problem in the situations we have encountered).

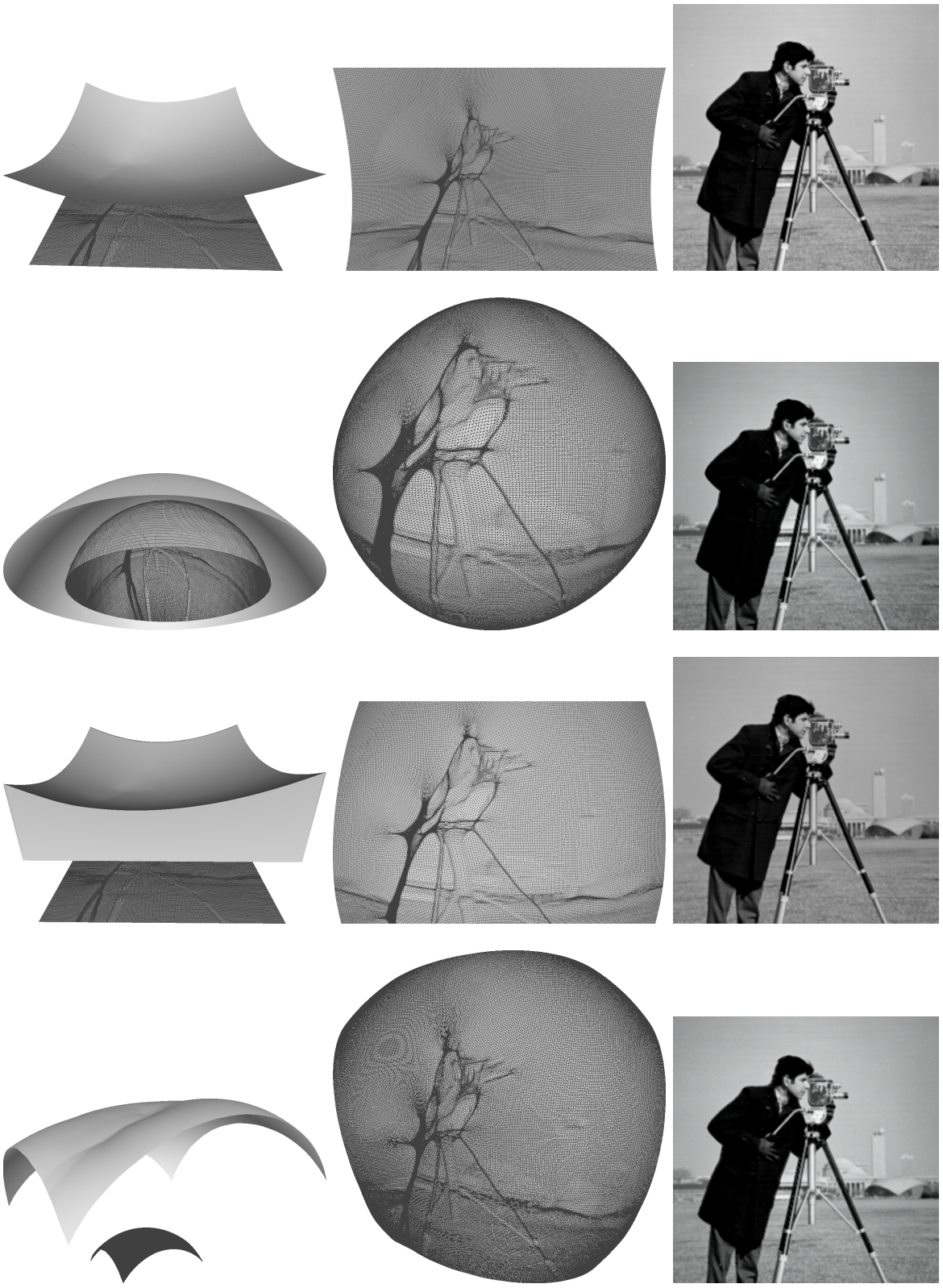

Figure 2 – **Four non-imaging problems solved with Algorithm 5**. *From left to right:* Visibility diagram on $X_\rho$ (wireframe) with the optical component $\mathcal{R}$, Triangulation $\mathcal{R}_T$ of $\mathcal{R}$; forward simulation using LuxRender. *From top to bottom:* Convex Collimated Source Mirror; Concave Point Source Mirror; Concave Collimated Source Lens; Point Source Lens (with the union of ellipsoids).

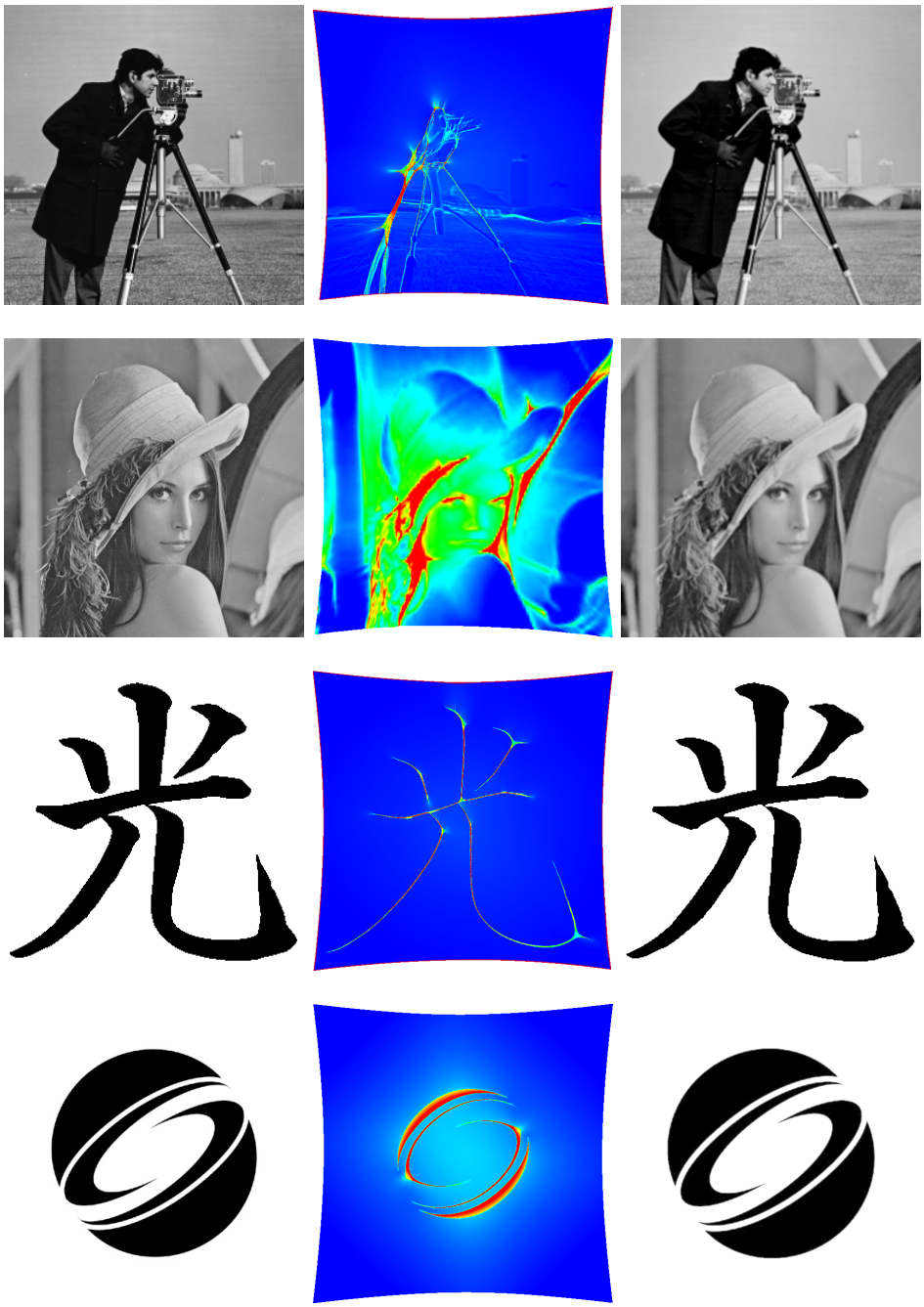

Figure 3 – **Convex Collimated Source Mirror** problem with a uniform light source for different target distributions. *From left to right:* target distribution, mean curvature plot of the mirror, forward simulation using LuxRender. Dimensions of the images from top to bottom: 256x256, 256x256, 300x300, 400x400.

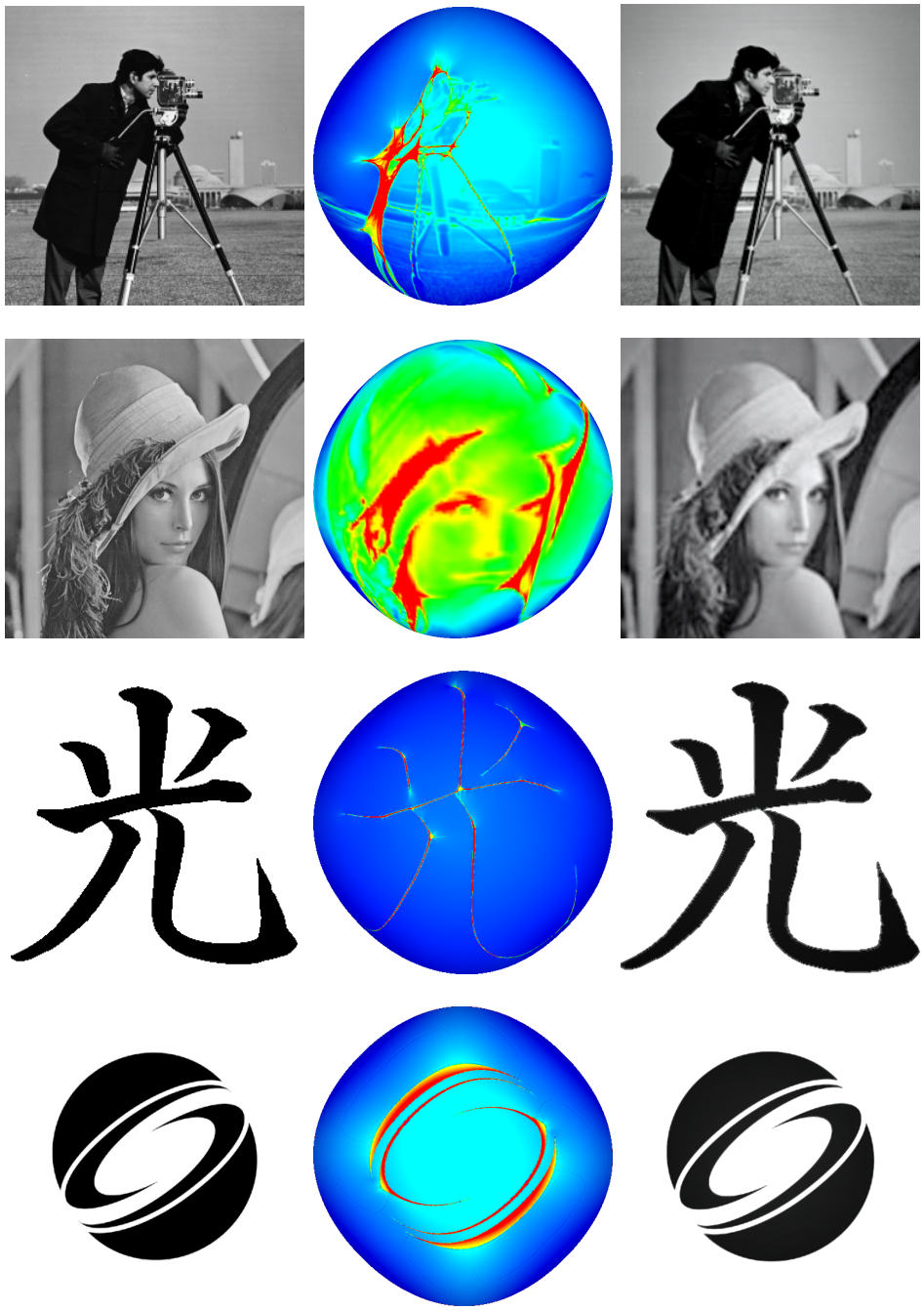

Figure 4 – **Concave Point Source Mirror** problem for a uniform point light source with different target distributions. *From left to right:* target distribution, mean curvature plot of the mirror (top view), forward simulation using LuxRender. Dimensions of the images from top to bottom: 256x256, 256x256, 300x300, 400x400.

Figure 5 – **Triangulation $\mathcal{R}_T$ for a non-uniform light source.** *From left to right*: non uniform collimated light source; mesh of the lens for this non-uniform light; mesh of the lens for a uniform light source.

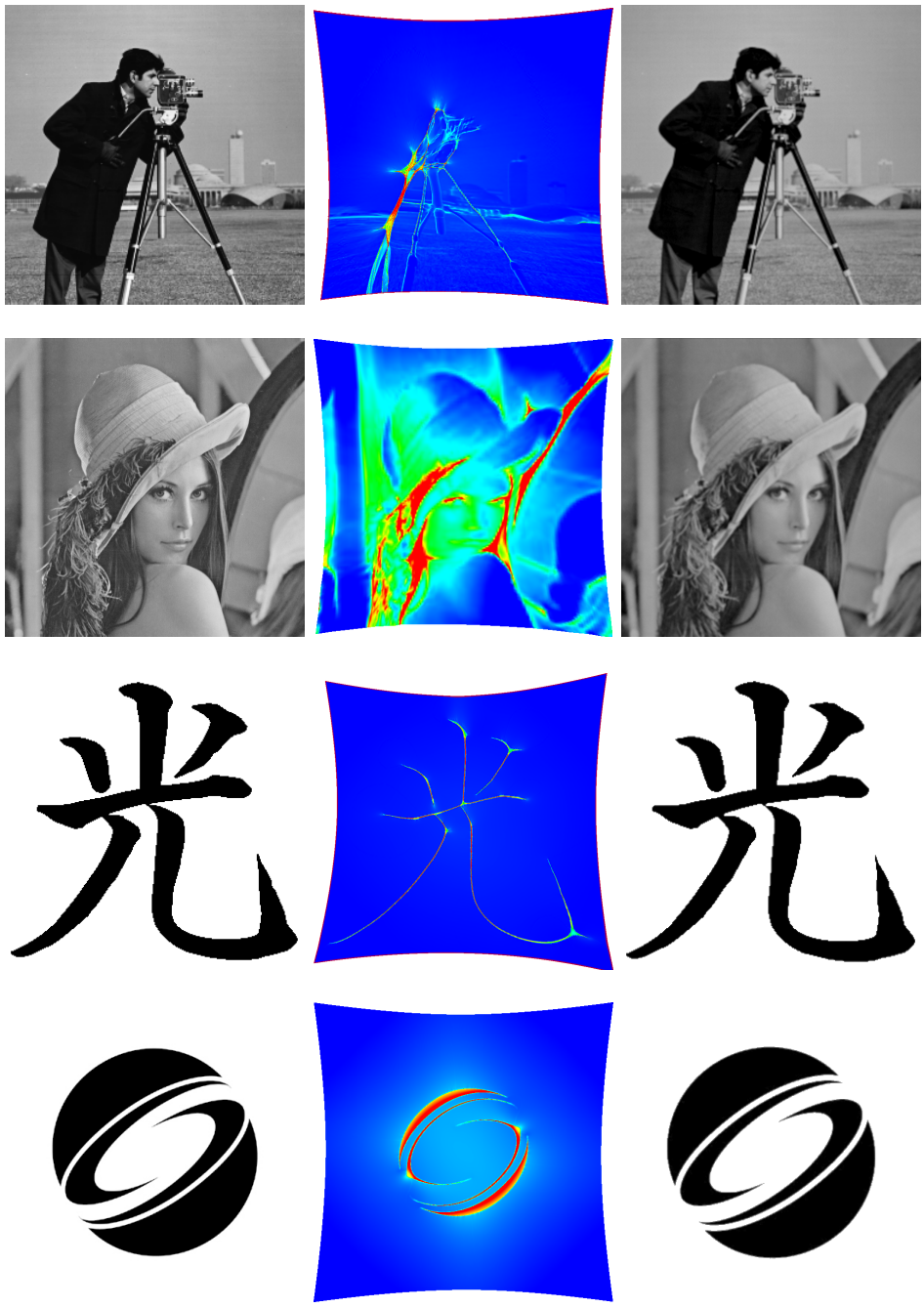

Figure 6 – **Concave Collimated Source Lens** with a uniform light source for different target distributions. *From left to right:* target distribution, mean curvature plot of the lens (top view), forward simulation using LuxRender. Dimensions of the images from top to bottom: 256x256, 256x256, 300x300, 400x400.

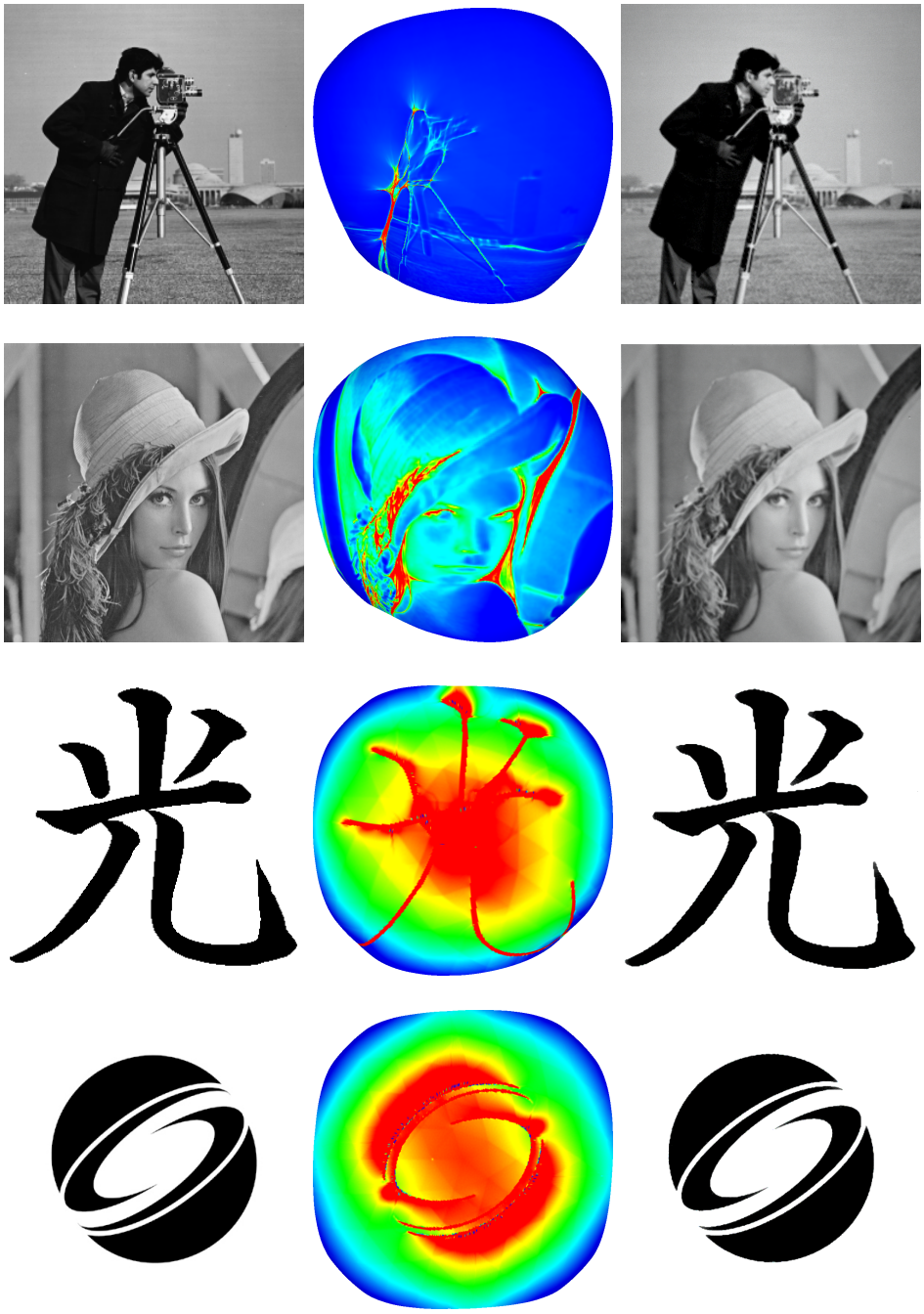

Figure 7 – **Point Source Lens** with a uniform light source for different target distributions. The lens surface is the boundary of the union of filled ellipsoids, hence is not convex, nor concave. *From left to right:* target distribution, mean curvature plot of the lens (top view), forward simulation using LuxRender. Dimensions of the images from top to bottom: 256x256, 256x256, 300x300, 400x400.

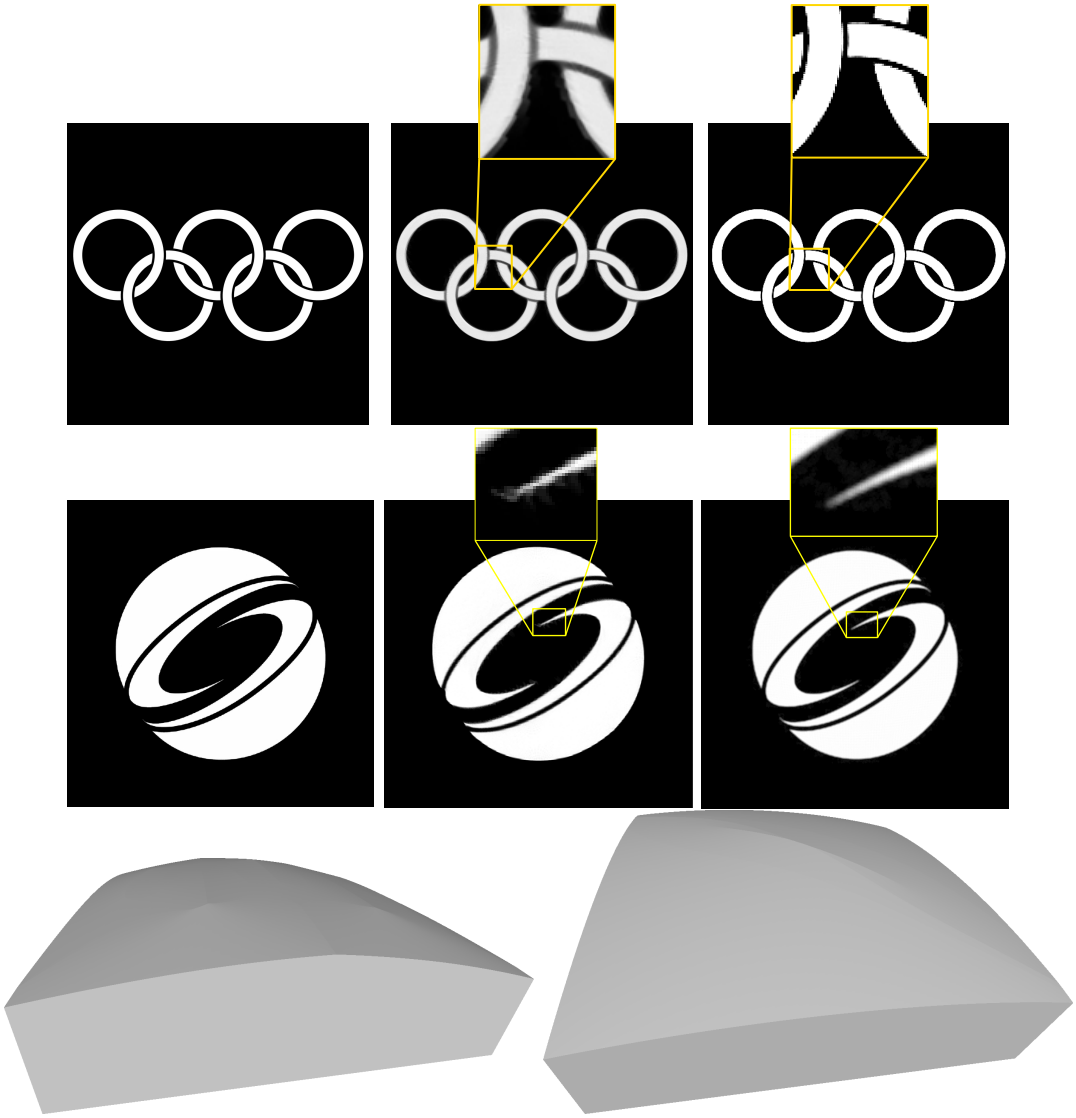

Figure 8 – **Comparison with [Sch+14]** *From left to right:* target distribution; images obtained by [Sch+14] and taken from their article; our forward simulation using LuxRender. Last row: meshes of the two corresponding convex lenses: RINGS (left) and SIGGRAPH (right).

Figure 9 – **Stability under rotation of the lens.** LuxRender renderings in the (CS/Lens) setting for the CAMERAMAN target while rotating the lens with respect to the direction of the collimated light source (0 degree / 5 degrees / 15 degrees).

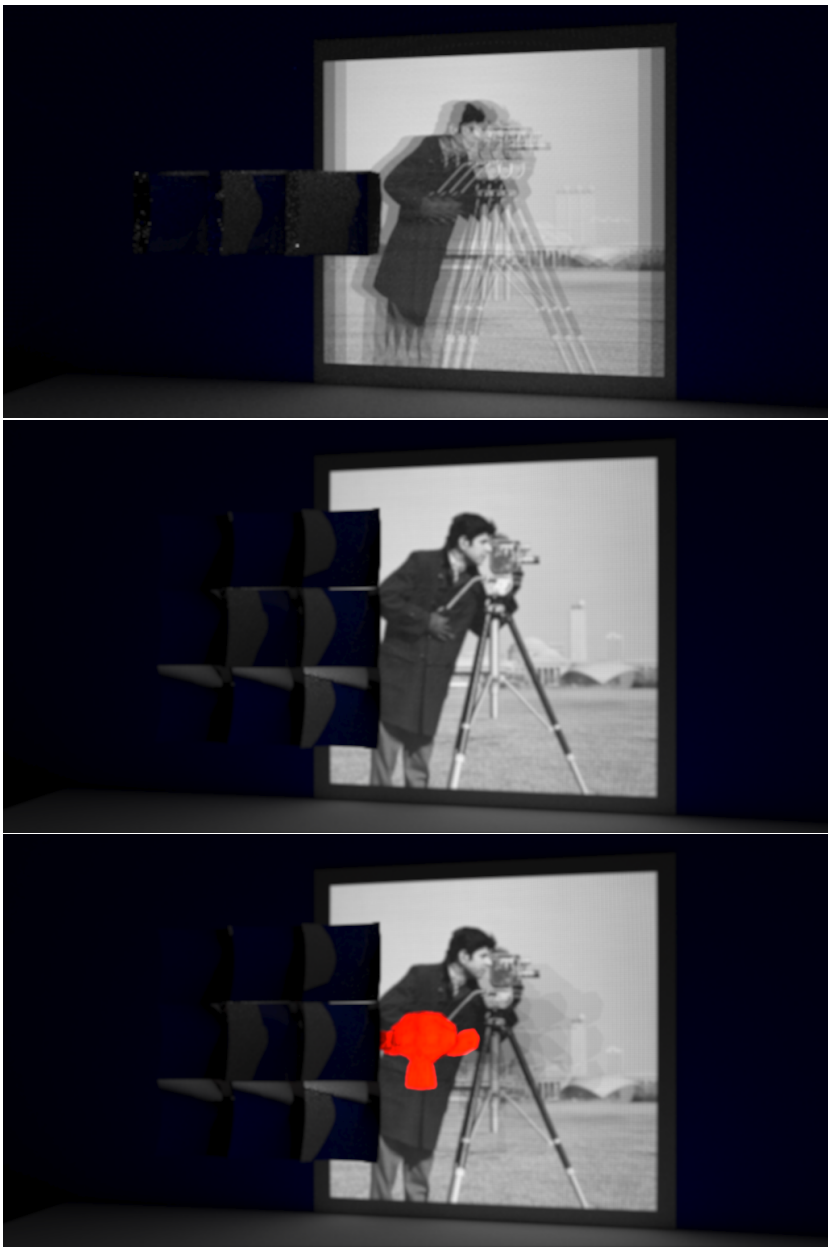

Figure 10 – **Pillows and differences between FF and NF.** *Top*: The lens is composed of three pillows that solve the FF problem. Remark in the last image the shift between the three projected images. *Middle*: A lens composed of nine pillows (each of them solving the NF problem) that refracts a uniform collimated light source; *Bottom*: The same lens with an obstacle in red.

Figure 11 – **Fabricated lenses for a collimated light source.** From left to right: experimental setup, zoom on the target screen. From top to bottom: CAMERAMAN, HIKARI, EINSTEIN'S SIGNATURE targets. Images are focused on a screen at 2 meters for the first two rows and 1 meter for the last one.


Figure 12 – **Fabricated mirror for a collimated light source.** HIKARI target. The image is focused on a screen located at 1 meter.

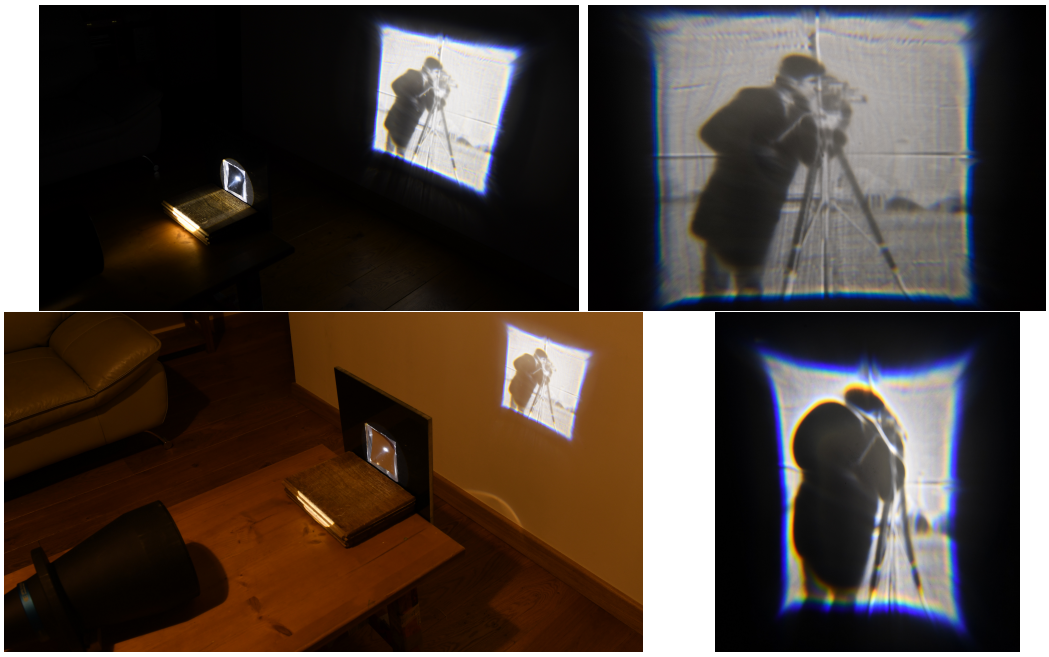

Figure 13 – **Stability with respect to the depth of the focus plane.** The lens of CAMERAMAN is designed to focus at a distance of 2 meters. The target screen is at different depths, top: 1 meter; bottom: 50cm (left), 25cm (right).

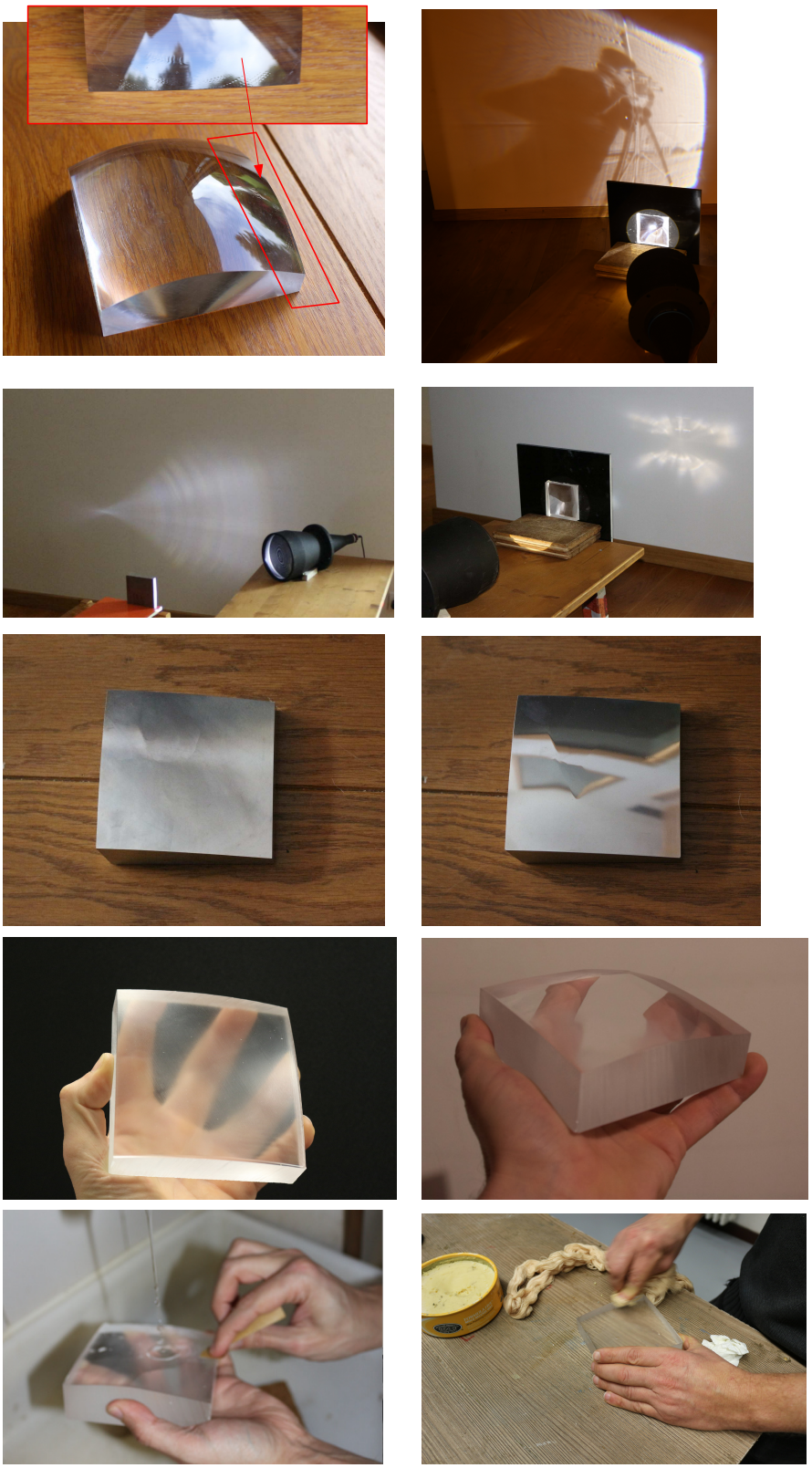

Figure 14 – **Fabrication process.** *First row:* lens for the CAMERAMAN target with a zoom on milling errors (left). The lens is rotated (20 degrees) around the axis of the light source (right). *Second row:* Collimated light projected after reflection or refraction on the screen (rough CAMERAMAN mirror and rough EINSTEIN'S SIGNATURE)) *Third row:* Rough mirror / sandpapered and polished mirror. *Fourth row:* Rough lens / sandpapered and polished lens. *Fifth row:* Sandpapering with water / Polishing by hand

Chapter 4

# Initialization procedures for optimal transport algorithms

## Contents



We now look at an important aspect of computational optimal transport, namely the choice of the initial weight vector $\psi^0$. In the semi-discrete setting, we described the damped Newton's method in Chapter 1 and saw that a necessary condition for the convergence of this algorithm is that all Laguerre cells must have positive mass at every stage of the algorithm and in particular at the beginning. We also saw in the previous chapter that this condition can be hard to satisfy in practice, for instance when the support of the source measure is small (when building *pillows* for example, see Section 3.3). When the cost function is quadratic, it is easy to ensure that all the Laguerre cells are not empty, see Section 4.1.1, but there can still be Laguerre cells that have zero mass. Thus, one must develop other ways of finding correct initial weights.

Let us remark that this initialization problem is also important in other numerical methods for optimal transport and in particular in the discrete setting. Indeed, the auction algorithm (see Algorithm 1) as well as the Sinkhorn-Knopp algorithm (see Algorithm 2) can be very slow in some cases if we don't use scaling techniques. We now look more precisely at the latter case and denote by $\epsilon$ the regularization parameter. In practice, when $\epsilon$ is not too small, the algorithm behaves well and there are no initialization issues as it suffices to take for $\varphi^0$ and $\psi^0$ constant vectors equal to 0. However if $\epsilon$ is small, and in some settings, this choice can lead to numerical instabilities. Consider the case where the source and target measures have disjoint supports that are far away one from the other and suppose that we choose constant vectors

equal to 0 for the initial Kantorovich potentials $\varphi^0$ and $\psi^0$. In that case, there will exist points $x$ and $y$ such that $c(x, y)$ will be large so that the update formula (Sinkhorn/Update) will involve dividing by numbers close to 0. This problem can be seen as an initialization problem and can be handled using scaling techniques on $\epsilon$, see [SS13].

In this chapter, we investigate different initialization strategies to find initial weights in the semi-discrete setting. Let us remark that this is an ongoing work. In Section 4.1, we describe three methods to initialize semi-discrete optimal transport algorithms and in Section 4.2, we illustrate the different methods on numerous examples to show their effectiveness and robustness.

## 4.1 Initialization procedures

We detail here three procedures that can be used to initialize Algorithm 4 for solving the semi-discrete optimal transport for the cost functions used in this thesis. We detail each method and prove its convergence. Numerical examples can be found in Section 4.2. We recall that $\mu$ denotes a probability measure on a source domain $X$ and $\nu$ a probability measure on a target domain $Y$.

### 4.1.1 Local perturbation method

We saw that, for the cost functions considered in this thesis (quadratic cost and cost functions related to inverse problems in optics) the Laguerre diagram can be seen as the intersection between a Power diagram and the support $X$ of the source measure $\mu$, see Section 3.1. We will see how one can leverage this formulation to develop a method based on local perturbations to find a good initialization. The next proposition describes how we can choose weights $\psi^0$ such that all the Laguerre cells $(\mathrm{Lag}_i(\psi^0))_{1\leqslant i\leqslant N}$ are *non-empty* i.e. contains at least one point.

Proposition 56. *Let $X \subset \mathbb{R}^d$ be a compact set, $Y = \{y_1, \dots, y_N\} \subset \mathbb{R}^d$ be a point set and $\psi_i^0 = -d_X(y_i)^2$. Then*

$$\emptyset \neq \{x \in X \mid d_X(y_i) = \|x - y_i\|\} \subset \mathrm{Lag}_i(\psi^0)$$

*where $d_X(y_i) = \min_{x\in X} \|x - y_i\|$.*

*Proof.* Let $i \in \{1, \dots, N\}$ and $x \in X$ such that $d_X(y_i) = \|x - y_i\|$, then for $j \in \{1, \dots, N\}$

$$\begin{aligned}\|x - y_j\|^2 + \psi_j^0 &= \|x - y_j\|^2 - d_X(y_j)^2 \\ &\geqslant d_X(y_j)^2 - d_X(y_j)^2 = 0 = \|x - y_i\|^2 + \psi_i^0.\end{aligned}$$

Thus $x \in \mathrm{Lag}_i(\psi^0)$. □

With this choice of initial weights, some Laguerre cells can still have zero mass. This

happens for instance when the intersection of a Power cell with the source domain $X$ is reduced to a point. We therefore need a stronger result on how to choose initial weights such that the Laguerre cells also have *non empty interiors*. In the remaining of this section, we show how one can use an iterative method to find such weights.

**Setting.** We now suppose that $\mu$ is a regular simplicial measure supported on a simplex soup $X$, as in Definition 1.7. We also suppose that $Y$ is in generic position with respect to $X$, as in Definition 1.9. We denote by $Z(\psi)$ the set of indices of Laguerre cells with zero mass for weights $\psi \in \mathbb{R}^N$ (see the left image of Figure 1 for an illustration). More precisely,

$$Z(\psi) := \{i \in \{1, \dots, N\} \mid G_i(\psi) = 0\}.$$

We also denote by $\mathbf{1}_A$, for a set of indices $A \subset \{1, \dots, N\}$, the vector of $\mathbb{R}^N$ whose $i$-th entry is equal to 1 if $i \in A$ and 0 otherwise.

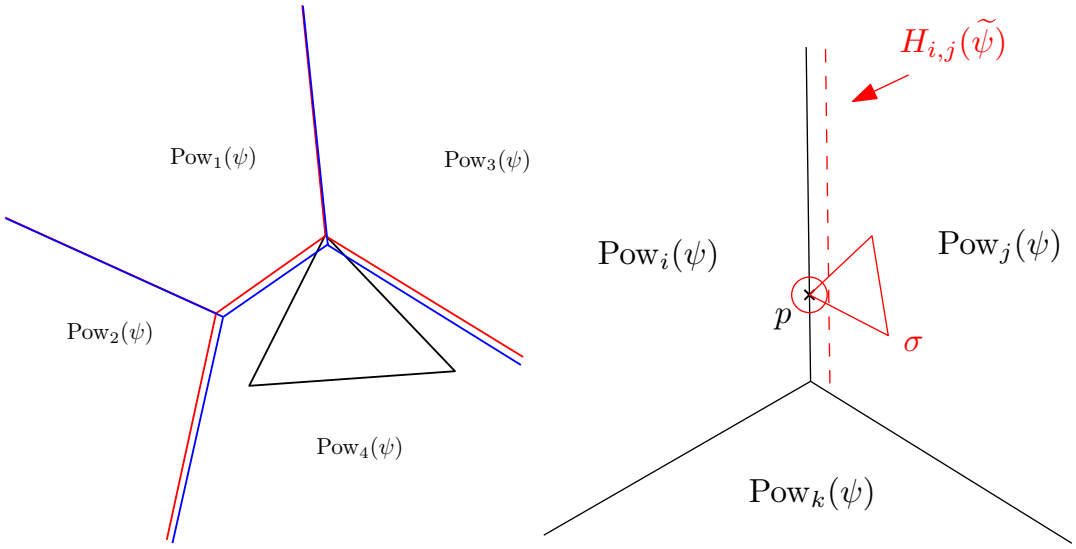


Figure 1 – **Illustration of the Local Perturbation method**. Left: $\mu$ is supported on the black triangle. The boundaries of the Power cells are drawn in red and blue. For the red Power diagram, $Z(\psi^0) = \{1, 2, 3\}$. By applying a small perturbation to $\psi^0$, we obtain weights $\psi^1$ and the corresponding blue Power diagram where $Z(\psi^1) = \{2\}$; Right: illustration of Lemma 59.

**Iterative local perturbation.** We now present an iterative method to find weights $\psi$ such that $Z(\psi) = \emptyset$. Roughly speaking, at each iteration, we show that we can find a small decrement $\epsilon > 0$ such that $\widetilde{\psi} = \psi - \epsilon \mathbf{1}_{Z(\psi)}$ satisfies $\mathrm{Card}(Z(\widetilde{\psi})) < \mathrm{Card}(Z(\psi))$. Repeating this process, we end up with weights for which all the Laguerre cells have positive mass. The effect of one iteration is summarized in the following proposition.

Proposition 57. *Let $\mu$ be a regular simplicial measure supported on a simplex soup $X$, $Y$ be a point cloud in generic position with respect to $X$. Let $\psi \in \mathbb{R}^N$ be a vector such that $Z(\psi) \neq \emptyset$*

*and for every* $i \in \{1, \dots, N\}$, $\mathrm{Pow}_i(\psi) \cap X \neq \emptyset$. *Then there exists* $\epsilon_0 > 0$ *such that for* $\epsilon < \epsilon_0$

$$\mathrm{Card}(Z(\widetilde{\psi})) < \mathrm{Card}(Z(\psi)) \text{ and } \forall i \in \{1, \dots, N\},\ \mathrm{Pow}_i(\widetilde{\psi}) \cap X \neq \emptyset$$

*where* $\widetilde{\psi} = \psi - \epsilon \mathbf{1}_{Z(\psi)}$.

We start by defining, for $i \neq j$, the halfspace $H_{i,j}(\psi)$ by

$$H_{i,j}(\psi) = \{x \in \mathbb{R}^d \mid \|x - y_i\|^2 + \psi_i \leqslant \|x - y_j\|^2 + \psi_j\}.$$

Let us remark that we have $\mathrm{Pow}_i(\psi) \subset H_{i,j}(\psi)$ for any $j \neq i$. We will also need the following lemmas.

Lemma 58. *Let* $i \in \{1, \dots, N\}$, $\epsilon > 0$, *and* $\widetilde{\psi} = \psi - \epsilon \mathbf{1}_{\{i\}}$, *then the distance between* $\partial H_{i,j}(\psi)$ *and* $\partial H_{i,j}(\widetilde{\psi})$ *is given by*

$$d(\partial H_{i,j}(\psi), \partial H_{i,j}(\widetilde{\psi})) = \frac{\epsilon}{2\|y_i - y_j\|}.$$

Let us remark that $H_{i,j}(\psi) \subset H_{i,j}(\widetilde{\psi})$ meaning that $\partial H_{i,j}(\widetilde{\psi})$ moves closer to $y_j$.

Lemma 59. *Let* $\mu$ *be a regular simplicial measure supported on* $X$, $p$ *a point in* $X$ *and* $\psi \in \mathbb{R}^N$, *we define the set of empty cells containing* $p$ *by*

$$Z_p(\psi) = \{i \in Z(\psi) \text{ and } p \in \mathrm{Pow}_i(\psi) \cap X\}.$$

*If* $Z_p(\psi) \neq \emptyset$ *and* $\widetilde{\psi} = \psi - \epsilon \mathbf{1}_{Z(\psi)}$ *for* $\epsilon > 0$, *there exists* $r > 0$ *such that*

$$B(p, r) \subset \bigsqcup_{j \in Z_p(\psi)} \mathrm{Pow}_j(\widetilde{\psi})$$

*where* $B(p, r)$ *denotes the ball of center* $p$ *and radius* $r$ *and* $\bigsqcup$ *the disjoint union.*

An illustration of this lemma can be found in the right image of Figure 1.

*Proof.* We take a point $p \in X$, a weight vector $\psi \in \mathbb{R}^N$, $i \in Z_p(\psi)$, $\epsilon > 0$ and define $\widetilde{\psi} = \psi - \epsilon \mathbf{1}_{Z(\psi)}$. We also take $j \notin Z_p(\psi)$.

By definition, we have $\mathrm{Pow}_i(\psi) = \bigcap_{k \neq i} H_{i,k}(\psi)$, thus Lemma 58 implies that if we choose $r < \frac{\epsilon}{2\|y_i - y_j\|}$ then $B(p, r) \subset H_{i,j}(\widetilde{\psi})$. Furthermore, if we choose

$$r = \frac{\epsilon}{4 \min\limits_{i \in Z_p(\psi), j \notin Z_p(\psi)} \|y_i - y_j\|},$$

then we get

$$\forall j \notin Z_p(\psi),\ B(p, r) \subset \bigcup_{i \in Z_p(\psi)} H_{i,j}(\widetilde{\psi}).$$

Taking the complement that we denote by $X^{\complement}$ (for a set $X \subset \mathbb{R}^d$), we get

$$\forall j \notin Z_p(\psi),\ B(p,r)^{\complement} \supset \left( \bigcup_{i \in Z_p(\psi)} H_{i,j}(\widetilde{\psi}) \right)^{\complement} = \bigcap_{i \in Z_p(\psi)} H_{i,j}(\widetilde{\psi})^{\complement} = \bigcap_{i \in Z_p(\psi)} H_{j,i}(\widetilde{\psi}) \supset \mathrm{Pow}_j(\widetilde{\psi}).$$

This means

$$\bigcap_{j \notin Z_p(\psi)} \mathrm{Pow}_j(\widetilde{\psi}) \subset B(p,r)^{\complement}.$$

Finally, taking the complement again and using the fact that the Power diagram is a partition of $\mathbb{R}^d$, we obtain the intended result i.e.

$$B(p,r) \subset \bigcup_{j \notin Z_p(\psi)} \mathrm{Pow}_j(\widetilde{\psi})^{\complement} = \bigcup_{j \in Z_p(\psi)} \mathrm{Pow}_j(\widetilde{\psi}).$$

□

*Proof of Proposition 57.* We take $\psi \in \mathbb{R}^N$ such that $Z(\psi) \neq \emptyset$ and for every $i$, $\mathrm{Pow}_i(\psi) \cap X \neq \emptyset$. According to Theorem 18, $G$ is of class $\mathcal{C}^1$. In particular it is continuous, so we can find $\epsilon_0 > 0$ such that for $\epsilon < \epsilon_0$ and $\widetilde{\psi} = \psi - \epsilon \mathbf{1}_{Z(\psi)}$, we have: $G_i(\psi) > 0 \implies G_i(\widetilde{\psi}) > 0$. Furthermore, if $i \in Z(\psi)$ then $\mathrm{Pow}_i(\psi) \subset \mathrm{Pow}_i(\widetilde{\psi})$. We conclude that if $\mathrm{Pow}_i(\psi) \cap X$ is not empty then $\mathrm{Pow}_i(\widetilde{\psi}) \cap X$ stays not empty.

We now take $p \in X$ such that $Z_p(\psi) \neq \emptyset$ and $\epsilon < \epsilon_0$, then Lemma 59 gives the existence of $r > 0$ such that

$$B(p,r) \subset \bigsqcup_{j \in Z_p(\psi)} \mathrm{Pow}_j(\widetilde{\psi}).$$

We know that $p$ belongs to a simplex $\sigma$. Since $\mu$ is a regular simplicial measure, the density $\mu_\sigma$ with respect to the $\dim(\sigma)$-dimensional Haussdorf measure on $\sigma$ is bounded from below, so that $\mu(B(p,r) \cap \sigma) > 0$. Thus $0 < \mu(B(p,r) \cap \sigma) \leqslant \sum_{j \in Z_p(\psi)} \mu(\mathrm{Pow}_j(\widetilde{\psi}) \cap \sigma)$. This means that there exists $j \in Z_p(\psi)$ such that $\mu(\mathrm{Pow}_j(\widetilde{\psi}) \cap \sigma) > 0$ i.e. $j \notin Z(\widetilde{\psi})$. Thus we found a Laguerre cell $\mathrm{Lag}_j(\psi)$ that gained mass i.e. $\mathrm{Card}(Z(\widetilde{\psi})) < \mathrm{Card}(Z(\psi))$. □

The Local Perturbation method is detailed in Algorithm 7. Since we have no bounds on $\epsilon$, we use the following heuristic: we assume we are given a maximal decrement and if, for this choice, the number of empty cells does not decrease then we halve it and try again. The next proposition explains the convergence of this algorithm.

Proposition 60. *Let $\mu$ be a regular simplicial measure supported on a simplex soup $X$, $Y \subset \mathbb{R}^d$ a point cloud in generic position with respect to $X$. Then, Algorithm 7 converges in a finite number of steps.*

*Proof.* We simply iterate the result of Proposition 57. Indeed, the proposition tells us that we can find a sequence of weights $(\psi^k)_k$ such that $(\mathrm{Card}(Z(\psi^k)))_k$ is strictly decreasing while maintaining the fact that all the Laguerre cells are not empty. □

**Algorithm 7:** Local perturbation method

**Input** A regular simplicial measure $\mu$,
A finitely supported measure $\nu = \sum_{1 \leqslant i \leqslant N} \nu_i \delta_{y_i}$,
A maximal decrement $\epsilon > 0$,
A family of weights $\psi^0 \in \mathbb{R}^N$ such that $\forall i,\ \mathrm{Lag}_i(\psi^0) \neq \emptyset$
**Output** A family of weights $\psi$ such that $\forall i,\ G_i(\psi) > 0$.
**Initialization** $\psi \leftarrow \psi^0, p \leftarrow \mathrm{Card}(Z(\psi)^0)$
**while** $p > 0$ **do**
  $p_{cur} \leftarrow N$
  $\epsilon_{cur} \leftarrow \epsilon$
  **while** $p_{cur} > p$ **do**
    $\psi_{cur} \leftarrow \psi - \epsilon_{cur} \mathbf{1}_{Z(\psi_{cur})}$
    $p_{cur} \leftarrow \mathrm{Card}(Z(\psi_{cur}))$
    **if** $p_{cur} < p$ **then**
      $\psi \leftarrow \psi_{cur}$
      $p \leftarrow \mathrm{Card}(Z(\psi))$
      **break**
    **else**
      $\epsilon_{cur} \leftarrow \epsilon_{cur}/2$
    **end**
  **end**
**end**

### 4.1.2 Interpolation method

We now detail another method which can work with every cost function as long as an algorithm to solve optimal transport exist for this cost. This covers all the problems studied in this thesis: quadratic cost and cost functions related to inverse problems in optics. The core of the method is a linear interpolation between two source measures: the initial source measure $\mu$ and the normalized Lebesgue measure $\lambda_P$ supported on a *bigger* domain $P$ containing $X \cup Y$. Note that this measure can be defined on a set that has a higher dimension than the support of $\mu$. For instance, when $X$ is a surface embedded in $\mathbb{R}^3$, then $\lambda_P$ can be the normalized Lebesgue measure supported on a bounding cube containing $X \cup Y$; when $X$ is a 2D domain, $\lambda_P$ can be the normalized Lebesgue measure supported on a sufficiently large rectangle containing $X \cup Y$. We define for $t \in [0, 1]$ an interpolating measure

$$\mu_t = t\lambda_P + (1 - t)\mu$$

in such a way that $\mu_1 = \lambda_P$ and $\mu_0 = \mu$. It is easy to see that choosing constant weights is sufficient to ensure that all the Laguerre cells for $\mu_1 = \lambda_P$ have positive mass.

The method then consists in iteratively decreasing $t$ and solving the optimal transport problem between $\mu_t$ and $\nu$, refer to Algorithm 8 for more details. In this algorithm, we denote by `SOLVE_OT`$(\mu, \nu, \eta, \psi^0)$ a function that solves the optimal transport between $\mu$ and $\nu$ with a numerical error $\eta$ starting from weights $\psi^0$ (which can for instance be Algorithm 4).

**Algorithm 8:** Linear interpolation method

**Input** A measure $\mu$ supported on $X$,
A finitely supported measure $\nu = \sum_{1 \leqslant i \leqslant N} \nu_i \delta_{y_i}$,
A set $P$ that contains $X \cup Y$,
A tolerance $\eta > 0$,
Weights $\psi^0 \in \mathbb{R}^N$ such that $\forall i,\ \lambda_P(\mathrm{Lag}_i(\psi^0)) > 0$.
**Output** A family of weights $\psi$ such that $\forall i,\ G_i(\psi) > 0$.
**Initialization** $k := 0$ and $t := 1$
**while** $t > t_{min} := \min_i \nu_i - \eta, 0$ **do**
  Define $\mu_t = t\lambda_P + (1-t)\mu$
  $\psi^{k+1} \leftarrow$ `SOLVE_OT`$(\mu_t, \nu, \eta, \psi^k)$
  $t \leftarrow t/2$
  $k \leftarrow k + 1$
**end**

The stopping criterion is justified by the following proposition.

Proposition 61. *For $t < \min_i \nu_i - \eta$ where $\eta$ is the numerical error of Algorithm 4 then for every $i \in \{1, \dots, N\}$, $\mu(\mathrm{Lag}_i(\psi)) > 0$.*

To prove this proposition, we will need the next lemma.

Lemma 62. *Let $\mu$ be a probability measure defined on $X$, $\lambda_P$ the normalized Lebesgue measure on $P$ and $\mu_t = t\lambda_P + (1-t)\mu$, for $t \in ]0, 1[$ be a probability measure on $P$. Then*

$$\forall A \subset P,\ \mu_t(A) > t \implies \mu(A) > 0.$$

*Proof.* We take $0 < t < 1$ and suppose that $\mu_t(A) > t$ and $\mu(A) = 0$. Then since $\mu_t(A) = t\lambda_P(A)$, we get $\lambda_P(A) > 1$ which is not possible since $\lambda_P$ is a probability measure over $P$. Thus $\mu(A) > 0$. □

*Proof of Proposition 61.* For a numerical error $\eta$ and $i \in \{1, \dots, N\}$, at the end of the optimal transport algorithm (see Algorithm 4) between two measures $\mu_t$ and $\nu = \sum_{i=1}^N \nu_i \delta_{y_i}$, we have $|\mu_t(\mathrm{Lag}_i(\psi)) - \nu_i| \leqslant \eta$. Thus $\mu_t(\mathrm{Lag}_i(\psi)) \geqslant \nu_i - \eta \geqslant \min_i \nu_i - \eta$. So if $t < \min_i \nu_i - \eta$ then $\mu_t(\mathrm{Lag}_i(\psi)) > t$ and we can apply Lemma 62 with $A = \mathrm{Lag}_i(\psi)$, so that $\mu(\mathrm{Lag}_i(\psi)) > 0$. □

### 4.1.3 Rescaling method

The third method we look at is a classical trick, that we call *rescaling*. Let us remark that this procedure only works for the quadratic cost. The idea is to first translate and/or rescale the target point cloud $Y$ such that it is included in the support of the source measure $X$. More precisely, we find a translation vector $t \in \mathbb{R}^d$ and a scalar $\lambda > 0$ such that the point cloud $Z = \{z_1, \dots, z_N\}$ defined by $z_i = \lambda y_i + t$ is such that $z_i \in X$ and we look at the optimal transport between $(X, \mu)$ and $(Z, \nu)$ for the quadratic cost.

In the following, we denote by $\overline{X}$ the *centroid* of a domain $X \subset \mathbb{R}^d$. If $X = \{x_1, \dots, x_N\}$ is finite then $\overline{X} = \frac{1}{N}\sum_{i=1}^N x_i$; if it is a continuous domain equipped with some measure $\mu$ then $\overline{X} = \frac{\int_X x \mathrm{d}\mu(x)}{\int_X \mathrm{d}\mu(x)}$. We also denote by $\mathrm{bbox}(X)$ an axis-aligned bounding box of a domain $X \subset \mathbb{R}^d$. Finally, the diameter of a compact set $X$ is denoted by $\mathrm{diam}(X) = \max_{x,y\in X} \|x - y\|$.

We propose two choices for the translation vector $t$ and the scaling factor $\lambda$:

1. $(t, \lambda) = \left(\overline{X} - \overline{Y}, \frac{\mathrm{diam}(X)}{\mathrm{diam}(Y)}\right)$,
2. $(t, \lambda) = \left(\overline{\mathrm{bbox}(X)} - \overline{\mathrm{bbox}(Y)}, \frac{\mathrm{vol}(\mathrm{bbox}(X))}{\mathrm{vol}(\mathrm{bbox}(Y))}\right)$.

REMARK 63. *These choices are not guaranteed to work for any domain $X$ and point cloud $Y$ as they heavily depend on the geometry of $X$ and $Y$. For instance, when $X$ is disconnected or non-convex, such values for $t$ and $\lambda$ does not always guarantee that $Z = \lambda Y + t \subset X$. A possible solution (at least when both sets have the same underlying dimension) would be to find the biggest box enclosed in $X$ and find the parameters $t$ and $\lambda$ to fit $Y$ inside this box.*

The following proposition details the relation between the Laguerre cells of the two point clouds $Y$ and $Z$.

PROPOSITION 64. *Given a point set $Y = \{y_1, \dots, y_N\} \subset \mathbb{R}^d$, $\lambda > 0$ and $t \in \mathbb{R}^d$, if we define the point set $Z = \{z_1, \dots, z_N\} \subset \mathbb{R}^d$ by $z_i = \lambda y_i + t$, then there is the following relation between the Laguerre cells of the two sets*

$$\mathrm{Lag}_{z_i}(\psi) = \mathrm{Lag}_{y_i}(\varphi)$$

*where $\varphi_i = \frac{\psi_i}{\lambda} + 2\langle t \mid y_i\rangle + (\lambda - 1)\|y_i\|^2$ for $i \in \{1, \dots, N\}$.*

*Proof.* Let us take $x \in \mathrm{Lag}_{z_i}(\psi)$ for a vector of weights $\psi \in \mathbb{R}^N$ and $j \in \{1, \dots, N\}$, we have

$$\begin{aligned}
\|z_j\|^2 - \|z_i\|^2 &= \langle z_j - z_i \mid z_j + z_i\rangle \\
&= \lambda\langle y_j - y_i \mid \lambda(y_j + y_i) + 2t\rangle \\
&= \lambda^2(\|y_j\|^2 - \|y_i\|^2) + 2\lambda\langle t \mid y_j - y_i\rangle \\
&= \lambda\left[\|y_j\|^2 - \|y_i\|^2 + (\lambda - 1)(\|y_j\|^2 - \|y_i\|^2) + 2\langle t \mid y_j - y_i\rangle\right].
\end{aligned}$$

Thus

$$\begin{aligned}
x \in \mathrm{Lag}_{z_i}(\psi) &\iff \forall j,\ \|x - z_i\|^2 + \psi_i \leqslant \|x - z_j\|^2 + \psi_j \\
&\iff \forall j,\ -2\langle x \mid z_i - z_j\rangle \leqslant \psi_j - \psi_i + \|z_j\|^2 - \|z_i\|^2 \\
&\iff \forall j,\ -2\langle x \mid y_i - y_j\rangle \leqslant \varphi_j - \varphi_i + \|y_j\|^2 - \|y_i\|^2 \\
&\iff \forall j,\ \|x - y_i\|^2 + \varphi_i \leqslant \|x - y_j\|^2 + \varphi_j \iff x \in \mathrm{Lag}_{y_i}(\varphi).
\end{aligned}$$

□

An easy consequence of this proposition is the following result.

COROLLARY 65. *$T_\psi$ is an optimal transport map between $(X, \mu)$ and $(Z, \nu)$ for the quadratic cost if and only if $T_\varphi$ is an optimal transport map between $(X, \mu)$ and $(Y, \nu)$ for the quadratic cost where $\varphi$ are the weights defined in Proposition 64.*

It then suffices to solve the optimal transport between $(X, \mu)$ and $(Z, \nu)$. Solving the optimal transport between $(X, \mu)$ and $(Z, \nu)$ can be easier since $Z \subset X$. Indeed, in this case, choosing constant weights is sufficient to ensure that the Laguerre cells are non-empty. If there are still empty cells, we can apply one of the two previous methods. In practice, we observe that there are less empty cells with $(Z, \nu)$ than with $(Y, \nu)$.

## 4.2 Numerical results

In this section, we compare and illustrate the pros and cons for the different methods we presented in the previous section. We recall that the work presented in this chapter is an ongoing work.

### 4.2.1 Illustrations of the convergence of the three methods

We start by illustrating the different methods when both the source and target measures are supported on 2D domains. In Figure 2, we display the evolution of the number of empty Laguerre cells for the LOCAL PERTURBATION method in different settings, the maximal decrement is chosen to be $\epsilon = 10^{-2}$ and the number of iterations corresponds to the number of times the outer loop of Algorithm 7 is executed. One can observe that the number of empty Laguerre cells is strictly decreasing. Let us also remark that the settings considered in Figure 2 are bad candidates for the LOCAL PERTURBATION method because a lot of points in $Y$ are "projected" onto the same point on $X$. More precisely, there are a lot of points $y_i$ such that $\mathrm{Pow}_i(\psi^0) \cap X$ is the same point (where $\psi^0$ are the weights detailed in Proposition 56). Furthermore, in the last row of Figure 2, the support of $\mu$ is disconnected. In particular, $\mu$ is not regular and is not covered by Proposition 57, meaning that we do not have guarantees on the convergence of the algorithm in this case. In particular, one Laguerre cell could be "stuck" inside one connected component.

In Figure 3, we detail the INTERPOLATION method on one example where the source measure is supported on $[-1, 0] \times [-1, 1]$ and the target point cloud is composed of random points sampled in $[-1, 1]^2$. We see that the number of empty Laguerre cells is also strictly decreasing. At the end of the procedure, all the Laguerre cells intersect the support of the source measure. In Figure 4, we display the initial, final Laguerre cells for different settings for the INTERPOLATION method. Finally, in Figure 5, we illustrate how the RESCALE method works on one example. Let us stress that the heuristics used to choose $t$ and $\lambda$ can not work when the support of the source measure is disconnected, non-convex or more generally when the two sets do not have the same topology.

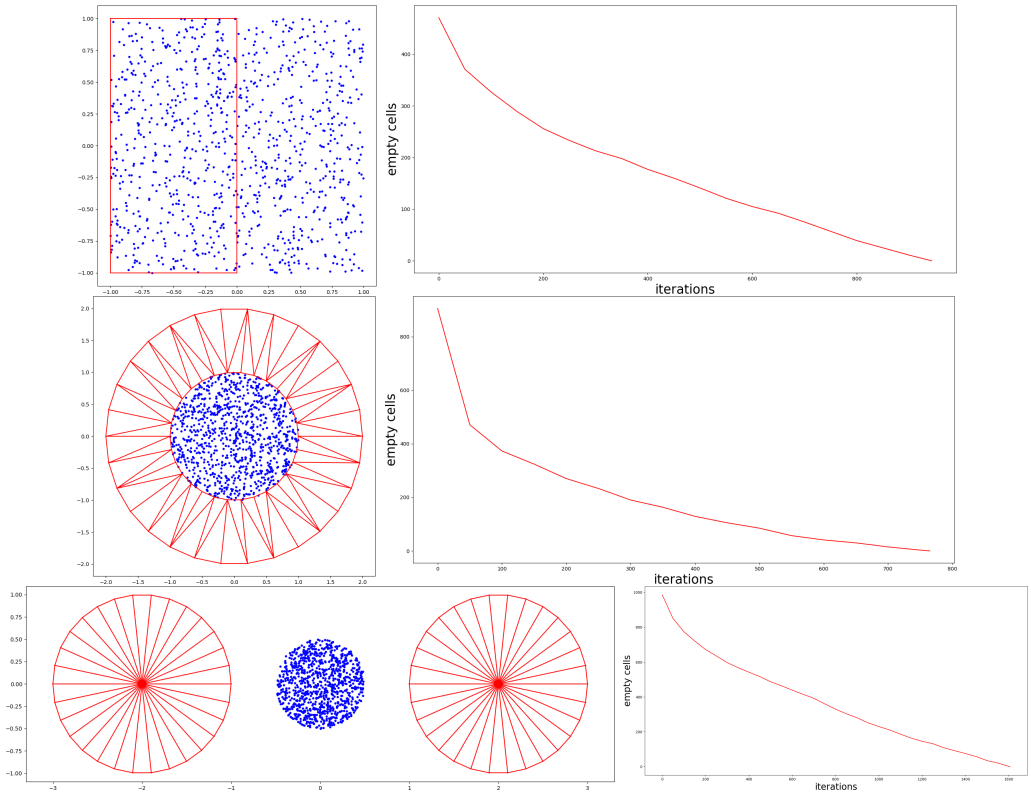


Figure 2 – **Local Perturbation method in 2D.** Both measures are uniform. Left: setting (the source measure is supported on the red domain, the target measure on the blue point cloud); Right: evolution of the number of empty Laguerre cells (for $\epsilon = 10^{-2}$), an iteration correponds to an update of the vector $\psi$. From top to bottom: left part of a square; ring; two disks.

We now show that we can also use these initialization procedures when the source measure is supported on a triangulated surface in $\mathbb{R}^3$ and the target measure is a weighted point cloud in $\mathbb{R}^3$. In Figure 6, we display the evolution of the number of empty Laguerre cells for three examples of triangulated surfaces for the Local Perturbation method. Let us remark that the number of empty Laguerre cells is strictly decreasing. We can also note that there is less iterations than in Figure 2. This is because for a triangulated surface, the number of points $y_i$ that are "projected" on the same point is finite thus avoiding the drawbacks present in the 2D examples. As in Figure 2, for the last example, since the support of $\mu$ is not connected we do not have guarantees of convergence but it still works in practice. In Figure 7, we display the initial and final Laguerre cells obtained during the Interpolation procedure. Let us note that for a triangulated surface, the Rescale method can not work since it is impossible to move, with only a translation vector and a scaling factor, a point cloud onto a surface.

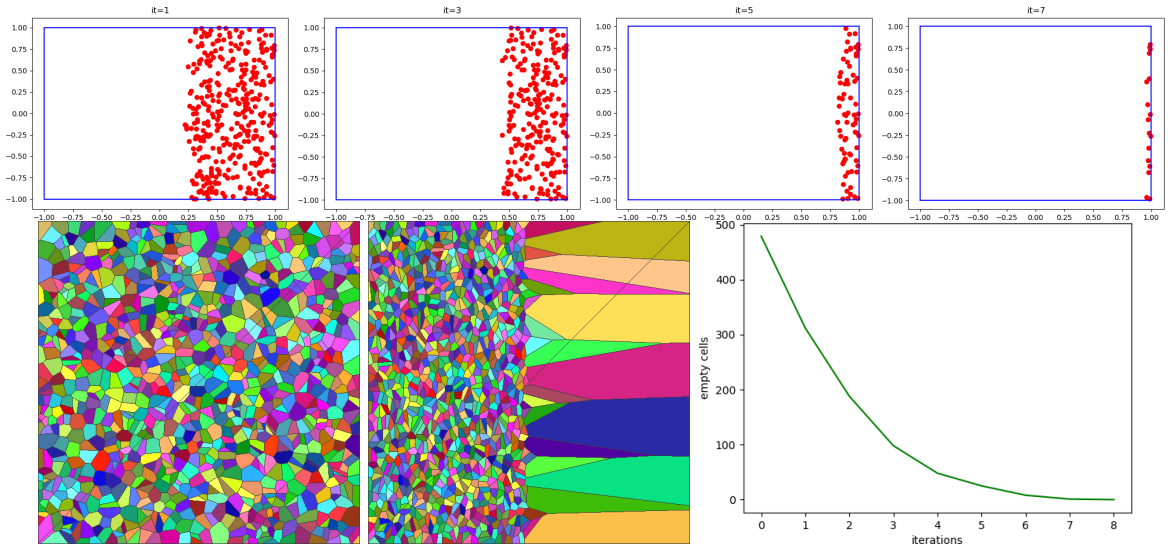


Figure 3 – **Detailed illustration of the INTERPOLATION method in 2D.** The setting corresponds to the first row of Figure 2. Top: we show for different iterations (1, 3, 5 and 7) the points corresponding to empty Laguerre cells in red. Bottom: initial Laguerre cells; final Laguerre cells; evolution of the number of empty Laguerre cells.

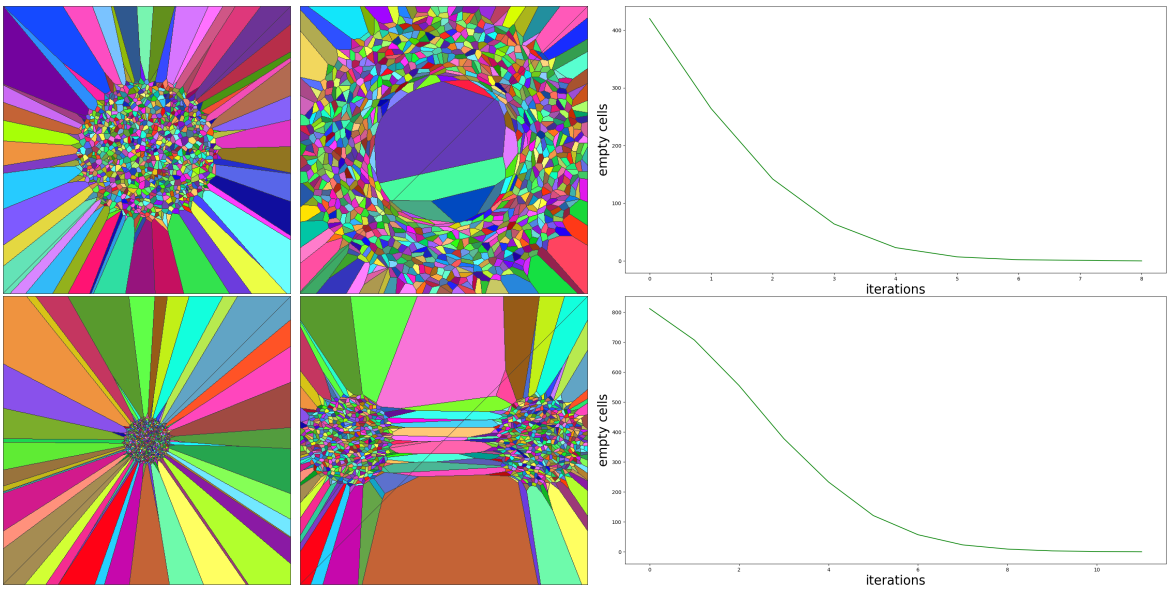


Figure 4 – **Illustrations of the INTERPOLATION method in 2D.** The settings correspond to the second and third rows of Figure 2. From left to right: initial Laguerre cells; final Laguerre cells; evolution of the number of empty Laguerre cells.

### 4.2.2 Performance

We now briefly look at the performance of the three methods we presented on the semi-discrete. All tests have been done on a laptop with a 3.6 GHz i7 CPU. We compare the running times of

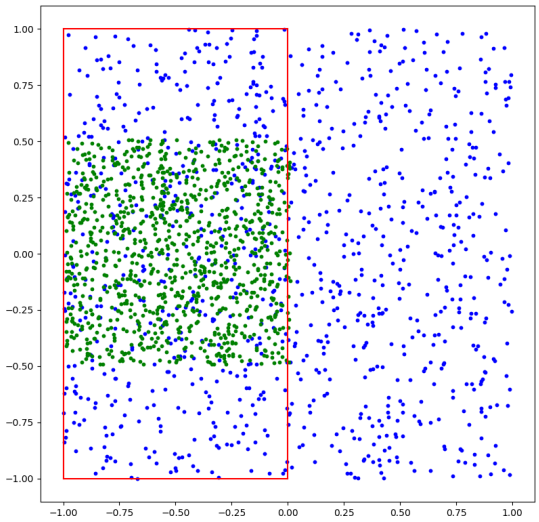


Figure 5 – **Rescale method in 2D.** The source measure is supported on the red domain, the target measure on the blue point cloud and the rescaled point cloud is in green.

the three methods as well as the number of secondary optimal transport problems that we need to solve. The results can be found in Table 4.1. We can also see that in most of the examples, the Local Perturbation method takes more time than the Interpolation one. This is due to the fact that the number of iterations is much greater in the Local Perturbation method (because $\epsilon$ is small), and also because the Laguerre diagram is computed more times: 4104 vs 106 times for instance in the setting corresponding to the first row. Let us also stress that the Rescale is inapplicable in most of the settings because of the simplicity of the choice of the translation vector $t$ and the scaling factor $\lambda$ that can not handle complex geometries.

| Setting | Local Perturbation | | Interpolation | | Rescale | |
|---|---|---|---|---|---|---|
| | time (s) | # OT | time (s) | # OT | time (s) | # OT |
| square | 75 | 0 | 4 | 9 | 0.4 | 1 |
| ring | 56 | 0 | 8 | 9 | X | X |
| two disks | 175 | 0 | 4 | 11 | X | X |
| hemisphere | 52 | 0 | 20 | 9 | X | X |
| torus | 17 | 0 | 33 | 7 | X | X |
| two spheres | 14 | 0 | 9 | 5 | X | X |

Table 4.1 – Running time and number of secondary optimal transport problems using the three initialization procedures. An X denotes that the method can not be used.

### 4.2.3 Application to non-imaging optics

We now show how we can use the methods described in the previous section and in particular the Interpolation method to solve non-imaging optics problems. In particular, we will look at the case of pillows detailed in Section 3.3. The setting is the following: we have a uniform collimated light source supported on $X = [-1, 0] \times [-1, 1]$ the target is the cameraman image

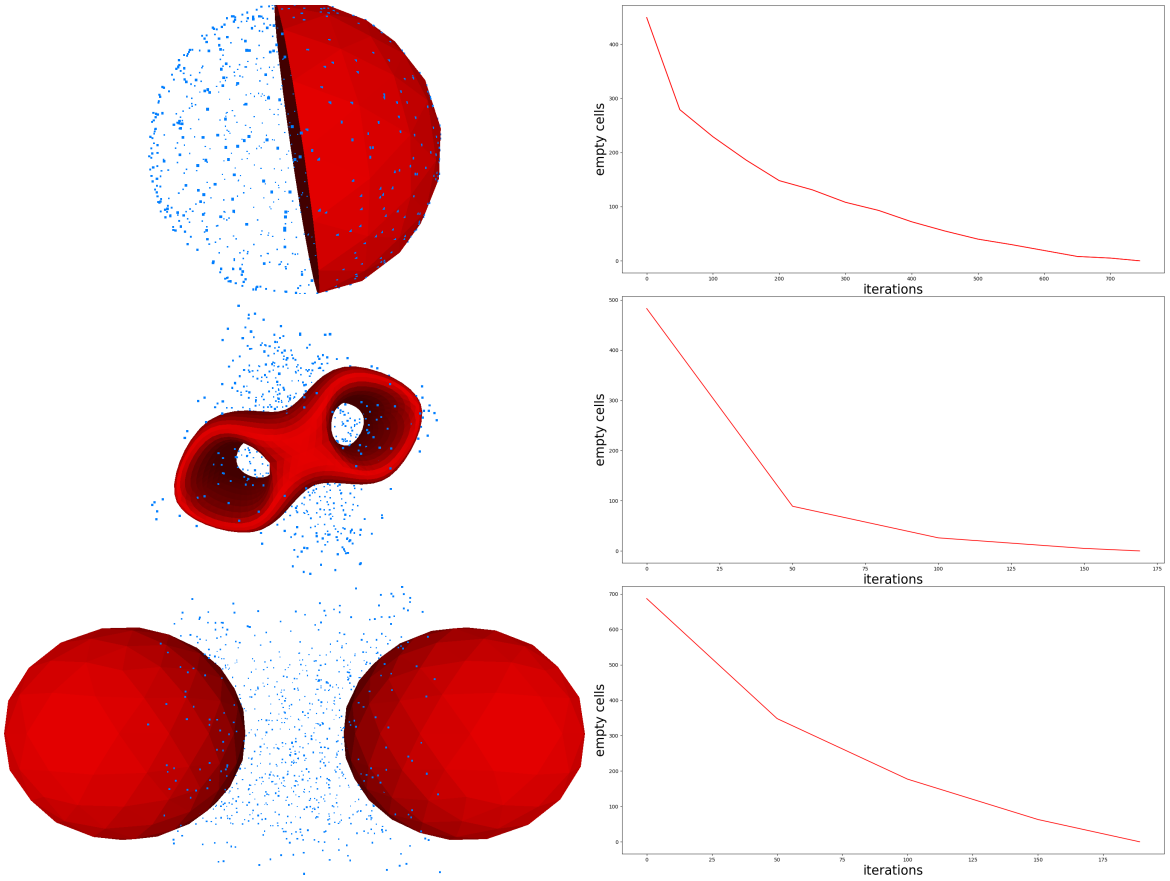


Figure 6 – **Illustrations of the Local Perturbation method for a triangulated surface in $\mathbb{R}^3$.** Left: setting (the source measure is supported on the red domain and the target one on the blue point cloud); Right: evolution of the number of empty Laguerre cells. From top to bottom: $\mu$ is supported on the right hemisphere and the target point cloud is sampled on the whole sphere $\mathbb{S}^2$; $\mu$ is supported on a torus and the target point cloud is sampled inside $[-1,1]^3$; $\mu$ is supported on two disconnected spheres and the target point cloud is sampled inside $[-1,1]^3$.

discretized with 1024 Dirac masses. In this case, we can not choose constant weights to get good initial weights since with this choice, approximately half of the Power cells do not intersect $X$, see the top row of Figure 8. One can remark that at the end of the Interpolation procedure all the Laguerre cells intersect the support of the light source.

### 4.2.4 Pros and cons

We finish this section by summarizing the pros and cons of each method in Table 4.2. We make some observations on this table:

- the Local Perturbation method is relatively fast since we only have to compute squared distances (to get $\psi^0$) and the areas of the Laguerre cells i.e. $G_i(\psi)$ for $i \in$

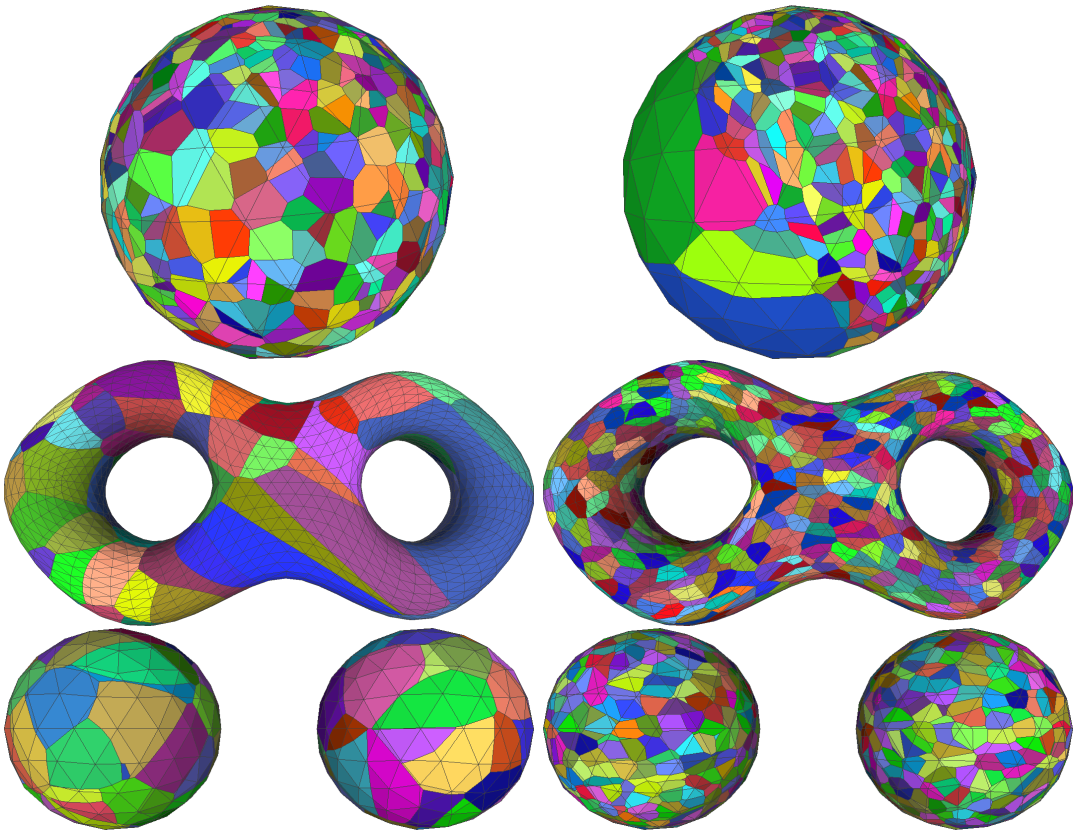

Figure 7 – **Illustrations of the INTERPOLATION method for a triangulated surface in $\mathbb{R}^3$.** We display the initial and the final Laguerre cells. Both the source and target measures are uniform for the three examples of Figure 6.

$\{1, \dots, N\}$;

- the INTERPOLATION method is potentially slow since there can be many different optimal transport problems to solve but can handle any cost function;
- the RESCALE method is fast since there is only one optimal transport to solve but it can not be used in most of the cases because of the geometry of the source and target domains.

| Method | Pros | Cons |
|---|---|---|
| LOCAL PERTURBATION | fast; guarantees | quadratic cost; many iterations |
| INTERPOLATION | any cost; guarantees | slow |
| RESCALE | fast; guarantees | quadratic cost; same dimension |

Table 4.2 – Pros and cons of each initialization method. "guarantees" means that we have theoretical guarantees on the convergence of the method.

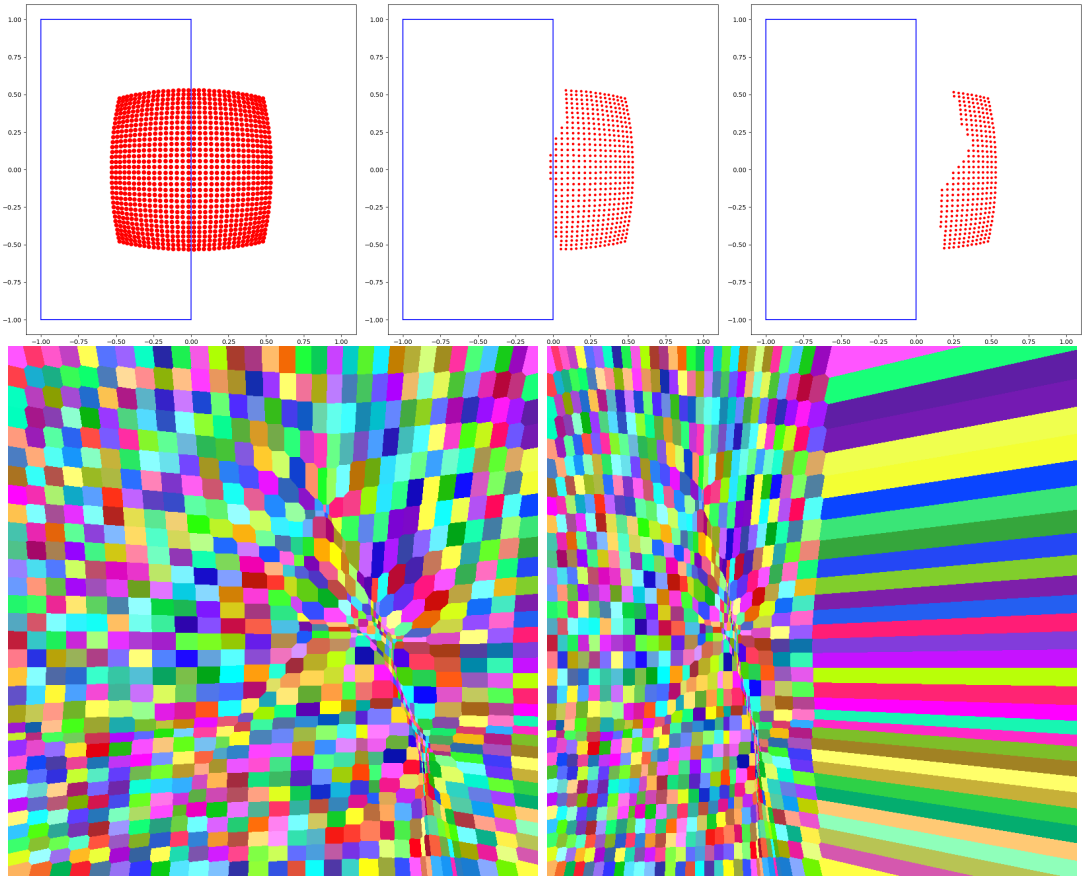

Figure 8 – **Illustrations of the Interpolation method for a non-imaging optics problem.** Top: in blue the support of the light source and in red the points $p_i$ corresponding to empty Laguerre cells; Bottom: initial and final Laguerre cells;

# Conclusion and perspectives

In this thesis, we looked at the relations between optimal transport and optical component design. We showed that many inverse problems arising in optics can be seen as an instance of an optimal transport problem between the light source represented by a probability measure supported on a planar or spherical domain and a target illumination at infinity represented by a probability measure on the sphere, for different cost functions. This allowed us to use a very efficient numerical method by combining the damped Newton's algorithm with results coming from computational geometry and create a common framework for solving optical component design problems. We also looked at another important aspect of optimal transport algorithms namely the choice of the initial weight vectors.

In the future, we first want to see if we can extend the initialization strategies mentioned in the last chapter to discrete optimal transport. Another perspective would be to look at the so-called *Generated Jacobian Equations* and try to see if some of the methods developed in this thesis can be adapted in this setting. This could be interesting since such equations encompass other optical component design problems such as the near-field setting. The main difficulty is the additional non-linearity added by such equations which makes the study of solutions more complex. We also want to see if we can adapt the methods developed in this thesis for other non-convex inverse problems. For instance, we want to consider the optical component design problems we presented where we replace the ideal light source by an extended one. The first step would be to find a setting in which this problem is well-posed. Another example is in seismic imaging and more precisely in the *full waveform inversion* framework. In this setting, one wants to measure the misfit between a predicted and recorded seismic signal. It has recently been shown that choosing a Wasserstein distance to measure the misfit between the two signals "convexifies" the error function and thus helps in minimizing it. We want to see if introducing semi-discrete optimal transport can improve the minimization of the error. The main difficulty is because these signals are by nature very oscillatory, the initial weights need to be chosen carefully since we can easily end up in a case where many Laguerre cells have zero mass. We believe that we can develop some heuristics to speed up the initialization procedures in this specific setting.

# Bibliography


[AHA98] Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov. "Minkowski-type theorems and least-squares clustering." In: *Algorithmica* 20.1 (1998), pp. 61–76 (cit. on pp. 13, 20, 24).

[And+15] Julien André et al. "Far-field reflector problem under design constraints." In: *International Journal of Computational Geometry & Applications* 25.02 (2015), pp. 143–162 (cit. on p. 54).

[BB00] Jean-David Benamou and Yann Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem." In: *Numerische Mathematik* 84.3 (2000), pp. 375–393 (cit. on p. 19).

[BCM16] Jean-David Benamou, Francis Collino, and Jean-Marie Mirebeau. "Monotone and Consistent discretization of the Monge-Ampere operator." In: *Mathematics of computation* 85.302 (2016), pp. 2743–2775 (cit. on p. 20).

[Ber81] D. P. Bertsekas. "A new algorithm for the assignment problem." In: *Mathematical Programming* 21.1 (1981), pp. 152–171 (cit. on p. 17).

[Ber88] Dimitri P Bertsekas. "The auction algorithm: A distributed relaxation method for the assignment problem." In: *Annals of operations research* 14.1 (1988), pp. 105–123 (cit. on p. 17).

[BFO12] Jean-David Benamou, Brittany D Froese, and Adam M Oberman. "Numerical solution of the optimal transportation problem via viscosity solutions for the Monge-Ampere equation." In: *CoRR, abs/1208.4873* 2 (2012) (cit. on pp. 17, 20).

[BFO14] Jean-David Benamou, Brittany D Froese, and Adam M Oberman. "Numerical solution of the optimal transportation problem using the Monge–Ampere equation." In: *Journal of Computational Physics* 260 (2014), pp. 107–126 (cit. on p. 20).

[BM92] Paul J Besl and Neil D McKay. "Method for registration of 3-D shapes." In: *Robotics-DL tentative*. International Society for Optics and Photonics. 1992, pp. 586–606 (cit. on p. 49).

[Bre91] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions." In: *Communications on pure and applied mathematics* 44.4 (1991), pp. 375–417 (cit. on p. 16).

[Car+17] Guillaume Carlier et al. "Convergence of entropic schemes for optimal transport and gradient flows." In: *SIAM Journal on Mathematical Analysis* 49.2 (2017), pp. 1385–1418 (cit. on p. 18).

[CBC77] F. Cork, D.F. Bettridge, and P.C. Clarke. *Method of and mixture for aluminizing a metal surface*. US Patent 4,009,146. 1977 (cit. on p. 55).

[CCSM10] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. "Boundary measures for geometric inference." In: *Foundations of Computational Mathematics* 10.2 (2010), pp. 221–240 (cit. on p. 62).

[CGA16] CGAL Project. *CGAL User and Reference Manual*. 4.9. CGAL Editorial Board, 2016 (cit. on p. 43).

[CGH08] Luis A Caffarelli, Cristian E Gutiérrez, and Qingbo Huang. "On the regularity of reflector antennas." In: *Annals of mathematics* 167.1 (2008), pp. 299–323 (cit. on pp. 4, 10, 57).

[CGS10] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. "From Knothe's transport to Brenier's map and a continuation method for optimal transport." In: *SIAM Journal on Mathematical Analysis* 41.6 (2010), pp. 2554–2576 (cit. on p. 39).

[CKO99a] L Caffarelli, S. Kochengin, and VI Oliker. "On the numerical solution of the problem of reflector design with given far-field scattering data." In: *Contemporary Mathematics* 226 (1999), pp. 13–32 (cit. on pp. 57, 64).

[CKO99b] Luis A Caffarelli, Sergey A Kochengin, and Vladimir I Oliker. "Problem of Reflector Design with Given Far-Field Scattering Data." In: *Monge Ampère Equation: Applications to Geometry and Optimization: NSF-CBMS Conference on the Monge Ampère Equation, Applications to Geometry and Optimization, July 9-13, 1997, Florida Atlantic University*. Vol. 226. American Mathematical Soc. 1999, p. 13 (cit. on p. 23).

[CMT15] Pedro Machado Manhães de Castro, Quentin Mérigot, and Boris Thibert. "Far-field reflector problem and intersection of paraboloids." In: *Numerische Mathematik* (2015), pp. 1–23 (cit. on pp. 13, 24, 57, 73, 82, 86).

[CO08] L. A. Caffarelli and V. Oliker. "Weak solutions of one inverse problem in geometric optics." In: *Journal of Mathematical Sciences* 154.1 (2008), pp. 39–49 (cit. on pp. 2, 8, 57, 59, 63, 67).

[Cut13] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport." In: *Advances in Neural Information Processing Systems*. 2013, pp. 2292–2300 (cit. on pp. 13, 18).

[DH15] Gerwin Damberg and Wolfgang Heidrich. "Efficient freeform lens optimization for computational caustic displays." In: *Optics express* 23.8 (2015), pp. 10224–10232 (cit. on p. 54).

[Dig+14] Julie Digne et al. "Feature-preserving surface reconstruction and simplification from defect-laden point sets." In: *Journal of mathematical imaging and vision* 48.2 (2014), pp. 369–382 (cit. on p. 27).

[FCR09] Florian R Fournier, William J Cassarly, and Jannick P Rolland. "Designing freeform reflectors for extended sources." In: *Nonimaging Optics: Efficient Design for Illumination and Solar Concentration VI*. Vol. 7423. International Society for Optics and Photonics. 2009, p. 742302 (cit. on p. 58).

[FDL10] Manuel Finckh, Holger Dammertz, and Hendrik PA Lensch. "Geometry construction from caustic images." In: *Computer Vision–ECCV 2010*. Springer, 2010, pp. 464–477 (cit. on pp. 54, 56).

[Fey+17] Jean Feydy et al. "Optimal Transport for Diffeomorphic Registration." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 291–299 (cit. on p. 19).

[FFL16] Zexin Feng, Brittany D Froese, and Rongguang Liang. "Freeform illumination optics construction following an optimal transport map." In: *Applied optics* 55.16 (2016), pp. 4301–4306 (cit. on p. 57).

[Fro+15] Charlie Frogner et al. "Learning with a Wasserstein loss." In: *Advances in Neural Information Processing Systems*. 2015, pp. 2053–2061 (cit. on p. 13).

[GH09] Cristian E Gutiérrez and Qingbo Huang. "The refractor problem in reshaping light beams." In: *Archive for rational mechanics and analysis* 193.2 (2009), pp. 423–443 (cit. on pp. 2, 8, 54, 57, 70, 71, 73, 75, 77).

[GH14] Cristian E Gutiérrez and Qingbo Huang. "The near field refractor." In: *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*. Vol. 31. 4. Elsevier. 2014, pp. 655–684 (cit. on p. 58).

[GK17] Nestor Guillen and Jun Kitagawa. "Pointwise estimates and regularity in geometric optics and other generated Jacobian equations." In: *Communications on Pure and Applied Mathematics* 70.6 (2017), pp. 1146–1220 (cit. on pp. 2, 8, 58).

[GM17] Thomas O Gallouët and Quentin Mérigot. "A Lagrangian scheme à la Brenier for the incompressible Euler equations." In: *Foundations of Computational Mathematics* (2017), pp. 1–31 (cit. on p. 13).

[GM96] Wilfrid Gangbo and Robert J McCann. "The geometry of optimal transportation." In: *Acta Mathematica* 177.2 (1996), pp. 113–161 (cit. on p. 20).

[GO03] T. Glimm and V. Oliker. "Optical design of single reflector systems and the Monge–Kantorovich mass transfer problem." In: *Journal of Mathematical Sciences* 117.3 (2003), pp. 4096–4108 (cit. on p. 57).

[Goe+12] Fernando de Goes et al. "Blue noise through optimal transport." In: *ACM Transactions on Graphics* 31.6 (2012), p. 171 (cit. on pp. 13, 24–26, 45, 57).

[Goe+15] Fernando de Goes et al. "Power particles: an incompressible fluid solver based on power diagrams." In: *ACM Trans. Graph.* 34.4 (2015), pp. 50–1 (cit. on pp. 25, 26).

[GT13] Cristian E Gutiérrez and Federico Tournier. "The parallel refractor." In: *From Fourier Analysis and Number Theory to Radon Transforms and Geometry*. Springer, 2013, pp. 325–334 (cit. on pp. 2, 8, 57, 58).

[Gu+13] Xianfeng Gu et al. "Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge-Ampere equations." In: *arXiv preprint arXiv:1302.5472* (2013) (cit. on p. 59).

[Gut12] Cristian E Gutiérrez. *The Monge—Ampère Equation*. Vol. 42. Springer Science & Business Media, 2012 (cit. on p. 20).

[Kan58] Leonid Kantorovitch. "On the translocation of masses." In: *Management Science* 5.1 (1958), pp. 1–4 (cit. on p. 15).

[Kis+12] Thomas Kiser et al. *Architectural caustics–controlling light with geometry.* Citeseer, 2012 (cit. on pp. 1, 7, 56).

[KMT16] Jun Kitagawa, Quentin Mérigot, and Boris Thibert. "A Newton algorithm for semi-discrete optimal transport." In: *arXiv preprint arXiv:1603.05579* (2016) (cit. on pp. 21, 22, 24, 27, 28, 31, 39, 45, 71–73, 84, 86).

[KO97] Sergey A Kochengin and Vladimir I Oliker. "Determination of reflector surfaces from near-field scattering data." In: *Inverse Problems* 13.2 (1997), p. 363 (cit. on p. 58).

[Lév15] Bruno Lévy. "A numerical algorithm for $L^2$ semi-discrete optimal transport in 3D." In: *ESAIM: Mathematical Modelling and Numerical Analysis* 49.6 (2015), pp. 1693–1715 (cit. on pp. 13, 24, 43, 86).

[Llo82] Stuart Lloyd. "Least squares quantization in PCM." In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137 (cit. on p. 56).

[LN89] D. C. Liu and J. Nocedal. "On the limited memory BFGS method for large scale optimization." In: *Mathematical programming* 45.1-3 (1989), pp. 503–528 (cit. on p. 45).

[Loe09] Grégoire Loeper. "On the regularity of solutions of optimal transportation problems." In: *Acta mathematica* 202.2 (2009), pp. 241–283 (cit. on pp. 28, 71, 72).

[Loe11] Grégoire Loeper. "Regularity of optimal maps on the sphere: The quadratic cost and the reflector antenna." In: *Archive for rational mechanics and analysis* 199.1 (2011), pp. 269–289 (cit. on p. 73).

[LR05] Grégoire Loeper and Francesca Rapetti. "Numerical solution of the Monge–Ampère equation by a Newton's algorithm." In: *Comptes Rendus Mathematique* 340.4 (2005), pp. 319–324 (cit. on p. 19).

[LS17] Bruno Levy and Erica Schwindt. "Notions of optimal transport theory and how to implement them on a computer." In: *arXiv preprint arXiv:1710.02634* (2017) (cit. on pp. 18, 20).

[Man+17] Manish Mandad et al. "Variance-Minimizing Transport Plans for Inter-surface Mapping." In: *ACM Transactions on Graphics* 36 (2017), p. 14 (cit. on p. 27).

[Mér11] Quentin Mérigot. "A multiscale approach to optimal transport." In: *Computer Graphics Forum.* Vol. 30. 5. Wiley Online Library. 2011, pp. 1583–1592 (cit. on pp. 13, 24, 57).

[Mér13] Quentin Mérigot. "A comparison of two dual methods for discrete optimal transport." In: *Geometric science of information.* Springer, 2013, pp. 389–396 (cit. on p. 18).

[Mir15] Jean-Marie Mirebeau. "Discretization of the 3D Monge-Ampere operator, between Wide Stencils and Power Diagrams." In: *arXiv preprint arXiv:1503.00947* (2015) (cit. on pp. 24, 39).

[MMT18a] Quentin Mérigot, Jocelyn Meyron, and Boris Thibert. "An algorithm for optimal transport between a simplex soup and a point cloud." In: *SIAM Journal on Imaging Sciences* 11.2 (2018), pp. 1363–1389 (cit. on pp. 4, 6, 10, 12, 26).

[MMT18b] Jocelyn Meyron, Quentin Mérigot, and Boris Thibert. "Light in Power: A General and Parameter-free Algorithm for Caustic Design." In: *SIGGRAPH Asia 2018 Technical Papers*. ACM. 2018, p. 224 (cit. on pp. 6, 12, 79).

[Mon81] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais." In: *Histoire de l'Académie Royale des Sciences de Paris* (1781) (cit. on p. 13).

[MTW05] Xi-Nan Ma, Neil S Trudinger, and Xu-Jia Wang. "Regularity of potential functions of the optimal transportation problem." In: *Archive for rational mechanics and analysis* 177.2 (2005), pp. 151–183 (cit. on pp. 27, 28, 53, 58, 71).

[OP89] VI Oliker and LD Prussner. "On the numerical solution of the equation and its discretizations, I." In: *Numerische Mathematik* 54.3 (1989), pp. 271–293 (cit. on pp. 13, 23).

[Pap+11] Marios Papas et al. "Goal-based Caustics." In: *Computer Graphics Forum*. Vol. 30. 2. Wiley Online Library. 2011, pp. 503–511 (cit. on p. 56).

[Pog64] Aleksei Vasil'evich Pogorelov. *Monge-Ampere equations of elliptic type*. P. Noordhoff, 1964 (cit. on pp. 2, 8, 59).

[PP05] Gustavo Patow and Xavier Pueyo. "A survey of inverse surface design from light transport behavior specification." In: *Computer Graphics Forum*. Vol. 24. 4. Wiley Online Library. 2005, pp. 773–789 (cit. on pp. 1, 7, 56).

[Pri+13] CR Prins et al. "A numerical method for the design of free-form reflectors for lighting applications." Technische Universiteit Eindhoven. 2013 (cit. on pp. 57, 60, 64).

[Ral81] Louis B Rall. "Automatic differentiation: Techniques and applications." In: *Springer* (1981) (cit. on p. 86).

[San15] Filippo Santambrogio. "Optimal transport for applied mathematicians." In: *Birkäuser, NY* (2015) (cit. on pp. 13, 15, 16, 59, 62).

[Sch+14] Yuliy Schwartzburg et al. "High-contrast computational caustic design." In: *ACM Transactions on Graphics (TOG)* 33.4 (2014), p. 74 (cit. on pp. 1, 7, 56, 57, 88, 97).

[SNA17] Maxime Sainlot, Vincent Nivoliers, and Dominique Attali. "Restricting Voronoi diagrams to meshes using corner validation." In: *Computer Graphics Forum*. Vol. 36. 5. Wiley Online Library. 2017, pp. 81–91 (cit. on pp. 43, 86).

[Sol+15] Justin Solomon et al. "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains." In: *ACM Transactions on Graphics (TOG)* 34.4 (2015), p. 66 (cit. on p. 19).

[SS13] Bernhard Schmitzer and Christoph Schnörr. "A hierarchical approach to optimal transport." In: *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer. 2013, pp. 452–464 (cit. on pp. 18, 104).

[Su+13] Zhengyu Su et al. "Area preserving brain mapping." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 2235–2242 (cit. on p. 24).

[Tho+16] Matthew Thorpe et al. "A Transportation $L^p$ Distance for Signal Analysis." In: *arXiv preprint arXiv:1609.08669* (2016) (cit. on p. 27).

[TPG16] Guillaume Tartavel, Gabriel Peyré, and Yann Gousseau. "Wasserstein loss for image synthesis and restoration." In: *SIAM Journal on Imaging Sciences* 9.4 (2016), pp. 1726–1755 (cit. on p. 13).

[Tru12] Neil S Trudinger. "On the local theory of prescribed Jacobian equations." In: *arXiv preprint arXiv:1211.4661* (2012) (cit. on p. 58).

[Vil03] C. Villani. *Topics in optimal transportation*. Vol. 58. American Mathematical Soc., 2003 (cit. on p. 16).

[Vil09] C. Villani. *Optimal transport: old and new*. Springer Verlag, 2009 (cit. on pp. 13, 17, 21, 72).

[Wan04] X.J. Wang. "On the design of a reflector antenna II." In: *Calculus of Variations and Partial Differential Equations* 20.3 (2004), pp. 329–341 (cit. on p. 57).

[Wan96] Xu-Jia Wang. "On the design of a reflector antenna." In: *Inverse problems* 12.3 (1996), p. 351 (cit. on pp. 2, 8, 73).

[Wey+09] Tim Weyrich et al. "Fabricating microgeometry for custom surface reflectance." In: *ACM Transactions on Graphics (TOG)* 28.3 (2009), p. 32 (cit. on p. 56).

[WMB+05] Roland Winston, Juan C Miñano, Pablo G Benitez, et al. *Nonimaging optics*. Academic Press, 2005 (cit. on pp. 1, 7, 55).

[Yue+12] Yonghao Yue et al. "Pixel art with refracted light by rearrangeable sticks." In: *Computer Graphics Forum*. Vol. 31. 2pt3. Wiley Online Library. 2012, pp. 575–582 (cit. on p. 56).

[Yue+14] Yonghao Yue et al. "Poisson-based continuous surface generation for goal-based caustics." In: *ACM Transactions on Graphics (TOG)* 33.3 (2014), p. 31 (cit. on pp. 1, 7, 56, 57).