The Optimal Transport problem and its application to dissipative partial differential equations

Gabriel Martine La Boissonière

Department of Mathematics and Statistics

McGill University

Montréal, Québec

April 2015

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

DEDICATION

I dedicate this thesis to Nastasija, without whom I wouldn't have made it so far, in both my personal and academic lives.

ACKNOWLEDGEMENTS

I would first like to offer my gratitude to my supervisor, Rustum Choksi, for his academic and financial support during my time at McGill. His supervision allowed me to grow into a well rounded researcher by answering my questions, from the most basic analytical subtleties to more complex PDE problems, and by offering me many great opportunities.

I am also grateful to Jean-Christophe Nave for his research ideas, suggestions and counseling, and to Adam Oberman for our multiple discussions and for his financial support. I also acknowledge generous support from the Fields Institute, L'Institut des Sciences Mathématiques and the McGill Department of Mathematics and Statistics.

Je remercie finalement ma mère, Zack, Tiberiu et Nastasija. Vous savez bien sûr combien je compte sur vous.

ABSTRACT

The optimal transport problem has found many applications in mathematics and physical sciences, in part due to the importance of the Wasserstein gradient flow. To appreciate this importance, we first introduce the optimal transport problem in the formulations of Monge and Kantorovich and present a numerical approach to the discrete equivalent problem. This numerical procedure is used to visualize optimal transport plans. We then prove the result of Gangbo and McCann that, under standard assumptions, there exists a unique optimal transport plan to problems involving strictly convex cost functions. This background allows us to build the Wasserstein gradient flow from its discretization, the Jordan-Kinderlehrer-Otto scheme. We use this procedure to justify that the Fokker-Planck equation is the Wasserstein gradient flow of a physical energy functional and conclude by briefly presenting similar applications to other dissipative equations.

RÉSUMÉ

La théorie du transport optimal est aujourd'hui appliquée dans plusieurs domaines des sciences physiques et mathématiques. Cette omniprésence s'explique en partie par la puissance de la descente de gradient par la métrique de Wasserstein. Pour apprécier l'importance de cette technique, on introduit le problème de Monge et de Kantorovich ainsi qu'une approche numérique et visuelle au problème discret. On montre le résultat de Gangbo et McCann qu'il n'existe qu'une unique solution aux problèmes avec un coût strictement convexe. On construit ensuite la descente de gradient de Wasserstein à partir de sa discrétisation, la méthode de Jordan, Kinderlehrer et Otto. On établit ainsi que l'équation de Fokker-Planck est la descente de gradient de Wasserstein d'une fonctionnelle avec une interprétation manifestement physique. On conclut par un bref sommaire des applications de cette descente de gradient aux équations de dissipation.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
Introduction

We begin with a very brief exposition of the history of the optimal transport problem then detail the organization of this thesis.

1.1 A Brief History of the Optimal Transport Problem

The first scientist to investigate ideas related to the optimal transport problem was the French mathematician Gaspard Monge in his 1781 paper entitled *Mémoire sur la théorie des déblais et des remblais* [1]. He considers the problem, hereafter called the Monge problem, of displacing a quantity of materials from several quarries (*déblais*) to construction sites (*remblais*) as "efficiently" as possible. Historically, this context was inspired by the military problem of efficiently assigning sand quarries to the construction of fortifications, a situation sketched in figure 1-1. To make the notion of efficiency precise, it is necessary to attribute a cost to the displacement of materials, usually as a function of the distance over which it must be carried. Monge simply assumed this cost to be the Euclidian distance between the material's initial position in a quarry and its final position in a construction site. Such an assignment from initial to final position, a "transport plan", is optimal if the cost required to effect the transportation is minimal over

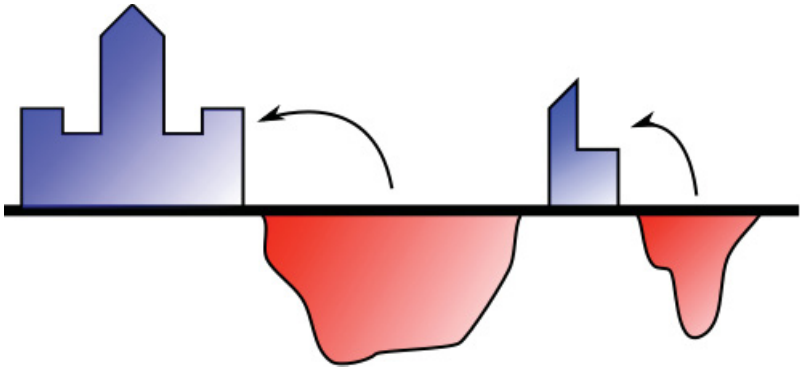

Figure 1–1: If sand is the material to be carried from quarries (red) to construction sites (blue), Monge's problem is to calculate the transport plans that minimize the expenditure in carrying the sand.

an admissible set of realizable transport plans. This terminology gives the name of this field of study.

This work of Monge was at first dismissed, but it was eventually rediscovered and built upon by the Russian economist Leonid Kantorovich in parallel to his work that lead to the foundation of linear programming and optimization, especially in an economics setting. His contribution to optimal transportation theory was to recognize that the minimization in the Monge problem could be mirrored by the maximization of a dual quantity that turned out to be much simpler to work with. Kantorovich's formulation of the optimal transport problem is usually called the Monge-Kantorovich problem in their honor [2].

After Kantorovich revived interest in the Monge problem, various mathematicians sought to obtain the prized results of existence, uniqueness and regularity for such problems. In 1987, Yann Brenier showed in [3] that there indeed exists a unique transport plan that minimizes the total cost associated to the Euclidian

distance *squared*. This work was expanded upon in 1995 by Wilfrid Gangbo and Robert McCann who showed in [4] and [5] that this result may be extended to any strictly convex or concave functions of distance, under some analytical restrictions.

In 1998, Felix Otto introduced a discrete variational method that deeply connected the optimal transport problem to gradient flows and dissipative partial differential equations, those evolution equations in which an energy decreases in time. In [6] and [7], he showed that it is possible to write certain dissipative equations, notably the Fokker-Planck equation, as the gradient flow of a physically motivated energy functional with respect to a metric arising from optimal transportation theory. This spawned many recent developments in studying physical equations with such techniques. Indeed, optimal transport is now a very active field of research which has found diverse applications in many scientific fields of study, including statistical physics, fluid dynamics, biology, economics and image processing.

1.2 Organization of This Thesis

The remainder of this thesis is organized as follows. In **chapter 2**, we describe the basic Monge problem, its limitations from an analytical point of view and its generalizations to the more amenable problem of optimal transport. We then present the intimately related Wasserstein distance and some of its properties.

In **chapter 3**, we develop intuition into the problem's nature by solving a discrete version of it. We then obtain a visual representation of the discrete optimal transport plans in certain important cases.

In **chapter 4**, we introduce Brenier's theorem and then review Gangbo and McCann's 1995 paper [4] and expand on several aspects of their proof that optimal transport problems have a unique solution if the cost is a strictly convex function of Euclidian distance. This proof relies on showing that the minimization and maximization of Monge and Kantorovich are equivalent.

In **chapter 5**, we briefly introduce the notion of gradient flows and present the Fokker-Planck equation. With these preliminaries, we review Jordan, Kinderlehrer and Otto's 1998 paper [7] and highlight some aspects of their proof that the Fokker-Planck equation is the gradient flow of a physical energy functional with respect to the Wasserstein distance. We conclude this final chapter with a formal introduction to the Otto calculus and a few other examples of Wasserstein or Wasserstein-like gradient flows.

CHAPTER 2
The Monge Problem and its Generalizations

We begin this chapter by introducing a very simple version of the Monge problem. We will show that this simple description must be generalized if we are to obtain any uniqueness result. We close the chapter by introducing the Wasserstein distance, which we will use extensively in Chapter 5.

2.1 The Basic Monge Problem

We now make precise the ideas introduced earlier and properly define the Monge problem and related terminology. Consider two compact sets A and B of $\mathbb{R}^n$ that have the same volume. The Monge problem can be stated rather simply: what is the map s mapping A to B that, among a set of admissible maps, minimizes the integral

$$\mathcal{C} = \int_A |x - s(x)| dx \ ? \tag{2.1}$$

In words, the sets A and B correspond to the *déblais* and *remblais* which we shall also call the source and sink or initial and final densities. For now, we consider A to be uniformly filled with materials or "mass", and this mass must be displaced into B; a map s that effects this displacement is called a transport plan. Since it is a priori reasonable to expect that total costs, whether in money or energy, are linear in the total transportation distance, a transport plan that

minimizes the integral above may be understood as efficient or optimal. We then call $\mathcal{C}$ a total cost while we will call the distance $|x - s(x)|$ a cost density[1] . The procedure of finding an optimal map s will be called the Monge problem, and later, the optimal transport problem.

So far, our description lacks generality in two ways. First, a linear cost does not correspond to a wide variety of actual problems. Another important issue with this linear cost is that it makes the Monge problem highly degenerate: many optimal transport plans may equally minimize the total cost. Second, our choice that the mass to be transported be uniformly distributed over a set is very restrictive. We will spend the next few sections fixing these issues, but we now spend some time discussing the important insight that Monge had in regards to this problem.

The original work by Monge [1] essentially begins by asking the question above in two and three dimensions. Monge considers discrete mass elements, sand scoops for example if moving sands from the *déblais* to the *remblais*. His work is based on geometrical intuition and drawings, see figure 2-1, in moving these scoops from point to point. It is then not rigorous mathematics from our modern point of view, but several results that Monge obtained are valid and provide good intuition into the nature of transport plans, of which we mention only two:

[1] If the total cost is to be measured in energy and the material to be transported is assumed to be mass, the cost density should be measured in Joules per kilogram. These units and notions may however be adapted to any applied context: energy and charge or price and quantity are other examples.

- *Optimal transport plans must map points to points using straight lines.* This is intuitively obvious: it is never more efficient to displace the *same* mass element from x_a to x_b then from x_b to x_c than from x_a to x_c directly. This can also be seen to follow from the triangle inequality: $|x_a - x_c| \leq |x_a - x_b| + |x_b - x_c|$.
- *If the straight lines connecting x_a to x_c and x_b to x_d cross and are not collinear, it is more efficient to connect x_a to x_d and x_b to x_c.* This is again intuitively obvious and likewise follows from the triangle inequality, which is strict whenever the points considered do not lie on the same line.

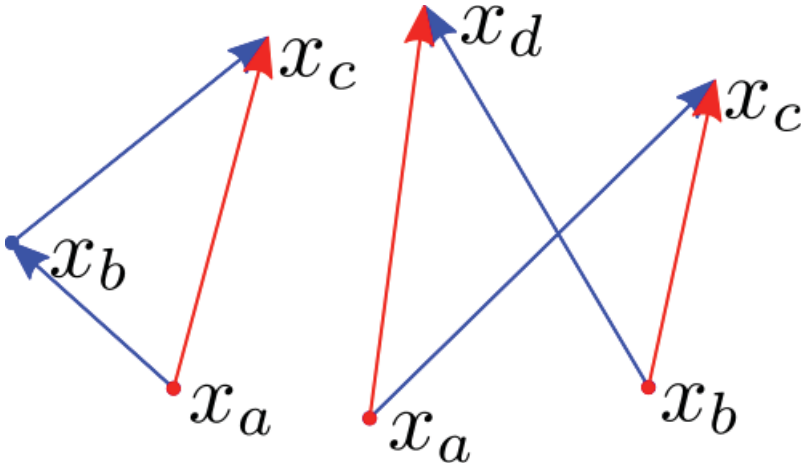

Figure 2–1: This diagram illustrates the two properties described by Monge. The transport plans in red are geometrically more optimal than the ones in blue.

In the following, we will generalize Monge's problem greatly and potentially lose most of Monge's insight. However, the first of the previous two points is so important, especially from a numerical point of view, that we will adapt our

assumptions to ensure its validity. The second point is less important and in fact, we will give a counterexample under a specific relaxation of the problem.

2.2 A First Generalization and the Degeneracy of the Monge Problem

To weaken our notion of efficiency, fix a general cost function $c(x, y)$: $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. This cost represents the expense in moving mass from position x to y. In the Monge problem, $c(x, y) = |x - y|$, but we want to consider forms of the cost function that have useful applications. It will be sufficient for our purpose to assume that c is a non-negative, finite and continuous function on $\mathbb{R}^n \times \mathbb{R}^n$. Moreover, we will assume that c depends only on the distance between x and y, such that we can write $c(x, y) = h(|x - y|)$ for a monotone function h with the required properties. The monotonicity condition is required in order to ensure that optimal transport plans still map points to points with straight lines. Indeed, "curving" the path may then never decrease the total cost $\mathcal{C}$, so that without loss of generality, only the coordinates x and y are needed to compute a cost. Later on, especially to invert h, we shall often write $c(x, y) = h(x - y)$ with the understanding that h is radially symmetric for analytical simplicity.

As promised earlier, let us now demonstrate that given the linear cost, $|x - y|$, the Monge problem turns out to admit multiple minimizers. In one dimension, let the initial mass distribution A consist of m unit mass packets of length less than 1 and lined up at the integer positions $\{1, 2, ..., m\}$. Similarly, let the final mass distribution B consists of similar packets lined up at the positions $\{2, 3, ..., m + 1\}$. Consider the following transport plans:

- s_1: Move all mass packets by one unit to the right.
- s_2: Move the mass packet at position 1 to position $m + 1$.

The total cost for these transport plans can be calculated as the number of packets transported times the transport distance, or $m \cdot 1$ and $1 \cdot m$ respectively. The linearity of the cost $|x - y|$ allows for other equivalent transport plans: intermediates between s_1 and s_2, splitting packets in two, and even allowing packets to "invert" upon themselves, as shown in figure 2-2. In this problem, there are in fact an infinite number of minimizers since the packets can be split in arbitrary subpackets.

The existence of these optimal transport plans precludes any sort of nice regularity theory, not to mention uniqueness, for the Monge problem. The simplest remedy to this issue is to break the linearity of $|x - y|$ by considering the costs $c(x, y) = |x - y|^p$ for p positive but not equal to 1. Using these cost functions and the previous two transport plans, the new total costs are $m \cdot 1^p$ and $1 \cdot m^p$ for s_1 and s_2 respectively. For the same reason, the other previously equivalent transport plans will no longer have the same total cost, such that one can find a unique optimal transport plan: s_1 if $p > 1$ and s_2 if $p < 1$.

This method of lifting the degeneracy of the Monge problem is generic: we will prove in chapter 4 that given reasonable assumptions on the densities, a cost function of the form $c(x, y) = h(x - y)$ for h strictly convex will be sufficient to show the existence and uniqueness of optimal transport plans. A slight modification of the proof will extend the result for h strictly concave, albeit we will not need this result.

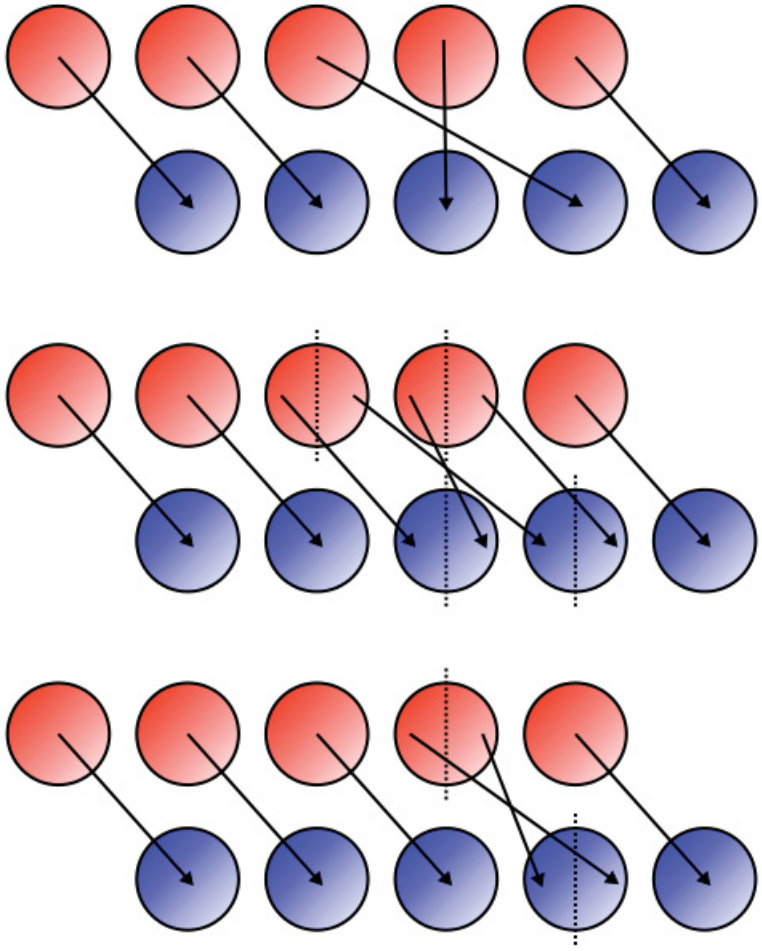

Figure 2–2: Minimizers of the given Monge problem with the initial and final distributions drawn in red and blue respectively. The initial and final distributions are not drawn on the same line for clarity purposes. From top to bottom; arbitrary packets are moved, packets are split in two and a packet is inverted upon itself. Note that the later examples do not violate Monge's second result on optimal transport plans since all points are collinear in one dimension.

2.3 The Modern Monge Problem

The second generalization we will carry out is to replace the assumption that the source and sink densities are the characteristic functions of sets. We will assume that f and g are non-negative functions with compact support, corresponding to "continuous" source and sink mass densities respectively. Without loss of generality and in the interest of simplicity, we require that f and g have the same total mass of 1. Note in particular that the support of f and g correspond to the sets A and B in the basic Monge problem and that we no longer require their volume to be the same. As we will eventually want to apply our developments to partial differential equations, PDEs, it will be useful to let the functions f and g be suitably weakly defined to apply the tools of elliptic theory.

In order of "weakness", the classes of functions we will study are the following:

- f and g belong to $L^1(\mathbb{R}^n)$.
- f and g are probability measures on $\mathbb{R}^n$.
- f and g correspond to the marginals of a coupling measure on $\mathbb{R}^n \times \mathbb{R}^n$.

We remark that another avenue for generalization would be to use a different metric space instead of $\mathbb{R}^n$. There have been many developments in the literature when the underlying metric space is a Riemannian manifold, see for example [8], but we will not explore this generalization here.

The densities are L^1 functions

Let us assume that f and g are non-negative Lebesgue measurable functions with unit norm, such that $f, g \in L^1(\mathbb{R}^n)$ and $||f|| = ||g|| = 1$. We define a transport plan as a Borel map $s : \mathbb{R}^n \to \mathbb{R}^n$ such that for every Borel set $B \subset \mathbb{R}^n$,

$$\int_B g(y)dy = \int_{s^{-1}(B)} f(x)dx \ . \tag{2.2}$$

This condition makes precise the notion that the mass of g found in some set B originates from the mass of f on the set $s^{-1}(B)$. Notice that this definition may be rewritten in terms of the characteristic function of B, χ_B. In particular, since $\chi_B(s(x))$ is non-zero only when $x \in s^{-1}(B)$, it follows that

$$\int_{\mathbb{R}^n} \chi_B(y)g(y)dy = \int_{\mathbb{R}^n} \chi_B(s(x))f(x)dx \ . \tag{2.3}$$

Using the fact that characteristic functions span the continuous functions, the following result holds:

Proposition 2.1. *Suppose s is a transport plan, then for any Borel continuous function a,*

$$\int_{\mathbb{R}^n} a(y)g(y)dy = \int_{\mathbb{R}^n} a(s(x))f(x)dx \ . \tag{2.4}$$

In the formalism where f and g are L^1 functions, we will characterize transport plans using the previous proposition. We denote by $S(f,g)$ the set of transport plans from f to g. We are now ready to write our first version of the optimal transport problem.

Definition 2.2. *Consider a non-negative, finite and continuous cost function $c(x,y)$. The **total cost** $\mathcal{C} : S(f,g) \to [0,\infty)$ of moving mass from f to g with respect to c is given by*

$$\mathcal{C}(s) = \int_{\mathbb{R}^n} c(x, s(x)) f(x) dx \ . \tag{2.5}$$

*We say that $t \in S(f,g)$ is an **optimal transport plan** if it is such that*

$$\mathcal{C}(t) = \inf_{s \in S(f,g)} \mathcal{C}(s) \ . \tag{2.6}$$

The optimal transport problem in these terms is then to obtain an optimal transport plan, if such a map exists. Further, we can question the uniqueness of such a plan. We note in passing that given our assumption that $c(x,y) = h(|x-y|)$, it is intuitively clear that if an optimal transport plan t were to exist, then the analogous problem of moving mass "backward" from g to f,

$$\inf_{s \in S(g,f)} \int_{\mathbb{R}^n} c(y, s(y)) g(y) dy \ , \tag{2.7}$$

would produce at least an optimal transport plan $\hat{t}$ that corresponds to the inverse map of t.

We will prove that in the case where the cost function is *strictly* convex, then an optimal transport plan t exists and is unique. Similarly, the backward problem will admit a unique optimal transport plan $\hat{t}$ such that $\hat{t} \circ t = \mathrm{Id}$ on the support of f, and $t \circ \hat{t} = \mathrm{Id}$ on the support of g. While our present formalism is a good generalization of Monge's problem, it will be useful later on to view f and g as measures.

The densities are probability measures

Let us now assume that f and g are probability measures on $\mathbb{R}^n$ that are absolutely continuous with respect to the Lebesgue measure. We will denote this set by $\mathcal{P}(\mathbb{R}^n)$. Using the Radon-Nikodym theorem, we note that if $\mu \in \mathcal{P}(\mathbb{R}^n)$, there exists a Lebesgue measurable function $m \in L^1(\mathbb{R}^n)$, such that

$$\mu(B) = \int_B m(x)dx = \int_B m(dx) \ \ \forall B \text{ Borel} . \tag{2.8}$$

We will always label m as μ by abuse of notation. The new notion of a transport plan can be given in terms of the measure theoretic pushforward:

Definition 2.3. *The* ***pushforward*** *of the measure f by the Borel map $s : \mathbb{R}^n \to \mathbb{R}^n$ is written $s_\# f$ and satisfies, for all Borel sets $B \in \mathbb{R}^n$,*

$$s_\# f(B) = f(s^{-1}(B)) . \tag{2.9}$$

The map s is called a transport plan if $s_\# f = g$ as measures.

The class $S(f,g)$ and the total cost $\mathcal{C}$ are defined as in the previous section. The truly powerful version of the optimal transport problem may be stated in terms of couplings.

The densities are couplings

This final flavor of optimal transport has been an active area of research in the past decade, mainly because it generalizes readily to situations where f and g are Dirac masses. The principal idea is to view f and g not as separate elements, but as two sides of the same object.

Definition 2.4. *A **coupling** γ between f and g in $\mathcal{P}(\mathbb{R}^n)$ is a measure on $\mathbb{R}^n \times \mathbb{R}^n$ such that for all Borel sets $B \in \mathbb{R}^n$,*

$$\gamma(B \times \mathbb{R}^n) = f(B) \text{ and } \gamma(\mathbb{R}^n \times B) = g(B) \ . \tag{2.10}$$

The set of all couplings between f and g is denoted by $\Gamma(f,g)$.

We also say that the coupling γ has marginals f and g.

The new set $\Gamma(f,g)$ may be regarded as the relaxation of $S(f,g)$. Indeed, let $s \in S(f,g)$ and define the measure $\tau(A \times B) = ((Id \times s)_{\#} f)(A \times B)$, then clearly $\tau \in \Gamma(f,g)$. However, it is easy to find couplings that cannot be written in terms of elements of $S(f,g)$. With this new notion of transport plans as couplings, we must slightly change the total cost function:

Definition 2.5. *Consider a non-negative, finite and continuous cost function* $c(x,y) = h(|x-y|)$. *The* ***total cost*** $\mathcal{C} : \Gamma(f,g) \to [0,\infty)$ *of moving mass from* f *to* g *with respect to* c *is given by*

$$\mathcal{C}(\gamma) = \int_{\mathbb{R}^n \times \mathbb{R}^n} c(x,y)\gamma(dx \times dy) \ . \tag{2.11}$$

We say that $\tau \in \Gamma(f,g)$ *is an* ***optimal coupling*** *if it is such that*

$$\mathcal{C}(\tau) = \inf_{\gamma \in \Gamma(f,g)} \mathcal{C}(\gamma) \ . \tag{2.12}$$

This problem is clearly a relaxation of our optimal transport problem, and is often called the Monge-Kantorovich problem. We can ask two important questions at this point: is it true that

$$\inf_{s \in S(f,g)} \mathcal{C}(s) = \inf_{\gamma \in \Gamma(f,g)} \mathcal{C}(\gamma) \tag{2.13}$$

and if so, can the optimal coupling be written in terms of an optimal transport plan? As Gangbo and McCann show in [5], this is true in particular for strictly convex costs.

2.4 The Wasserstein Distance

We close this chapter by introducing the main tool that will be used in the final chapter. In the coupling framework, let us fix the cost function to be $|x-y|^2$. Let us also define a new space of measures and the metric we will work with:

Definition 2.6. *The* ***space of probability measures with finite second moment*** *is the set*

$$\mathcal{P}_2(\mathbb{R}^n) = \left\{ \mu \in \mathcal{P}(\mathbb{R}^n) \;\middle|\; M[\mu] = \int_{\mathbb{R}^n} |x|^2 \mu(dx) < \infty \right\} . \tag{2.14}$$

Definition 2.7. *The (2)-****Wasserstein distance*** *between two elements μ and ν in $\mathcal{P}_2(\mathbb{R}^n)$ is*

$$d_W(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} |x-y|^2 \gamma(dx \times dy) \right)^{1/2} . \tag{2.15}$$

The p-Wasserstein distances are those that use the p-Euclidian norm instead, and are similarly defined on spaces of probability measures with finite p^{th} moment. We will not consider these and simply refer to the 2-Wasserstein distance as the Wasserstein distance. The following result justifies the choice of $\mathcal{P}_2(\mathbb{R}^n)$.

Proposition 2.8. *The Wasserstein distance is well defined on $\mathcal{P}_2(\mathbb{R}^n)$.*

Proof. Let γ be the trivial coupling $\gamma(A \times B) = \mu(A)\nu(B)$. Using that $|x-y|^2 \leq |x|^2 + |y|^2 + 2|x||y|$,

$$\begin{aligned} d_W^2(\mu,\nu) &\leq \int_{\mathbb{R}^n \times \mathbb{R}^n} |x-y|^2 \gamma(dx \times dy) \\ &\leq \int_{\mathbb{R}^n \times \mathbb{R}^n} \left(|x|^2 + |y|^2 + 2|x||y| \right) \mu(dx)\nu(dy) . \end{aligned} \tag{2.16}$$

The first two terms equal $M[\mu]$ and $M[\nu]$ respectively since

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} |x|^2 \mu(dx)\nu(dy) = \int_{\mathbb{R}^n} |x|^2 \mu(dx) \int_{\mathbb{R}^n} \nu(dy) = M[\mu] \tag{2.17}$$

by definition. The remaining term is then simplified using Jensen's inequality:

$$\begin{aligned} d_W^2(\mu,\nu) &\leq M[\mu] + M[\nu] + 2\int_{\mathbb{R}^n} |x|\mu(dx) \int_{\mathbb{R}^n} |y|\nu(dy) \\ &= M[\mu] + M[\nu] + 2\left(\left(\int_{\mathbb{R}^n} |x|\mu(dx)\right)^2 \left(\int_{\mathbb{R}^n} |y|\nu(dy)\right)^2\right)^{1/2} \\ &\leq M[\mu] + M[\nu] + 2\left(\left(\int_{\mathbb{R}^n} |x|^2\mu(dx)\right) \left(\int_{\mathbb{R}^n} |y|^2\nu(dy)\right)\right)^{1/2} \\ &= M[\mu] + M[\nu] + 2\sqrt{M[\mu]M[\nu]} < \infty \end{aligned} \tag{2.18}$$

■

We will not prove the following results, referring instead the reader to [9].

Theorem 2.9. *The pair $(\mathcal{P}_2(\mathbb{R}^n), d_W)$ forms a complete metric space.*

Proposition 2.10. *The infimum in the definition of d_W is achieved.*

We will obtain a specialized version of the previous proposition as a corollary of the main theorem of the next chapter. This theorem will in addition give that the infimizer, hence the minimizer, of d_W is unique.

CHAPTER 3
Intuitive Optimal Transport and a Visualization Procedure

From an applied point of view, the optimal transport problem is unlike many similar association problems in that it is not easy to visually come up with "good approximations" to the optimal transport plan. For this reason, we build a numerical method to easily visualize the flow of mass according to a given transport plan. To do this, we must introduce a dynamical interpolation technique and a numerical method to compute discrete transport plans. We then combine these ideas to gain intuition into the optimal transport problem. We do not attempt to rigorously justify our developments in this chapter. The general methodology presented here was suggested by Adam Oberman in private communications [10].

3.1 The Applications of Convex and Concave Costs

Let us first spend some time investigating the domains of application of optimal transport plans with respect to costs that are concave or convex functions of the Euclidian distance. The intuition we present will be expanded upon when we present the numerical simulations at the end of this chapter.

Convex cost functions

Convex cost functions seem to arise naturally in a physical setting that depends very much on geometric factors. In an electrostatic context for example, it is possible to prepare charge configurations that give rise to electric potentials of the form $V = k|x|^p$ for some positive integer exponent p and a constant k. The energy of a charge q in this field is given by $U = qV$ and therefore, the work in displacing this charge over space may be likened to the cost $c(x, y) = k|x - y|^p$. The total work needed to displace a distribution of charges into another is then given by the optimal transport total cost $\mathcal{C}$ associated to $c(x, y)$. In particular, the Monge cost arises naturally in the well-known infinite charged plate problem while the quadratic cost arises inside of an infinitely long uniformly charged cylinder [11]. The quadratic cost also arises when comparing the kinetic energy $T = \frac{1}{2}m|v|^2$ of particles with mass m and velocity v: the work necessary to change the kinetic energy of a particle from u to v may be associated to the cost $c(u, v) = |u - v|^2$.

The convex cost is unfavorable to large displacements, therefore, the optimal transport plans will intuitively tend to displace small packets of mass over small distances.

Concave cost functions

In an economics setting, concave costs arise naturally when production costs grow slower than profits in a process known as the "economies of scale". Suppose that the cost c is measured in expense per mass or units sold, the total cost can then be viewed as a measure of the total expense in producing and selling goods. In many situations, the profit in selling a number of goods increases linearly

with this number while the expenses in producing this number of goods grows slowly. This can be accounted for by various mechanisms, including "rebates" in buying bulk raw materials and the fact that after an initial investment, production facilities may be operated over a large range of production quotas for the same expenses. The overall result in balancing loss and gains makes it generally more efficient to produce a large number of goods where demand exists. This also translates to the transport of goods.

In contrast with the convex setting, the optimal transport plans corresponding to concave cost functions will intuitively tend to displace large packets of mass over long distances. As seen in the previous chapter when lifting the degeneracy of the Monge problem, it becomes more optimal to effect few large displacements than several small ones due to the concavity of the cost.

3.2 The Interpolation

In the optimal transport problem, the total cost is minimized by an optimal association between sources and sinks s. To visualize this pairing, it would be intuitive to let a mass element at an initial position x in the source distribution f move onto its assigned position y in g. The issue here is that there is no intrinsic notion of "dynamics" in the problem, however, we accept that optimal transport plans must map points to points with straight lines. To implement some sort of dynamics, we therefore simply let the mass move at constant velocity $|x - y|/t$ over a unit of time t. This is a natural choice that allows us to visualize the transformation of f into g.

To implement this artificial dynamics, let $\rho(t)$ be a probability density distribution for all $t \in [0,1]$ such that $\rho(0) = f$ and $\rho(1) = g$. For which choice of $\rho(t)$ will the simple dynamics described above arise? An incorrect approach would be the trivial interpolation $\rho(t) = (1-t)f + tg$. This is not an appropriate flow since mass may flow infinitely fast. The correct interpolation may be seen to be

$$\rho(t) = ((1-t)Id + ts)_{\#} f \ . \tag{3.1}$$

Indeed, $\rho(0) = Id_{\#}f = f$ and $\rho(1) = s_{\#}f = g$. To see why this push-forward provides the correct dynamics, suppose f contains a localized mass m in a very small set B around x. If s is sufficiently continuous, nearly all the mass m will be moved into g in a small set B' around $s(x)$. While the push-forward itself is not linear, let us track the motion of the small set B. At time t, the small set will be around the point $((1-t)Id + ts)(x) = (1-t)x + ts(x)$ which gives the correct behavior: the small set B is moving at constant velocity, in a straight line, from x to $s(x)$.

This interpolation procedure defines a density that can be visualized in time. To apply this procedure numerically, we must first translate the optimal transport problem into discrete terms.

3.3 The Optimal Transport Problem on a Grid

The discretization procedure we now describe was communicated to us by Adam Oberman [10]. For simplicity, we consider the optimal transport problem on an interval $\Omega \subset \mathbb{R}$. We take f and g to be continuous on Ω and the cost

function to be strictly convex or concave and continuous. We are now interested in computing a numerical approximation for the optimal transport plan t.

To discretize the problem, we begin by discretizing the domain Ω into N equal subintervals. The idea is that on each subinterval, there lies a mass $f_j = f(x_j)$ localized at x_j, without loss of generality, the left endpoint of the interval. This is justified by our assumption that f is continuous. The mass elements must be moved in such a way as to become the corresponding masses g_i, localized at x_i. In general, the values of f_j and g_i will not be equal, so that it is necessary to allow the masses to split and reform. This splitting and matching can be expressed as the $N \times N$ matrix P whose non-negative element P_{ij} corresponds to the mass transferred from x_j to x_i:

$$P = \begin{array}{c} \\ g_1 \\ g_2 \\ \vdots \\ g_i \\ \vdots \\ g_N \end{array} \begin{array}{c} \begin{array}{ccccccc} f_1 & f_2 & \cdots & f_j & \cdots & f_N \end{array} \\ \begin{bmatrix} P_{11} & P_{12} & & & & P_{1N} \\ P_{21} & P_{22} & & & & \\ & & \ddots & & & \vdots \\ & & & P_{ij} & & P_{iN} \\ & & & & \ddots & \vdots \\ P_{N1} & & \cdots & P_{Nj} & \cdots & P_{NN} \end{bmatrix} \end{array}$$

To implement the requirement that P be a "discrete transport plan", the total mass leaving x_j should equal f_j and the total mass arriving at x_i should equal g_i.

We then have:

$$\begin{aligned} \sum_i P_{ij} &= f_j \\ \sum_j P_{ij} &= g_i \\ P_{ij} &\geq 0 \end{aligned} \tag{3.2}$$

Any matrix P satisfying these three conditions will be called a discrete transport plan. Note that the discrete class $P_N(f, g)$ consisting of such matrices is convex[1] and thus much simpler than its continuous equivalent $S(f, g)$. It is clear that the maps P need not be one-to-one or onto in general because the mass elements will need to be split or assembled to deconstruct or reconstruct f and g respectively.

In a similar fashion, the approximate cost matrix C can be calculated by letting $C_{ij} = c(x_i, x_j)$, such that the total cost of the plan P is given by

$$\mathcal{C}(P) = \sum_{ij} C_{ij} P_{ij} \ . \tag{3.3}$$

To obtain an optimal discrete transport plan T, the quantity $\mathcal{C}(P)$ must be minimized over $P_N(f, g)$. The discrete optimal transport problem can then be

[1] The set $P_N(f, g)$ is convex in the sense that for matrices P^1 and P^2 in $P_N(f, g)$, $\alpha P^1 + (1 - \alpha)P^2 \in P_N(f, g)$ for any $\alpha \in [0, 1]$.

written as the convex optimization procedure:

$$\begin{aligned} &\text{Minimize } \sum_{ij} C_{ij} P_{ij} \\ &\text{subject to} \\ &\qquad \sum_{i} P_{ij} = f_j \\ &\qquad \sum_{j} P_{ij} = g_i \\ &\qquad P_{ij} \geq 0 \end{aligned} \tag{3.4}$$

The procedure outputs a matrix T which can be used to visualize approximations to the continuous optimal transport plan t. It is straightforward to implement the above convex optimization procedure numerically using a convex optimization routine.

To extend the procedure to higher dimensions, note that the convex optimization problem only requires the indices i and j to refer to sinks (g_i) and sources (f_j). It is conceptually easy to label arbitrary non-intersecting regions $R_i \subset \Omega \subset \mathbb{R}^n$ represented by some $x_i \in R_i$. The element P_{ij} will carry some mass f_j from the region R_j to the region R_i with a cost approximated by $C_{ij} = c(x_i, x_j)$.

At this point, it is outside of the scope of this work to guarantee that the original continuous optimal transport problem is well approximated by our current discrete construction. It seems however reasonable to expect that the discrete optimal plans T should converge in some sense to the continuous optimal transport plan t.

3.4 The Visualization Procedure

We now wish to apply the interpolation technique discussed above to a discrete optimal transport problem in an attempt to visualize it and gain valuable intuition. In this section, we assume that we are given a discrete optimal transport plan matrix T with the corresponding spatial labels x_i and x_j. In one dimension, we describe the numerical procedure that uses the dynamic interpolation to represent T visually.

Since this interpolation is not dynamic in nature, we must calculate intermediate densities as described earlier. The number of these intermediate densities may be chosen to be sufficiently high to produce a "movie" of the motion. Unfortunately, we cannot reproduce a movie in paper format so we limit our time resolution to a few time steps only.

To describe the visualization procedure in 1d, we first define several variables then write the numerical procedure as pseudo code. The following variables are used

- $x_{\text{left}}, x_{\text{right}}$: The borders of the interval on which the transportation problem is defined.
- N_x: The number of discrete positions in space.
- N_t: The number of intermediate densities, or time steps, to create.
- T: The matrix whose elements are the masses m_i carried from x_i to y_i.
- R: A refinement multiple to create a finer grid.

Since the discrete optimal transport plans are computed with coarse grids for efficiency reasons, a finer grid is needed to create a smooth interpolation in

space. The ratio of the grid spacing of these two grids is dictated by R. With these variables, we now describe the numerical procedure that prepares an array of densities that may be animated:

- Set up a discrete grid using $x_{\text{left}}, x_{\text{right}}$ and N_x to match the setup of T.
- Set up the finer grid using the same interval but $N_x \times R$ as the number of subintervals.
- Extract the elements m_i, x_i, y_i from T, discarding empty elements.
- Create an empty time array of size N_t to accumulate densities on the finer grid.
- Begin a loop over the times t and select the corresponding intermediate density.
    - For each mass element m_i, calculate the interpolated position

$$z_i = (1-t)x_i + ty_i \ . \tag{3.5}$$

    - For each mass element m_i, add m_i to the selected density at the position z_i of the finer grid.

Finally, it only remains to animate the time array of densities. We implemented this numerical procedure and a few sample animations are presented in the following section.

3.5 Sample Results and Discussion

We now present a few chosen examples with relevant cost functions and briefly analyze the output to expand on the intuition formulated in the first section of this chapter. In the following simulations, we use the interval $[x_{\text{left}}, x_{\text{right}}] = [-1, 1]$, subdivided in $N = 200$ subintervals with a grid refinement multiple of $R = 20$. The initial density is $f(x) = 1 + \sin(8x)$ and the final density is $g(x) = e^{-x}$ after an appropriate rescaling. This choice of densities is interesting because the two densities overlap and because the initial density presents three "source lumps" while the final density is gently varying. As will be seen later, this choice highlights the distinguishing aspects of concave and convex costs. Eight images are generated ($N_t = 8$), including the initial and final densities.

Convex cost functions

Let us begin with the discrete optimal transport plan generated with $c(x_i, x_j) = |x_i - x_j|^2$, a discrete version of the Wasserstein distance squared. The sequence of plots is printed chronologically in figure 3-1 from left to right and from top to bottom, and this at equal intervals. The red and blue curves correspond to f and g respectively while the shaded gray region corresponds to the computed mass density. Note that the lack of mass around 1 and the spurious oscillations are caused by the numerical smoothing procedure. These effects are purely visual and do not contribute to our interpretation below.

It is clear from the snapshots in figure 3-1 that the mass initially located in the three peaks of f is displaced to the left, almost continuously. This is because the cost being convex favors many local mass transfers. We can then give the

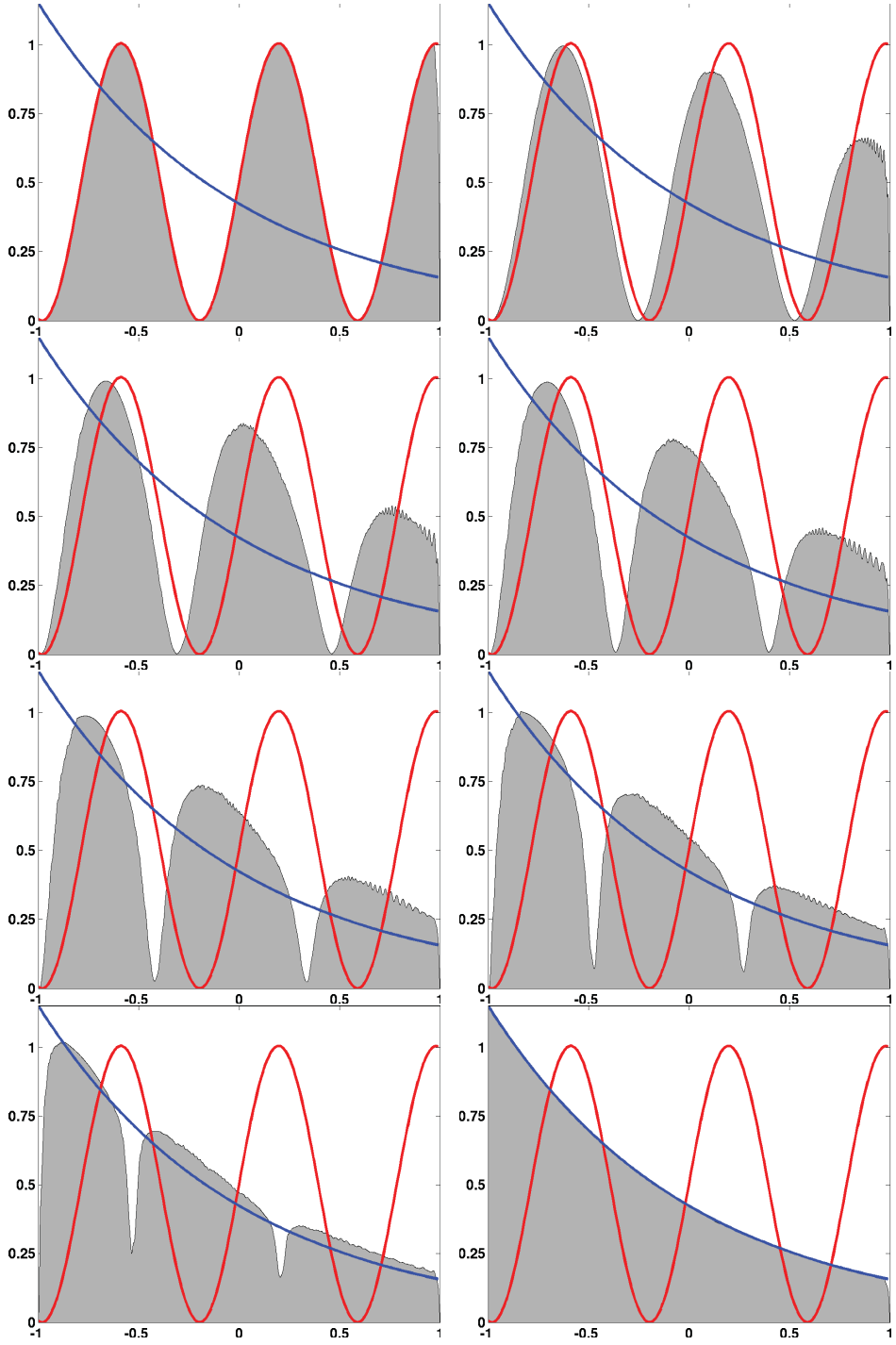

Figure 3–1: An example of a discrete optimal transport plan corresponding to the square of the Euclidian distance.

general intuition that convex cost functions give rise to transport plans that continuously deform f into g while preserving the "ordering" of the mass packets. Note indeed that the mass common to f and g *is* displaced, even though this a priori unnecessary displacement incurs a cost. This behavior may be likened to dissipative processes like diffusion which smoothly change the state of a system to another. This is an important aspect of our later developments, and the intuition gained in this example may be regarded as justifications to use the Wasserstein distance to explain physical processes.

Concave cost functions

We now present the discrete optimal transport plan generated with $c(x_i, x_j) = \sqrt{|x_i - x_j|}$.

The snapshots in figure 3-2 present the potential difficulty in dealing with concave cost functions. Two general behavior may be observed: the displacement of large mass packets to the left over long distances and the displacement of small mass packets to the right over short distances. Note especially how the rightmost peak in f is split in three parts to fill the three troughs of f. One may also see these shipments overtaking others and further, it is more optimal to use some mass from the middle peak to fill the rightmost compared to filling this trough with mass from the third peak. This reveals that concave costs may give rise to optimal transport plans that cross in different ways; the "ordering" of the packets is then not necessarily preserved. Another important aspect that is found is that the mass common to f and g does *not* move. This is impossible in the continuous version of the problem since that mass packets may not split.

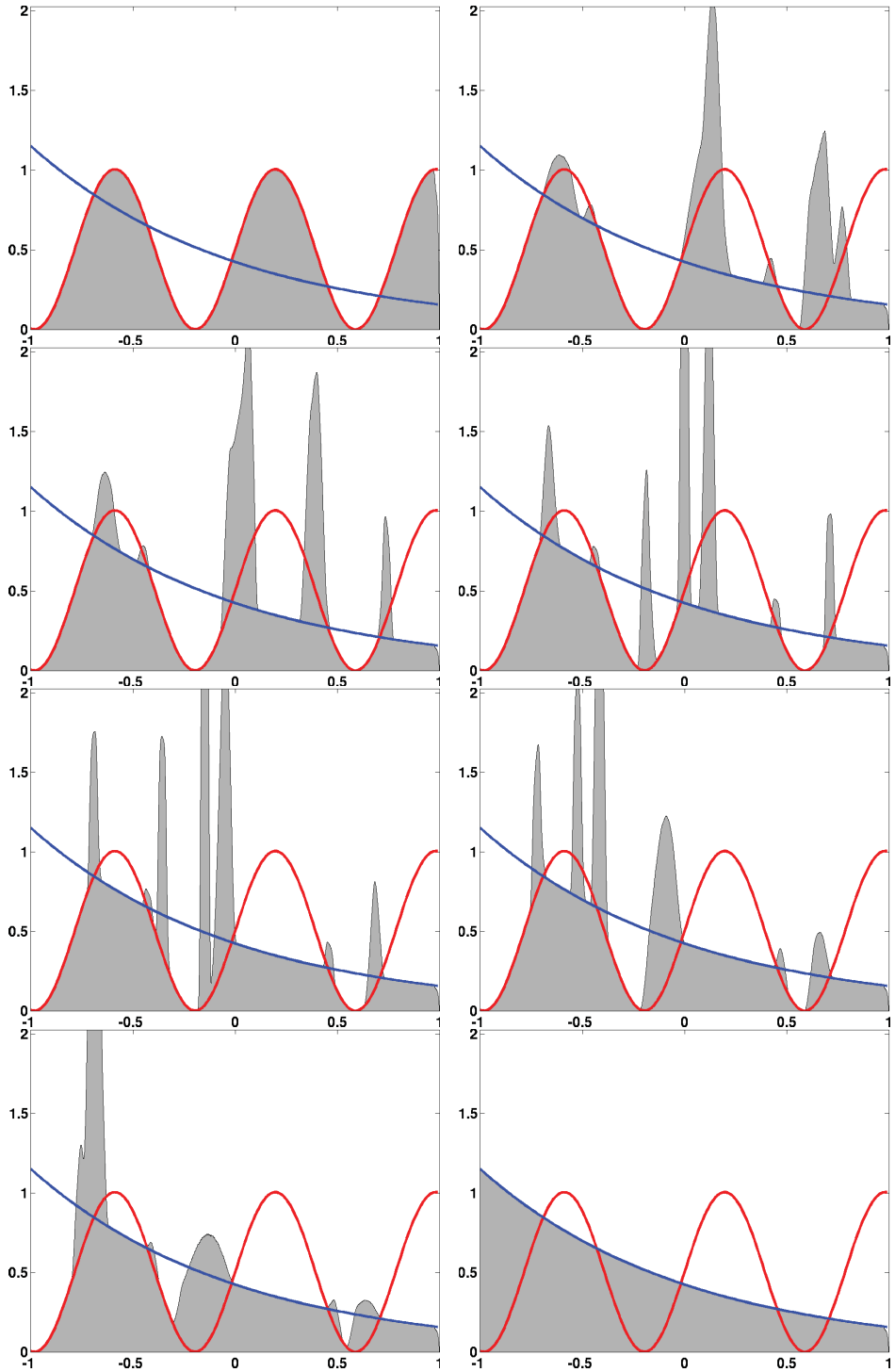

Figure 3–2: An example of a discrete optimal transport plan corresponding to the square root of the Euclidian distance.

It is interesting to note however that the existence and uniqueness proof of the next chapter may be modified to account for concave cost functions with exactly this additional requirement that the mass common to f and g be removed from both distributions.

The numerical tool developed in this chapter is useful to give a firsthand intuition on the behavior of optimal transport plans, especially when the cost function is changed. This intuition carries over at least heuristically to the continuous problem, to which we now return.

CHAPTER 4
An Existence and Uniqueness Result

We prove in detail the result of Gangbo and McCann in [4] that there exists a unique optimal transport plan when the cost function is strictly convex, under suitable conditions on the initial and final densities. To do so, we begin by introducing the prototypical result of Brenier. We then introduce the Kantorovich dual problem which will allow us to compute an optimal transport plan.

4.1 Brenier's Theorem

The result which we now present was essential to the development of optimal transportation theory. In fact, the existence and uniqueness result which we prove later in this chapter is but a generalization of Brenier's ideas. The main result of Brenier in [3] is that there exist unique polar factorizations and rearrangements of vector fields. This is terminology to state that under certain conditions, a vector field may be written as the composition of the gradient of a convex function (a one-to-one map) and of a mass-preserving map. The formulation of the theorem is as follows:

Theorem 4.1. *Let ϕ be a Borel, bounded and Lebesgue integrable map from a compact set K to $\mathbb{R}^n$. Denote by S the set of mass-preserving maps from K to*

itself, for example, in the sense of chapter 2. If the Lebesgue measure $\mu(\phi^{-1}A)$ *is* 0 *whenever the set* A *has measure* 0*, there exist maps* g and u *such that*

- $g \in S$
- u *is Lipschitz and convex in a neighborhood of* K
- $\phi = (Du) \circ g$
- g *is the unique projection of* ϕ *on* S *in the* L^2 *sense*

In steps similar to our next developments, Brenier shows that the Monge-Kantorovich problem with a *quadratic* cost may be molded into the shape of his general theorem. It can be shown that g plays the role of the optimal transport plan and can be written as $Dv \circ \phi$ where v is the Legendre transform of u, which is convex. This immediately gives the existence and uniqueness of optimal transport plans while optimality holds since g is a projection in the L^2 sense: it is exactly given by the optimal transport problem with a quadratic cost. This theorem was generalized to strictly convex or concave costs by Gangbo and McCann, as we show later.

Let us give a formal example of the usefulness of Brenier's theorem to PDEs. Recall that a transport plan is a map s satisfying, for any continuous function a,

$$\int_{\mathbb{R}^n} a(y)g(y)dy = \int_{\mathbb{R}^n} a(s(x))f(x)dx \ . \tag{4.1}$$

Brenier's theorem gives the existence of a transport plan t for quadratic optimal transport problems. Moreover, $t = Dv$ for some convex function v.

Formally, the change of variables $y = Dv(x)$ gives

$$\int_{\mathbb{R}^n} a(Dv(x))g(Dv(x))\det(D^2v(x))dx = \int_{\mathbb{R}^n} a(Dv(x))f(x)dx \tag{4.2}$$

where the determinant of the Hessian D^2v is at least non-negative since v is convex. Therefore, in an appropriately weak sense, v satisfies the famous Monge-Ampère PDE

$$g(Dv(x))\det(D^2v(x) = f(x)\ . \tag{4.3}$$

From results in optimal transportation theory, it is therefore possible to imply properties of PDEs: existence, uniqueness, convexity of solutions, and so on. This is the main subject of chapter 5: combining functionals, optimal transport and PDEs.

4.2 Assumptions for the Existence and Uniqueness Result

We now make certain assumptions on the densities and the cost to prove a limited existence and uniqueness result. We isolate these assumptions for clarity.

Assumptions on the densities

First, let us work in the L^1 formalism. We assume that the source and sink densities f and g are $L^1(\mathbb{R}^n)$ probability densities with compact support U and V respectively. Since we eventually want to obtain an optimal transport plan t, it is clear that modifying t outside of U or at any point where f vanishes will have no impact on the total cost $\mathcal{C}$. For this reason, we will say that a statement holds f-almost everywhere if it holds almost everywhere where f is positive.

Assumptions on the cost

Second, we will work with a uniformly continuous cost function $c : \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty)$ of the form $c(x, y) = h(x - y)$. Moreover, and more importantly, we assume that c is a strictly convex function of $|x - y|$ and $C^1(\mathbb{R}^n \times \mathbb{R}^n)$.

In their paper, McCann and Gangbo show that our result also holds for strictly concave functions of $|x-y|$ if $U \cap V$ is empty. This can be further relaxed if the mass that is common to f and g is constrained to remain in the same position by using the densities

$$f^* = \max\{f - g, 0\} \text{ and } g^* = \max\{g - f, 0\} \ . \tag{4.4}$$

These new densities will then have disjoint supports $U^* = U \backslash (U \cap V)$ and $V^* = V \backslash (U \cap V)$ respectively.

Another relaxation is to work with costs that are not C^1, in which case the notion of c-convexity must be used extensively. This approach is used in Gangbo and McCann's second paper [5].

4.3 The Kantorovich Dual Problem

We now present a problem that is "dual" in some sense to the optimal transport problem. The infimizer of the later problem is difficult to find because of the complexity of $S(f, g)$. The idea is to construct a simpler functional that is always less than the total cost, but whose maximum value equals the infimum of $\mathcal{C}$.

The particular case in which this holds true will be used to construct a transport plan t that will turn out to be the (unique) minimizer of $\mathcal{C}$.

Let us define the class of objects and the functional which we will use.

Definition 4.2. *The **Lipschitz class** with respect to the cost $c(x, y)$ is the set*

$$Lip_c = \{(u, v) \in C(\mathbb{R}^n) \times C(\mathbb{R}^n) \mid u(x) + v(y) \leq c(x, y) \ \forall (x, y) \in U \times V\} \ . \quad (4.5)$$

The name "Lipschitz" is justified by the fact that the elements u and v in the above definition are Lipschitz over the compact sets U or V since they can be bounded by the continuous function c. This property will eventually allow us to differentiate u and v.

Definition 4.3. *The **dual functional** $J(u, v) : Lip_c \to (-\infty, \infty)$ to $\mathcal{C}$ is defined as*

$$J(u, v) = \int_{\mathbb{R}^n} u(x) f(x) dx + \int_{\mathbb{R}^n} v(y) g(y) dy \ . \quad (4.6)$$

The dual problem will be to find, if it exists, a maximizer $(\psi, \phi) \in Lip_c$ to $J(u, v)$, such that

$$J(\psi, \phi) = \sup_{(u,v) \in Lip_c} J(u, v) \ . \quad (4.7)$$

An important element of the proof will be to show that such a maximizer has convex properties with respect to the cost function. The notion of "c-convexity" generalizes the familiar notion of convexity while the Fenchel transform generalizes

the Legendre transform. We will only require basic definitions for our purpose, but more details can be found in [5].

Definition 4.4. *The* ***Fenchel transform*** $u^c : V \to \mathbb{R}$ *of* $u : U \to \mathbb{R}$ *is*

$$u^c(y) = \inf_{x \in U} \left(c(x, y) - u(x) \right) \ . \tag{4.8}$$

The ***Fenchel transform*** $v^c : U \to \mathbb{R}$ *of* $v : V \to \mathbb{R}$ *is*

$$v^c(y) = \inf_{y \in V} \left(c(x, y) - v(y) \right) \ . \tag{4.9}$$

We will use the notation $u^{cc} = (u^c)^c$ to mean the "double" Fenchel transform of u.

Proposition 4.5. *Suppose* $u \in C(U)$*, then* $(u, u^c) \in Lip_c$.

Proof. Let us first show that the Fenchel transform of a continuous function is continuous. For any $y_1, y_2 \in V$:

$$\begin{aligned} |u^c(y_1) - u^c(y_2)| &= \left| \inf_{x \in U} \left(c(x, y_1) - u(x) \right) - \inf_{x \in U} \left(c(x, y_2) - u(x) \right) \right| \\ &\leq \sup_{x \in U} |c(x, y_1) - u(x) - c(x, y_2) + u(x)| \\ &= \sup_{x \in U} |c(x, y_1) - c(x, y_2)| \end{aligned} \tag{4.10}$$

Since c is jointly continuous in x and y over the compact domain $U \times V$, u^c must be continuous over V. Now, fix $(x, y) \in U \times V$ and choose $z = x$ such that

$$u(x) + u^c(y) = u(x) + \inf_{z \in U} (c(z, y) - u(z)) \leq u(x) + (c(x, y) - u(x)) = c(x, y) \tag{4.11}$$

which gives that $(u, u^c) \in Lip_c$. ■

An equivalent inequality can be obtained for v^c by symmetry. The string of inequalities in the previous proof can then be used to bound the Lipschitz constant of the Fenchel transforms with that of the cost function. In particular, both members of the pair $(u^{cc}, u^c) \in Lip_c$ are Lipschitz whenever $u \in C(U)$. The following lemma makes clear the usefulness of the Fenchel transform in our context:

Lemma 4.6. *Suppose* $(u, v) \in Lip_c$*, then* $J(u, v) \leq J(u^{cc}, u^c)$.

Proof. Consider first the difference

$$\begin{aligned} J(u, u^c) - J(u, v) &= \int_U (u(x) - u(x)) f(x) dx + \int_V (u^c(y) - v(y)) g(y) dy \\ &= \int_V \inf_{z \in U} (c(z, y) - u(z)) - v(y) g(y) dy \ . \end{aligned} \tag{4.12}$$

Since $v(y)$ is a constant with respect to the infimum and $u(z) + v(y) \leq c(z, y)$,

$$\begin{aligned} J(u, u^c) - J(u, v) &= \int_V \inf_{z \in U} (c(z, y) - u(z) - v(y)) g(y) dy \\ &\geq \int_V \inf_{z \in U} (c(z, y) - c(z, y)) g(y) dy = 0 \ . \end{aligned} \tag{4.13}$$

By symmetry, the same can be done to show that $J(u, v) \leq J(v^c, v)$, such that $J(u^{cc}, u^c) \leq J(u, v)$. ■

This result implies that if there exist a maximizer of the functional J, there exists at least one maximizer of the form (ψ, ϕ) where $\psi = \psi^{cc}$ and $\phi = \psi^c$. A

function ψ satisfying these requirements is called "c-convex" in analogy with the fact that a function that is its own double-Legendre transform is convex.

We are now ready to show that, under our assumptions, the dual problem admits at least one maximizer (ψ^{cc}, ψ^c). We begin by showing that the supremum of J is bounded, hence we can find a maximizing sequence (ψ_n, ϕ_n). This sequence will be shown to satisfy the requirements of the Arzela-Ascoli theorem, such that at least a subsequence converges to a maximizer $(\psi, \phi) \in Lip_c$. We follow the demonstration given in Guillaume Carlier's lecture notes [12].

Theorem 4.7. *There exists a maximizer $(\psi^{cc}, \psi^c) \in Lip_c$ such that*

$$J(\psi^{cc}, \psi^c) = \mu = \sup_{(u,v)\in Lip_c} J(u, v) \ . \tag{4.14}$$

Proof. First, μ is non-negative since $(0, 0) \in Lip_c$ and $J(0, 0) = 0$.

Fix $(u, v) \in Lip_c$, then for all $w \in V$, we may write $u(x) \leq c(x, w) - v(w)$, such that

$$\int_U u(x)f(x)dx \leq \int_U c(x, w)f(x)dx - v(w) \int_U f(x)dx \leq C_f - v(w) \ , \tag{4.15}$$

where C_f is the (finite) supremum over w of the integral of $c(x, w)$ against f. A similar expression can be written for v with another constant C_g:

$$\int_V v(y)g(y)dy \leq \int_V c(z, y)g(y)dy - u(z) \int_V g(y)dy \leq C_g - u(z) \tag{4.16}$$

Using that $u(z) + v(w) \leq c(z, w)$ and that c is non-negative,

$$J(u, v) \leq C_f + C_g - (u(z) + v(w)) \leq C_f + C_g - c(z, w) \leq C_f + C_g \tag{4.17}$$

so that $\mu \leq C_f + C_g < \infty$. Now, let $\{(\psi_n, \phi_n)\}_{n\in\mathbb{N}} \subset Lip_c$ be such that, without relabeling, $\psi_n = \psi_n^{cc}$, $\phi_n = \psi_n^c$ and

$$\lim_{n\to\infty} J(\psi_n, \phi_n) = \mu \ . \tag{4.18}$$

Since $J(u - \lambda, v + \lambda) = J(u, v)$ for any pair $(u, v) \in Lip_c$ and any constant λ, we may further assume that $\min_U \psi_n = 0$.

We now want to show that the maximizing sequence above satisfies the requirements of the Arzela-Ascoli theorem, namely uniform boundedness and equicontinuity. Let $\omega_c : (0, \infty) \to \mathbb{R}$ be the modulus of continuity of the cost function over $U \times V$:

$$\omega_c(t) = \sup_{|x_1-x_2|+|y_1-y_2|\leq t} |c(x_1, y_1) - c(x_2, y_2)| \tag{4.19}$$

Since c is assumed to be jointly continuous on $U \times V$, $\lim_{t\to 0} \omega_c(t) = 0$. Moreover, from the calculation in equation (4.10), we may obtain the bound

$$|\phi_n(y_1) - \phi_n(y_2)| \leq \sup_{x\in U} |c(x, y_1) - c(x, y_2)| \leq \omega_c(|y_1 - y_2|) \ . \tag{4.20}$$

The same calculation can be repeated with the double Fenchel transform ψ_n^{cc} such that both sequences ψ_n and ϕ_n are uniformly bounded by the modulus of continuity of the cost function. The sequence (ψ_n, ϕ_n) is therefore equicontinuous. Moreover, since $\psi_n(x_0) = 0$ for at least some $x_0 \in U$,

$$0 \leq |\psi_n(x)| = |\psi_n(x) - \psi_n(x_0)| \leq \omega_c(|x - x_0|) \leq M \tag{4.21}$$

for all $x \in U$, where M is the modulus of continuity of c evaluated at the diameter of U. Since ψ_n is uniformly bounded, and the infimum of the Fenchel transform preserves uniform boundedness, (ψ_n, ϕ_n) is uniformly bounded.

The two conditions of the Arzela-Ascoli theorem are met, so there exists a subsequence of $(\psi_n, \phi_n) \subset Lip_c$ that converges uniformly to a pair $(\psi, \phi) \in Lip_c$ for which $J(\psi, \phi) = \mu$. Finally, $\mu = J(\psi, \phi) \leq J(\psi^{cc}, \psi^c) \leq \mu$ which shows that there exists a maximizer $(\psi^{cc}, \psi^c) \in Lip_c$ to J. ■

We are now ready to use the maximizer $(\psi, \phi) = (\psi^{cc}, \psi^c)$ in the next section.

4.4 Existence and Uniqueness for Strictly Convex Costs

We first present a result that will allow us to construct a transport plan from the pair (ψ, ϕ).

Lemma 4.8. *Let $W \subset \mathbb{R}^n$ be open and $f : W \to \mathbb{R}$ be Lipschitz continuous on W, then f is differentiable almost everywhere in W.*

This theorem is often called Rademacher's theorem, theorem 3.1.6 of [13]. We will apply this lemma to show that ψ can be differentiated almost everywhere in the support of f, U. We may now prove the main result of this chapter.

Theorem 4.9. *Let f and g be $L^1(\mathbb{R}^n)$ probability densities with compact support U and V respectively. Let $c : \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty)$ be of the form $c(x, y) = h(x-y)$ where h is a strictly convex $C^1(\mathbb{R})$ function of $|x - y|$. Then there exists a transport plan*

$t \in S(f,g)$ that minimizes the total cost $\mathcal{C}$ over $S(f,g)$. Moreover, this optimal transport plan is f-almost everywhere unique.

Proof. In order to use Rademacher's theorem in what follows, let W be the smallest open domain containing both U and V and let (ψ, ϕ) be as in Theorem 4.7. The cost c is Lipschitz on W because it is continuous and W is bounded. Since the modulus of continuity of c bounds the Lipschitz constants of ψ and ϕ, both functions ψ and ϕ are also Lipschitz in W. Rademacher's theorem then guarantees that ψ is almost everywhere differentiable in $U \subset W$ with gradient $\nabla\psi$. This gradient is a Borel map since its set of discontinuities has measure zero.

The gradient of h is continuous and one-to-one since $h \in C^1(\mathbb{R}^n)$ and strictly convex. Denoting this gradient by ∇h, the inverse $(\nabla h)^{-1}$ must also be continuous by the open mapping theorem.

We now construct a transport plan from (ψ, ϕ). Since $\psi(x) = \inf_{y \in V}(c(x,y) - \phi(y))$ and c and ϕ are continuous, the infimum must be attained at some value y for all $x \in U$. Fix x where ψ is differentiable, then for some y,

$$\psi(x) = h(x-y) - \phi(y) \ . \tag{4.22}$$

From the previous justifications, one sees that $\nabla\psi(x) = \nabla h(x-y)$, and using the continuous inverse of ∇h, we can write

$$y = t(x) = x - (\nabla h)^{-1}\nabla\psi(x) \ . \tag{4.23}$$

The map t can be seen to be Borel and defined everywhere on U. Note that this implies that for almost every x, a *unique* point $y \in V$ is such that $\psi(x) + \phi(y) = c(x, y)$.

Let us now show that t is a transport plan by using the fact that the first variation of J vanishes at (ψ, ϕ) since it is an extremizer. Let a be a continuous function and ϵ a small parameter and define the functions:

$$\begin{aligned} v_\epsilon(y) &= \phi(y) + \epsilon a(y) \\ u_\epsilon(x) &= v_\epsilon^c(x) = \inf_{y \in V} (c(x, y) - \phi(y) - \epsilon a(y)) \end{aligned} \tag{4.24}$$

For all x where ψ is differentiable, only the choice $y = t(x)$ produces the infimum in the definition of $\psi(x) = \phi^c(x)$, such that $\psi(x) = c(x, t(x)) - \phi(t(x)$. The continuity of a and the smallness of ϵ means that the infimum in the definition of u_ϵ will be attained near $y = t(x)$. The error between the true $u_\epsilon(x)$ and the value of $c(x, y) - \phi(y) - \epsilon a(y)$ evaluated at $t(x)$ can therefore only be superlinear in ϵ:

$$u_\epsilon(x) = c(x, t(x)) - \phi(t(x)) - \epsilon a(t(x))) + o(\epsilon) = \psi(x) - \epsilon a(t(x)) + o(\epsilon) \tag{4.25}$$

Note now that $(u_0, v_0) = (\psi, \phi)$ is a critical point of J, so

$$\mathrm{grad}_a J(\psi, \phi) = \lim_{\epsilon \to 0} \frac{J(u_\epsilon, v_\epsilon) - J(\psi, \phi)}{\epsilon} = 0 \tag{4.26}$$

independently of the function a.

Let us compute directly the limit

$$\begin{aligned}\operatorname{grad}_a J(\psi,\phi) &= \lim_{\epsilon\to 0}\left(\int_U \frac{u_\epsilon(x)-\psi(x)}{\epsilon}f(x)dx + \int_V \frac{v_\epsilon(y)-\phi(y)}{\epsilon}g(y)dy\right)\\ &= \lim_{\epsilon\to 0}\left(\int_U \frac{-\epsilon a(t(x))+o(\epsilon)}{\epsilon}f(x)dx + \int_V \frac{\epsilon a(y)}{\epsilon}g(y)dy\right)\\ &= -\int_U a(t(x))f(x)dx + \int_V a(y)g(y)dy = 0\end{aligned} \tag{4.27}$$

which is exactly the condition for t to be a transport plan by Proposition 2.1.

Consider now any pair $(u,v) \in Lip_c$ and any transport plan $s \in S(f,g)$. The continuity of v gives that

$$\begin{aligned}J(u,v) &= \int_U u(x)f(x)dx + \int_V v(y)g(y)dy\\ &= \int_U u(x)f(x) + \int v(s(x))f(x)dx = \int_{\mathbb{R}^n}(u(x)+v(s(x)))f(x)dx\\ &\le \int_{\mathbb{R}^n} c(x,s(x))f(x)dx = \mathcal{C}(s)\end{aligned} \tag{4.28}$$

where the equality holds if $(u,v) = (\psi,\phi)$ and $s = t$. Therefore, $\sup_{Lip_c} J = \inf_{S(f,g)} \mathcal{C}$ such that indeed, $\mathcal{C}(t) \le \mathcal{C}(s)$ for all $s \in S(f,g)$. This proves the duality of J and $\mathcal{C}$ and that t is an optimal transport plan.

To finally show that t is almost everywhere unique on U, suppose that t^* is another optimal transport plan such that $\mathcal{C}(t^*) = \sup_{Lip_c} J$. Without requiring that (ψ,ϕ) be unique, it remains nonetheless true that $J(\psi,\phi) = \mathcal{C}(t^*)$ such that

$$\sup_{Lip_c} J = \int_{\mathbb{R}^n}(\psi(x)+\phi(t^*(x)))f(x)dx = \int_{\mathbb{R}^n} c(x,t^*(x))f(x)dx = \inf_{S(f,g)} \mathcal{C} \tag{4.29}$$

which similarly holds for t. Therefore, f almost-everywhere, $\psi(x)+\phi(t(x)) = c(x,t(x))$ and $\psi(x)+\phi(t^*(x)) = c(x,t^*(x))$. Since for all x there exists a unique

y with the property that $\psi(x) + \phi(y) = c(x, y)$, it follows that $y = t(x) = t^*(x)$ f-almost everywhere, concluding the proof. ■

A corollary is that the optimal transport problem is symmetric under the exchange of f and g.

Proposition 4.10. *Under the assumptions of the previous theorem, there exists an optimal transport plan $\hat{t} \in S(g, f)$ that minimizes the total cost $\mathcal{C}$ over $S(g, f)$. This transport plan is g-almost everywhere unique and is such that $\hat{t} \circ t = Id$ f-almost everywhere and $t \circ \hat{t} = Id$ g-almost everywhere.*

Proof. It is clear that there exists a unique optimal transport plan $\hat{t} \in S(g, f)$ by interchanging f and g, without changing J, in the previous theorem. By symmetry, $\hat{t}$ will be the only map such that $\psi(\hat{t}(y)) + \phi(y) = c(\hat{t}(y), y)$ g-almost everywhere. Fixing $x = \hat{t}(y)$, the only solution to $\psi(x) + \phi(y) = c(x, y)$ is $y = t(x)$. Therefore, $\hat{t} \circ t(x) = x$ f-almost everywhere and the result holds by symmetry. ■

4.5 The Connection With the Monge-Kantorovich Problem

We have thus shown that given a strictly convex cost function, the infimum problem

$$\inf_{s \in S(f,g)} \int_{\mathbb{R}^n} c(x, s(x)) f(x) dx \tag{4.30}$$

is actually a minimum problem with a unique solution t. We now investigate the significance of t under the relaxation of the optimal transport problem to the Monge-Kantorovich formulation. Namely, is it true that t directly gives rise to a

coupling τ, that uniquely minimizes the problem

$$\inf_{\gamma \in \Gamma(f,g)} \int_{\mathbb{R}^n \times \mathbb{R}^n} c(x,y)\gamma(dx \times dy) \ ? \tag{4.31}$$

This question was answered by Gangbo and McCann in [5]. To further extend the result to measures in $\mathcal{P}_2(\mathbb{R}^n)$ of unbounded support, extra growth requirements on the cost function are required, in particular, superlinearity away from the origin. When $c(x,y) = |x-y|^2$, these requirements are satisfied and the following theorem holds:

Theorem 4.11. *Suppose that the initial and final mass distributions f and $g \in \mathcal{P}_2(\mathbb{R}^n)$ and that the cost function is $c(x,y) = |x-y|^2$, then given the map $t(x) = x - (\nabla h)^{-1}\nabla\psi(x)$ obtained in Theorem 4.9, the coupling*

$$\tau(A \times B) = ((Id \times s)_{\#} f)(A \times B) \tag{4.32}$$

is the unique (Borel) optimal coupling to the Monge-Kantorovich problem.

In other words, at least for certain cost functions, only the optimal transport plans may achieve the lowest total cost in the Monge-Kantorovich problem. The proof of this result requires extensive background on c-convexity to obtain a non-empty set of optimal couplings and a c-convex function ψ. This set is a singleton since the gradient of the maximizers ψ of J can be shown to be unique.

This result is extremely important because it not only states that the Wasserstein distance d_W is uniquely attained, but it also explicitly characterizes the form of the connection between f and g.

We have now solved the optimal transport problem, at least in the case of the Wasserstein distance. Other avenues for investigation in the properties of optimal transport plans would be to understand their geometry and regularity. The modern developments in these issues are reviewed in McCann and Guillen's series of lecture notes [8].

In the following chapter, we will use the result of this chapter applied to the Wasserstein distance to highlight an important relationship between this metric and the Fokker-Planck equation and other dissipative PDEs.

CHAPTER 5
The Jordan-Kinderlehrer-Otto Scheme and Dissipative Equations

Let us summarize the previous developments in preparation for this final chapter. In chapters 1 and 2, we have introduced the optimal transport formalism and its generalizations. Most importantly, we introduced the Wasserstein distance, a metric on the space of probability densities $\mathcal{P}_2(\mathbb{R}^n)$ that computes an "energy cost" in transforming one probability density into another with respect to the quadratic cost. In chapter 3, we have given a visual interpretation to the problem and in chapter 4, we have proved that the optimal transport problem has a unique solution, in particular for the quadratic cost. From this analysis and the correspondence between the "flavors" of optimal transport problems, the Wasserstein distance becomes extremely powerful because it intrinsically guarantees the existence and uniqueness of transport plans on the space of probability densities.

The strength of the metric space $(\mathcal{P}_2(\mathbb{R}^n), d_W)$ rests upon four major ideas:

- The space of probability densities is relevant in virtually all fields of physical sciences due to the statistical treatment of particles in statistical mechanics and to the wavefunction formalism in quantum mechanics.
- The existence and uniqueness theorem guarantees that there is a unique optimal way to move particles from one configuration to the other with the energy cost given by the Wasserstein distance.

- As seen intuitively, the Wasserstein optimal transport plans continuously transform densities into others similarly to physical dissipative processes.
- Indeed, as we will see in this chapter, the gradient flow of the *negative* entropy in the space $(\mathcal{P}_2(\mathbb{R}^n), d_W)$ is the diffusion equation.

The last point in particular makes one wonder if the space $(\mathcal{P}_2(\mathbb{R}^n), d_W)$ might not be natural for dissipative systems, in which entropy plays an important role. This is true as we will see later for the Fokker-Planck equation: the gradient flow with respect to $(\mathcal{P}_2(\mathbb{R}^n), d_W)$ of the free energy functional of an ensemble of particles in an external field is the Fokker-Planck equation. One might then hope the Wasserstein distance to be a general tool to connect physical energy principles to dissipative equations of motion, similarly to the Lagrangian formalism, but this simple idea does not seem to be directly applicable.

A different but general question might be: in what space will a gradient flow link a physical energy principle to a dissipative equation of motion? We will briefly present at the end of the chapter a few such pairs linked through a gradient flow with respect to the Wasserstein distance or to a related "metric". This might seem to indicate that dissipative equations and optimal transportation theory are deeply related.

Before we continue, let us point out that many heuristic arguments in this chapter are based on the application of well-known thermodynamic principles, notably the Second Law of thermodynamics: entropy is always increasing for spontaneous processes. For our purpose, we only need this concept to justify that physically reasonable dynamics should follow from maximizing the dissipation of

free energy, the difference between the energy of the system and its entropy. If the system is described by a probability density function ρ, the relevant negative entropy is the Gibbs or Shannon entropy

$$S : \rho \mapsto \int \rho \ln(\rho) \ . \tag{5.1}$$

This choice arises from physical considerations and the interested reader is directed to Landau's Statistical Physics course [14] for more developments.

To make the previous ideas clear, we will begin by briefly reviewing the main ideas of gradient flow theory and the Fokker-Planck equation. We will then present the result of Jordan, Kinderlehrer and Otto that this equation is a gradient flow with respect to the Wasserstein distance. Finally, we will briefly introduce similar results that link Wasserstein and Wasserstein-like metrics to dissipative equations of motion.

5.1 Gradient Flows

The notion of a gradient flow can be described in simple terms. Take a smooth energy functional F over a metric space (X, d) satisfying the boundedness condition $F : X \to [0, \infty)$. Given an initial point $x_0 \in X$, is it possible to find a "natural" path in X that carries x_0 into a minimum of F? Precisely, does there

exist a map $x(t) \subset X$ satisfying

$$\begin{aligned} x(0) &= x_0 \\ \lim_{t\to\infty} F[x(t)] &= \inf_{z\in X} F[z] \\ \lim_{t\to\infty} x(t) &= \text{a ``minimizer'' of } F \text{ in } X \end{aligned} \tag{5.2}$$

If such a map exists, it is called a gradient flow for F with initial point x_0. Note that a major issue with this formulation is that it is possible for $x(t)$ to become "stuck" in a local minimum of F or even diverge if F does not grow far away from the origin. These issues may be avoided if the functional F is explicitly required to be convex and lower semicontinuous, as described by Evans in the Gradient Flows section of [15]. In the calculus of variations framework, this condition ensures that there exists a single point x that satisfies the Euler-Lagrange equations. Alternatively, x is the only point in X where the first variation of F vanishes. An intuitive diagram of a gradient flow is presented in figure 5-1.

A very simple example of this problem would be to set $X = \mathbb{R}^n$ and let $x(t)$ evolve according to the gradient of F. The procedure is best explained by using discrete time steps: let $h > 0$ be a small time increment and consider the sequence of points $\{x(i)\}_{i\in\mathbb{N}}$ defined by:

$$\begin{aligned} x(0) &= x_0 \\ x(i+1) &= x(i) - h\nabla F[x(i)] \end{aligned} \tag{5.3}$$

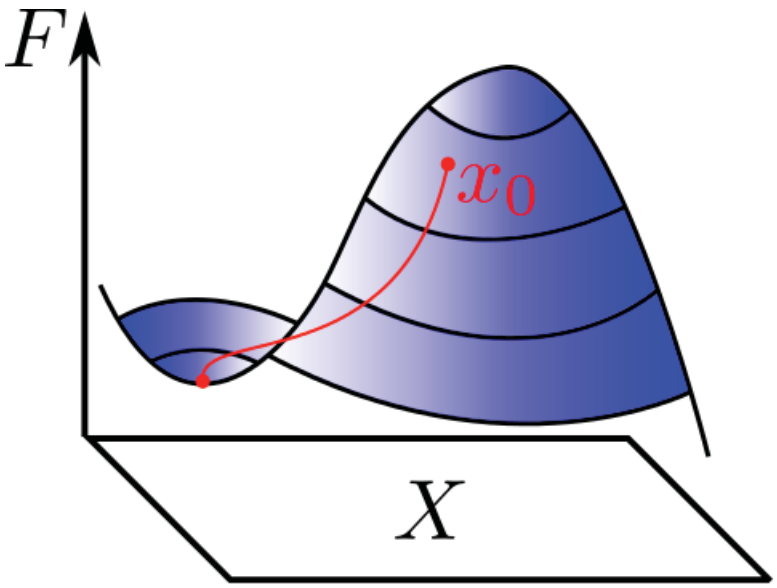

Figure 5–1: In this illustration, the plane represents the metric space X and the graph of F is drawn in blue. The red curve is the gradient flow starting at x_0 and moving into a *local* minimum of F.

In words, $x(i+1)$ is determined by moving from $x(i)$ into the direction where F is decreasing the fastest. If we now take the continuous time interpolation of the sequence $x(i)$

$$x(t) = x(i) \text{ where } i = \lfloor t/h \rfloor \ , \tag{5.4}$$

where $\lfloor \cdot \rfloor$ denotes the floor function, then it is intuitively clear that as $h \to 0$, $x(t)$ becomes a continuous curve in X that may satisfy the properties of a gradient flow. Using the numerical analogy further, the combination $\frac{x(i+1)-x(i)}{h}$ can be thought of as the first-order approximation to the time derivative of x, such that the gradient flow may also be understood as a solution of the PDE

$$\frac{dx}{dt} = -\nabla F[x(t)] \tag{5.5}$$

with $x(0) = x_0$. Note that the steady-states of this equation correspond to the extremizers of F. In this basic setting, the numerical method is aptly named a "gradient descent" and is used to locate local or global minima of functions in $\mathbb{R}^n$, depending on the choice of x_0.

When the underlying space is an infinite dimensional space of functions, say $X = L^2(\mathbb{R}^n)$ or $X = H^{-1}(\mathbb{R}^n)$, the same ideas can be used with the slight modification that the "gradient" of the functional F must be replaced by an appropriate derivative: the functional derivative, denoted by grad$F$. The gradient flow in these spaces is then given by a solution of the analogous PDE

$$\frac{\partial u}{\partial t} = -\text{grad}F[u(t)] \tag{5.6}$$

where $u(t) \subset X$ for $t \in [0, \infty)$ and $u(0) = u_0$. Without exposing too many details, the functional derivative corresponds to the Euler-Lagrange equations of the functional *with respect to the chosen metric*. This last statement is of crucial importance: the same functional can give rise to different gradient flows depending on the chosen metric space. The ubiquitous example is that of the Allen-Cahn and Cahn-Hilliard equations, well-known models for smoothing and homogenization processes. Consider the functional F_W acting on $u(t) : \mathbb{R}^n \to [0, 1]$, with the initial spatial distribution $u(0) = u_0$,

$$F_W : u \mapsto \int_{\mathbb{R}^n} \frac{1}{2}|\nabla u|^2 + W(u)dx \tag{5.7}$$

for some potential W. This potential is typically a double well with minima at 0 and 1, $u^2(1-u)^2$ for example. The following result holds:

Proposition 5.1. *The gradient flow of F_W with respect to the $L^2(\mathbb{R}^n)$ metric is a solution of*

$$\frac{\partial u}{\partial t} = \nabla^2 u - W'(u) \tag{5.8}$$

while the gradient flow of F_W with respect to the $H^{-1}(\mathbb{R}^n)$ metric is a solution of

$$\frac{\partial u}{\partial t} = -\nabla^2 \left(\nabla^2 u - W'(u)\right) \tag{5.9}$$

both with the initial condition $u(0) = u_0$.

These two PDEs behave very differently and model completely different physical situations. In particular, the steady states of these PDEs are *not* necessarily the same and must indeed model different smoothing and homogenization processes. The reader is directed to [16] for explicit calculations and exact definitions.

From this last example, it must be appreciated that a gradient flow is a construction that relates three kinds of objects:

- An energy functional
- An underlying metric space
- A dissipative partial differential equation

If two of these objects are fixed, an interesting problem is to obtain the third that completes the gradient flow procedure. Writing a PDE from a functional and

a metric is well understood, but the other directions have not been investigated as much.

With this brief introduction to gradient flows, we now wish to consider a special case of the given problem where the energy functional is the free energy of a system of particles and the PDE is the Fokker-Planck equation. We will see in the next few sections that the metric that connects the two is the Wasserstein distance by using a discrete gradient flow formalism.

5.2 The Fokker-Planck Equation and the Jordan-Kinderlehrer-Otto Scheme

In this section and the following, we review the approach of Jordan, Kinderlehrer and Otto in [7] to show that the Wasserstein distance produces physically relevant gradient flows. This property was realized by Otto in the context of pattern formation in magnetic fluids [6], and was generalized as the Jordan-Kinderlehrer-Otto scheme in [7]. In essence, the scheme relies on the construction of a discrete gradient flow with respect to the Wasserstein distance. This discrete flow converges in an appropriate sense to a limit that solves a certain PDE. We then say that this limit *is* the Wasserstein gradient flow of the functional, and that this gradient flow solves this PDE. In the spirit of the original paper, we will consider a standard free energy functional to write the Fokker-Planck equation as a gradient flow. To do so, we will first introduce the necessary objects required in the proof, then highlight its most important aspects. In the next section, we will

present the formal Otto Calculus, which is a direct but a priori formal method to compute Wasserstein gradient flows.

The Fokker-Planck equation

The Fokker-Planck or Smoluchowski equation models the behavior of a large number of weakly interacting particles in an external potential. An individual particle will experience random interaction forces leading to diffusion and an external force due to the potential. The qualitative behavior of the system will then be an approximate superposition of diffusion and the motion of a single particle in the same potential. Because of this, the Fokker-Planck equation can be thought of as a stochastic generalization of the deterministic single particle dynamics of classical or quantum mechanics to an ensemble of particles.

The ensemble of particles is represented by a time varying probability density function $\rho : \mathbb{R}^n \to [0, \infty)$, where $\rho(x)$ corresponds to the probability of finding a particle at position x. Formally, given a volume element dV and a region R of $\mathbb{R}^n$, $\int_R \rho(x)dV$ corresponds to the number of particles found in R. For our purpose, we will assume that ρ is unit normalized and has finite second moment. We will then work with the space $\mathcal{P}_2(\mathbb{R}^n)$; recall from equation (2.14):

$$\mathcal{P}_2(\mathbb{R}^n) = \left\{ \rho \in \mathcal{P}(\mathbb{R}^n) \;\middle|\; M[\rho] = \int_{\mathbb{R}^n} |x|^2 \rho(x) dx < \infty \right\}$$

Recall that elements in $\mathcal{P}_2(\mathbb{R}^n)$ may be represented by an $L^1(\mathbb{R}^n)$ probability density function. The additional requirement that $M[\rho]$ be finite is necessary to ensure that the free energy of the system (or its partition function) is finite.

The external potential will be a smooth function $\Psi : \mathbb{R}^n \to [0, \infty)$ satisfying the requirement that

$$|\nabla \Psi| \leq C(\Psi + 1) \tag{5.10}$$

for some constant C. Note that the energy of a particle at position x is given by $\Psi(x)$.

The last ingredient is temperature, or the "inverse temperature" $\beta > 0$, which controls the timescale of diffusion processes. Let then $\rho(t)$ be the time evolution of the particle density $\rho \in \mathcal{P}_2(\mathbb{R}^n)$; the Fokker-Planck equation may be written in the form

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla \Psi) + \frac{1}{\beta} \nabla^2 \rho \tag{5.11}$$

with an appropriate initial condition.

The first term on the right corresponds to the dynamics given by the potential while the other term corresponds to diffusion. Note that this equation degenerates into the diffusion equation when Ψ is constant, and into the single particle dynamics when the temperature vanishes or $\beta \to \infty$.

The development in the remaining of this section will be to show that the solutions to (5.11) can be written as the Wasserstein gradient flow of the free energy functional

$$F[\rho] = \int_{\mathbb{R}^n} \rho(x) \Psi(x) dx + \frac{1}{\beta} \int_{\mathbb{R}^n} \rho(x) \ln(\rho(x)) dx \ . \tag{5.12}$$

Intuitively, F has the thermodynamic interpretation of "Free Energy = Energy - Temperature $\times$ Entropy" if the energy and (negative) entropy are defined

by:

$$\begin{aligned} E[\rho] &= \int_{\mathbb{R}^n} \rho(x)\Psi(x)dx \\ S[\rho] &= \int_{\mathbb{R}^n} \rho(x)\ln(\rho(x))dx \end{aligned} \tag{5.13}$$

The free energy functional may be written as $F[\rho] = E[\rho] + \frac{1}{\beta}S[\rho]$ which is ubiquitous in statistical mechanics. To complete this heuristic interpretation, the steady states of the Fokker-Planck equation correspond to local or global minimizers of F. Given necessary growth conditions, a formal calculation using the calculus of variations shows that the minimizer of F corresponds to the Gibbs distribution obtained in statistical mechanics:

$$\rho_{\text{steady}}(x) = \frac{1}{Z}e^{-\beta\Psi(x)} \text{ where } Z = \int_{\mathbb{R}^n} e^{-\beta\Psi(x)}dx \tag{5.14}$$

It is standard to call $e^{-\beta\Psi(x)}$ the Boltzmann factor and Z the partition function of the system. In equilibrium, $\frac{\partial\rho}{\partial t} = 0$ and it is easy to verify that $\rho_{\text{steady}}(x)$ solves the Fokker-Planck equation.

Without loss of generality, we now let $\beta = 1$. We have now laid the setup of our next developments and now turn to the discrete scheme that will become our gradient flow.

The discrete scheme

We now come to the main section of this chapter: the Jordan-Kinderlehrer-Otto or JKO scheme. The scheme is a reformulation of the gradient descent approach (5.3) adapted to the Wasserstein distance to compute the gradient flow

of an energy functional F. Let our underlying space be $(\mathcal{P}_2(\mathbb{R}^n), d_W)$ and take an initial probability density $\rho_0 \in \mathcal{P}_2(\mathbb{R}^n)$. Define the small parameter $h > 0$ corresponding to the length of a time step. The JKO scheme is an iterative algorithm that computes the sequence $\{\rho_i^h\}_{i\in\mathbb{N}^*}$ using the formula

$$\rho_{i+1}^h = \text{argmin}_{\rho\in\mathcal{P}_2(\mathbb{R}^n)} \left\{ \frac{1}{2} d_W^2(\rho_i^h, \rho) + hF[\rho] \right\} . \tag{5.15}$$

We also define the interpolation $\rho^h(t) \subset \mathcal{P}_2(\mathbb{R}^n)$ for $t \in [0, \infty)$ to be

$$\rho^h(t) = \rho_i^h \text{ where } i = \lfloor t/h \rfloor . \tag{5.16}$$

The interpolation $\rho^h(t)$ can then be interpreted as a discrete gradient descent of F with respect to the Wasserstein distance. We will highlight details of the proof that the limit $\rho(t) = \lim_{h\to 0} \rho^h(t)$ is actually a continuous gradient flow that solves an equation of motion. For now, let us examine the JKO scheme further and formally justify why it can be thought of as a gradient descent and why it is well-behaved.

First, since a step of the JKO scheme is a minimization problem, subtracting the constant $-hF[\rho_i]$ will not change the minimizer, such that the step can be rewritten in the intuitive form

$$\rho_{i+1}^h = \text{argmin}_{\rho\in\mathcal{P}_2(\mathbb{R}^n)} \left\{ \frac{1}{2} d_W^2(\rho_i^h, \rho) - h(F[\rho_i] - F[\rho]) \right\} . \tag{5.17}$$

It is now clear how this formulation can be linked to the usual gradient descent approach: the difference on the right hand side corresponds to an approximation of the "gradient" of F while the Wasserstein distance corresponds to the

adaptation of the difference $|\rho_i - \rho|$ to the metric we are currently using. With the usual interpretation that the Wasserstein distance squared evaluates the work in displacing mass from ρ_i to ρ and that the difference $F[\rho_i] - F[\rho]$ corresponds to a gain in free energy from displacing the mass, the JKO scheme can be thought of as minimizing the difference "work - gain in free energy = -dissipation" from the laws of thermodynamics. Minimizing the negative dissipation is equivalent to maximizing it, therefore, the JKO scheme has the nice physical interpretation of sequentially maximizing the dissipation, or maximizing the rate of energy dissipation in a system. This is essentially what a gradient descent is doing since at each step, it evolves in the direction where functional decreases the fastest.

With this justification, it should be reasonable to expect that the interpolation (5.16) should behave, in the limit, as a continuous gradient flow, provided that certain well-posedness conditions are met. Of main interest is that the JKO scheme should not wander "outside" of the chosen class $\mathcal{P}_2(\mathbb{R}^n)$: if no maximizer of the dissipation could be found in $\mathcal{P}_2(\mathbb{R}^n)$. This could happen if either d_W or the difference in F would not be well-defined, or if each step would increase $M[\rho_i]$ without bound. From the results we have obtained about the Wasserstein distance, the first problem is a non-issue since this distance is well-defined on $\mathcal{P}_2(\mathbb{R}^n)$. The other issues must explicitly be verified to not happen: i.e., M and F must be bounded for all finite times[1] .

[1] This bound needs not remain valid in the other limit as *time* grows to infinity.

Another potential issue with the method is that the interpolation $\rho(t)$ should be a solution to some equation of motion if the gradient flow approach is to be successful at linking a functional to dynamics. This turns out to be guaranteed in some cases by the existence and uniqueness of the optimal transport plans associated to the Wasserstein distance given $L^1(\mathbb{R}^n)$ probability densities. This is where our developments on optimal transport become of quintessential importance.

For simplicity of exposition, we now strictly turn our attention to the Fokker-Planck equation and the functional that we have defined in the previous section. The following steps are general in that they can be applied to a variety of problems, provided that the quantities of interest are well-defined. Let us now argue that the JKO scheme is well-defined for the Fokker-Planck equation, omitting a few steps detailed in [7].

Theorem 5.2. *Suppose that $\rho_0 \in \mathcal{P}_2(\mathbb{R}^n)$, $h > 0$ is a given constant and F is as in (5.12), then there exists a unique sequence $\{\rho_i^h\}_{i\in\mathbb{N}} \subset \mathcal{P}_2(\mathbb{R}^n)$ defined by the scheme (5.15).*

Sketch of proof. The sequence ρ_i is obtained recursively, therefore, it will be uniquely defined if and only if each step of the scheme satisfies this property. In other words, we only need to show that there exists a unique $\rho_1 \in \mathcal{P}_2(\mathbb{R}^n)$ that solves the minimization problem

$$\mu = \inf_{\rho\in\mathcal{P}_2(\mathbb{R}^n)} \delta[\rho] = \inf_{\rho\in\mathcal{P}_2(\mathbb{R}^n)} \left(\frac{1}{2} d_W^2(\rho_0, \rho) + hF[\rho]\right) . \tag{5.18}$$

As in the existence proof of a maximizer to the functional J, we must build a minimizing sequence $\rho_k \in \mathcal{P}_2(\mathbb{R}^n)$ such that $\delta[\rho_k] \to \mu$ and show that this sequence has a limit $\rho_1 \in \mathcal{P}_2(\mathbb{R}^n)$ with $\delta[\rho_1] = \mu$. A minimizing sequence can be found if δ is bounded below. The non-negativity of Ψ and ρ implies that

$$F[\rho] \geq S[\rho] = \int_{\mathbb{R}^n} \rho(x) \ln(\rho(x)) dx \geq \int_{\mathbb{R}^n} \min\{0, \rho(x) \ln(\rho(x))\} dx \ . \tag{5.19}$$

For δ to be bounded below on $\mathcal{P}_2(\mathbb{R}^n)$, the Wasserstein distance must grow faster than the negative part of the entropy.

Note from the triangle inequality that

$$|y|^2 \leq (|y - x| + |x|)^2 = |y - x|^2 + |x|^2 + 2|x||x - y| \ . \tag{5.20}$$

Young's inequality applied to $|x|$ and $|x - y|$ gives that $|x||x - y| \leq |x|^2/2 + |x - y|^2/2$. Combining these inequalities, we have that $|x - y|^2 \geq |y|^2/2 - |x|^2$. Now fix $\rho \in \mathcal{P}_2(\mathbb{R}^n)$ and let τ be the optimal coupling between ρ_0 and ρ such that

$$\begin{aligned} d_W^2(\rho_0, \rho) &= \int_{\mathbb{R}^n \times \mathbb{R}^n} |x - y|^2 \tau(dx \times dy) \\ &\geq \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n} |y|^2 \tau(dx \times dy) - \int_{\mathbb{R}^n \times \mathbb{R}^n} |x|^2 \tau(dx \times dy) \\ &= \frac{1}{2} \int_{\mathbb{R}^n} |y|^2 \rho(y) dy - \int_{\mathbb{R}^n} |x|^2 \rho_0(x) dx = \frac{1}{2} M[\rho] - M[\rho_0] \ . \end{aligned} \tag{5.21}$$

We must thus show that the negative entropy of ρ grows sublinearly in $M[\rho]$ such that $\delta[\rho]$ may not become arbitrarily negative. It can indeed be shown that, bounding the growth of $-z \ln(z)$ by any fractional power of z,

$$S[\rho] \geq -C(M[\rho] + 1)^\alpha \tag{5.22}$$

for some positive constant C and an exponent $\alpha \in (\frac{n}{n+2}, 1)$ for all $\rho \in \mathcal{P}_2(\mathbb{R}^n)$. From these two arguments, it follows that

$$\delta[\rho] \geq \frac{1}{4}M[\rho] - \frac{1}{2}M[\rho_0] - C(M[\rho]+1)^\alpha \tag{5.23}$$

is bounded below for every $\rho \in \mathcal{P}_2(\mathbb{R}^n)$, it is thus possible to find a minimizing sequence $\rho_k \in \mathcal{P}_2(\mathbb{R}^n)$ for δ.

It now remains to show that this minimizing sequence converges to some $\rho_1 \in \mathcal{P}_2(\mathbb{R}^n)$. This follows by ensuring that:

$$\begin{aligned} F[\rho_1] &\leq \liminf F[\rho_k] \\ d_W^2(\rho_0, \rho_1) &\leq \liminf d_W^2(\rho_0, \rho_k) \end{aligned} \tag{5.24}$$

such that $\delta[\rho_1] \leq \liminf \delta[\rho_k] = \mu$, showing that ρ_1 solves the minimization problem.

To show that ρ_1 uniquely minimizes δ, suppose that $\rho_1^* \in \mathcal{P}_2(\mathbb{R}^n)$ is another minimizer different from ρ_1. Notice that $\mathcal{P}_2(\mathbb{R}^n)$ is convex in the sense that for $\alpha \in [0,1]$, $\rho_\alpha = \alpha\rho_1 + (1-\alpha)\rho_1^* \in \mathcal{P}_2(\mathbb{R}^n)$. Note that E is linear and that the function $z\ln(z)$ is strictly convex making S a strictly convex functional. Also, recall from Theorem 2.9 that $(\mathcal{P}_2(\mathbb{R}^n), d_W)$ is a metric space, therefore, the square metric $d_W^2(\rho, \cdot)$ is a convex function. Since δ is the sum of a linear, a convex and a strictly convex functional on $\mathcal{P}_2(\mathbb{R}^n)$, it is itself strictly convex so that

$$\delta[\rho_\alpha] < \alpha\delta[\rho_1] + (1-\alpha)\delta[\rho_1^*] = \alpha\mu + (1-\alpha)\mu = \mu \tag{5.25}$$

for any $\alpha \in (0,1)$, which is a contradiction. Hence, each step of the JKO scheme admits a unique minimizer such that there exists a unique sequence $\rho_i^h \subset \mathcal{P}_2(\mathbb{R}^n)$ as claimed. ■

From the sequence provided by the theorem, it is possible to construct an interpolation for every $h > 0$ using (5.16). Note especially that the JKO scheme does not "get stuck" in local minimas because δ has a nice convex structure. Taking the limit as $h \to 0$, a continuous function of time can be obtained. This is the main result of Otto's work in [6] and [7] which may be summarized as follows in our present context:

- The limit interpolation $\rho(t)$ exists as a weak limit in $L^1(\mathbb{R}^n)$ for finite times.
- This limit is actually strong in $L^1(\mathbb{R}^n)$ for *all* times.
- This limit uniquely solves the Fokker-Planck equation and satisfies the initial condition that $\rho(0) = \rho_0$.

Therefore, the solution to the Fokker-Planck equation, $\rho(t)$, is the gradient flow of the functional F with respect to the Wasserstein distance. In particular, taking Ψ to be constant, the gradient flow of the standard entropy $\int \rho \ln \rho$ is the diffusion equation. In this example, note that the limit of the solution to the diffusion equation is 0 in an unbounded domain, so that the second moment of the interpolation $\rho(t)$ may *not* remain bounded as time grows to infinity.

We will not reproduce the proof of these statements as this is done in extensive detail in the previously cited papers and relies principally on combining the existence of optimal transport plans of the Wasserstein distance with usual

techniques in regularity theory. Instead, we now describe a technique to compute Wasserstein gradient flows without first using a discrete scheme and taking interpolation limits.

5.3 The Otto Calculus

The discrete approach we have followed previously is lengthy but analytically sound and instructive. However, it would be incredibly useful to have a formalism that allows one to calculate the Wasserstein gradient flow heuristically. Such a formal approach exists and is called the "Otto Calculus" in honor of Otto's work. We briefly mention the main result of this formalism based on Villani's exposition in [17].

To keep the notation simple, we consider energy functionals operating on densities defined on $\mathbb{R}^n$ of the form

$$F[\rho] = \int_{\mathbb{R}^n} U(\rho(x))dx \tag{5.26}$$

with U the energy density associated to the particle density. Note that this framework is valid not only in $\mathbb{R}^n$ but in a more general differential geometry setting. The following result is the basis of the Otto calculus:

Theorem 5.3. *With our usual notation, the Wasserstein gradient flow of F is given by*

$$\frac{\partial \rho}{\partial t} = -gradF[\rho] = \nabla \cdot (\rho \nabla U'(\rho)) \tag{5.27}$$

where U' is the derivative of U with respect to its argument.

A trivial calculation immediately shows that this holds for the Fokker-Planck equation. We shall not attempt to justify this formula, referring the reader to [17] for further discussion.

5.4 Other Wasserstein Gradient Flows

In this final chapter, we have argued that the Fokker-Planck equation naturally arises as the gradient flow of the Fokker-Planck functional F with respect to the 2-Wasserstein distance, the Otto calculus providing a formalism to simplify the calculations. We now present three other examples of such gradient flows, one of which makes use of a *modified* "Wasserstein distance".

The porous medium equation

After the publication of the JKO scheme in [6] and [7], Otto published [18] in which he undertakes the study of the porous medium equation

$$\frac{\partial \rho}{\partial t} = \nabla^2(\rho^m) \tag{5.28}$$

with the usual notation and a positive number m. It can be shown that m must be chosen to be greater than the maximum of $1 - \frac{1}{n}$ and $\frac{n}{n+2}$ for the equation to make sense. This equation introduces a non-linearity in the usual diffusion equation that favors "fast" diffusion and homogenization if $m > 1$ and "slow" diffusion if $m < 1$. This modification of the usual linear equation models systems in which the

speed of diffusion depends on the density, for example, when the diffusion of gas molecules is obstructed by a solid porous medium. Similarly to the Fokker-Planck equation, it is possible to write the porous medium equation as the Wasserstein gradient flow of the functional, for $m \neq 1$,

$$F_{\text{porous}}[\rho] = \frac{1}{m-1} \int_{\mathbb{R}^n} \rho^m dx \ . \tag{5.29}$$

Indeed, first observe that $\nabla(\rho^m) = m\rho^{m-1}\nabla\rho$. Since $U(\rho) = \frac{1}{m-1}\rho^m$,

$$\nabla U' = \nabla\left(\frac{m}{m-1}\rho^{m-1}\right) = \frac{m}{m-1}(m-1)\rho^{m-2}\nabla\rho = m\rho^{m-2}\nabla\rho \tag{5.30}$$

The Otto calculus gives that the gradient flow of F_{porous} with respect to the Wasserstein distance is

$$\frac{\partial\rho}{\partial t} = \nabla\cdot(\rho\nabla U') = \nabla\cdot\left(m\rho^{m-1}\nabla\rho\right) = \nabla\cdot\nabla(\rho^m) = \nabla^2(\rho^m) \ . \tag{5.31}$$

Therefore, the porous medium equation arises as the Wasserstein gradient flow of F_{porous}. This result may be shown rigorously by using techniques similar to those presented in this chapter, but adapted to a Riemannian geometry setting. Such a reformulation is necessary to decouple the energetic and entropic driving forces in the problem which has the advantage of making the rigorous analysis simpler and more transparent. The interested reader is directed to Otto's rigorous proof and exposition in [18].

Aggregation equations

Another example of Wasserstein gradient flow comes about in the context of aggregation equations [19] of the form

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla K * \rho) \tag{5.32}$$

where K is the translation independent interaction kernel in the energy functional

$$F_{\text{aggregation}}[\rho] = \int_{\mathbb{R}^n \times \mathbb{R}^n} K(x-y)\rho(x)\rho(y)dxdy \ . \tag{5.33}$$

These aggregation equations model the behavior of strongly interacting particles. Indeed, the above free energy is the obvious generalization of summing the interaction energies (and self energies) between discrete particles. Of particular importance are the equations whose interaction kernel may be written as the sum of attractive and repulsive potentials of the form

$$K(x) = \frac{1}{q}|x|^q - \frac{1}{p}|x|^p \tag{5.34}$$

where $-N < p < q$. Such models have been applied to colloidal suspensions, swarm formation and other problems in pattern formation systems.

Again using the Otto calculus,

$$U(\rho(y)) = \int_{\mathbb{R}^n} K(x-y)\rho(x)\rho(y)dx \implies U'(\rho) = K * \rho \ . \tag{5.35}$$

Given some weak regularity conditions, $\nabla(K * \rho) = \rho * \nabla K$, such that the Wasserstein gradient flow of $F_{\text{aggregation}}$ is indeed an aggregation equation. The reader is directed to [19] and [20] for rigorous calculations and related results.

The Vlasov-Poisson-Fokker-Planck system

A similar kind of gradient flow can be found in the work of Huang and Jordan [21], where the equation of motion describes a plasma in a very particular regime. For simplicity, the plasma consists of electrons at position x with velocity u; it can then be described with the density distribution $p(x, u)$ in the phase space $\mathbb{R}^{2n}$. The electrons interact through the electromagnetic interaction which can be split into two components:

- A "long range" self-consistent electromagnetic field.
- Brief and "short range" collision events.

The magnetic interaction is neglected, further simplifying the first field to the electric field $E(x) = -\nabla_x \phi(x)$. The second contribution accounts for the brief events in which a pair of electrons interact at close range. Under some conditions on the temperature of the plasma, these collisions may be assumed to be inefficient, meaning that the momenta of the particles are almost unchanged after each collision. This then leads to the slow diffusion of $p(x, u)$ in the *velocity* argument, characterized by a damping constant β and a diffusion constant σ.

The system of equations that describe the behavior of the electron density distribution in this context is the Vlasov-Poisson-Fokker-Planck (VPFP) system[2] :

$$\begin{aligned} \frac{\partial p(x,u)}{\partial t} + u \cdot \nabla_x p(x,u) + \nabla_u \cdot ((E(x) - \beta u)p(x,u)) &= \sigma \nabla_u^2 p(x,u) \\ \nabla_x^2 \phi(x) &= \int_{\mathbb{R}^n} p(x,u)du \\ \lim_{|x| \to \infty} \phi(x) &= 0 \end{aligned} \tag{5.36}$$

The free energy of the system may be written as $F_{\text{VPFP}} = H - \frac{\sigma}{\beta} S$ where

$$\begin{aligned} H[p] &= \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n} \left(\phi(x) + |u|^2 \right) p(x,u) dx du \\ S[p] &= - \int_{\mathbb{R}^n \times \mathbb{R}^n} p(x,u) \ln(p(x,u)) dx du \end{aligned} \tag{5.37}$$

are the energy and entropy of the system. For further discussion on plasma physics and the regime of application of the VPFP system, the reader is directed to [22].

From the mathematical point of view, the insight provided by Huang and Jordan was to recognize that the VPFP system could be written as the gradient flow of the energy functional F_{VPFP} with respect to a distance functional similar to the Wasserstein distance. This new distance is the total cost of the optimal transport plan between the densities p_1 and p_2 given the cost density

$$c(x,u;y,v) = \frac{1}{2}|u-v|^2 + \frac{1}{2h^2}|(y - hv) - (x + hu)|^2 \ . \tag{5.38}$$

[2] Vlasov for the plasma physicist Anatoly Vlasov; Poisson for the electrostatic Poisson equation; Fokker-Planck for the diffusion dynamics in the electrostatic potential.

This cost measures the work in changing the kinetic energy of particles with velocity distribution u to v in the first term, and the spread in the particle positions after one time-step h in the second term. This cost is obviously much more difficult to work with than the usual, simple, quadratic cost of the Wasserstein distance, yet it is "adapted" to the physical problem. The overall procedure to prove these claims rigorously is analogous to the Jordan-Kinderlehrer-Otto scheme but involves more subtleties due to the complexity of the new cost function.

The physical relevance of Wasserstein gradient flows

The natural question at this point is to wonder whether the Wasserstein distance, or indeed optimal transport total costs, have physical relevance or not. Since it is possible to express the diffusion equation as the L^2 gradient flow of the Dirichlet energy, there are surely other combinations of metric and energy functional that give rise to the equations discussed above. Assuming that the equation of motion is somehow fundamental since it is experimentally observable, there are then two opposing points of view: which of the metric and energy functional is also fundamental.

If we consider the metric to be fundamental, both the Gibbs or Shannon entropy *and* the Dirichlet energy could be called "entropy" since the later also gives rise to the diffusion equation in the L^2 metric. In this framework, if we accept that the Wasserstein distance is the "correct" metric to describe dissipative equations, we must conclude that F_{VPFP} is not correct even though it is derived from physical principles.

On the other hand, if energy functionals are to be fundamental, the metric that correctly produces the equation of motion from a functional must have physical meaning. This reasoning implies that the modified cost function presented above must have deep physical relevance to plasma systems.

As usual in mathematical physics, the previous points of view are not necessarily contradictory, the "truth" often being a complex superposition. We choose to leave this questioning for further investigations and adopt the conservative position that the Wasserstein distance, and optimal transport total costs in general, turn out to be very convenient tools for the analysis of many dissipative PDEs, but it remains to be shown that they have deep physical meaning.

CHAPTER 6
Conclusion

In this thesis, we have reviewed two very important subjects: optimal transportation theory and gradient flow theory. We have sequentially generalized the basic Monge problem to the modern Monge-Kantorovich formulation and introduced the related Wasserstein distance. To gain intuition into the nature of optimal transport plans, we presented a numerical formalism to obtain discrete approximations to these maps and visualize the displacement of mass. We then reviewed the important result of existence and uniqueness for optimal transport problems with strictly convex costs. We briefly introduced the gradient flow formalism and connected this variational method to optimal transport with the Jordan-Kinderlehrer-Otto scheme. Finally, we introduced several Wasserstein gradient flows for dissipative equations, notably the Fokker-Planck equation.

While there exist many reviews of subjects presented in this thesis, we believe to have succeeded in combining such ideas to elaborate a concise overview of key results in optimal transportation theory, leading to its connection to gradient flows and mathematical physics. We hope that our work will prove useful to students and researchers who wish to tackle the optimal transport problem and its multiple applications. In particular, we hope that other mathematicians will take an interest in the problem of connecting physical energy principles to PDEs through Wasserstein-like metrics and gradient flows.

REFERENCES

[1] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématiques et de Physique pour la même année*, pp. 666–704, 1781.

[2] S. T. Rachev, "The Monge-Kantorovich mass transference problem and its stochastic applications," *Theory of Probability & Its Applications*, vol. 29, no. 4, pp. 647–676, 1985.

[3] Y. Brenier, "Décomposition polaire et réarrangement monotone des champs de vecteurs," *Comptes Rendus de l'Académie des Sciences-Serie I-Mathématique*, vol. 305, no. 19, pp. 805–808, 1987.

[4] W. Gangbo and R. J. McCann, "Optimal maps in Monge's mass transport problem," *Comptes Rendus de l'Académie des Sciences-Serie I-Mathématique*, vol. 321, no. 12, pp. 1653–1658, 1995.

[5] W. Gangbo and R. J. McCann, "The geometry of optimal transportation," *Acta Mathematica*, vol. 177, no. 2, pp. 113–161, 1996.

[6] F. Otto, "Dynamics of labyrinthine pattern formation in magnetic fluids: a mean-field theory," *Archive for Rational Mechanics and Analysis*, vol. 141, no. 1, pp. 63–103, 1998.

[7] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the Fokker-Planck equation," *SIAM Journal on Mathematical Analysis*, vol. 29, no. 1, pp. 1–17, 1998.

[8] R. J. McCann and N. Guillen, "Five lectures on optimal transportation: geometry, regularity and applications," *Analysis and geometry of metric measure spaces: lecture notes of the séminaire de Mathématiques Supérieures (Montréal)*, pp. 145–180, 2011.

[9] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions," *The Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.

[10] A. Oberman. Personal communication.

[11] J. D. Jackson, *Classical Electrodynamics*. Wiley, 3 ed., 1998.

[12] G. Carlier, "Optimal transportation and economic applications," *New mathematical models in economics and finance: lecture notes of the Institute for Mathematics and its Applications (Minneapolis)*, 2010.

[13] H. Federer, *Geometric measure theory*. Springer, 1969.

[14] L. D. Landau and E. M. Lifshitz, *Statistical Physics: Second Revised and Enlarged Edition*, vol. 5 of *Course of Theoretical Physics*. Pergamon Press, 2 ed., 1969.

[15] L. C. Evans, *Partial Differential Equations*, vol. 19 of *Graduate studies in mathematics*. American Mathematical Society, 2 ed., 2010.

[16] C. Cowan, "The Cahn-Hilliard equation as a gradient flow," Master's thesis, Department of Mathematics-Simon Fraser University, 2005.

[17] C. Villani, *Optimal transport: old and new*. Grundlehren der mathematischen Wissenschaften, Springer, 2009.

[18] F. Otto, "The geometry of dissipative evolution equations: the Porous Medium equation," *Communications in Partial Differential Equations*, vol. 26, no. 1-2, pp. 101–174, 2001.

[19] R. Choksi, R. C. Fetecau, and I. Topaloglu, "On minimizers of interaction functionals with competing attractive and repulsive potentials," in *Annales de l'Institut Henri Poincaré (C) Non Linear Analysis*, Elsevier, 2014.

[20] J. A. Carrillo, M. DiFrancesco, A. Figalli, T. Laurent, and D. Slepčev, "Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations," *Duke Mathematical Journal*, vol. 156, no. 2, pp. 229–271, 2011.

[21] C. Huang and R. Jordan, "Variational formulations for Vlasov-Poisson-Fokker-Planck systems," *Mathematical Methods in the Applied Sciences*, vol. 23, no. 9, pp. 803–843, 2000.

[22] R. J. Goldston and P. H. Rutherford, *Introduction to Plasma Physics*. Institute of Physics Publishing, 1995.