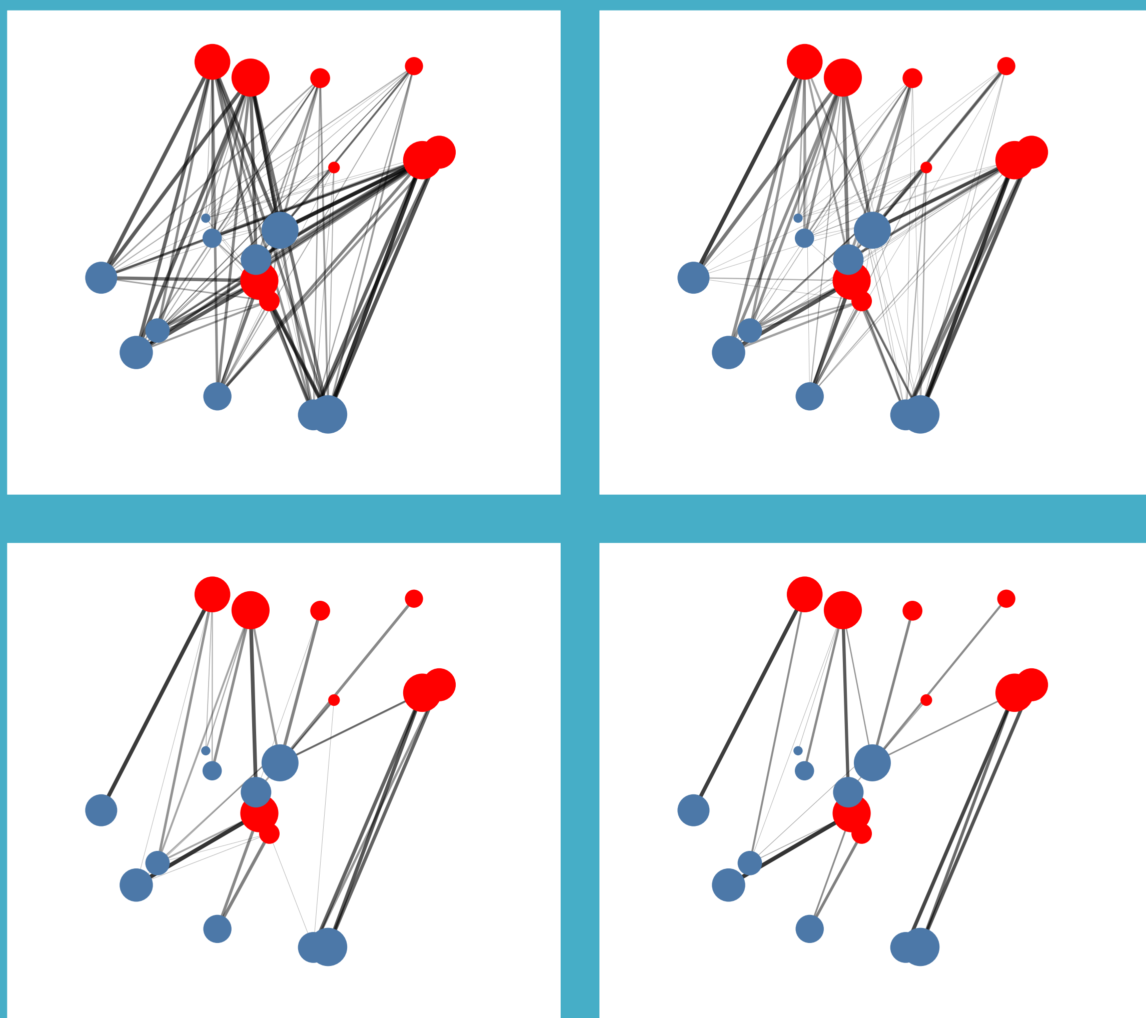
Introdução à Teoria de Transporte Ótimo

primeira edição



davi barreira &
joão miguel machado

Introdução à Teoria de Transporte Ótimo

by

Davi Barreira, João Miguel Machado

TBD

First edition, 2021

ISBN XYZ

Published by TBD

Preface

The field of Optimal Transport has grown quite substantially in recent years [1], and going through the theory in order understand how it can be applied to real world problems can be a challenging task for researchers not acquainted with the field. Hence, we have filtered the main theoretical results necessary for mathematically inclined researchers that are interested in learning Optimal Transport.

These notes are mainly based on the book "Optimal Transport for Applied Mathematicians" by Santambrogio [7]. We do not focus on proving the measurability of the sets, functions and maps, although it can be indeed shown that the ones presented here are indeed measurable.

As prerequisites to properly understand this text, we advise some knowledge of Measure Theory and Functional Analysis. Although, we've tried to either prove or state all the necessary results.

[1] Villani [10] is roughly a thousand pages of theoretical results on OT.

Contents

Notation

The following symbols are used in the text without always recalling their meaning.

- $\mathcal{M}(X), \mathcal{M}_+(X)$: Space of finite measures and finite positive measures on X, respectively.
- $\mathcal{P}(X), \mathcal{P}_p(X)$: Space of probability measures and space of probability measures with pth finite moment, respectively.
- $\mathbb{1}_A(x)$: Indicator function of set A, i.e. $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 otherwise.
- $\mathbf{1}_n$: n dimensional vector of ones.
- id: Identity operator, i.e. $id(x) = x$.
- $\oplus$: For $\phi : X \to \mathbb{R}$, $\psi : Y \to \mathbb{R}$, then $(\phi \oplus \psi)(x, y) = \phi(x) + \psi(y)$.
- π_X: Projection operator on X, i.e. for $\pi_X : X \times Y \to X$, then $\pi_X(x, y) = x, \ \forall (x, y) \in X \times Y$.
- $\Pi(\mu, \nu)$: Coupling of measures μ and ν.
- $\mathbb{R}_+$: Positive real number greater or equal than 0.
- $\overline{\mathbb{R}}$: Real numbers extended to include $+\infty$ and $-\infty$.
- $C(X)$: Set of functions $f : X \to \mathbb{R}$, where f is continuous.
- $C_b(X)$: Set of functions $f : X \to \mathbb{R}$, where f is continuous and bounded.

- $C_0(X)$: Set of functions $f : X \to \mathbb{R}$, where f is continuous and goes to zero at infinity.
- $C_c(X)$: Set of functions $f : X \to \mathbb{R}$, where f is continuous and has compact support.
- $\mu_n \rightharpoonup \mu$: Measure μ_n converges weakly to μ.
- $OT_c(\mu, \nu)$: Optimal Transport cost between measures μ and ν for a ground cost function c.
- $OT_{c,\varepsilon}(\mu, \nu)$, $\overline{OT}_{c,\varepsilon}(\mu, \nu)$: Entropic Optimal Transport distance and the Entropic Optimal Transport cost between measures μ and ν for a ground cost function c, respectively.
- W_p, $W_{p,\varepsilon}$, $S_{c,\varepsilon}$, SW, GW: Wasserstein distance, Entropic Wasserstein distance, Sinkhorn divergence, Sliced-Wasserstein distance and Gromov-Wasserstein distance, respectively.
- KL: Kullback-Leibler divergence.

List of Theorems

List of Figures

Chapter 1

Fundamentals of Optimal Transport

This chapter focuses on presenting some of the fundamentals of Optimal Transport Theory. Many current applications of Optimal Transport are related to the use of OT based metrics, such as the Wasserstein distance. Thus, we start this chapter with the basic notions of Optimal Transport, proving some results related to existence and duality, and we finish with an introduction to the Wasserstein distance. After finishing this chapter, one can already comprehend many of the applications in, for example, the field of Machine Learning.

1.1 A Brief Introduction

Before delving into formal definitions, theorems and proofs, let's give an informal overview of what is Optimal Transport, what are the main results we are interested in and how they relate to possible applications.

Optimal Transport theory main subject of study is the problem of optimally transporting quantities from one configuration to another given a cost function. Although it may seem like a very narrow subject, this seemly simple problem has a plethora of variations and can be significantly hard not only to solve, but to even prove that a solution exists.

The origin of the field of Optimal Transport is usually attributed to Gaspard Monge (1746-1818), a French mathematician, who was interested in the problem of "what is the optimal way to transport soil extracted from one location and move to another where it will be used, for example, on a construction?" [1][10]. Monge studied this problem restricting the transportation assignment to deterministic maps, i.e. the soil extracted from location x should be moved entirely to an specific location y (see Figure 1.1), a condition that is known as "non-mass splitting". Monge also considered that the cost of transportation was proportional to the distance traveled (e.i. $c(x, y) = |x - y|$), but different cost functions can be used.

Although it has been considered the founding problem of Optimal Transport, the Monge Problem is not actually the most common formulation when it comes to applications in Machine Learning. The formulation most used when referring to the Optimal Transport problem is actually due to Leonid Kantorovich (1912-1986), a Russian mathematician. Kantorovich proposed a relaxation of the non-mass splitting condition, such that the optimal transportation solution could now transport the mass "excavated" from x to many locations (see Figure 1.2).

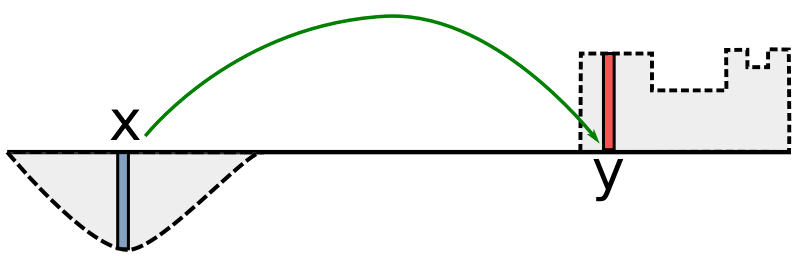


Figure 1.1: The figure illustrates the original Monge Problem, where all the mass is excavated from location x is transported to a deterministic location y. The transport assignment map is represented by the arrow in green.

[1]This is not a quote from Monge.

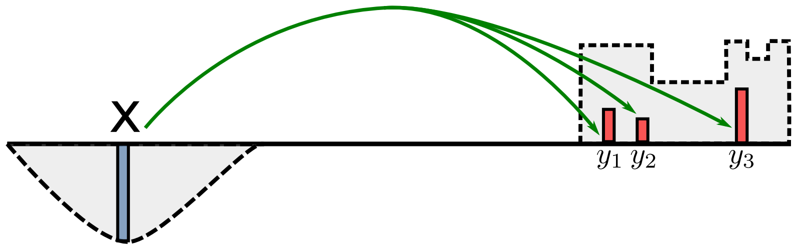


Figure 1.2: The figure illustrates the Optimal Transport Problem with the Kantorovich relaxation. The transportation assignment now can split the mass in blue, transporting it to many positions.

The transportation assignment that solves the Monge Problem is called the Optimal Transport **map**, while the solution to the Kantorovich Problem is called the Optimal Transport **plan**. As we will show in the following sections, if the Monge Problem has a solution so does the Kantorovich Problem, but the contrary is not always true. From here on out, every time we refer to the OT problem, we'll be implicitly referring to the Kantorovich formulation, unless stated otherwise.

Although the original OT problem is about soil excavation, we can apply it to abstract mathematical objects such as probability distributions. Consider two 1-dimensional probability distributions μ and ν, and define an Optimal Transport problem where the objective is to transport distribution μ to ν with $c(x,y) = |x-y|^p$ for $p \in [1, +\infty)$. Note that, if the OT problem has a solution, then there exists a minimum total cost. This minimum cost of transporting μ to ν is known as the Wasserstein distance ($W_p(\mu,\nu)$). The use of the Wasserstein distance to measure the discrepancy between probability distributions is one of the main applications of OT on Machine Learning.

If we want to use the Wasserstein, then many questions have to be answered:

- Does the transport plan exists?
- If the transport plan exists, how does one obtains it and then calculates the Wasserstein distance?
- If the Wasserstein distance between two probability distributions goes to zero, does this imply convergence in probability?

The field of Optimal Transport has addressed these types of questions, thus the importance of understanding the theory before using it on real applications.

We end this brief introduction to OT with a description of the contents addressed in each of the following section:

(i) **Monge & Kantorovich** - We formally define the Monge Problem, the Kantorovich Problem and the notion of *relaxation*. Then, we prove that for compact spaces with continuous cost functions, the Kantorovich Problem is a relaxation of the Monge Problem if the starting distribution μ is atomless;

(ii) **On the Existence of Transport Plans** - This section focuses on proving the existence of solutions to the Optimal Transport problem. We first prove the existence for compact metric spaces with continuous cost functions, which helps us prove the more general existence theorem for Polish spaces with lower semi-continuous cost functions;

(iii) **Duality Results** - The Kantorovich Problem admits a dual formulation, which, under some conditions, yields the same optimal cost as the primal formulation (i.e. strong duality). This section focuses on formally introducing the dual problem and proving the strong duality. We start from more restricted conditions which helps us prove the more general cases. We finish the section with the celebrated Kantorovich-Rubinstein Duality Theorem, which is used in Machine Learning applications such as WGANs;

(iv) **Wasserstein Distance** - We define the Wasserstein and show that it is formally a metric (4.1.1). Next, we prove that the convergence of probability measures under the Wasserstein distance is equivalent to convergence in distribution. We end the section with some comments on the properties of the Wasserstein distance and why it is useful to fields like Machine Learning.

1.2 Monge & Kantorovich

Let's start by providing some definitions that will be used throughout this section.

Definition 1.2.1. Given $(\Omega, \mathcal{F})$ where $\mathcal{F}$ is a σ-algebra, then, $\mu : \mathcal{F} \to [0, +\infty]$ is a measure if:

i) $\mu(\varnothing) = 0$

ii) $(A_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ with $A_j \cap A_i = \varnothing,\ \forall i, j \in \mathbb{N} \implies \mu(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$

We say that μ is a probability measure if besides the two properties above, we also have $\mu(\Omega) = 1$.

Definition 1.2.2. We call $\mathcal{P}(X)$ the space of probability measures defined on $(X, \mathcal{F})$, where the σ-algebra $\mathcal{F}$ is implicit and usually refers to the Borel σ-algebra.

Definition 1.2.3. (Pushforward) Let $(X, \mathcal{F})$ and $(Y, \mathcal{G})$ be measurable spaces, $T : X \to Y$ a measurable map and $\mu \in \mathcal{P}(X)$. We call $T_\#\mu$ the pushforward of μ, where:

$$T_\#\mu(B) = \mu(T^{-1}(B)), \quad \forall B \in \mathcal{G} \tag{1.1}$$

Theorem 1.2.1. *Let $T : X \to Y$ be a measurable map between $(X, \mathcal{F}, \mu)$ and $(Y, \mathcal{G})$. Then, $T_\#\mu$ is a measure on $(Y, \mathcal{G})$ and $\forall f$ measurable and integrable with respect to $T_\#\mu$ one has:*

$$\int_Y f dT_\#\mu = \int_X f \circ T d\mu \tag{1.2}$$

Proof. Let f_n be a simple positive measurable function. Hence

$$f_n(y) = \sum_{i=0}^{N} a_i \mathbb{1}_{A_i}(y) \;\; \therefore \int_Y f_n \; dT_\# \mu = \sum_{i=0}^{N} a_i T_\# \mu(A_i) = \sum_{i=0}^{N} a_i \mu(T^{-1}(A_i))$$

$$(f_n \circ T)(x) = \sum_{i=0}^{N} a_i \mathbb{1}_{A_i}(T(x)) = \sum_{i=0}^{N} a_i \mathbb{1}_{T^{-1}(A_i)}(x)$$

$$\therefore$$

$$\int_X f_n \circ T \; d\mu = \sum_{i=0}^{N} a_i \mu(T^{-1}(A_i))$$

Hence, $\int_X f_n \circ T \; d\mu = \int_Y f_n \; dT_\# \mu$.

Now, for a positive integrable measurable function f, there exists a sequence of positive simple functions such that $f_n \uparrow f$. Then, by the Monotone Convergence Theorem,

$$\begin{aligned}\int_Y f \; dT_\# \mu = \int_Y \lim_{n \to +\infty} f_n \; dT_\# \mu = \lim_{n \to +\infty} \int_Y f_n \; dT_\# \mu = \\ = \lim_{n \to +\infty} \int_X f_n \circ T \; d\mu = \int_Y f \; dT_\# \mu\end{aligned}$$

If f is non-positive, just use the same argument by splitting the negative and positive portions of the function.

□

With these definitions, we can enunciate the so called Monge Problem, which is known as the motivating problem that gave birth to the field of Optimal Transport.

Definition 1.2.4. (Monge Problem) Given two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a cost function $c : X \times Y \to [0, +\infty]$, solve:

$$(MP) \qquad \inf \left\{ \int_X c(x, T(x)) d\mu \quad : \quad T_\# \mu = \nu \right\} \tag{1.3}$$

In the Monge Problem, no mass can be split. Therefore, one can easily come up with situations in which there is no solution to the problem, as shown in 1.3. A viable solution T to MP is called a **Transport Map**.

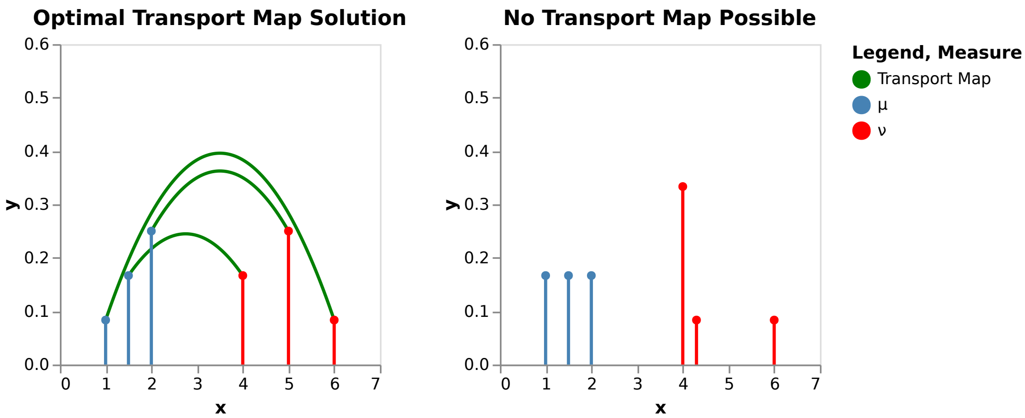


Figure 1.3: Example of two Optimal Transport Problems. On the left, there exists an optimal transport plan, while on the right there is no possible solution.

The Monge Problem is hard to solve, and, as we stated, it might not have a solution. Hence, this problem can be relaxed, becoming the so called Kantorovich Problem. This relaxation consists of allowing mass to be split, thus making the set of possible solutions larger. Before stating the Kantorovich Problem, let's introduce some more definitions.

Definition 1.2.5. (Projection and Marginal) Let $\gamma \in \mathcal{P}(X \times Y)$ and $\pi_x : X \times Y \to X$ such that $\pi_x(x, y) = x, \forall (x, y) \in X \times Y$. Hence, we say that π_x is the projection operator on X. We then call $(\pi_x)_{\#}\gamma = \mu$ the marginal distribution of γ with respect to X.

Equivalently, if for every measurable set $A \subset X$, we have $\gamma(A \times Y) = \mu(A)$, then μ is the marginal of γ with respect to X.

Corollary 1.2.1. *Given $\gamma \in \mathcal{P}(X \times Y)$, μ and ν are the marginals in X and Y, respectively $\iff$ For every f, g integrable measurable non-negative functions, we have*

$$\int_{X\times Y} f + g \; d\gamma = \int_X f d\mu + \int_Y g d\nu$$

Proof. $\implies$) Note that $(f \circ \pi_x)(x, Y) = f(\pi_x(x, Y)) = f(x)$, therefore,

$$\int_{X\times Y} f(x) \; d\gamma = \int_{X\times Y} f \circ \pi_x(x, y) \; d\gamma \underset{Theo.1}{=} \int_X f \; d(\pi_x)_{\#}\gamma = \int_X f \; d\mu$$

$\Longleftarrow$) If for all integrable measurable non-negative functions f, g we have

$$\int_{X\times Y} f + g \; d\gamma = \int_x f d\mu + \int_Y g d\nu$$

Then, for any $A \subset X$ measurable, make $f(x) = \mathbb{1}_A(x)$ and $g(y) = 0$. Hence,

$$\gamma(A \times Y) = \int_{X\times Y} \mathbb{1}_{A\times Y}(x, y) \; d\gamma = \int_{X\times Y} \mathbb{1}_A(x) \; d\gamma = \int_X \mathbb{1}_A(x) \; d\mu = \mu(A)$$

$\square$

Definition 1.2.6. (Coupling) Let (X, μ) and (Y, ν) be probability spaces. For $\gamma \in \mathcal{P}(X \times Y)$, we say that γ is a coupling of (μ, ν) if $(\pi_x)_\#\gamma = \mu$ and $(\pi_y)_\#\gamma = \nu$. Also, we call $\Pi(\mu, \nu)$ the set of **Transport Plans**:

$$\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) \; : \; (\pi_x)_\#\gamma = \mu \quad \text{and} \quad (\pi_y)_\#\gamma = \nu\} \tag{1.4}$$

Finally, we can state the Kantorovich Problem.

Definition 1.2.7. (Kantorovich Problem) Given two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a cost function $c : X \times Y \to [0, +\infty]$, solve:

$$(KP) \qquad \inf \left\{ \int_{X\times Y} c(x, y) d\gamma \; : \; \gamma \in \Pi(\mu, \nu) \right\} \tag{1.5}$$

One can prove that indeed every time the Monge Problem has a solution, so will the Kantorovich Problem. More than that, the minimal cost of both problems will indeed coincide. Note that when the Monge Problem has a solution $T : X \to Y$, then $\gamma = (id, T)_\#\mu$ is a solution to the Kantorovich Problem.

We stated in the beginning of this section that (KP) was a relaxed version of (MP). Let's now formalize this concept.

Definition 1.2.8. (Lower Semi-Continuity) A function $f : X \to \mathbb{R}$ is lower semi-continuous (l.s.c) if

$$\forall x \in X, \; f(x) \leq \liminf_{n\to+\infty} f(x_n) \tag{1.6}$$

Definition 1.2.9. (Relaxation) Given a metric space X and functional $F : X \to \mathbb{R} \cup \{+\infty\}$ bounded below. We call $\bar{F} : X \to \mathbb{R} \cup \{+\infty\}$ a of relaxation of F if:

$$\bar{F}(x) := \inf \left\{ \liminf_n F(x_n) \ : \ x_n \to x \right\} \tag{1.7}$$

Hence, $\bar{F}$ is the maximal functional G where G is lower semi-continuous and $G \leq F$.

Below in Figure 1.4 we present an example of a relaxation with the aim of improving the intuition regarding the definition. Note that, as a consequence of this definition, $\inf_x F = \inf_x \bar{F}$. Therefore, if we can prove that Kantorovich Problem is a relaxation of the Monge Problem, we would get that inf (KP) = inf (MP)

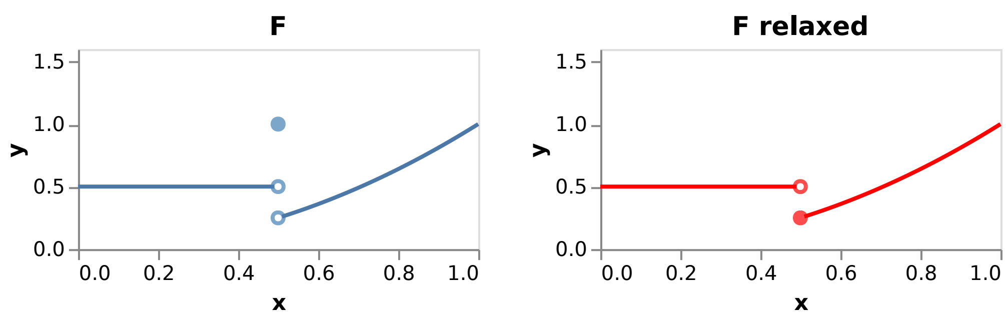


Figure 1.4: Example of a function F and it's relaxation.

To prove that indeed (KP) is a relaxation of (MP) under some conditions, we use the following theorem, for which the complete proof can be found on Santambrogio [7].

Theorem 1.2.2. *(Santambrogio 1.32) Let $\Omega \subset \mathbb{R}^d$ compact, with $c : \Omega \times \Omega :\to [0, +\infty]$ continuous and $\mu \in \mathcal{P}(\Omega)$ atomless (i.e., for every $x \in \Omega$, we have $\mu(\{x\}) = 0$). Then, the set of plans $\gamma_T = (id, T)_{\#}\mu$ induced by the map T is dense in $\Pi(\mu, \nu)$.*

We can now prove the following:

Theorem 1.2.3. *For $\Omega \subset \mathbb{R}^d$ compact, $c : \Omega \times \Omega :\to [0, +\infty]$ continuous and $\mu \in \mathcal{P}(\Omega)$ atomless. Then, (KP) is a relaxation of (MP).*

Proof. First, let's restate the Monge Problem as

$$\inf\{J(\gamma)\ :\ \gamma \in \Pi(\mu,\nu)\}$$

Where

$$J(\gamma) = \begin{cases} K(\gamma) = \int_\Omega c(x, T(x))\ d\mu = \int_{\Omega\times\Omega} c\ d\gamma_T, & \text{if } \gamma = \gamma_T \\ +\infty & \text{otherwise} \end{cases}$$

Note that indeed minimizing J is equal to minimizing the Monge Problem, since we only consider the transport plans γ_T that coincide with the cost when using a transport map T.

For $K(\gamma) = \int_{\Omega\times\Omega} c\ d\gamma$, we can show that K is continuous with respect to weak convergence (see 4.1.2), since

$$\gamma_n \rightharpoonup \gamma \iff \forall f \text{ continuous}, \int f d\gamma_n \to \int f d\gamma \implies$$

$$\implies K(\gamma_n) = \int_{\Omega\times\Omega} c\ d\gamma_n \to K(\gamma), \text{ for } c \text{ continuous.}$$

Also, by the definition of J, for any $\gamma \in \Pi(\mu,\nu)$, then $K(\gamma) \le J(\gamma)$.

By Theorem 1.2.2, for any $\gamma \in \Pi(\mu,\nu)$ we can create a sequence of $\gamma_{T_n} \rightharpoonup \gamma$. And by the continuity of K with respect to weak convergence, we have that $J(\gamma_{T_n}) = K(\gamma_{T_n}) \to K(\gamma)$. Therefore:

$$\forall \gamma \in \Pi(\mu,\nu), \exists (\gamma_{T_n})\ :\ \liminf_{n\to+\infty} J(\gamma_{T_n}) = K(\gamma)$$

Hence,

$$\inf\{\liminf_{n\to+\infty} J(\gamma_n)\ :\ \gamma_n \to \gamma\} \le K(\gamma) \le J(\gamma)$$

We can conclude that

$$\inf\{\liminf_{n\to+\infty} J(\gamma_n)\ :\ \gamma_n \to \gamma\} = K(\gamma)$$

□

1.3 On the Existence of Transport Plans

As stated before, it is not trivial to know when the Monge Problem indeed has a solution. It is easier to work with the Kantorovich Problem. In this section we present some results that relate to the existence of Optimal Transport Plans for the Kantorovich Problem.

Theorem 1.3.1. *(Santambrogio 1.4) Let X and Y be compact metric spaces. Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c : X \times Y \to [0, +\infty]$, if c is continuous, then (KP) admits a solution.*

Proof. We begin by using the notion of weak convergence to characterize continuity of functions defined on probability measures.

Note that since c is continuous and $(X \times Y)$ is compact, then c is continuous and bounded. Also, $K(\gamma) = \int_{X\times Y} c \; d\gamma$ is continuous with respect to weak convergence, since $\gamma_n \rightharpoonup \gamma$, if, and only if, for every f continuous and bounded function, it's true that $\int f \; d\gamma_n \to \int f \; d\gamma$.

Now, let's **show that** $\Pi(\mu, \nu)$ **is compact**. Take $\gamma_n \in \Pi(\mu, \nu)$. Note that γ_n is tight (4.1.3), because $(X \times Y)$ is compact. Then, by Prokhorov Theorem 4.1.3, $\exists \gamma_{n_k} \rightharpoonup \gamma$.

Take $\phi(x) \in C(X)$ and $\psi(y) \in C(Y)$. Therefore,

$$\int \phi(x) \; d\mu \underset{Cor. 1.2.1}{=} \int \phi(x) \; d\gamma_{n_k} \to \int \phi(x) \; d\gamma$$
$$\int \psi(y) \; d\nu \underset{Cor. 1.2.1}{=} \int \psi(y) \; d\gamma_{n_k} \to \int \psi(y) \; d\gamma$$

We conclude that $\gamma \in \Pi(\mu, \nu)$, which implies that $\Pi(\mu, \nu)$ is compact. Finally, since $K(\cdot)$ is continuous with respect to weak convergence and defined on a compact set, it attains a minimum. In other words, there exists a transport plan γ that minimizes the Kantorovich Problem. □

Before going into the next theorem, let's prove a small result.

Lemma 1.3.1. *Let (X, d) be a metric space and $f_k : X \to \mathbb{R}$ be l.s.c and bounded from below for every $k \in \mathbb{N}$. Then, $f = \sup_k f_k$ is also l.s.c and bounded from below.*

Proof. Since $f_k > L$, then $\sup_k f_k > L$, thus f is bounded from below.

Next, since f_k is l.s.c, therefore for $x_n \to x$:

$$f_k(x) \leq \lim_j \inf_{n \geq j} f_k(x_n) \implies \sup_k f_k(x) \leq \sup_k \lim_j \inf_{n \geq j} f_k(x_n)$$

Note that $\inf_{n\geq j} f_k(x_n) \leq \sup_k \inf_{n\geq j} f_k(x_n)$, hence

$$\lim_j \inf_{n\geq j} f_k(x_n) \leq \lim_j \sup_k \inf_{n\geq j} f_k(x_n) \implies \sup_k \lim_j \inf_{n\geq j} f_k(x_n) \leq \lim_j \sup_k \inf_{n\geq j} f_k(x_n)$$

Also, note that $\inf_{n\geq j} f_k(x_n) \leq \inf_{n\geq j} \sup_k f_k(x_n)$, hence

$$\sup_k \inf_{n\geq j} f_k(x_n) \leq \inf_{n\geq j} \sup_k f_k(x_n) \implies \lim_j \sup_k \inf_{n\geq j} f_k(x_n) \leq \lim_j \inf_{n\geq j} \sup_k f_k(x_n)$$

We conclude that $\sup_k f(x) \leq \lim_j \inf_{n\geq j} \sup_k f_k(x_n)$. So f is l.s.c.

□

Theorem 1.3.2. *(Santambrogio 1.5) Let X and Y be compact metric spaces. Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c : X \times Y \to [0, +\infty]$, if c is lower semi-continuous bounded from below, then (KP) admits a solution.*

Proof.

This proof follows the same ideas from the proof of Theorem 1.3.1. The only thing we need to prove is that $K(\gamma)$ is l.s.c with respect to weak convergence.

Let's use that for $c : X \to \mathbb{R} \cup \{+\infty\}$ bounded from below, then, c is l.s.c if and only if there exists a sequence of $k-$Lipschitz functions c_k such that $\forall x \in X$, $\sup_k c_k(x) = c(x)$.

Since c is indeed l.s.c and bounded from below, then we know that $c = \sup_k c_k$, and by the Monotone Convergence Theorem,

$$K(\gamma) = \int c \; d\gamma = \int \sup_k c_k \; d\gamma = \sup_k \int c_k \; d\gamma$$

Note that we also know that c_k are Lipschitz, hence, they are also all continuous and bounded. This implies that $K_k(\gamma) = \int c_k \; d\gamma$ is also bounded and continuous with respect to weak convergence. Therefore, $K(\gamma) = \sup_k K_k(\gamma)$, which implies that $K(\gamma)$ is l.s.c and bounded. By the Weierstrass's Theorem, we conclude that there exists a transport plan γ that minimizes the Kantorovich Problem. □

Theorem 1.3.3. *(Santambrogio 1.7) Let X and Y be Polish (complete and separable) metric spaces. Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c : X \times Y \to [0, +\infty]$, if c is lower semi-continuous then (KP) admits a solution.*

Proof.

Let's prove that $\Pi(\mu, \nu)$ is compact. To do this, we prove that $\Pi(\mu, \nu)$ is tight (4.1.3), and therefore, by Prokhorov's Theorem (i) 4.1.3, it is pre-compact. Once this is done, the proof follows in the same manner as Theorem 1.3.1.

Note that since μ and ν are probability measures, then, the families $\{\mu\}$ and $\{\nu\}$ each containing only one element are pre-compact (actually, compact). Since X is Polish, we can use Prokhorov (ii) 4.1.3, to conclude that μ and ν are tight. Hence, for $\epsilon > 0, \exists K_X \subset X$ and $K_Y \subset Y$ both compacts, such that $\mu(X \setminus K_X), \nu(Y \setminus K_Y) < \epsilon/2$.

Next, note that

$$(X \times Y) \setminus (K_X \times K_Y) \subset (X \setminus K_X \times Y) \cup (X \times Y \setminus K_Y)$$

Therefore, for any $\gamma_n \in \Pi(\nu, \mu)$ we obtain

$$\gamma_n((X \times Y) \setminus (K_X \times K_Y)) \leq \gamma_n((X \setminus K_X) \times Y) + \gamma_n(X \times (Y \setminus K_Y))$$

Finally, note that $\gamma_n(A \times Y) = \mu(A)$. Hence,

$$\gamma_n((X \times Y) \setminus (K_X \times K_Y)) \leq \mu(X \setminus K_X) + \nu(Y \setminus K_Y) < \epsilon$$

Which shows that every sequence $\gamma_n \in \Pi(\mu, \nu)$ is tight, concluding our proof.

$\square$

1.4 Duality of the Kantorovich Problem

In this section we deal with Duality Theorems regarding the Kantorovich Problem. Under some conditions, the original Kantorovich Problem (Primal) is equivalent to a Dual formulation, where instead of minimizing transport plans, one seeks to maximize potentials. Hence, we'll begin this section by introducing the notion of the Dual Problem, and then we'll prove the equivalence between the Dual and the Primal, starting from more restrictive conditions (e.g. compact spaces) and moving to more general conditions (e.g. Polish spaces). We finish the section with the celebrated Kantorovich-Rubinstein's Duality Theorem.

Before introducing the Dual Problem, we need the following result:

Lemma 1.4.1. *The Kantorovich Problem (1.2.7) is equivalent to:*

$$\inf_{\gamma \in \mathcal{M}_+(X \times Y)} \int_{X \times Y} c(x,y) d\gamma + \sup_{(\phi,\psi) \in B} \int_X \phi(x)\ d\mu + \int_Y \psi(y)\ d\nu - \int_{X \times Y} \phi(x) + \psi(y)\ d\gamma \tag{1.8}$$

Where $B := \{\phi \in C_b(X) \text{ and } \psi \in C_b(Y)\}$.

Proof. Let's suppose that $\gamma \notin \Pi(\mu,\nu)$. Then, without lost of generality, $\exists A\ :\ \mu(A) \neq \gamma(A, Y)$. Hence, can make $\phi(x) = M$ in A and null elsewhere. So,

$$\int_A \phi\ d\mu - \int_A \phi\ d\gamma = M(\mu(A) - \gamma(A,Y))$$

Since we can make M arbitrarily large or small, we conclude that

$$\sup_{(\phi,\psi) \in B} \int_X \phi(x)\ d\mu + \int_Y \psi(y)\ d\nu - \int_{X \times Y} \phi(x) + \psi(y)\ d\gamma = +\infty$$

This implies that for $\gamma \notin \Pi(\mu,\nu)$, equation (1.8) is $+\infty$. If $\gamma \in \Pi(\mu,\nu)$, then we return to

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times Y} c\ d\gamma$$

With this, we proved that the argument that minimizes equation (1.8) must be inside $\{\gamma \in \Pi(\mu,\nu)\}$, which is the original Kantorovich Problem. □

With (KP) reformulated, the Dual Problem consists of exchanging the order of the inf and the sup:

- **Primal** [2]:

$$\inf_{\gamma\in\mathcal{M}_+(X\times Y)} \sup_{(\phi,\psi)\in B} \int_{X\times Y} c\, d\gamma + \int_X \phi\, d\mu + \int_Y \psi\, d\nu - \int_{X\times Y} \phi\oplus\psi\, d\gamma \tag{1.9}$$

- **Dual**:

$$\sup_{(\phi,\psi)\in B} \inf_{\gamma\in\mathcal{M}_+(X\times Y)} \int_{X\times Y} c\, d\gamma + \int_X \phi\, d\mu + \int_Y \psi\, d\nu - \int_{X\times Y} \phi\oplus\psi\, d\gamma \tag{1.10}$$

Note that in the Dual formulation, we can rewrite it as:

$$\sup_{(\phi,\psi)\in B} \int_X \phi\, d\mu + \int_Y \psi\, d\nu - \inf_{\gamma\in\mathcal{M}_+(X\times Y)} \int_{X\times Y} c - (\phi\oplus\psi)\, d\gamma \tag{1.11}$$

If there exists an A such that for all $\forall(x,y)\in A,\ \phi(x)+\psi(y)\geq c(x,y)$, then $\inf_\gamma \int c-(\phi\oplus\psi)\, d\gamma = -\infty$ since we can choose any $\gamma\in\mathcal{M}_+(X\times Y)$.

Therefore, we can formally state the Dual Problem as:

Definition 1.4.1. Given $\mu\in\mathcal{P}(X)$, $\nu\in\mathcal{P}(Y)$ and a cost $c: X\times Y\to\mathbb{R}_+$. The Dual Problem is given by

$$\text{(DP)}\qquad \sup\left\{\int_X \phi\, d\mu + \int_Y \psi\, d\nu\ : \phi\in C_b(X)\ , \psi\in C_b(y)\ ,\ \phi\oplus\psi\leq c\right\} \tag{1.12}$$

We call **Weak Duality** if (DP) $\leq$ (KP), and we call **Strong Duality** if (DP) $=$ (KP). One can easily prove that for (KP), the Weak Duality is always true. The more interesting question is "When does one have Strong Duality?".

Lemma 1.4.2. *The Dual Problem for the Kantorovich Problem always satisfies the Weak Duality, i.e.* (DP) $\leq$ (KP)*.*

Proof. Since $\phi\oplus\psi\leq c$. Therefore,

$$\int_X \phi\, d\mu + \int_Y \psi\, d\nu = \int_{X\times Y} \phi\oplus\psi\, d\gamma \leq \int_{X\times Y} c(x,y)\, d\gamma$$

$\square$

Before starting the proof of duality, we must introduce the concepts of c-transform and c-Cyclical monotonicity.

[2] $(\phi\oplus\psi)(x,y) = \phi(x)+\psi(y)$

Definition 1.4.2. (c-Transform) Given $f : X \to \overline{\mathbb{R}}$, and $c : X \times Y \to \overline{\mathbb{R}}$, the c-transform of f is:

$$f^c(y) := \inf_x c(x,y) - f(x) \tag{1.13}$$

Function f^c is also called the c-conjugate of f. Moreover, we say that f is c-concave if $\exists\ g : Y \to \overline{\mathbb{R}}$ such that $g^c(x) = f(x)$.

Note that the c-transform is a generalization of the Legendre-Fenchel transform, which is defined as:

$$f^*(y) := \sup_x x \cdot y - f(x) \tag{1.14}$$

Lemma 1.4.3. *Let $c : X \times Y \to \overline{\mathbb{R}}$ be uniformly continuous. Define two functions $\phi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ Therefore, ϕ^c and ψ^c have the same modulus of continuity* [3] *as c.*

Proof. By Theorem 4.1.5, there exists a modulus of continuity ω, such that

$$|c(x,y) - c(x',y')| \leq \omega(d(x,x') + d(y,y'))$$

Observe that for $g_x(y) = c(x,y) - \phi(x)$

$$|g_x(y) - g_x(y')| = |c(x,y) - c(x,y')| \leq \omega(d(x,x) + d(y,y')) = \omega(d(y,y'))$$

Hence, g_x has modulus of continuity ω. Now, using the Inf-Sup Inequality 4.2.1

$$\begin{aligned} |\inf_x g_x(y) - \inf_x g_x(y')| = |\phi^c(y) - \phi^c(y')| \leq \sup_x |g_x(y) - g_x(y')| = \\ = \sup_x |c(x,y) - c(x,y')| \leq \omega(d(y,y')) \end{aligned}$$

Using the same argument for ψ^c, we showed that both $c-$transforms have the same modulus of continuity.

□

With the definition of c-transforms and the lemma above, we can prove the following theorem:

Theorem 1.4.1. *(Santambrogio 1.11)*

[3] Check Theorem 4.1.5 for the definition of modulus of continuity

For X and Y compact metric spaces, and $c : X \times Y \to \overline{\mathbb{R}}$ continuous. Then, the Dual Problem has a solution (ϕ, ϕ^c) for ϕ $c-$concave. Hence

$$\max(\text{DP}) = \max_{\phi \in c-conc.(X)} \int_X \phi \, d\mu + \int_Y \phi^c \, d\nu \tag{1.15}$$

Proof. Let (ϕ_n, ψ_n) be a maximizing sequence of the Dual problem. Note that the c-transforms always improve the Dual Problem, since $\phi_n \oplus \psi_n \leq c$, which implies that

$$\begin{aligned} \phi_n^c(y) &:= \inf_x c(x,y) - \phi_n(x) \geq \psi_n(y) \\ \psi_n^c(x) &:= \inf_y c(x,y) - \psi_n(y) \geq \phi_n(x) \\ \int_X \phi_n \, d\mu + \int_Y \psi_n \, d\nu &\leq \int_X \phi_n \, d\mu + \int_Y \phi_n^c \, d\nu \end{aligned}$$

Hence, the sequence (ϕ_n, ϕ_n^c) is also maximizing.

Since $X \times Y$ is compact, the cost c is uniformly continuous. Therefore, by Lemma 1.4.3, the $c-$transforms of ϕ_n and ψ_n are bounded by the same modulus of continuity ω as the cost function c.

Instead of using

$$\psi_n^c(x) = \inf_y c(x,y) - \psi(y)$$

We will use

$$\psi_n^c(x) := \inf_y c(x,y) - \phi_n^c(y) = \phi_n^{cc}(x)$$

This sequence is still maximizing, since

$$\begin{aligned} \phi_n^c(y) = \inf_x c(x,y) - \phi_n(x) \geq \psi_n(y) &\implies \phi_n(x) + \phi_n^c(y) \leq c(x,y) \\ &\implies \psi_n^c(x) = \inf_y c(x,y) - \phi_n^c(y) \geq \phi_n(x) \end{aligned}$$

Therefore, for a maximizing sequence (ϕ_n, ψ_n), we can instead take the maximizing sequence $(\psi_n^c, \phi_n^c) = (\phi_n^{cc}, \phi_n^c)$.

Our goal now is to use the Àrzela-Ascoli Theorem (4.1.6), so we can take a subsequence converging uniformly. To use the theorem, we'll show that our sequence (ψ_n^c, ϕ_n^c) is Equicontinuous (see Definition 4.1.5) and Equibounded (see definition 4.1.6).

First, note that (ψ_n^c, ϕ_n^c) is in fact Equicontinuous, since for any $\epsilon > 0$, we can take $\delta > 0$ such that $d(y,y') < \delta \implies w(d(y,y')) < \epsilon$ and $|\phi_n^c(y) - \phi_n^c(y')| \leq w(d(y,y')) < \epsilon$, for every $n \in \mathbb{N}$.

Next, let's prove that the sequence is Equibounded. Taking the supremum of the inequality, we obtain

$$\sup_{y,y'} |\phi_n^c(y) - \phi_n^c(y')| \leq \sup_{y,y'} w(d(y,y')) = w(\mathrm{diam}(Y))$$

The equality in the equation above is true because the function ω is increasing, and the set Y is compact. Again, the same argument works for ψ_n^c.

Next, realize that we can add and subtract constants from the Dual Problem without modifying the results:

$$\int_X \psi_n^c \; d\mu + \int_Y \phi_n^c \; d\nu = \int_X \psi_n^c + C_n \; d\mu + \int_Y \phi_n^c - C_n \; d\nu$$

Let's take $C_n = \min_y \phi_n^c(y)$. We now change the sequence of functions to $(\psi_n^c + C_n, \phi_n^c - C_n)$, which preserves the maximizing property. Note that $\min_y \phi_n^c - C_n = 0$. Hence,

$$\sup_{y,y'} |\phi_n^c(y) - \phi_n^c(y')| = \max_y \phi_n^c(y) - \min_y \phi_n^c(y) = \max_y \phi_n^c(y) \leq \omega(\mathrm{diam}(Y))$$

Also, for any $x \in X$:

$$\psi_n^c(x) = \inf_y c(x,y) - \phi_n^c(y) \in [\min_y \; c(x,y) - \omega(\mathrm{diam}(Y)), \max_y \; c(x,y) \;]$$

With this, we showed that the sequence is Equibounded. Therefore, since we are on a compact set and the sequence (ψ_n^c, ϕ_n^c) is both Equicontinuous and Equibounded, we can apply the Àrzela-Ascoli Theorem 4.1.6. Thus, we can obtain a subsequence $(\psi_{n_k}^c, \phi_{n_k}^c)$ that converges uniformly to (ψ, ϕ). As a consequence of this uniform convergence

$$\int_X \psi_{n_k}^c \; d\mu + \int_Y \phi_{n_k}^c \; d\nu \to \int_X \phi \; d\mu + \int_Y \psi \; d\nu$$

With this, we proved that there exists a pair of functions (ϕ, ψ) that are the limits of a maximizing sequence and that satisfy the constraint (i.e. $\phi(x) + \psi(y) \leq c(x,y)$), hence, the Dual Problems has a solution. Also, since $\phi^c \geq \psi$, then (ϕ, ϕ^c) is also an optimal solution for the Dual, and this maximization problem can be restricted to searching in c-concave functions, i.e.:

$$\max(\mathrm{DP}) = \max_{\phi \in c-conc.(X)} \int_X \phi \; d\mu + \int_Y \phi^c \; d\nu$$

$\square$

When Strong Duality is true, the functions ϕ, ψ that maximize the Dual Problem are called the **Kantorovich Potentials**. We haven't yet proved that $\max(\mathrm{DP}) = \min(\mathrm{KP})$, the theorem above only gave us an idea of how the solution of the Dual Problem looks-like. Before proving our first theorem on Strong Duality, we'll need a bit more definitions and results.

Definition 1.4.3. (Cyclic Monotonicity) For $c : X \times Y \to \overline{\mathbb{R}}$, a set $\Gamma \subset X \times Y$ is called c-cyclical monotone (c-CM) if $\forall n \in \mathbb{N}$ and $(x_i, y_i) \in \Gamma$ for $i \in \{1, ..., n\}$

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) \tag{1.16}$$

Where $\sigma(i)$ is a permutation of the indexes.

Note that this is a stronger property than monotonicity, since for $n = 2$ and $c(x, y) = \langle x, y \rangle$, if Γ is c-CM, then monotonicity is satisfied:

$$\langle x_1, y_1 \rangle + \langle x_2, y_2 \rangle \leq \langle x_1, y_2, \rangle + \langle x_2, y_1 \rangle \tag{1.17}$$

Definition 1.4.4. For X a separable metric space, we define the support of a a measure μ as

$$\mathrm{spt}\ \mu := \bigcap \{A \; : \; A \text{ is closed and } \mu(X \setminus A) = 0\} \tag{1.18}$$

We can now give an overview of the proof of first Strong Duality Theorem. The proof consists of showing that for an optimal plan γ, its support $\mathrm{spt}(\gamma)$ is c-CM and that for a c-CM set there exists a c-concave function $\phi(x)$ such that $\phi(x) + \phi^c(y) = c(x, y)$ for $(x, y) \in \mathrm{spt}(\gamma)$. Hence, this would prove that

$$\int_{X \times Y} c(x, y) \; d\gamma = \int_X \phi(x) \; d\mu + \int_Y \phi^c(y) d\nu \tag{1.19}$$

Theorem 1.4.2. *(Santambrogio 1.37) If $\Gamma \neq \varnothing$ and is c-CM with $c : X \times Y \to \mathbb{R}$. Then, there exists a c-concave function $\phi : X \to \mathbb{R} \cup \{-\infty\}$ (different than the constant value $-\infty$) such that*

$$\Gamma \subset \{(x, y) \ : \ \phi(x) + \phi^c(y) = c(x, y)\} \tag{1.20}$$

In other words, $\forall x, y \in \Gamma,\ c(x, y) = \phi(x) + \phi^c(y)$.

Proof. Fix a point $(x_0, y_0) \in \Gamma$. For $x \in X$, let

$$\begin{aligned} \phi(x) := \inf\{&c(x, y_n) - c(x_n, y_n) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + ...+ \\ &+ c(x_1, y_0) - c(x_0, y_0) \ : \ n \in \mathbb{N}, (x_i, y_i) \in \Gamma \ \forall i = 1, ..., n\} \end{aligned}$$

$$\begin{aligned} \psi(y) := -\inf\{&-c(x_n, y) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + ...+ \\ &c(x_1, y_0) - c(x_0, y_0)) \ : \ n \in \mathbb{N}, (x_i, y_i) \in \Gamma \ \forall i = 1, ..., n, y_n = y\} \end{aligned}$$

Note that if $y \notin (\pi_y)(\Gamma)$, then there is no $(x_n, y) = (x_n, y_n) \in \Gamma$. Therefore,

$$\psi(y) = -\inf\{\varnothing\} = -\infty$$

This implies that $\psi(y) > -\infty \iff y \in (\pi_y)(\Gamma)$. Note that:

$$\begin{aligned} \psi^c(x) &= \inf_y c(x, y) - \psi(y) = \inf_{y \in (\pi_y)(\Gamma)} c(x, y) - \psi(y) \\ &= \inf_{y \in (\pi_y)(\Gamma)} c(x, y) + \inf\{-c(x_n, y) + ... + +c(x_1, y_0) - c(x_0, y_0)) : \\ &\qquad\qquad n \in \mathbb{N}, (x_i, y_i) \in \Gamma \ \forall i = 1, ..., n, y_n = y\} \\ &= \phi(x) \end{aligned}$$

Hence, $\phi(x)$ is c-concave, and $\phi(x)$ is not constantly equal to $-\infty$, since for $x = x_0$, we have

$$\begin{aligned} &c(x_0, y_n) + (\sum_{i=0}^{n-1} c(x_{i+1}, y_i)) - \sum_{i=0}^{n} c(x_i, y_i) \geq 0 \\ \implies &\phi(x_0) = \inf\{c(x_0, y_n) + (\sum_{i=0}^{n-1} c(x_{i+1}, y_i)) - \sum_{i=0}^{n} c(x_i, y_i)\} \geq 0 \end{aligned}$$

Note that the inequality above is true due to the fact that Γ is c-CM.

Now, the only thing left to prove is that $\phi(x) + \phi^c(y) = c(x,y)$ for every $(x,y) \in \Gamma$. First, note that for $\epsilon > 0$ and $(x,y) \in \Gamma$, then:

$$\phi(x) = \psi^c(x) = \inf_y c(x,y) - \psi(y) = \inf_{y \in (\pi_y)(\Gamma)} c(x,y) - \psi(y) \implies$$
$$\exists \bar{y} \in (\pi_y)(\Gamma) \ : \ \phi(x) + \epsilon > c(x,\bar{y}) - \psi(\bar{y})$$

Also, note that from the definition of ψ, we have:

$$-\psi(y) \leq -c(x,y) + c(x,\bar{y}) - c(\bar{x}_n,\bar{y}) + ... - c(\bar{x}_0,\bar{y}_0) \ : \forall i, (\bar{x}_i,\bar{y}_i) \in \Gamma$$

Since this is true for any chain on Γ starting on $\bar{y}$, it's true for the infimum, therefore:

$$-\psi(y) \leq -c(x,y) + c(x,\bar{y}) - \psi(\bar{y}) \leq -c(x,y) + \phi(x) + \epsilon$$

Since the ϵ was arbitrary, we can conclude that $c(x,y) \leq \phi(x,y) + \psi(x)$. But, we also know that

$$\begin{aligned}
\phi^c(y) = \psi^{cc}(y) &= \inf_x c(x,y) - \phi(x) \\
&= \inf_x c(x,y) - \inf_y c(x,y) - \psi(y) \\
&\geq \inf_x c(x,y) - c(x,y) + \psi(y) \\
&= \psi(y)
\end{aligned}$$

Hence, $\phi(x) + \phi^c(y) \geq \phi(x) + \psi(y) \geq c(x,y)$.

Lastly, one would need to show that this ϕ is indeed measurable. The general proof is complicated, but, if we assume that c is uniformly continuous, then, we know that c-transforms are continuous (this was shown in Theorem 1.4.1). Since $\phi = \psi^c$, then, ϕ is continuous, therefore, it is measurable if we consider the Borel σ-algebra. □

Theorem 1.4.3. *(Santambrogio 1.38) If γ is an optimal transport plan for cost c continuous, then spt γ is c-CM.*

Proof. The proof consists in supposing that spt γ is not c-CM. Then, we construct a $\tilde{\gamma} \in \Pi(\mu,\nu)$ such that $\int_{X \times Y} c(x,y) \, d\tilde{\gamma} < \int_{X \times Y} c(x,y) \, d\gamma$, which contradicts the optimality of γ.

Check Santambrogio [7] for the complete proof. □

With these results, we can prove the first Strong Duality theorem.

Theorem 1.4.4. *For X and Y compact metric spaces, and $c : X \times Y \to \overline{\mathbb{R}}$ continuous. Then,* $\max(\text{DP}) = \min(\text{KP})$*, and DP admits a solution* (ϕ, ϕ^c).

Proof. Using Theorem 1.3.1, we obtain that $\exists \gamma \in \Pi(\mu, \nu)$ such that it minimizes the Kantorovich Problem, therefore, by Theorem 1.4.3, sptγ is c-CM.

By Proposition 1.4.1, we know that a solution to the Dual Problem can be found in the set of c-concave functions. Using 1.4.2, we can assert that there is a set of c-concave functions such that $\phi(x)+\phi^c(y) = c(x, y)$ for every $(x, y) \in \text{spt } \gamma$. Since $X \times Y$ is compact, then c is uniformly compact, which implies that ϕ and ϕ^c are continuous and bounded.

Hence, since we already know that $\max(\text{DP}) \leq \min(\text{KP})$, we conclude that $\max(\text{DP}) = \min(\text{KP})$. $\square$

Theorem 1.4.5. *For X and Y Polish spaces and $c : X \times Y \to \mathbb{R}$ uniformly continuous and bounded. Then, (DP) admits a solution (ϕ, ϕ^c) and* $\max(\text{DP}) = \min(\text{KP})$.

Proof. First, note that since X and Y are Polish and c is continuous, one can use Theorem 1.3.3 and affirm that exists an optimal solution γ to (KP).

By the same arguments used on the proof of Theorem 1.4.4, we stablish that spt γ is c-CM, and that ϕ, ϕ^c are continuous functions such that $\forall (x, y) \in$ spt γ, $\phi(x) + \phi^c(y) = c(x, y)$.

In the Dual Problem, the admissible functions ϕ and ψ must be continuous and bounded. Hence, we just need to prove that the ϕ and ϕ^c are indeed bounded. Note that, since c is bounded, then, $|c| \leq M \in \mathbb{R}$ and

$$\phi^c(y) = \inf_x c(x, y) - \phi(x) \leq \inf_x M - \phi(x) = M - \sup_x \phi(x)$$

Note that in 1.4.2, we showed that ϕ is not constantly $-\infty$. Therefore,

$$-\infty < L < \sup_x \phi(x) \implies \phi^c(y) \leq M - \sup_x \phi(x) \leq M - L$$

Similarly, since $\phi = \psi^c$ and $\phi^c(y) \geq \psi(y)$ (shown in 1.4.3), then:

$$\begin{aligned} \phi(x) = \inf_y c(x, y) - \psi(y) \geq -M - \sup_y \psi(y) &\geq -M - \sup_y \phi^c(y) \\ &\geq -M - M + L \end{aligned}$$

Hence, we obtained an upper bound for ϕ^c and a lower bound for ϕ. Now, we obtain an upper bound for ϕ and a lower bound for ϕ^c using a similar argument and relying on the fact that $\sup \psi(y) > L > -\infty$:

$$\phi(x) = \inf_y c(x,y) - \psi(y) \leq M - \sup_y \psi(y) \leq M - L$$
$$\phi^c(x) = \inf_x c(x,y) - \phi(x) \geq -M - \sup_x \phi(x) \geq -M - M - L$$

Finally, using the same arguments as Theorem 1.4.4, we conclude that $\max(\text{DP}) = \min(\text{KP})$ and that (ϕ, ϕ^c) are a solution for the Dual Problem. $\square$

One cost that is of special interest is the quadratic cost $\frac{1}{2}|x-y|^2$. Note that this cost is neither bounded nor uniformly continuous for non-compact metric spaces. Hence, the previous theorems do not address it. But one can still prove that Strong Duality is true for such case.

Theorem 1.4.6. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, with $c(x,y) = \frac{1}{2}|x-y|^2$. Suppose that $\int |x|^2 d\mu, \int |y|^2 d\nu < +\infty$ [4] . Instead of the original Dual Problem, consider the following formulation:*

$$(\text{DP}') \quad \sup \left\{ \int_{\mathbb{R}^d} \phi \; d\mu + \int_{\mathbb{R}^d} \psi \; d\nu \;\; : \phi \in L^1(\mu) \;, \psi \in L^1(\nu) \;, \; \phi \oplus \psi \leq c \right\} \tag{1.21}$$

Therefore, (DP') admits a solution (ϕ, ψ) and $\max(\text{DP}') = \min(\text{KP})$.

Proof. First, in the same way as the proof of Theorem 1.4.5, (KP) has an optimal solution γ with spt γ that is c-CM and $\forall (x,y) \in$ spt γ we have $\phi(x) + \psi(y) = c(x,y)$. We also have that $-\psi(y) = -\phi^c(y) = \sup_x -\frac{|x-y|^2}{2} + \phi(x)$. Note that, for $h(x) := \frac{|x|^2}{2} - \phi(x)$

$$h^*(y) := \sup_x \langle x, y \rangle - h(x) = \sup_x \langle x, y \rangle - \frac{|x|^2}{2} + \phi(x) =$$
$$\frac{|y|^2}{2} + \sup_x -\frac{|x-y|^2}{2} + \phi(x) = \frac{|y|^2}{2} - \psi(y)$$

[4] This is Theorem 1.40 in Santambrogio [7], but note that there is a small typo in the book, where it states $\int |x|^2 dx, \int |y|^2 dy < +\infty$ instead of the correct $\int |x|^2 d\mu, \int |y|^2 d\nu < +\infty$.

Therefore, $h(x)$ is equal to the Legendre-Fenchel transform of $\frac{|y|^2}{2} + \psi(y)$, which implies that h is convex l.s.c. The same argument can be used to show that $\frac{|y|^2}{2} - \psi(y)$ is also convex l.s.c.

Since $\frac{|x^2|}{2} - \phi(x)$ is convex, there exists a supporting hyperplane, hence, it is bounded from below by a linear function, which implies that

$$\begin{aligned} \frac{|x^2|}{2} - \phi(x) \geq \alpha\langle x, y\rangle + \beta \implies \phi(x) &\leq \frac{|x^2|}{2} - \alpha\langle x, y\rangle - \beta \\ \implies \int_{\mathbb{R}^d} \phi(x)\ d\mu &\leq \int_{\mathbb{R}^d} \frac{|x^2|}{2} - \alpha\langle x, y\rangle - \beta\ d\mu < +\infty \end{aligned}$$

The same argument can be made for ψ, which means that $\phi_+ \in L^1(\mu)$ and $\psi_+ \in L^1(\nu)$. Due to the fact that $\phi(x) + \psi(y) = c(x, y)$ in the support of γ, then

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \phi \oplus \psi\ d\gamma = \int_{\mathbb{R}^d \times \mathbb{R}^d} c\ d\gamma \geq 0$$

Which implies that the negative portions of ϕ and ψ are also integrable, leading us to conclude that $\phi \in L^1(\mu)$ and $\psi \in L^1(\nu)$.

Finally, by the same arguments as the previous theorems, we prove that $\max(\mathrm{DP}') = \min(\mathrm{KP})$.

□

A stronger result can be proven regarding the duality of KP. We'll present it here without a proof.

Theorem 1.4.7. *(Santambrogio 1.42) For X and Y Polish spaces and $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ l.s.c and bounded from below. Then,* $\sup(\mathrm{DP}) = \min(\mathrm{KP})$*. Note that in this theorem, one cannot guarantee the existence of the (ϕ, ψ) that maximize the Dual Problem.*

If the cost $c(x, y)$ is actually a distance metric (Def. 4.1.1), then we can prove the following result:

Theorem 1.4.8. *Let X be a metric space, and $c : X \times X \to \mathbb{R}$, where c is a distance metric. Therefore, a function $f : X \to \mathbb{R}$ is c-concave if and only if it is Lipschitz continuous with a constant less than 1 with respect to the distance c. We call $Lip_1^{(c)}$ this set of Lipschitz functions with constant less than 1. Moreover, $f^c = -f$.*

Proof.

$\Longrightarrow$) Let $f : X \to \mathbb{R}$ be a c-concave function. Hence, $\exists\ g : X \to \overline{\mathbb{R}}$ such that

$$f(x) := \inf_y c(x, y) - g(y)$$

Using the triangle inequality of the cost, we get:

$$c(x, y) \leq c(x, z) + c(z, y) \implies \sup_y c(x, y) - c(y, z) \leq c(x, z)$$
$$c(y, z) \leq c(y, x) + c(x, z) \implies \sup_y c(y, z) - c(x, y) \leq c(x, z)$$
$$\therefore$$
$$\sup_y |c(y, z) - c(x, y)| \leq c(x, z)$$

Therefore,

$$|f(x) - f(z)| = |\inf_y \{c(x, y) - g(y)\} \ - \ \inf_y \{c(z, y) - g(y)\}| \leq$$
$$\underset{4.2.1}{\leq} \sup_y |c(x, y) - c(z, y)| \leq c(x, z)$$

$\Longleftarrow$) Let $f \in \mathrm{Lip}_1^{(c)}$. Using the Lipschitz inequality,

$$f(x) - f(y) \leq c(x, y) \implies f(x) \leq \inf_y c(x, y) + f(y)$$

But note that $f(x) = c(x, x) + f(x) \geq \inf_y c(x, y) - f(y)$. This implies that $f(x) = \inf_y c(x, y) + f(y)$. Hence, $f(x) = g^c(x)$, where $g(y) = -f(y)$. Which proves that f is c-concave, and $f = (-f)^c$. Finally, note that $-f$ is also Lip_1, therefore, the same argumentation leads to $-f = f^c$. □

Lastly, using Theorems 1.4.7 and 1.4.8, one obtains the famous Kantorovich-Rubinstein Duality:

Theorem 1.4.9. *(Kantorovich-Rubinstein)*

Let (X, d) be a Polish space with metric d, and cost function $c(x, y) = d(x, y)$. Then, for $\mu, \nu \in \mathcal{P}(X)$, the Kantorovich Problem is equivalent to

$$\sup \left\{ \int_X \phi \ d\mu - \int_X \phi \ d\nu \ \ : \phi \in Lip_1(X) \right\} \tag{1.22}$$

1.5 Wasserstein Distance

In this section we focus on how the minimal transport cost can be used as a distance metric in the space of probability measures. Let's assume that (X, d) is a Polish metric space, d is a lower semi-continuous metric on this space and $p \in [1, +\infty)$.

Definition 1.5.1. (Probability space with p-Moments)

$$\mathcal{P}_p(X) := \{\mu \in \mathcal{P}(X) : \int_{X \times X} d(x,y)^p \; d\mu(x) d\mu(y) < +\infty\} \tag{1.23}$$

Note that this is equivalent to the set of probability measures such that $\int_X d(x, x_0) \; d\mu < +\infty$ for every $x_0 \in X$. The proof of this statement can be found in Garling [3] Proposition 21.1.1.

Definition 1.5.2. (Wasserstein Distance)

Let (X, d) be a Polish metric space, with $c : X \times X \to \mathbb{R}$ such that $c(x, y) = d(x, y)^p$, and $p \in [1, +\infty)$. For $\mu, \nu \in \mathcal{P}_p(X)$, the Wasserstein Distance is given by:

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times X} d(x,y)^p \; d\gamma \right)^{1/p} \tag{1.24}$$

Note that the restriction to $\mu, \nu \in \mathcal{P}_p(X)$ is necessary for W_p to be a distance metric. Moreover, for $p = 1$, then $c(x, y) = d(x, y)$ is a metric on X, therefore, for X Polish, one can use Kantorovich-Rubinstein's Duality Theorem 1.4.9 to obtain:

$$W_1(\mu, \nu) = \sup_{\phi \in Lip_1} \int_X f d(\mu - \nu) \tag{1.25}$$

Let's prove that W_p indeed is a metric on $\mathcal{P}_p(X)$.

Lemma 1.5.1. *(Gluing Lemma)*

Let (X, d) be a metric space. For $\mu, \nu, \rho \in \mathcal{P}(X)$ and $\gamma^+ \in \Pi(\mu, \rho)$, $\gamma^- \in \Pi(\rho, \nu)$. Then, $\exists \; \sigma \in \mathcal{P}(X \times X \times X)$ such that $(\pi_{x,y})_{\#}\sigma = \gamma^+$. $(\pi_{y,z})_{\#}\sigma = \gamma^-$.

Proof. First, use disintegration (Def. 4.1.4) with respect to $f = \pi_y$ to obtain γ_y^+ and γ_y^-. We know that such disintegration exists and is essentially unique since X is Polish (see Theorem 4.1.4). Note that disintegrated measures are actually defined on $X \times \{y\} \subset X \times X$, but, by abuse of notation, we'll consider that they are measures on X, and y is only an index.

Therefore, make $\sigma = \gamma_y^+ \otimes \rho \otimes \gamma_y^-$, and let $\phi : X \times X \to \mathbb{R}$ be a measurable function. Hence:

$$\begin{aligned}
\int_{X\times X\times X} \phi(x,y)\ d\sigma &\overset{\text{Fubini}}{=} \int_X \int_X \int_X \phi(x,y)\ d\gamma_y^+(x) \otimes \rho(y) \otimes \gamma_y^-(z) \\
&\overset{\text{Indep.}}{=} \int_X d\gamma_y^-(z)\ \int_X \int_X \phi(x,y)\ d\gamma_y^+(x) \otimes \rho(y) \\
&\overset{\text{Disint.}}{=} \int_X d\gamma_y^-(z)\ \int_{X\times X} \phi(x,y)\ d\gamma^+(x,y) \\
&= \int_{X\times X} \phi(x,y)\ d\gamma^+(x,y)
\end{aligned}$$

Since $\phi(x,y)$ is arbitrary, then by Corollary 1.2.1, we can conclude that $(\pi_{x,y})_\# \sigma = \gamma^+$. By the same argument, we obtain $(\pi_{y,z})_\# \sigma = \gamma^-$, which concludes our proof.

□

Proposition 1.5.1. *$W_p(\cdot,\cdot)$ is a metric on $\mathcal{P}_p(X)$.*

Proof. Let's prove each of the three properties that categorize a metric.

i) $d(x,y) = 0 \iff x = y$.

If $\mu = \nu$, then $(id, id)_\# \mu = \gamma$, hence $\int_{X\times X} d(x,y)^p\ d\gamma = \int_{X\times X} d(x,x)^p\ d\mu = 0$.

If $W_p(\mu,\nu) = 0$, then $\int_{X\times X} d(x,y)^p\ d\gamma = 0$. Therefore, γ is concentrated on $\{x = y\}$, otherwise, there would exist a set $A \times B$ such that $\gamma(A\times B) > 0$ and $x \neq y$. Therefore $\int_X d(x,y)^p d\gamma > 0$.

Since γ is concentrated on $\{x = y\}$, then for any set Borel set $K \subset X$:

$$\gamma(K) = \int_{X\times X} \mathbb{1}_K(x,y)\, d\gamma = \int_{x=y} \mathbb{1}_K(x,y)\, d\gamma = \int_{x=y} \mathbb{1}_K(x)\, d\mu = \int_{x=y} \mathbb{1}_K(y)\, d\nu$$

We can conclude that $\mu(K) = \nu(K)$ for every Borel set K, therefore $\mu = \nu$ almost everywhere.

ii) $d(x,y) = d(y,x)$.

$$W_p(\mu,\nu) = \left(\int_{X\times X} d(x,y)^p d\gamma\right)^{1/p} = \left(\int_{X\times X} d(y,x)^p d\gamma\right)^{1/p} = W_p(\nu,\mu)$$

iii) $d(x,z) \leq d(x,y) + d(y,z)$.

Let $\mu,\nu,\rho \in \mathcal{P}_p(X)$, and $\gamma^+ \in \Pi(\mu,\rho)$, $\gamma^- \in \Pi(\rho,\nu)$ are the optimal transport plans for the respective measures. Using the Gluing Lemma 1.5.1, we know that there exists a measure $\sigma \in \mathcal{P}(X\times X\times X)$, where $(\pi_{x,y})_{\#}\sigma = \gamma^+$ and $(\pi_{y,z})_{\#}\sigma = \gamma^-$. Also, let $\gamma := (\pi_{x,z})_{\#}\sigma$. Hence,

$$\begin{aligned}
W_p(\mu,\nu) \quad &\leq \left(\int_{X\times X} d(x,z)^p \ d\gamma\right)^{1/p} = \left(\int_{X\times X} d(x,z)^p \ d(\pi_{x,z})_{\#}\sigma\right)^{1/p} \\
&\underset{Thm. 1.2.1}{=} \left(\int_{X\times X\times X} d(x,z)^p \ d\sigma\right)^{1/p} \\
&\leq \int_{X^3} (d(x,y) + d(y,z))^p \ d\sigma \\
&= ||d\circ(\pi_{x,y})(x,y,z) - d\circ(\pi_{y,z})(x,y,z)||_{L^p(\sigma)} \\
&\underset{4.2.2}{\leq} ||d\circ(\pi_{x,y})(x,y,z)||_{L^p(\sigma)} + ||d\circ(\pi_{y,z})(x,y,z)||_{L^p(\sigma)} \\
&= \left(\int_{X^3} d(x,y)^p \ d\sigma\right)^{1/p} + \left(\int_{X^3} d(y,z)^p \ d\sigma\right)^{1/p} \\
&= \left(\int_{X^2} d(x,y)^p \ d\gamma^+\right)^{1/p} + \left(\int_{X^2} d(y,z)^p \ d\gamma^-\right)^{1/p} \\
&= W_p(\mu,\rho) + W_p(\rho,\nu)
\end{aligned}$$

Which proves the triangle inequality for the Wasserstein distance.

□

Definition 1.5.3. (Wasserstein Space) For a Polish space X, we call $\mathcal{P}_p(X)$ a Wasserstein space if it is endowed with the p-Wasserstein metric. Note that is also common to see this space symbolized by $\mathcal{W}_p(X)$.

Proposition 1.5.2. *For a bounded Polish space X, $p \in [1,+\infty)$, $\mu,\nu \in \mathcal{P}_p(X)$ and $C \in \mathbb{R}_+$, then*

$$W_1(\mu,\nu) \leq W_p(\mu,\nu) \leq C W_1(\mu,\nu)^{1/p} \tag{1.26}$$

Proof. Let $p \leq q \in [1, +\infty)$ and $\gamma \in \Pi(\mu, \nu)$. Hence, $\phi(x) = x^{q/p}$ is a convex function, so by Jensen's inequality:

$$\phi\left(\int d(x,y)^p d\gamma\right)^{1/q} = \left(\int d(x,y)^p d\gamma\right)^{1/p} \leq \left(\int \phi(d(x,y)^p) d\gamma\right)^{1/q}$$
$$= \left(\int (d(x,y)^q) d\gamma\right)^{1/q}$$

This implies that $W_p(\mu,\nu) \leq W_q(\mu,\nu)$, when $p \leq q$. In particular, $W_1(\mu,\nu) \leq W_p(\mu,\nu)$ for $p \geq 1$.

Now, since X is bounded, then $d(x,y) \leq \sup_{x,y\in X} d(x,y) = d(X)$. Hence,

$$d(x,y)^p \leq d(X)^{p-1} d(x,y)$$
$$\therefore$$
$$\left(\int d(x,y)^p d\gamma\right)^{1/p} \leq \left(\int d(x,y) d\gamma\right)^{1/p} d(X)^{\frac{p-1}{p}}$$

Therefore, we conclude that $W_p(\mu,\nu) \leq d(X)^{\frac{p-1}{p}} W_1(\mu,\nu)^{1/p}$

□

Next, let's present some of the topological properties of such space. A first thing to note is that for probability spaces, the notion of weak convergence can be made more strict with the following lemma:

Lemma 1.5.2. *For a space of probability measures, we say that μ_n converges weakly to μ, i.e. $\mu_n \rightharpoonup \mu \iff \forall f \in C_c(X), \int f\ d\mu_n \to \int f\ d\mu$, where $C_c(X)$ is the space of continuous functions with compact support. Note that $C_c(X) \subset C_0(X) \subset C_b(X)$.*

Proof.
$\implies$) If $\mu_n \rightharpoonup \mu$, then $f \in C_c(X) \subset C_b(X)$, hence $\int f d\mu_n \to \int f d\mu$.

$\impliedby$) Suppose that $\forall f \in C_c(X),\ \int f d\mu_n \to \int f d\mu$. Hence, note that for any constant M, $\int f + M d\mu_n = \int f d\mu_n + C \to \int f d\mu + C$. Take $g \in C_b(X)$ and make $g' = g + C \geq 0$ and $g' \mathbb{1}_{[-k,k]} = f_k \in C_c(X)$. Which implies that $f_k \uparrow g'$. Now,

$$\left|\int g d\mu_n - \int g d\mu\right| = \left|\int g' d\mu_n - \int g' d\mu\right|$$
$$\leq \left|\int g' d\mu_n - \int f_k d\mu_n\right| + \left|\int f_k d\mu_n - \int f_k d\mu\right| + \left|\int f_k d\mu - \int g' d\mu\right|$$

Since $f_k \in C_c(X)$, then for n big enough, $\left|\int f_k d\mu - \int f_k d\mu_n\right| < \epsilon$. Therefore,

$$\left|\int g d\mu_n - \int g d\mu\right| \leq \left|\int g' d\mu_n - \int f_k d\mu_n\right| + \epsilon + \left|\int f_k d\mu - \int g' d\mu\right|$$

Since $f_k \uparrow g'$, then, by the Monotone Convergence Theorem,

$$\lim_{k\to+\infty} \left|\int g' d\mu_n - \int f_k d\mu_n\right| = 0$$
$$\lim_{k\to+\infty} \left|\int f_k d\mu - \int g' d\mu\right| = 0$$
$$\therefore$$

$$\lim_{k\to+\infty} \left|\int g d\mu_n - \int g d\mu\right| = \left|\int g d\mu_n - \int g d\mu\right| \leq \epsilon$$

$\square$

Theorem 1.5.1. *Let (X,d) be a Polish compact space with $\mu_n, \mu \in P_p(X)$ and $p \in [1,+\infty)$, then $W_p(\mu_n,\mu) \to 0 \iff \mu_n \rightharpoonup \mu$.*

Proof.

$\Longrightarrow$) Let $W_p(\mu_n,\mu) \to 0$. Since X is compact and c is a continuous function, by Theorem 1.3.1 the Kantorovich Problem has a solution. Also, by Theorem 1.4.4, we obtain that max(DP) = min(KP). First, we prove for $p = 1$. In this case, using the Lipschitz version of DP:

$$W_1(\mu,\nu) = \max\left\{\int_X \phi \, d\mu - \int_X \phi \, d\nu \ : \phi \in Lip_1(X)\right\} \to 0$$

This implies that for any $f \in \mathrm{Lip}_1$, $\int f d\mu_n \to \int f d\mu$. Note that, by linearity, the same is true for any Lipschitz function. Since X is compact, then Lipschitz functions are dense on $C(X)$ (see Theorem 4.1.7), which leads us to conclude that $\mu_n \rightharpoonup \mu$ (by Portmanteau 4.1.1). Now, by Proposition 1.5.2, the same is valid for any $p \geq 1$.

$\Longleftarrow$) Let $\mu_n \rightharpoonup \mu$. Define a subsequence μ_{n_k} such that $\lim_k W_1(\mu_{n_k},\mu) = \limsup_n W_1(\mu_n,\mu)$. By the same arguments already used, we know that for each μ_{n_k} there is a $\phi_{n_k} \in \mathrm{Lip}_1$ such that $W_1(\mu_{n_k},\mu) = \int_X \phi_{n_k} d(\mu_{n_k} - \mu)$.

For an arbitrary $\epsilon > 0$, make $\delta = \epsilon$. Since ϕ_{n_k} is 1-Lipschitz, if $d(x, y) < \delta$, then $|\phi_{n_k}(x) - \phi_{n_k}(y)| \leq d(x, y) < \epsilon$, $\forall k \in \mathbb{N}$. Therefore, the sequence is Equicontinuous.

Also, for $x_0 \in X$, we can make $\phi'_{n_k}(x) := \phi_{n_k}(x) - \phi_{n_k}(x_0)$. Note that these functions are 1-Lipschitz and still satisfy $W_1(\mu_{n_k}, \mu) = \int_X \phi'_{n_k} d(\mu_{n_k} - \mu)$. Hence, let's use ϕ'_{n_k} as our new subsequence. In this case,

$$|\phi'_{n_k}(x)| = |\phi_{n_k}(x) - \phi_{n_k}(x_o)| \leq d(x, x_o) \leq d(X) < +\infty$$

This implies that this sequence of ϕ'_{n_k} is Equibounded. With this, we can use Arzelà-Ascoli Theorem (4.1.6) to obtain a sub-subsequence that converges uniformly to a $\phi \in \text{Lip}_1(X)$. Replace and relabel the original subsequence, obtaining:

$$\begin{aligned}
&W_1(\mu_{n_k}, \mu) = \int_X \phi_{n_k} d(\mu_{n_k} - \mu) \\
&= \left| \int_X \phi_{n_k} d\mu_{n_k} + \int_X \phi d\mu_{n_k} - \int_X \phi d\mu_{n_k} + \int_X \phi d\mu - \int_X \phi d\mu - \int_X \phi_{n_k} d\mu \right| \\
&\leq \underbrace{\left| \int_X \phi_{n_k} d\mu_{n_k} - \int_X \phi d\mu_{n_k} \right|}_{\text{Goes to 0,due to } \phi_{n_k} \to_u \phi} + \underbrace{\left| \int_X \phi d\mu - \int_X \phi_{n_k} d\mu \right|}_{\text{Goes to 0,due to } \phi_{n_k} \to_u \phi} + \underbrace{\left| \int_X \phi d\mu_{n_k} - \int_X \phi d\mu \right|}_{\text{Goes to 0,due to } \mu_{n_k} \rightharpoonup \mu}
\end{aligned}$$

Therefore $\limsup_n W_1(\mu_n, \mu) \leq 0 \implies W_1(\mu_n \mu) \to 0$. To conclude the proof for any $p \in [1, +\infty)$, we use Proposition 1.5.2:

$$0 \leq W_p(\mu_n, \mu) \leq C W_1(\mu_n, \mu)^{1/p} \leq 0$$

$\square$

Theorem 1.5.2. *For $X \subset \mathbb{R}^n$, $\mu_n, \mu \in \mathcal{P}_p(X)$, $x_0 \in X$ and d is metric on X, then*

$$W_p(\mu_n, \mu) \to 0 \iff \int_X d(x, x_0)^p d\mu_n \to \int_X d(x, x_0)^p d\mu \text{ and } \mu_n \rightharpoonup \mu \quad (1.27)$$

Proof.

$\implies$) Let $W_p(\mu_n, \mu) \to 0$. Since X is Polish, and c is a continuous function, by Theorem 1.3.3 the Kantorovich Problem has a solution. Also, by Theorem 1.4.7, we obtain that sup(DP) = min(KP). We know that $W_p(\mu_n, \mu) \geq W_1(\mu_n, \nu) \geq 0$, hence, using the Lipschitz version of the Dual Problem for W_1:

$$\sup \left\{ \int_X \phi \, d\mu_n - \int_X \phi \, d\mu \ : \phi \in Lip_1(X) \right\} \to 0$$

This implies that for any $f \in \mathrm{Lip}_1, \int f d\mu_n \to \int f d\mu$. Note that, by linearity, the same is true for any Lipschitz function, not only Lip_1. Finally, since Lipschitz functions are dense on $C_c(X)$ (see Theorem 4.1.7), we can use Lemma 1.5.2 to conclude that $\mu_n \rightharpoonup \mu$.

To prove the other condition (i.e. $\int_X d(x, x_0)^p d\mu_n \to \int_X d(x, x_0)^p d\mu$), define δ_{x_0} as a measure with mass on x_0. Which means that the optimal transport plan γ_n is in $\Pi(\mu_n, \delta_{x_0})$. This implies that $\gamma_n(x, y) = 0$ for any $y \neq x_0$. Therefore

$$\begin{aligned} W_p(\mu_n, \delta_{x_0})^p = \int_{X \times X} d(x, y)^p d\gamma_n &= \int_{X \times \{x_0\}} d(x, y)^p d\gamma_n \\ &= \int_X d(x, x_0)^p d\mu_n \to W_p(\mu, \delta_{x_0})^p = \int_X d(x, x_0)^p d\mu \end{aligned}$$

Where we used the fact that $W(\mu_n, \delta_{x_0}) \to W(\mu, \delta_{x_0})$, which is true since $W(\mu_n, \delta_{x_0}) - W(\mu, \delta_{x_0}) \leq W(\mu_n, \mu)$.

$\Longleftarrow$) Consider now that $\mu_n \rightharpoonup \mu$ and Define $\pi_R : X \to \overline{\mathrm{B}(R)}$, which is the projection on the closed ball with radius R centered at x_0. Since $W_p(\cdot, \cdot)$ is a metric, we have:

$$W_p(\mu_n, \mu) \leq W_p(\mu_n, (\pi_R)_\# \mu_n) + W_p((\pi_R)_\# \mu_n, (\pi_R)_\# \mu) + W_p((\pi_R)_\# \mu_n, \mu)$$

For sake of clarity in the proof, let's define, without loss of generalization, that $d(x, x_0) = |x|$ and $d(x, y) = |x - y|$. Now, note that $|x - \pi_R(x)| = |x| - |x| \wedge R$ and that the plan $(id, \pi_R)_\# \mu$ is a possible solution to the OT Problem of transporting μ to $(\pi_R)_\# \mu$. Therefore:

$$\begin{aligned} W_p(\mu, (\pi_R)_\# \mu)^p \leq \int_{X \times X} |x - y|^p d(id, \pi_R)_\# \mu &= \int_{(id, \pi_R)^{-1}(X \times X)} |x - \pi_R(x)|^p d\mu \\ = \int_X |x - (x \wedge R)|^p d\mu &= \int_{B(R)^c} (|x| - R)^p d\mu \end{aligned}$$

And using the same arguments:

$$W_p(\mu_n, (\pi_R)_\#\mu_n)^p \leq \int_{B(R)^c} (|x| - R)^p d\mu_n$$

Now, note that

$$\int_X |x|^p - (|x| \wedge R)^p d\mu = \int_{B(R)} |x|^p - |x|^p d\mu + \int_{B(R)^c} |x|^p - R^p d\mu \leq \int_{B(R)^c} |x|^p d\mu$$

Since $\mu_n, \mu \in \mathcal{P}_p(X)$, we know that $\int_X |x|^p d\mu = C < +\infty$ and $\int_X |x|^p d\mu_n = C < +\infty$ then

$$\int_{B(R)^c} |x|^p d\mu = C - \int_{B(R)} |x|^p d\mu \quad \therefore \quad \lim_{R \to 0} \int_{B(R)^c} |x|^p = 0$$

Using that $(|x| - R)^p \leq |x|^p - (|x| \wedge R)^p$ for every $x \in B(R)^c$, we get

$$W_p(\mu_n, (\pi_R)_\#\mu)^p \leq \int_{B(R)^c} (|x| - R)^p d\mu_n \leq \int_{B(R)^c} |x|^p - R^p d\mu_n \leq \int_{B(R)^c} |x|^p$$

Now, note that since $\int |x|^p \mu_n \to \int |x|^p d\mu$ and that $(|x| \wedge R)$ is continuous and bounded,

$$\begin{aligned} \lim_n W_p(\mu_n, (\pi_R)_\#\mu_n) &\leq \lim_n \int_{B(R)^c} (|x| - R)^p d\mu_n \\ &\leq \lim_n \int_{B(R)^c} |x|^p - R^p d\mu_n = \int_{B(R)^c} |x|^p - R^p d\mu \leq \int_{B(R)^c} |x|^p d\mu \end{aligned}$$

Hence,

$$\begin{aligned} \lim_R \lim_n (W_p(\mu_n, (\pi_R)_\#\mu_n) &\leq \lim_R \int_{B(R)^c} |x|^p d\mu = 0 \\ \lim_R (W_p(\mu, (\pi_R)_\#\mu) &\leq \lim_R \int_{B(R)^c} |x|^p d\mu = 0 \end{aligned}$$

Lastly, note that since $\overline{B(R)}$ is compact, then we can use Theorem 1.5.1 to stablish that

$$\lim_n W_p((\pi_R)_\#\mu_n, (\pi_R)_\#\mu) = 0$$

We can then conclude that

$$\begin{aligned}\limsup_n W_p(\mu_n, \mu) &\leq \lim_R \limsup_n (W_p(\mu_n, (\pi_R)_\# \mu_n) \\ &\qquad + W_p((\pi_R)_\# \mu_n, (\pi_R)_\# \mu) \\ &\qquad + W_p((\pi_R)_\# \mu_n, \mu)) \\ &= 0\end{aligned}$$

□

The Theorem above was proved for $X \subset \mathbb{R}^d$, but a more general result can be proven for Polish spaces. Such result is presented below without a proof. The proof can be found in Villani [10] under Theorem 6.9.

Theorem 1.5.3. *For (X, d) a Polish metric space, $\mu_n, \mu \in \mathcal{P}_p(X)$ and $x_0 \in X$. Then*

$$W_p(\mu_n, \mu) \to 0 \iff \int_X d(x, x_0)^p d\mu_n \to \int_X d(x, x_0)^p d\mu \text{ and } \mu_n \rightharpoonup \mu \quad (1.28)$$

Let's just put some words on these last two theorems we introduced. We showed that the p-Wasserstein distance metrizes weak convergence of probability measures in the space $\mathcal{P}_p(X)$, with (X, d) a Polish space. Such property is very useful and is not present in many other commonly used distances such as Total Variation and the Kullback-Leibler Divergence.

Yet, there are many other ways to metrize weak convergence, such as Prokhorov's distance and bounded Lipschitz distance. So, besides this *metrization*, Villani [10] gives the following reasons that make W_p such an interesting metric:

(i) It's definition makes it a natural choice in OT problems;

(ii) The distance has a rich duality, especially for $p = 1$;

(iii) Since it's defined with an infimum, it is easy to bound from above;

(iv) Wasserstein distances incorporate information of the ground geometry.

For applications in Data Science, the equivalence with weak convergence and the incorporation of the ground geometry are probably the most attractive characteristics. Figure 1.5 highlights how W_p takes into account the underlying geometry compared to the Kullback-Leibler divergence, which does not.

Villani [10] also points out that:

> As a geenral rule, the W_1 distance is more flexible and easier to bound, while the W_2 distance better reflects geometric features (at least for problems with a Riemannian flavor), and is better adapted when there is more structure; it also scales better with the dimension. Results in W_2 distance are usually stronger, and more difficult to establish, than results in W_1 distance.

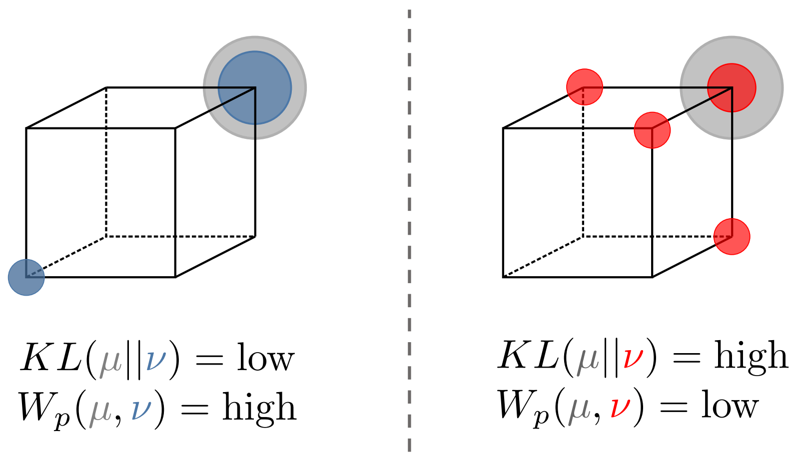


Figure 1.5: Comparison between Wasserstein distance and KL Divergence, based on Montavon et al. [5]. On the left, there is a large overlap between the two distributions, but a large geometrical distance for a portion. On the right, there is much less overlap, but the whole distribution is geometrically closer. These two cases clearly highlight how W_p incorporates geometrical information while KL doesn't.

Before finishing our initial exposition on the Wasserstein distance, let's prove some more relevant results.

Corollary 1.5.1. *(Continuity of W_p) Let (X, d) be a Polish metric space, and $p \in [1, +\infty)$, then W_p is continuous on $P_p(X)$, i.e. if $\mu_k \rightharpoonup \mu$, $\nu_k \rightharpoonup \nu$, and $\int d(x_0, x)^p d\mu_k(x) \to \int d(x_0, x)^p d\mu$, $\int d(x_0, x)^p d\nu_k(x) \to \int d(x_0, x)^p d\nu$, then*

$$W_p(\mu_k, \nu_k) \to W_p(\mu, \nu) \tag{1.29}$$

Proof. Just note that

$$W_p(\mu_k, \nu_k) \leq W_p(\mu_k, \mu) + W_p(\mu, \nu) + W_p(\nu_k, \nu)$$

Hence, taking the limit and using Theorem 1.5.3

$$\lim_{k \to +\infty} W_p(\mu_k, \nu_k) \leq W_p(\mu, \nu)$$

We can perform the same steps, but now for the reverse inequality

$$W_p(\mu, \nu) \leq W_p(\mu, \mu_k) + W_p(\mu_k, \nu_k) + W_p(\nu_k, \nu)$$

And again, taking the limit, we conclude that

$$W_p(\mu_k, \nu_k) \to W_p(\mu, \nu).$$

$\square$

Note that if we didn't have $\int d(x_0, x)^p d\mu_k(x) \to \int d(x_0, x)^p d\mu$ and $\int d(x_0, x)^p d\nu_k(x) \to \int d(x_0, x)^p d\nu$ in the hypothesis, we would instead conclude that

$$W_p(\mu, \nu) \leq \liminf W_p(\mu_k, \nu_k).$$

The last results that we want to prove is that for a Polish space X, then the p-Wasserstein space is also complete and separable. To prove this, we first need the following non-trivial lemma.

Lemma 1.5.3. *(Villani [10] 6.14 - Cauchy sequences in W_p are tight)*
Let X be a Polish space, $p \geq 1$ and $(\mu_n)_{n \in \mathbb{N}}$ a Cauchy sequence in $(P_p(X), W_p)$. Then (μ_n) is tight.

Finally, we can now prove the following theorem.

Theorem 1.5.4. *(Villani [10] 6.18)*
Let X be a ***complete*** *and* ***separable*** *metric space and $p \in [1, +\infty)$. Then $(P_p(X), W_p)$ is also* ***complete*** *and* ***separable****. Moreover, any probability measure on $P_p(X)$ can be approximated by a sequence of probability measures with finite support.*

1.6 Optimal Transport Problems with Exact Solution

In many cases, obtaining the exact solution to an OT problem might not be possible, thus requiring the use of methods to approximate the real solution. Yet, there are situations where it's possible to obtain the exact OT plan. This section explores some of these situations.

1.6.1 1D Optimal Transport

For $\mu, \nu \in P_p(\mathbb{R})$, the Wasserstein has a closed form solution, which relies on the pseudoinverse of the cumulative distribution function.

Definition 1.6.1. Let $\mu \in \mathcal{P}(\mathbb{R})$. The cumulative distribution function (CDF) is

$$F_\mu(x) := \mu((-\infty, x]) \tag{1.30}$$

Note that F_μ is a nondecreasing and right-continuous function.

Definition 1.6.2. Given a nondecreasing and right-continuous function $F : \mathbb{R} \to [0, 1]$, its pseudoinverse is

$$F^{-1}(x) := \inf\{y \in \mathbb{R} \ : \ F(y) \geq x\} \tag{1.31}$$

After introducing these definitions, we can present the formula for computing the Wasserstein distance (Remark 2.30 on Peyré et al. [6]):

$$W_p(\mu, \nu)^p = \int_0^1 |F_\mu^{-1}(x) - F_\nu^{-1}(x)|^p dx \tag{1.32}$$

Note that for $p = 1$ and μ, ν both atomless, then there exists an optimal Monge map $T = F_\nu^{-1} \circ F_\mu$.

For the discrete 1-D distributions, an even simpler algorithm can be devised. Let $\mu = \sum_i^n = 1 u_i \delta_{x_i}$ and $\nu = \sum_j^m = 1 v_i \delta_{v_j}$, where $x_1 \leq x_2 \leq ... \leq x_n$ and $y_1 \leq y_2 \leq ... \leq y_m$. Consider that each position x_i has mass u_i and each position y_j has capacity v_j. The optimal transport plan consists of moving particle x_i to the closest position y_j, until capacity v_j is filled.

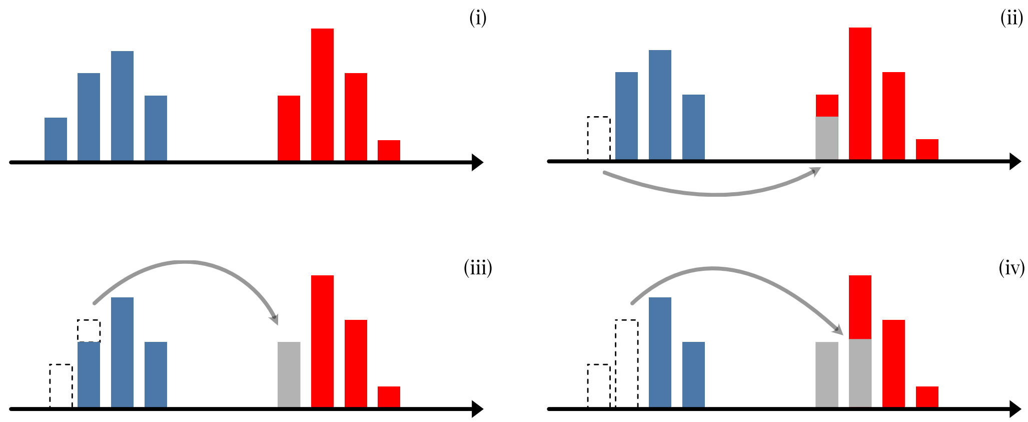


Figure 1.6: Illustration of the algorithm for optimally transporting distribution μ in blue to distribution ν in red.

1.6.2 Transport Between Discrete Measures

Let μ be a finite discrete probability measure, hence

$$\mu := \sum_{i=1}^{n} u_i \delta_{x_i} \tag{1.33}$$

Where $\mathbf{x} = (x_1, ..., x_n) \in \mathbb{R}^{n\times d}$ represent the location of each mass particle $i \in \{1, ..., n\}$. Vector $\mathbf{u} \in \mathbb{R}^{n\times 1}$, with components $u_i \in (0, 1]$, is the vector of weights, representing the amount of "mass" of each particle. Hence, discrete measures might be represented by a vector $\mathbf{x}$ of positions, and $\mathbf{u}$ of weights.

Now, suppose that both μ and ν are discrete measures. Let $\mathbf{u} \in \mathbb{R}^{n\times 1}$ and $\mathbf{v} \in \mathbb{R}^{m\times 1}$ represent the vector of weights, and $\mathbf{x} \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^{m\times d}$ represent the positions of each particle of μ and ν, respectively. In this scenario, the Optimal Transport Problem might be reformulated as the following. The cost function $c(x, y)$ can be substituted by a cost matrix $\mathbf{C} \in \mathbb{R}^{n\times m}$, where

$$\mathbf{C}_{i,j} := c(x_i, y_j), \quad i \in \{1, ..., n\},\ j \in \{1, ..., m\} \tag{1.34}$$

Any transport plan γ can be written as a matrix $\mathbf{P} \in \mathbb{R}_+^{n\times m}$, such that $\mathbf{P}_{i,j}$ represents the amount of mass flowing from particle i to particle j. Since $\gamma \in \Pi(\mu, \nu)$, the set of possible transport plans can be written as:

$$\mathbf{U}(\mathbf{u}, \mathbf{v}) := \left\{ \mathbf{P} \in \mathbb{R}_+^{n\times m} \ :\ \mathbf{P}\mathbf{1}_m = \mathbf{u},\ \mathbf{P}^{\mathrm{T}}\mathbf{1}_n = \mathbf{v} \right\} \tag{1.35}$$

Where $\mathbf{1}_n$ is a vector with n components, each equal to 1. In words, the sum of each row of $\mathbf{P}$ must be equal to $\mathbf{u}$ and the sum of each column must be equal to $\mathbf{v}$.

The Kantorovich Problem can be written as:

$$\text{(KP-Disc.)} \qquad \mathrm{L}_{\mathbf{C}}(\mathbf{u},\mathbf{v}) := \min_{\mathbf{P}\in\mathbf{U}(\mathbf{u},\mathbf{v})} \langle \mathbf{C},\mathbf{P}\rangle = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{u},\mathbf{v})} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{C}_{i,j}\mathbf{P}_{i,j} \tag{1.36}$$

The Dual Problem becomes:

$$\text{(DP-Disc.)} \qquad \mathrm{L}_{\mathbf{C}}(\mathbf{u},\mathbf{v}) := \max_{(\mathbf{f},\mathbf{g})\in\mathbf{R}(\mathbf{C})} \langle \mathbf{f},\mathbf{u}\rangle + \langle \mathbf{g},\mathbf{v}\rangle \tag{1.37}$$

Where

$$\mathbf{R}(\mathbf{C}) := \{(\mathbf{f},\mathbf{g}) \in \mathbb{R}^n \times \mathbf{R}^m \ : \ \forall (i,j) \in \{1,...,n\}\times\{1,...,m\},\ \mathbf{f}\oplus\mathbf{g} \leq \mathbf{C}\} \tag{1.38}$$

The Discrete Optimal Transport Problem is actually a Linear Programming (LP) problem. Hence, one can rearrange Equation (1.36) to the standard form of LP.

Definition 1.6.3. (Optimal Transport as standard LP)

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^{\mathrm{T}}\mathbf{p} \\ \text{subject to} \quad & \mathbf{A}\mathbf{p} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \\ & \mathbf{p} \geq 0 \end{aligned}$$

Where

$$\mathbf{p} := \begin{bmatrix} \mathbf{P}_{1,1} \\ \mathbf{P}_{2,1} \\ \vdots \\ \mathbf{P}_{n,1} \\ \mathbf{P}_{2,1} \\ \vdots \\ \mathbf{P}_{n,m} \end{bmatrix}, \quad \mathbf{c} := \begin{bmatrix} \mathbf{C}_{1,1} \\ \mathbf{C}_{2,1} \\ \vdots \\ \mathbf{C}_{n,1} \\ \mathbf{C}_{2,1} \\ \vdots \\ \mathbf{C}_{n,m} \end{bmatrix}, \quad \mathbf{A} := \begin{bmatrix} \mathbf{1}_n^{\mathrm{T}} \otimes \mathbf{I}_m \\ \mathbf{I}_n \otimes \mathbf{1}_m^{\mathrm{T}} \end{bmatrix},$$

Note that $\mathbf{I}_n$ stands for the identity matrix, and $\otimes$ is the tensor product.

Definition 1.6.4. (Optimal Transport Dual Problem as LP)

$$\begin{aligned} \text{minimize} \quad & \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}^{\mathrm{T}} \mathbf{h} \\ \text{subject to} \quad & \mathbf{A}^{\mathrm{T}} \mathbf{h} \leq \mathbf{c} \end{aligned}$$

Where $\mathbf{h} = [f_1, \dots, f_n, g_1, \dots, g_m]^{\mathrm{T}}$, with $\mathbf{c}$ and $\mathbf{A}$ the same as in the primal LP.

Since the Optimal Transport Problem is actually a Linear Programming Problem, therefore, one can use known methods for solving such problems, such as Simplex or Interior Point Method. An explanation on how to solve such LP problems in OT can be found in Chapter 3 of Peyré et al. [6].

Chapter 2

Benamou-Brenier Formulation and Geodesics in the Wasserstein Space

2.1 Derivation of the PDE by Benamou-Brenier Formula

In these notes we aim at proving the so called Benamou-Brenier formulation of Optimal Transport. Formally this result in summarized in the following formula.

$$\frac{1}{2}W_2^2(\rho_0,\rho_1) = \inf_{(\rho,v)} \left\{ \int_0^1 \int_\Omega |v(t,x)|^2 \, \mathrm{d}\rho_t(x)\mathrm{d}t : \begin{array}{c} \partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0 \\ \rho(0,\cdot)=\rho_0, \ \rho(1,\cdot)=\rho_1 \end{array} \right\} \tag{BB}$$

The present notes will the presentation from [1],[7]. There exists a huge literature concerning the topic, however the author believes that either the references present the results in a too informal way, or with a excessively heavy theoretical machinery. Our goal will be to present the following Theorem 2.1.1 in a precise way, in the sequel we will discuss the key points in the statement of the Theorem so that we are able to clarify the connection between the formula (BB) and the geodesics in the Wasserstein spaces $(\mathcal{P}_p(\Omega), W_p)$.

Theorem 2.1.1. *(Benamou-Brenier Formula) Let ρ_0, ρ_1 be probability measures in the space of probability measures with finite p-moments, $\mathcal{P}_p(\Omega)$, for*

$p > 1$. *Then the following characterization of the p-Wasserstein distance holds*

$$\frac{1}{p}W_p^p(\rho_0,\rho_1) = \inf_{(\rho,v)} \left\{ \int_0^1 \int_\Omega |v(t,x)|^p \,\mathrm{d}\rho_t(x)\mathrm{d}t : \begin{array}{c} (\rho_t)_{t\in[0,1]} \textit{ is A.C.}, \\ (\rho_t,v_t)_{t\in[0,1]} \textit{ solve } (\text{CE}), \\ \rho(0,\cdot)=\rho_0,\ \rho(1,\cdot)=\rho_1 \end{array} \right\},$$

where the inf is taken over family of probability measures $(\rho_t)_{t\in[0,1]}$ *which are* absolutely continuous *(A.C.) in the Wasserstein space* $(\mathcal{P}_p(\Omega), W_p)$. *Furthermore, the tuple of measures and velocities* $(\rho_t, v_t)_{t\in[0,1]}$ *are such that* $v_t \in L^p(\rho_t;\mathbb{R}^d)$ *and solve the* Continuity Equation (CE)

$$\partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0 \tag{CE}$$

in the weak sense.

Before the proof of Theorem (2.1.1) of this result, we need to understand the key concepts and the properties of the elements in its statement. Our objectives before setting out for the proof are threefold:

1. Understand what are AC curves in the general context of metric spaces, as this corresponds to the space where we are minimizing;

2. What means for $(\rho_t, v_t)_{t\in[0,1]}$ to solve (CE) and what properties can we expect from the solutions;

3. Which are the geodesics in $(\mathcal{P}_p(\Omega), W_p)$?

4. Draw a relation between the minimizer of the optimization problem given by the Benamou-Brenier formulation and the geodesics in the Wasserstein sapces.

2.1.1 A.C. curves in metric spaces (X,d)

In $\mathbb{R}^d$, we say a function $f : [0,T] \to \mathbb{R}^d$ is absolutely continuous if it is a.e. differentiable and its derivative belongs in $L^1\left([0,T],\mathbb{R}^d\right)$. If we are working in a general metric space (X,d), however, we cannot define A.C. curves this way, since to define an equivalent notion of derivative requires the structure of a vector space.[1]

[1]We will see however that we can still define the "norm" of the derivative.

However, for functions on euclidean spaces, we can always write that

$$|f(t) - f(s)| \leq \int_s^t |\dot{f}(r)| \mathrm{d}r,$$

as long as f has an derivative a.e. defined. This motivates the following definition.

Definition 2.1.1. Let (X, d) be a metric space and a curve over X, $\omega : [0, T] \to X$. We say that ω is an A.C. curve if there exists some $g \in L^1([0, T])$ such that, for $t > s$

$$d\left(\omega(t), \omega(s)\right) \leq \int_s^t g(r) \mathrm{d}r.$$

Let $\mathrm{AC}([0, T]; X)$, or merely $\mathrm{AC}(X)$, AC when there is no ambiguity concerning the metric space and/or time intervals, denote the space of all absolutely continuous curves from $[0, T]$ assuming values in X.

Remark 1. If g is a constant, then the curve ω is Lipschitz continuous.

As we have mentioned before, we can not define an derivative of ω as

$$\dot{\omega}(t) = \lim_{h \to 0} \frac{\omega(t + h) - \omega(t)}{h} \tag{2.1}$$

since the quantity $\omega(t + h) - \omega(t)$ does not make sense in a metric space. Instead, we can define the *metric derivative*

$$|\dot{\omega}|(t) := \lim_{h \to 0} \frac{d(\omega(t + h), \omega(t))}{h}. \tag{2.2}$$

Indeed, the class of A.C. curves always admits a metric derivative as we shall prove in the following theorem.

Theorem 2.1.2. *Let (X, d) be a separable, bounded, metric space. Then for all A.C. curves $\omega : [0, T] \to X$*

$$|\dot{\omega}|(t) := \lim_{h \to 0} \frac{d(\omega(t + h), \omega(t))}{h}.$$

exists $\mathcal{L}^1([0, T])$-a.e., the function $t \mapsto |\dot{\omega}|(t)$ is integrable and $|\dot{\omega}|(\cdot)$ is the minimal integral modulus of continuity of ω, which implies that

$$d(\omega(t), \omega(s)) = \int_s^t |\dot{\omega}|(r) \mathrm{d}r. \tag{2.3}$$

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a dense sequence in (X,d) and define $d_n(t) := d(x_n, \omega(t))$, for each $n \in \mathbb{N}$. Then, it is clear that $d_n : [0,T] \to \mathbb{R}_+$ is an AC curve since

$$\begin{aligned} |d_n(t) - d_n(s)| &= |d(x_n, \omega(t)) - d(x_n, \omega(s))| \\ &\le d(\omega(t), \omega(s))), \end{aligned}$$

so that it has the same modulus of continuity of ω. In particular, $d_n(\cdot)$ is a.e. differential in $[0,T]$. Setting

$$|\dot\omega(t)| := \sup_{n\in\mathbb{N}} d_n'(t),$$

we will prove that this supremum coincides with the metric derivative of ω. In other words, it a.e. coincides with the limit

$$\lim_{s\to t} \frac{d(\omega(s), \omega(t))}{|s-t|}.$$

It suffices to check that

$$\limsup_{s\to t} \frac{d(\omega(s), \omega(t))}{|s-t|} \le |\dot\omega(t)| \le \liminf_{s\to t} \frac{d(\omega(s), \omega(t))}{|s-t|}, \text{ for a.e. } t \in [0,T].$$

The first inequality is proven as follows: from the density of $(x_n)_{n\in\mathbb{N}}$ we have

$$\begin{aligned} d(\omega(s), \omega(t)) &= \sup_{n\in\mathbb{N}} |d_n(s) - d_n(t)| \\ &= \sup_{n\in\mathbb{N}} \int_s^t d_n'(r)\mathrm{d}r \\ &\le \int_s^t \left(\sup_{n\in\mathbb{N}} d_n'(r)\right) \mathrm{d}r \\ &= \int_s^t |\dot\omega(r)|\mathrm{d}r. \end{aligned}$$

By Lebesgue's differentiation theorem we have that for all Lebesgue points of $|\dot\omega(t)|$

$$\limsup_{s\to t} \frac{d(\omega(s), \omega(t))}{|s-t|} \le \lim_{s\to t} \frac{1}{|s-t|} \int_s^t |\dot\omega(r)|\mathrm{d}r = |\dot\omega(t)|.$$

The second inequality comes from the same observation that $d(\omega(s),\omega(t))) = \sup_{n\in\mathbb{N}} |d_n(s) - d_n(t)|$, implying that we can always write $d(\omega(s),\omega(t))) \geq |d_n(s) - d_n(t)|$ and therefore, dividing both sides by $|s-t|$, taking the liminf we have

$$\liminf_{s\to t} \frac{d(\omega(s),\omega(t))}{|s-t|} \geq \sup_{n\in\mathbb{N}} \liminf_{s\to t} \frac{|d_n(s) - d_n(t)|}{|s-t|} = \sup_{n\in\mathbb{N}} = |\dot{\omega}(t)| \text{ for a.e. } t \in [0,T].$$

□

We can also prove that

$$\text{Length}(\omega) := \sup\left\{\sum_{k=1}^{n} d(\omega(t_{k-1}),\omega(t_k)) : 0 = t_0 < t_1 < \cdots < t_n = 1\right\} = \int_0^1 |\dot{\omega}|(t)\mathrm{d}t \tag{2.4}$$

With this first discussion, we have the necessary tools to define the notion of *geodesics* and *constant speed geodesics* in metric spaces.

Definition 2.1.2. Given a metric space (X, d) and two points $x_0, x_1 \in X$, we say an A.C. curve is a geodesic if it attends the following infimum

$$\inf\left\{\int_0^1 |\dot{\omega}|(t)\mathrm{d}t : \begin{array}{c} \omega : [0,1] \to X \text{ is A.C.} \\ \omega(0) = x_0,\ \omega(1) = x_1. \end{array}\right\} \tag{2.5}$$

If ω is such a curve, we say it is a geodesic joining x_0 and x_1.

We are particularly interested in constant speed geodesics, *i.e.* geodesics with constant metric derivative ($|\dot{\omega}| \equiv \text{const}$). In fact, such geodesics are the minimizers, provided that they exist, of the following minimization problem

$$\min\left\{\int_0^1 |\dot{\omega}|^p(t)\mathrm{d}t : \begin{array}{c} \omega : [0,1] \to X \text{ is A.C.} \\ \omega(0) = x_0,\ \omega(1) = x_1. \end{array}\right\} \tag{2.6}$$

for $p > 1$. Indeed, we can easily check that the minimizers for (2.6) are constant speed geodesics using Jensen's inequality

$$\left(\min_{\omega\in\text{AC}} \text{Length}(\omega)\right)^p \leq \left(\int_0^1 |\dot{\omega}|(t)\mathrm{d}t\right)^p \leq \int_0^1 |\dot{\omega}|^p(t)\mathrm{d}t$$

where the second inequality is actually an equality if and only if the integrand $|\dot{\omega}|^p(\cdot)$ is constant. Hence taking the infimum over all $\omega \in \text{AC}$ on the r.h.s.

we get an lower bound with $(\min_{\omega\in\mathrm{AC}} \mathrm{Length}(\omega))^p$, which is attained for the curve with constant metric derivative having this same value. In particular, since this curve has minimal length, it is a geodesic with constant speed.

Hence we have proven that

$$\tilde{\omega} \in \arg\min_{\omega\in\mathrm{AC}} \int_0^1 |\dot{\omega}|^p(t)\mathrm{d}t \tag{2.7}$$

if and only if $\tilde{\omega}$ is a geodesic with constant speed.

2.1.2 The Continuity Equation

In this section we want to further precise what it means for a curve of probability measures to solve the PDE given by (CE). In fact, we will see that there are two equivalent definitions for the relaxed notion of solution to this PDE, which we will call *weak solutions* and *solutions in the sense of distributions*.

Definition 2.1.3 (Weak solution for (CE))**.** We say that a the tuple $(\mu_t, v_t)_{t\in[0,T]}$ solve (CE) in the weak sense if for all $\varphi \in C_c^1(\Omega)$

$$t \mapsto \int_\Omega \varphi(x)\mathrm{d}\mu_t(x), \text{ is an A.C. curve in time} \tag{2.8}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \varphi \mathrm{d}\mu_t = \int_\Omega \nabla\varphi \cdot v_t \mathrm{d}\mu_t \text{ for a.e. } t \in [0,T]. \tag{2.9}$$

Definition 2.1.4 (Solution for (CE) in the sense of distributions)**.** We say that a the tuple $(\mu_t, v_t)_{t\in[0,T]}$ solve (CE) in the sense of distributions *with fixed boundary conditions* if for all $\psi \in C^1([0,T]\times\Omega)$ it holds that

$$\int_0^1 \left(\int_\Omega \partial_t\psi \mathrm{d}\mu_t + \int_\Omega \nabla\psi\cdot v_t\mathrm{d}\mu_t\right)\mathrm{d}t = \int_\Omega \psi(1,\cdot)\mathrm{d}\mu_1 - \int_\Omega \psi(0,\cdot)\mathrm{d}\mu_0. \tag{2.10}$$

We say that a the tuple $(\mu_t, v_t)_{t\in[0,T]}$ solve (CE) in the sense of distributions *with free boundary conditions* if for all $\psi \in C_c^1((0,T)\times\Omega)$ it holds that

$$\int_0^1 \left(\int_\Omega \partial_t\psi \mathrm{d}\mu_t + \int_\Omega \nabla\psi\cdot v_t\mathrm{d}\mu_t\right)\mathrm{d}t = 0. \tag{2.11}$$

Let us discuss the physical meaning of (CE) and give an explicit solution for this PDE. Consider a velocity field $v = v(t,x)$ depending on time and space and let a particle move according to this field starting from the position x. Then, if $y_x(t)$ denotes the position of this particle at time t, it solves the following ODE

$$\begin{cases} \dot{y}_x(t) = v(t, y_x(t)), \\ y_x(0) = x. \end{cases} \tag{ODE}$$

Under suitable conditions, *e.g.* v being Lipschitz, (ODE) has a unique solution for all $t \in [0,T]$ and we can define the mapping

$$\begin{aligned} Y_t : \Omega &\to \Omega \\ x &\mapsto y_x(t). \end{aligned} \tag{2.12}$$

Define the family of measures $\mu_t := (Y_t)_{\#}\mu_0$.

Proposition 2.1.1. *The tuple $(\mu_t, v_t)_{t\in[0,T]}$, with $\mu_t : (Y_t)_{\#}\mu_0$, is a weak solution for the continuity equation (CE).*

Proof. It follows with a simple computation:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \varphi \mathrm{d}\mu_t = \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \varphi \mathrm{d}(Y_t)_{\#}\mu_0 &= \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \varphi\left(Y_t(x)\right)\mathrm{d}\mu_0 \\ &= \int_\Omega \nabla\varphi\left(Y_t(x)\right)\cdot \dot{Y}_t(x)\mathrm{d}\mu_0 = \int_\Omega \nabla\varphi\left(Y_t(x)\right)\cdot v_t(Y_t(x))\mathrm{d}\mu_0 \\ &= \int_\Omega \nabla\varphi\left(x\right)\cdot v_t(x)\mathrm{d}(Y_t)_{\#}\mu_0 \\ &= \int_\Omega \nabla\varphi\left(x\right)\cdot v_t(x)\mathrm{d}\mu_t \end{aligned}$$

□

2.1.3 Geodesics in $\mathcal{P}_p(\Omega)$ are solutions of the Continuity Equation

Now, let us answer the 3-rd question: ***Which are the geodesics in*** $(\mathcal{P}_p(\Omega), W_p)$***?*** More importantly, given two measures $\mu_0, \mu_1 \in \mathcal{P}_p(\Omega)$, can we always find a geodesic joining them? This is equivalent to asking if $(\mathcal{P}_p(\Omega), W_p)$ is a *geodesic space*.

If we can hope to construct an geodesic between any two measures in such space, we should try to use objects that are well defined for no matter the choice of such measures. Indeed, using Monge's optimal maps is not a good idea[2] since Monge's problem does not always admits a solution. On the other hand, the optimal transport plan from Kantorovitch's problem always exists.

With this idea in mind, given $\mu_0, \mu_1 \in \mathcal{P}_p(\Omega)$ for $p > 1$, let $\gamma \in \Pi(\mu_0, \mu_1)$ denote the optimal transport plan with the cost $c(x, y) = |x - y|^p$. Define the map π_t as

$$\begin{aligned} \pi_t : \Omega \times \Omega &\to \Omega \\ (x, y) &\mapsto (1-t)x + ty. \end{aligned}$$

Set $\mu_t := \pi_{t,\#}\gamma$.

Lemma 2.1.1. *The family $(\mu_t)_{t\in[0,1]}$ defined as $\mu_t := \pi_{t,\#}\gamma$, where γ is an optimal coupling between μ_0 and μ_1 is a constant speed geodesic joining μ_0 and μ_1.*

Proof. First let us show that

$$W_p(\mu_s, \mu_t) \le |t - s| W_p(\mu_0, \mu_1).$$

It is easy to check that the map

$$\gamma_{s,t} := (\pi_s, \pi_t)_\#\gamma \in \Pi(\mu_s, \mu_t).$$

Indeed, if π_X denotes the projection onto the first variable, we have that

$$\pi_{X,\#}\left((\pi_s, \pi_t)_\#\gamma\right) = \pi_X \circ (\pi_s, \pi_t)_\#\gamma = \pi_{s,\#}\gamma =: \mu_s.$$

So, since γ solves the Kantorovitch problem defining $W_p(\mu_0, \mu_1)$, we can always write

$$\begin{aligned} W_p^p(\mu_s, \mu_t) &\le \int_{\Omega\times\Omega} |x-y|^p \mathrm{d}\gamma_{s,t} = \int_{\Omega\times\Omega} |x-y|^p \mathrm{d}(\pi_s, \pi_t)_\#\gamma(x, y) \\ &= \int_{\Omega\times\Omega} |\pi_s(x, y) - \pi_t(x, y)|^p \mathrm{d}\gamma(x, y) \\ &= |t-s|^p \int_{\Omega\times\Omega} |x-y|^p \mathrm{d}\gamma(x, y) \\ &= |t-s|^p W_p^p(\mu_0, \mu_1). \end{aligned}$$

[2] Although we will see shortly that this particular case will have interesting properties liked to the continuity equation.

We can actually do better; it holds that $W_p(\mu_s, \mu_t) = |t-s| W_p(\mu_0, \mu_1)$. Using the triangle inequality we have that

$$\begin{aligned} W_p(\mu_0, \mu_1) &\leq W_p(\mu_0, \mu_s) + W_p(\mu_s, \mu_t) + W_p(\mu_t, \mu_1) \\ &\leq (s-0) W_p(\mu_0, \mu_1) + W_p(\mu_s, \mu_t) + (1-t) W_p(\mu_0, \mu_1). \end{aligned}$$

Which gives the converse inequality, *e.g.* $W_p(\mu_0, \mu_1) \leq |t-s| W_p(\mu_s, \mu_t)$.

In addition, if we compute the metric derivative of $(\mu_t)_{t\in[0,1]}$, we obtain

$$|\dot{\mu}|(t) = \lim_{h\to 0^+} \frac{W_p(\mu_{t+h}, \mu_t)}{h} = W_p(\mu_0, \mu_1),$$

for all $t \in [0,1]$. □

This shows that the Wasserstein space is a geodesic space, *i.e.* given two measures, we can always find a geodesic connecting them. However, we are particularly interested in the case where an optimal transport map between μ_0 and μ_1 exists, for instance if we take μ_0 to be absolutely continuous w.r.t. the Lebesgue measure.

If T denotes the optimal map for the transport problem between μ_0 and μ_1, then defining $T_t := (1-t)\mathrm{id} + tT$, and the measures $(\mu_t)_{t\in[0,1]}$, then analogous computations show that this family constitute a constant speed geodesic connecting μ_0, μ_1.

We claim that μ_t solves (CE) with an appropriated velocity field, let us try to compute it heuristically. From our previous computations, we can expect that the geodesic will have a constant velocity along the curve, so let's guess the speed will satisfy

$$v_t(T_t(x)) = T(x) - x, \tag{2.13}$$

that is, the velocity at any point in the geodesic at some time t will correspond to the difference of initial and final positions. Some direct properties are draw from the optimality of T, e.g.

$$W_p^p(\mu_0, \mu_1) = \int_\Omega |T(x) - x|^p \mathrm{d}\mu_0 = \int_\Omega |v_t(T_t(x))|^p \mathrm{d}\mu_0 = \int_\Omega |v_t(x)|^p \mathrm{d}\mu_t, \tag{2.14}$$

in particular, this curve satisfies

$$|\dot{\mu}|(t) = \|v_t\|^p_{L^p(\mu_t)} = W_p^p(\mu_0, \mu_1). \tag{2.15}$$

In addition, it is easy to check that (μ_t, v_t) solve (CE) using the characterization from 2.1.1 by just checking that the map $T_t(\cdot)$ corresponds to the flow (2.12) of the ODE with velocity field v_t defined above. Indeed, it reduces to checking

$$\frac{\mathrm{d}}{\mathrm{d}t}T_t(x) = T(x) - x =: v_t(T_t(x)). \tag{2.16}$$

Therefore, the velocity field we are looking for is $v_t(x) := (T - \mathrm{id}) \circ T_t^{-1}$, provided that T_t is invertible. It turns out that this is the case, regardless of the invertibility of the optimal map T, as long as the cost from the optimal transport problem that gives T is of the form $c(x, y) = h(x - y)$, with h strictly convex. This is the case for the p-Wasserstein distances with $p > 1$.

Lemma 2.1.2 ([7], Lemma 4.22)**.** *Let γ be an optimal transport plan between μ_0 and μ_1 for a transport cost $c(x, y) = h(x-y)$ where $h : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex function, and suppose that it is induced by a transport map T. Choose a representative of T such that $(x, T(x)) \in supp(\gamma)$ for all x. Then the map $x \mapsto (1-t)x + tT(x)$ is injective for $t \in (0, 1)$.*

Proof. Fill later. □

This way, the velocity field we wanted is well defined and we can summarize this discussion in the following Theorem.

Theorem 2.1.3. *Let $\mu_0, \mu_1 \in \mathcal{P}_p(\Omega)$ with $p > 1$ be such that there exists an optimal map T which realizes the optimal transport problem associated with $W_p^p(\mu_0, \mu_1)$. Setting $T_t := (1-t)id + tT$, then the family $(\mu_t, v_t)_{t\in[0,1]}$ given by*

$$\mu_t := T_{t,\#}\mu_0, \quad v_t := (T - id) \circ T_t^{-1} \tag{2.17}$$

solve (CE), correspond to a constant speed geodesic joining μ_0,μ_1 and satisfy the following

$$|\dot{\mu}|(t) = \|v_t\|^p_{L^p(\mu_t)} = W_p^p(\mu_0, \mu_1), \quad \textit{for all } t \in [0, 1]. \tag{2.18}$$

2.1.4 AC curves in $\mathcal{P}_p(\Omega)$ and the Benamou-Brenier formulation

Now we are almost in position to prove Theorem 2.1.1. In fact, given Theorem 2.1.3 it is no surprise that the Benamou-Brenier formula holds, using it our approach is to approximate any given measures with a absolutely continuous sequence that will admit an optimal transport map and use the velocities defined as above for these approximations.

There is still two major issues that need to be addressed:

1. The minimization is done over the space of AC curves, so in order for it to be equivalent to the minimization over solutions of the transport equation (CE), we need to establish a link between all the AC curves and (CE), namely prove the existence of a suitable velocity field for each AC curve in $\mathcal{P}_p(\Omega)$.

2. The functional
$$(\mu, v) \mapsto \frac{1}{p}\int_0^1 \int_\Omega |v(t,x)|^p \, \mathrm{d}\rho_t(x)\mathrm{d}t$$
is not jointly convex, hence we will have to introduce an change of variables $(\mu, v) \mapsto (\mu, E) := (\mu, \mu v)$ and study the following characterization
$$\frac{1}{p}\frac{|E|^p}{\mu^{p-1}} = \sup\left\{ a\mu + b\cdot E : \begin{array}{c} a\in\mathbb{R}, \\ b\in\mathbb{R}^d, \end{array} \quad a + \frac{1}{q}|b|^q \leq 0 \right\}.$$
This way, the new optimization problem will become
$$\inf_{(\rho,E)} \left\{ \mathcal{B}_p(\rho_t, E_t) : \begin{array}{c} (\rho_t)_{t\in[0,1]} \text{ is A.C.}, \\ (\rho_t, E_t)_{t\in[0,1]} \text{ solve (CE)}, \\ \rho(0,\cdot) = \rho_0, \; \rho(1,\cdot) = \rho_1 \end{array} \right\},$$
where $\mathcal{B}_p$, which stands for Benamou-Brenier, is a functional over $\mathcal{M}([0,1]\times\Omega)\times \mathcal{M}^N([0,1]\times\Omega)$ defined by
$$\mathcal{B}_p(\rho, E) := \sup\left\{ \int_\Omega a\mathrm{d}\rho + \int_\Omega b\cdot \mathrm{d}E : \begin{array}{l} (a,b)\in C([0,1]\times\Omega, \mathbb{R}\times\mathbb{R}^N), \\ a + \frac{1}{p}|b|^p \leq 0 \text{ pointwise.} \end{array} \right\}$$

We will start with the functional $\mathcal{B}_p$ since it will be used in the connection between the AC curves and (CE).

Lemma 2.1.3 ([7], Lemma 5.17)**.** *Let* $K_q := \left\{ (a,b)\in\mathbb{R}\times\mathbb{R}^N : a + \frac{1}{q}|b|^q \leq 0 \right\}$. *Then, for* $(t,x)\in\mathbb{R}\times\mathbb{R}^N$, *it holds that*
$$\sup_{a,b\in K_q} at + b\cdot x = \begin{cases} \frac{1}{p}\frac{|x|^p}{t^{p-1}}, & \text{for } t>0; \\ 0, & \text{for } t=0, x=0; \\ +\infty, & \text{if } t=0 \text{ and } x\neq 0, \text{ or if } t<0 \end{cases}$$

where $\dfrac{1}{p} + \dfrac{1}{q} = 1$.

Proof. **Case $t > 0$:** Let us fix some vector b and find the a which maximizes $at + b \cdot x$ given the constraint that $(a, b) \in K_q$. Since $t > 0$, the maximal a is given by $-\dfrac{1}{q}|b|^q$. Hence the supremum assumes the form

$$t\left(\sup_{b\in\mathbb{R}^N} b \cdot \frac{x}{t} - \frac{1}{q}|b|^q\right) = t\frac{1}{p}\left|\frac{|b|^p}{t^p}\right|,$$

where the second inequality is given by the explicit formula for the Legendre transform of $\dfrac{1}{p}|\cdot|$.

Case $t = 0, x = 0$: Trivially gives 0.

Case $t = 0$ and $x \neq 0$ or $t < 0$: If $t = 0$, then we can just take a sequence $b_n = nx$ and $a_n = -\dfrac{1}{q}|b|^q$, so that every the pair $(a_n, b_n) \in K_q$ and letting $n \to +\infty$, we get the desired result.

For $t < 0$, again we can just take a going to $+\infty$. □

Before establishing the desired relation between geodesics in the Wasserstein space and the solutions of the transport equation, we will state some more properties of the Benamou-Brenier funtional $\mathcal{B}_p$ in the following Proposition. Its proof is technical and can be skipped, however let's remark that these properties are one of the crucial ingredients of Theorem 2.1.4.

Proposition 2.1.2 (Properties of $\mathcal{B}_p$, [7])**.** *The functional $\mathcal{B}_p$ is convex and lower semi-continuous on the space $\mathcal{M}([0,1]\times\Omega)\times\mathcal{M}^N([0,1]\times\Omega)$ for the weak-* convergence. Moreover, the following properties hold:*

1. $\mathcal{B}_p(\rho, E) \geq 0$;
2. $\mathcal{B}_p(\rho, E) = \sup\left\{\int a\mathrm{d}\rho + \int b \cdot \mathrm{d}E : (a, b) \in L^\infty([0,1]\times\Omega; K_q)\right\}$;
3. *if both ρ and E are absolutely ´ continuous w.r.t. a same positive measure, we can write*

$$\mathcal{B}_p(\rho, E) = \int \frac{1}{p}\frac{|E(x)|^p}{\rho(x)^{p-1}}\mathrm{d}\lambda(x).$$

, where we identify ρ and E with their densities w.r.t.

4. *$\mathcal{B}_p(\rho, E) < +\infty$ only if $\rho \geq 0$ and $E \ll \rho$,* ´

5. *for $\rho \geq 0$ and $E \ll \rho$, there exists some velocity field v such that $E = v \cdot \rho$ and $\mathcal{B}_p(\rho, E) = \dfrac{1}{p}\displaystyle\int |v|^p \mathrm{d}\rho$;*

6. *If $\Omega = \mathbb{R}^N$, $\rho^\varepsilon = \rho \star \eta_\varepsilon$ and $E^\varepsilon = E \star \eta_\varepsilon$ (for standard even mollifying kernel η), then we have $\mathcal{B}_p(\rho^\varepsilon, E^\varepsilon) \leq \mathcal{B}_p(\rho, E)$.*

Theorem 2.1.4. *Let $(\mu_t)_{t\in[0,1]}$ be an AC curve in $(\mathcal{P}_p(\Omega), W_p)$ for $p > 1$ and $\Omega \subset \mathbb{R}^N$ compact. For a.e. $t \in [0,1]$ there is a vector field $v_t \in L^p(\mu_t; \mathbb{R}^N)$ such that*

1. *$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0$ in the weak sense;*

2. *For a.e. $t \in [0,1]$, it holds $\|v_t\|_{L^p(\mu_t)} \leq |\dot{\mu}|(t)$.*

Conversely, if $(\mu_t)_{t\in[0,1]}$ is a curve in $(\mathcal{P}_p(\Omega), W_p)$ such that

$$v_t \in L^p(\mu_t; \mathbb{R}^N), \text{ with } \int_0^1 \|v_t\|_{L^p(\mu_t)} \mathrm{d}t < +\infty \text{ and } \partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,$$

in the weak sense. Then $(\mu_t)_{t\in[0,1]}$ is an AC curve and $|\dot{\mu}|(t) \leq \|v_t\|_{L^p(\mu_t)}$.

Proof. **AC implies (μ_t, v_t) solves (CE):**

Let us begin with the with a AC curve. Up to a change of variables in time, we can assume it is a Lipschitz continuous curve. As we have discussed previously if $\mu_0, \mu_1 \ll \mathcal{L}^N$, then there is some optimal transport map with $\mu_1 = T_\# \mu_0$ and the interpolation $\mu_t := T_{t,\#}\mu_0$ with $T_t = ((1-t)\mathrm{id} + tT)$, such that (μ_t, v_t) solve (CE) with

$$v_t = (T - \mathrm{id}) \circ T_t^{-1}.$$

Therefore, the strategy for the proof is exploiting this connection between geodesics and the continuity equation and to ensure that we can always take the optimal transport map, we will use an approximation argument. So, take a mollifier η_k with compact support over $B(0, 1/k)$ and define

$$\eta^k_{i/k} := \eta_k \star \mu_{i/k}, \quad \text{for } i = 0, \ldots, j \text{ and } k \in \mathbb{N}.$$

Since the optimal transport map $T^{i,k}$ between $\mu^k_{i/k}$ and $\mu^k_{i+1/k}$ exists, we know from Theorem 2.1.3 that the constant speed geodesics are given by interpolations. The speed at a position $T^{i,k}(x)$ is

$$\frac{\Delta\text{displacement}}{\Delta t} = \frac{T^{i,k}(x) - x}{1/k} \Rightarrow v^{i,k} := k\left(T^{i,k} - \mathrm{id}\right)$$

and at some time t, since we have constant speed,

$$v^{i,k}_t = k\left(T^{i,k} - \mathrm{id}\right) \circ \left(T^{i,k}_t\right)^{-1}$$

where $T^{i,k}_t$ transports particles from time 0 to t and must be chosen such that:

1. it is invective so that $\left(T^{i,k}_t\right)^{-1}$ exists;
2. $\dfrac{\mathrm{d}}{\mathrm{d}t}T^{i,k}_t = v^{i,k}$, so that that $\mu^k_t := \left(T^{i,k}_t\right)_{\#}\mu^k_{i/k}$ solves (CE) with constant speed.

Clearly $T^{i,k}_t := (i+1-kt)\mathrm{id} + (kt-i)T^{i,k}$ satisfies such conditions. Therefore let us compute

$$\begin{aligned}
\left\|v^k_t\right\|^p_{L^p(\mu^k_t)} &= \int_\Omega |v^k_t|^p \mathrm{d}\mu^k_t = k^p \int_\Omega \left|(T^{i,k} - \mathrm{id}) \circ \left(T^{i,k}_t\right)^{-1}(x)\right|^p \mathrm{d}\mu^k_t(x) \\
&= k^p \int_\Omega \left|(T^{i,k} - \mathrm{id}) \circ \left(T^{i,k}_t\right)^{-1}(x)\right|^p \mathrm{d}\left(T^{i,k}_{t,\#}\mu^k_{i/k}\right)(x) \\
&= k^p \int_\Omega \left|T^{i,k} - \mathrm{id}\right|^p \mathrm{d}\mu^k_{i/k} \\
&= k^p W^p_p\left(\mu^k_{i/k}, \mu^k_{i+1/k}\right) = k^p W^p_p\left(\eta_k \star \mu^k_{i/k}, \eta_k \star \mu^k_{i+1/k}\right) \\
&\leq k^p W^p_p\left(\mu^k_{i/k}, \mu^k_{i+1/k}\right) \leq \int_{i/k}^{i+1/k} |\dot{\mu}|^p(t)\mathrm{d}t.
\end{aligned}$$

Where the last estimate comes from Jensen's inequality and from the definition of geodesics as the inf of AC curves over $[0,1]$ by taking $\omega(t) = \mu\left(\frac{t}{k} + \frac{i}{k}\right)$, which gives

$$k^p W^p_p\left(\mu^k_{i/k}, \mu^k_{i+1/k}\right) \leq k^p\left(\int_0^1 |\dot{\omega}|(t)\mathrm{d}t\right)^p \leq \int_{i/k}^{i+1/k} |\dot{\mu}|^p(t)\mathrm{d}t.$$

This gives an easy estimate for $\left\|v^k\right\|^p_{L^p(\Omega\times(a,b))}$, let i_a, i_b be such that $i_a \leq ka < i_a+1$ and the same for i_b. Then

$$\int_a^b \left\|v^k\right\|^p_{L^p\left(\mu_t^k\right)}\mathrm{d}t \leq \sum_{i=i_a}^{i_b}\int_{i/k}^{i+1/k}\left\|v^k\right\|^p_{L^p\left(\mu_t^k\right)}\mathrm{d}t \leq \int_a^b |\dot\mu|^p(t)\mathrm{d}t + \frac{2\mathrm{Lip}(\mu)}{k}.$$

Now define the momentum measures $E^k \in \mathcal{M}^N(\Omega\times[0,1])$ as

$$\int \phi\cdot\mathrm{d}E^k = \int_0^1\int_\Omega \phi(t,x)\cdot v_t^k(x)\mathrm{d}\mu_t^k(x)\mathrm{d}t$$

for all $\phi\in C(\Omega\times[0,1])$. Then we can estimate $\left\|E^k\right\|$ as follows

$$\left\|E^k\right\| = \int_0^1 \left\|v_t^k\right\|_{L^1(\mu_1^k)}\mathrm{d}t \leq \left(\int_0^1 \left\|v_t^k\right\|^p_{L^p(\mu_1^k)}\mathrm{d}t\right)^{1/p} \leq C.$$

Then from Banach-Alaoglu Theorem, we can assume, up to the extraction of a subsequence, that $E^k \overset{*}{\rightharpoonup} E$. In addition, we can also prove that

$$\mu^k \xrightarrow[k\to+\infty]{} \mu \text{ in } C([0,T];\mathcal{P}_p(\Omega))$$

with the following estimation

$$W_p(\mu_t^k,\mu_t) \leq W_p(\mu_t^k,\mu_{i/k}^k) + W_p(\mu_{i/k}^k,\mu_{i/k}) + W_p(\mu_{i/k}^k,\mu_t),$$

where $t\in[i/k, i+1/k]$. The first term can be estimated as

$$W_p(\mu_t^k,\mu_{i/k}^k) = \left|t-\frac{i}{k}\right|\left\|v^{i,k}\right\|_{L^p(\mu_{i/k}^k)} \leq \frac{C}{k}.$$

The same estimation follows for the second and third terms from properties of the convolution and from the Lipschitz continuity of the curve $(\mu_t)_{t\in[0,1]}$.

Therefore, as $\partial_t\mu^k + \nabla\cdot E^k = 0$, since (CE) is a linear PDE, standard weak convergence arguments give that $\partial_t\mu + \nabla\cdot E = 0$. So, from the weak-l.s.c. of the Benamou-Brenier functional, we have

$$\frac{1}{p}\int_0^1 \|v_t\|^p_{L^p(\mu_t)}\mathrm{d}t = \mathcal{B}_p(\mu,E) \leq \liminf_{k\to\infty}\mathcal{B}_p(\mu^k,E^k) = \frac{1}{p}\int_0^1 \left\|v_t^k\right\|^p_{L^p(\mu_t^k)}\mathrm{d}t < +\infty,$$

and from the properties of $\mathcal{B}_p$, we have that the limit measures are such that $E \ll \mu$ and therefore there exists some velocity field $v_t\in L^p(\mu_t)$ such that $E_t = v_t\mu$.

The only thing left to check is that $\|v_t\|_{L^p(\mu_t)} \leq |\dot{\mu}|(t)$. Indeed, using the l.s.c. of $\mathcal{B}_p$ from above, but now integrating over some interval (a,b), we have

$$\int_a^b \|v_t\|^p_{L^p(\mu_t)} \mathrm{d}t \leq \int_a^b |\dot{\mu}|(t) \mathrm{d}t.$$

Taking t_0 a Lebesgue point of both $t \mapsto \|v_t\|^p_{L^p(\mu_t)}$ and $t \mapsto |\dot{\mu}|(t)$, we obtain

$$\|v_t\|^p_{L^p(\mu_t)} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_{t_0-\varepsilon}^{t_0+\varepsilon} \|v_t\|^p_{L^p(\mu_t)} \mathrm{d}t \leq \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_{t_0-\varepsilon}^{t_0+\varepsilon} |\dot{\mu}|(t) \mathrm{d}t = |\dot{\mu}|(t).$$

AC implies (μ_t, v_t) solves (CE):

Given (μ, E) satisfying (CE) such that $E_t = v_t \mu_t$ for some velocity field $v_t \in L^p(\mu_t)$, by regularization techniques we can construct $(\mu_t^\varepsilon, v_t^\varepsilon)$ such that

$$\partial_t \mu_t^\varepsilon + \nabla \cdot (v_t^\varepsilon \mu_t^\varepsilon) = 0,$$

for each t, μ_t^ε is absolutely continuous w.r.t. $\mathcal{L}^N$ and this pair converges to (μ, v) as $\varepsilon \to 0$.

Then let T_s denote the optimal transport map from μ_0^ε to μ_s^ε, it is clear that

$$\gamma = (T_t, T_{t+h})_\# \mu_0^\varepsilon \in \Pi(\mu_t^\varepsilon, \mu_{t+h}^\varepsilon),$$

so that it holds

$$\begin{aligned}
W_p(\mu_t^\varepsilon, \mu_{t+h}^\varepsilon) \leq & \int_{\Omega \times \Omega} |x-y|^p \mathrm{d}\gamma = \int_\Omega |T_t(x) - T_{t+h}(x)|^p \, \mathrm{d}\mu_0^\varepsilon \\
\leq & |h|^{1-1/p} \left(\int_\Omega \int_t^{t+h} \left| \frac{\mathrm{d}}{\mathrm{d}s} T_s(x) \right|^p \mathrm{d}s \mathrm{d}\mu_0^\varepsilon \right)^{1/p} \\
= & |h|^{1-1/p} \left(\int_\Omega \int_t^{t+h} |v_s^\varepsilon(y_x(s))|^p \, \mathrm{d}s \mathrm{d}\mu_0^\varepsilon \right)^{1/p} \\
= & |h|^{1-1/p} \left(\int_\Omega \int_t^{t+h} |v_s^\varepsilon(x)|^p \, \mathrm{d}s \mathrm{d}\mu_s^\varepsilon \right)^{1/p}
\end{aligned}$$

where $y_x(t)$ is the solution of $\dot{y} = v_t^\varepsilon(y)$ at time t and initial condition x and we have used that since $\mu_t^\varepsilon = T_{t,\#}\mu_0^\varepsilon$ and $\mu_t^\varepsilon, \mu_t^\varepsilon$ solve (CE), then $T_t : x \mapsto y_x(t)$ from the uniqueness of the weak solution of the transport equation.

Since the regularized velocity field satisfies $\int_\Omega |v_t^\varepsilon(x)|^p \,\mathrm{d}\mu_t^\varepsilon \le \|v_t\|_{L^p(\mu_t)}$, we have proven that

$$\frac{W_p(\mu_t^\varepsilon, \mu_{t+h}^\varepsilon)}{|h|} \le \left(\frac{1}{|h|}\int_t^{t+h} \|v_t\|_{L^p(\mu_t)}\mathrm{d}t\right)^{1/p}.$$

Passing to the limit as $\varepsilon \to 0$, we recover $W_p(\mu_t, \mu_{t+h}$ and then as $h \to 0$, we obtain

$$|\dot{\mu}|\,(t) \le \|v_t\|_{L^p(\mu_t)}$$

for every t which is a Lebesgue point of $t \mapsto \|v_t\|_{L^p(\mu_t)}$. Therefore we conclude that the curve $(\mu_t)_{t\in[0,1]}$ is AC. $\square$

2.1.5 Proof of Benamou-Brenier Formula

With all the results from the previous section, we are in position to prove Theorem 2.1.1.

Proof of the Benamou-Brenier formula:

Starting with the characterization of constant speed geodesics in the Wasserstein space, we know that

$$\begin{aligned} W_p^p(\mu,\nu) &= \left(\min\left\{\int_0^1 |\dot{\rho}|(t)\mathrm{d}t : \begin{array}{c} \rho_t \text{ is AC,} \\ \rho_0=\mu,\ \rho_1=\nu. \end{array}\right\}\right)^p \\ &= \min\left\{\int_0^1 |\dot{\rho}|^p(t)\mathrm{d}t : \begin{array}{c} \rho_t \text{ is AC,} \\ \rho_0=\mu,\ \rho_1=\nu. \end{array}\right\}. \end{aligned}$$

However, from the relation between AC curves and the transport equation established in Theorem 2.1.4, this minimization can be rewritten as

$$W_p^p(\mu,\nu) = \min\left\{\mathcal{B}_p(\rho, E) : \begin{array}{c} \partial_t\rho_t + \nabla\cdot E_t = 0, \\ \rho_0=\mu,\ \rho_1=\nu. \end{array}\right\}.$$

In the minimization, we can take the (ρ, E) such that $\mathcal{B}_p(\rho, E) < +\infty$ without altering the minimum and for all such pairs, we know that $E \ll \rho$. In addition, for each ρ's, there exists some $v_t \in L^p(\rho_t, \mathbb{R}^N)$ such that $E_t = v_t\rho_t$ for a.e. $t \in [0,1]$. We conclude that

$$W_p^p(\mu,\nu) = \min\left\{\int_0^1 \|v_t\|^p_{L^p(\rho_t)}\mathrm{d}t : \begin{array}{c} \partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0, \\ \rho_0=\mu,\ \rho_1=\nu. \end{array}\right\}.$$

$\blacksquare$

Remark 2. Notice that, now that we know that any AC curve in the Wasserstein space must have a corresponding velocity field such that the pair solves (CE), the restriction of ρ_t being AC in Theorem (2.1.1) becomes redundant. It was kept so that the relation with (CE) and the set over which we minimize become evident.

2.1.6 Wasserstein Gradient Flows

In this section we will explore the classical theory of Wasserstein Gradient Flows and give the conceptual steps to prove the convergence of the JKO scheme (see the definition in (JKO) below) to the weak solution of a suitable evolution PDE.

Let $\mathcal{F}$ be a convex and l.s.c. functional over the space of probability measures $\mathcal{P}(\Omega)$. For simplicity, we assume Ω to be a compact subset of $\mathbf{R}^d$. We are interested in studying the following iterative scheme

$$\rho_{k+1}^\tau \in \underset{\rho\in\mathcal{P}(\Omega)}{\operatorname{argmin}}\, \mathcal{F}(\rho) + \frac{1}{2\tau} W_2^2(\rho, \rho_k^\tau), \quad \rho_0^\tau = \rho_0 \text{ given.} \tag{JKO}$$

Given this sequence, we want to check if a (suitable) interpolation of the obtained sequence $(\rho_k^\tau)_{k\in\mathbb{N}}$ converges to the appropriate notion of weak solution of the evolution PDE

$$\partial_t\rho - \operatorname{div}\left(\rho\nabla\frac{\delta\mathcal{F}}{\delta\rho}(\rho)\right) = 0, \quad \rho(0) = \rho_0. \tag{2.19}$$

In the seminal work [4] and following works ADD OTHER REFERENCES, one can divide the structure of the this convergence result in the following steps:

1 **Obtaining optimility conditions:** For each subproblem from (JKO), one characterizes the solution ρ_{k+1}^τ with an equation of a Euler-Lagrange equation of the form

$$\frac{\delta\mathcal{F}}{\delta\rho}(\rho_{k+1}^\tau) + \frac{\psi_{k+1}^\tau}{\tau} \equiv \text{const, on } \{\rho_{k+1}^\tau > 0\}.$$

2 **Interpolation in time:**

2.1 **Interpolations:** Definition of two different interpolations

$$\begin{aligned}(\rho^\tau, E^\tau) &: \text{ obtained by staircase interpolation} \\ \left(\tilde{\rho}^\tau, \tilde{E}^\tau\right) &: \text{ obtained by interpolation with geodesics}\end{aligned}$$

2.2 **Compactness:**

$$(\rho^\tau, E^\tau) \overset{*}{\rightharpoonup} (\rho, E)$$
$$\left(\tilde{\rho}^\tau, \tilde{E}^\tau\right) \overset{*}{\rightharpoonup} \left(\tilde{\rho}, \tilde{E}\right)$$

2.3 **Reconciliation:**

$$(\rho, E) = (\tilde{\rho}, \tilde{E}).$$

2.4 **Limit diffusion:**

$$\partial_t \tilde{\rho}^\tau + \nabla \cdot \tilde{E}^\tau = 0 \xrightarrow[\tau \to 0]{} \partial_t \rho + \nabla \cdot E = 0$$

3 **Characterization pf the limiting momentum E:** Use the optimality conditions to characterize the limit E as

$$E = -\rho \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho).$$

All the substeps from 2.1 to 2.4 are obtained analogously for no matter the choice of the functional $\mathcal{F}$, then steps 1 and 3, however, require a more *ad. hoc.* approach. For instance, obtaining the Euler-Lagrange equations can vary greatly; if we are dealing with the entropy functional

$$\mathcal{F}(\rho) = \int \rho \log \rho \mathrm{d}x,$$

then the solutions will always be absolutely continuous w.r.t. the Lebesgue measure, so that $\{\rho_k > 0\} = \Omega$ at every iteration. As we are interested in the case of the total variation functional TV, the level sets become a more critial issue.

Now our goal is to detail the interpolation step in a general framework, so that this modular analysis can be adapted to easier cases, *e.g.* the Fokker-Planck equation), or harder ones as in for the functional TV.

Chapter 3

Fluxo de Gradiente em Espaços de Wasserstein

Neste capítulo mostraremos como EDPs podem ser expressadas como fluxos de gradiente em um espaço de Wasserstein (i.e. espaço métrico de medidas de probabilidades com distância de Wasserstein)[1]. A exposição é focada em apresentar de maneira clara e sucinta o necessário para entendimento do assunto, sem provar os resultados mais refinados, o que tornaria o texto muito extenso e de difícil entendimento[2]. Assim, restringimos nossa exposição ao caso de $\mathbb{R}^n$. Note, porém, que muitas das definições utilizadas podem ser facilmente estendidas para espaços de Hilbert.

Seja uma função $F : \mathbb{R}^n \to \mathbb{R} \in C^1$, e $x_0 \in \mathbb{R}^n$, onde queremos descobrir $x(t)$ que resolve o seguinte sistema de equações:

$$\begin{cases} x'(t) = -\nabla F(x(t)),\ t > 0, \\ x(0) = x_0. \end{cases} \tag{3.1}$$

A solução $x(t)$ do sistema acima será uma curva iniciando em x_0 e se movendo na direção de menor gradiente, ou seja, a solução é dada pelo famoso algoritmo de descida de gradiente. Em outras palavras, a solução $x(t)$ caracteriza um fluxo de gradiente[3]. Um exemplo desse tipo de solução é mostrado na Figura 3.1

[1] A principal referência para este capítulo é Santambrogio [8].

[2] Fluxo de gradientes é uma área por si só, assim, sugerimos ao leitor nesta área que consulte Ambrosio et al. [1]

[3] Essa definição é informal. O conceito de fluxo de gradiente será formalizado na seção seguinte.

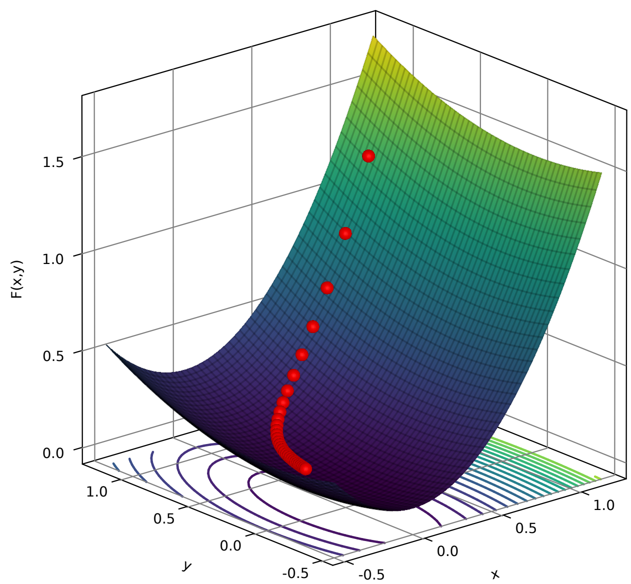


Figure 3.1: Exemplo de fluxo de gradiente onde $F(x, y) = x^2 + 0.2y^2$, com ponto inicial $(x(0), y(0)) = (1.0, 1.0)$. A superfície representa os valores de $F(x, y)$, enquanto que as esferas vermelhas representam a solução do fluxo de gradiente com um *time-step* de 0.1. Perceba que a "velocidade" do fluxo é proporcional ao gradiente, como pode ser visto pelo espaçamento decrescente entre cada uma das esferas.

Esse problema é simples quando estamos em espaços de dimensão finita e com funções diferenciáveis, porém, torna-se mais interessante e complexo quando começamos a considerar espaços de dimensão infinita como $\mathcal{P}_2(\mathbb{R}^n)$. Neste cenário, temos que repensar, por exemplo, a ideia de gradiente, já que não está mais claro que seria o gradiente quando $x(t) = \rho_t \in \mathcal{P}_2(\mathbb{R}^n)$. Além disso, F não é mais uma função de $\mathbb{R}^n$ em $\mathbb{R}$, mas um funcional atuando em medidas de probabilidade.

É interessante observar que, uma vez que consigamos reformular EDPs como fluxos de gradiente em Wasserstein, poderemos utilizar resultados obtidos nessa área para provar, por exemplo, existência e unicidade.

3.1 Introdução ao Fluxo de Gradiente

3.1.1 Definições Iniciais

Antes de formalizar a ideia de fluxo de gradiente, vamos introduzir alguns conceitos de análise convexa que são necessárias para tratar do assunto de maneira rigorosa.

Definition 3.1.1 (Subdiferencial)**.** Seja $f : \mathbb{R}^n \to (-\infty, +\infty]$ própria, ou seja, $f(x) \neq +\infty \forall x$. O subdiferencial de f é dado por:

$$\partial f(x) := \{p \in \mathbb{R}^n : f(y) \geq f(x) + \langle p, y - x\rangle, \forall y \in \mathbb{R}^n\}. \tag{3.2}$$

Se $p \in \partial f(x)$, então p é um subgradiente de f no ponto x.

A intuição por trás da definição de subdiferencial é ilustrada na Figura 3.2. Note que se a função f for convexa e diferenciável, teremos que $\partial f(x) = \{\nabla f(x)\}$. Porém, caso a função não seja convexa, não haverá essa garantia. Assim, é comum usar essa ideia de subdiferencial somente em funções convexas.

Uma primeira generalização dessa ideia de subdiferencial pode ser usada em funções que chamamos de semi-convexas, ou λ-convexas. Muitos resultados de fluxo de gradientes podem ser aplicados a essa classe maior de funções, que, como apontado por Santambrogio [8], cobrem vários casos de interesse (e.g. em conjuntos limitados, todas as funções $f \in C^2$ serão semi-convexas).

Definition 3.1.2 (λ-Convexidade)**.** Uma função $f : \mathbb{R}^n \to (-\infty, +\infty]$ é λ-convexa para um $\lambda \in \mathbb{R}$ se

$$g(x) := f(x) - \frac{\lambda}{2}|x|^2 \tag{3.3}$$

for convexa. Note que se f for λ-convexa com $\lambda = 0$, temos que a função é convexa. Se $\lambda < 0$, a noção é mais fraca que convexidade e implica que f tem sua parte negativa com crescimento no máximo quadrático. Finalmente, se $\lambda > 0$, a função é estritamente convexa e limitada inferiormente.

Definition 3.1.3 (λ-Subdiferencial)**.** Seja $f : \mathbb{R}^n \to (-\infty, +\infty]$ própria, e $\lambda \in \mathbb{R}$. O λ-subdiferencial de f é dado por

$$\nabla_\lambda f(x) := \{p \in \mathbb{R}^n : f(y) \geq f(x) + \langle p, y - x\rangle + \frac{\lambda}{2}|y - x|^2, \forall y \in \mathbb{R}^n\}. \tag{3.4}$$

Podemos generalizar ainda mais essa noção de subdiferencial com a seguinte definição.

Definition 3.1.4 (Subdiferencial Generalizado)**.** [4] Seja $f : \mathbb{R}^n \to (-\infty, +\infty]$ própria, ou seja, $f(x) \neq +\infty \forall x$. O subdiferencial generalizado de f é dado por:

$$\partial_G g(x) := \{p \in \mathbb{R}^n : \liminf_{t \to 0^+} \frac{f(x+tv) - f(x)}{t} \geq \langle p, v \rangle v \in \mathbb{R}^n\}. \tag{3.5}$$

Onde $t \to 0^+$ simboliza que t tendo a 0 pela direita. Note que $\nabla f(x) \subset \nabla_\lambda f(x) \subset \nabla_G f(x)$, com igualdade das três caso f seja convexa.

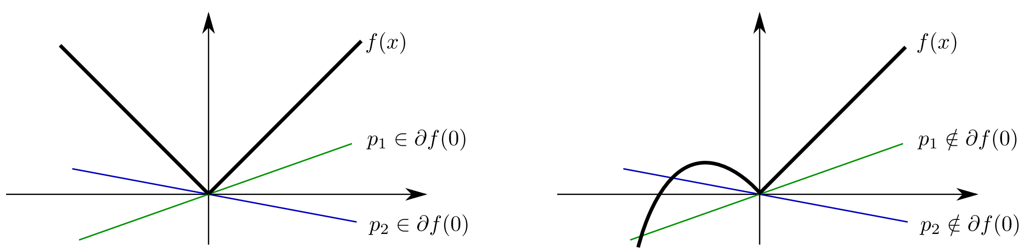


Figure 3.2: Exemplo de subdiferencial. Perceba que considerando o subdiferencial generalizado, teríamos $p_1 \in \partial_G f(0)$ e $p_2 \in \partial_G f(0)$ para as duas imagens.

Agora podemos definir de forma rigorosa o fluxo de gradiente.

Definition 3.1.5 (Fluxo de Gradiente)**.** Seja $f : \mathbb{R}^n \to \mathbb{R}$. Assim, dizemos que $x : (0, +\infty) \to \mathrm{Dom}(f)$ é um fluxo de gradiente de f se $x \in \mathrm{AC}_{\mathrm{loc}}((0, +\infty), \mathbb{R}^n)$ e

$$x'(t) \in -\nabla_G f(x(t)), \quad \text{para } \lambda\text{-a.e } t \in (0, +\infty), \tag{3.6}$$

onde λ é a medida de Lebesgue. Note que dizemos que x começa em x_0 se $\lim_{t \to 0^+} x(t) = x_0$.

[4] Ambrosio et al. [2] chama de Gateaux subdifferencial.

3.1.2 Resultados Básicos de Existência e Unicidade

Uma vez introduzida a ideia de fluxo de gradiente, vamos agora provar alguns resultados básicos relacionados ao sistema de equações

$$\begin{cases} x'(t) \in -\partial F(x(t)), \text{ para quase todo } t > 0, \\ x(0) = x_0. \end{cases} \tag{3.7}$$

Veja que agora a função F não é mais necessariamente derivável. Este cenário é bem menos restrito que o que apresentamos logo no início do capítulo.

Lemma 3.1.1. *Seja $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ convexa, $p_1 \in \partial f(x_1)$ e $p_2 \partial f(x_2)$. Então*

$$\langle p_1 - p_2, x_1 - x_2 \rangle \geq 0. \tag{3.8}$$

Proof. Pela definição de subdiferencial, temos que

$$\begin{aligned} p_1 \in \partial f(x_1) &\implies f(x_2) \geq f(x_1) + \langle p_1, x_2 - x_1 \rangle \\ p_2 \in \partial f(x_2) &\implies f(x_1) \geq f(x_2) + \langle p_2, x_1 - x_2 \rangle. \end{aligned}$$

Assim, somando as duas equações, temos que

$$f(x_2) + f(x_1) \geq f(x_1) + f(x_2) + \langle p_1, x_2 - x_1 \rangle \langle p_2, x_1 - x_2 \rangle$$

Rearranjando, obtemos

$$\begin{aligned} 0 &\geq \langle p_1, x_2 \rangle - \langle p_2, x_1 \rangle \langle p_2, x_1 \rangle - \langle p_2, x_2 \rangle \\ &\implies \langle p_1 - p_2, x_1 - x_2 \rangle \geq 0. \end{aligned}$$

□

Theorem 3.1.1. *Seja $F : \mathbb{R}^n \to \mathbb{R}$ **convexa**, $x_0 \in \mathbb{R}^n$, e x_1 e x_2 duas soluções de (3.7). Então,*

$$|x_1(t) - x_2(t)| \leq |x_1(0) - x_2(0)|, \ \forall t > 0. \tag{3.9}$$

Logo, a solução do sistema de equações é única.

Proof. Primeiro, faça $g(t) = \frac{(x_1(t)-x_2(t))^2}{2}$, e toma a derivada em t. Assim

$$g'(t) = \langle x_1(t) - x_2(t), x_1'(t) - x_2'(t) \rangle.$$

Usando o fato que $x_1'(t) \in \partial F(x_1(t))$ e $x_2'(t) \in \partial F(x_2(t))$, temos pelo lemma que acabamos de demonstrar que

$$g'(t) = \langle x_1'(t) - x_2'(t), x_1(t) - x_2(t) \rangle \geq 0.$$

Além disso, como ambas as soluções começam em x_0, temos que $g(0) = 0 \geq g(t)$, já que a derivada de $g(t)$ é sempre menor ou igual a zero. Portanto

$$\begin{aligned} g(t) = \frac{(x_1(t) - x_2(t))^2}{2} \leq 0 &\implies |x_1(t) - x_2(t)| \leq |x_1(0) - x_2(0)| = 0 \\ &\implies x_1(t) = x_2(t). \end{aligned}$$

□

Podemos estender o resultado do teorema acima para o caso mais geral onde F é λ-convexa. A demonstração é bastante parecida, porém, um pouco mais convoluta. Por conta disso optamos por apresentar os dois resultados de forma separada.

Theorem 3.1.2. *Seja $F : \mathbb{R}^n \to \mathbb{R}$ **λ-convexa**, $x_0 \in \mathbb{R}^n$. Então o sistema de equações (3.7) tem uma solução única. Além disso, se $\lambda > 0$, a taxa de convergência para o mínimo global de F é exponencial, ou seja, para $x^* = \operatorname{argmin}_{x \in \mathbb{R}^n} F(x)$*

$$|x(t) - x^*| \leq e^{-\lambda t} |x(0) - x^*|, \tag{3.10}$$

onde $x(t)$ é a solução.

Outra propriedade relevante dos fluxos de gradiente é que eles podem ser caracterizados por meio do chamado *Esquema de Minimização de Movimento*. A ideia é a seguinte, imagine que queremos resolver o problema de fluxo de gradiente descrito em (3.7). Uma forma de fazer isso seria discretizando a variável t, e aplicando algum método de descida em relação à F, já que $x(t)$ tem "velocidade" sempre na direção oposta do gradiente de F. Assim, denotaremos um *time-step* $\tau > 0$, e uma sequência (x_k^τ), onde $x_0^\tau = x(0)$, $x_1^\tau = x(\tau)$,..., $x_k^\tau = x(k\tau)$, ou seja, cada elementos da sequência representa um ponto da solução $x(t)$ no tempo discretizado.

O *Esquema de Minimização de Movimento* é definido pela seguinte iteração:

$$x_{k+1}^{\tau} \in \underset{x}{\operatorname{argmin}}\, F(x) + \frac{|x - x_k^{\tau}|^2}{2\tau}. \tag{3.11}$$

Note que se F for λ-convexa, o problema acima tem solução única. A primeira vista, o esquema de minimização acima pode parecer contra-intuitivo, entretanto, supondo que F é derivável, sabemos que a solução de (3.11) é obtida quando $\nabla(F(x) + \frac{|x-x_k^{\tau}|^2}{2\tau}) = 0$, assim,

$$-\nabla F(x_{k+1}^{\tau}) = \frac{x_{k+1}^{\tau} - x_k^{\tau}}{\tau}. \tag{3.12}$$

Ou seja, esse esquema de minimização é o famoso esquema implícito de Euler. Lembre-se da diferença entre o esquema implícito e o explícito de Euler:

$$\text{(Euler Implícito)} \qquad x_{k+1}^{\tau} = x_k^{\tau} - \tau \nabla F(x_{k+1}^{\tau}) \tag{3.13}$$

$$\text{(Euler Explicito)} \qquad x_{k+1}^{\tau} = x_k^{\tau} - \tau \nabla F(x_k^{\tau}) \tag{3.14}$$

Suponha que queremos achar a solução para o intervalo de tempo $[0, T]$. Assim, discretizamos esse intervalo em usando o passo τ. Uma vez que obtemos a sequência de pontos (x_k^{τ}), o objetivo agora é interpolar esses pontos de alguma forma a gerar uma solução aproximada de $x(t)$ para $t \in [0, T]$. A maneira mais óbvia seria por meio de uma interpolação linear, que denotaremos por $\tilde{x}^{\tau}(t)$. Outra seria fazendo $x^{\tau}(t) = x_{k+1}^{\tau}$, ou seja, uma função escada. Ambos métodos são ilustrados na Figura 3.3.

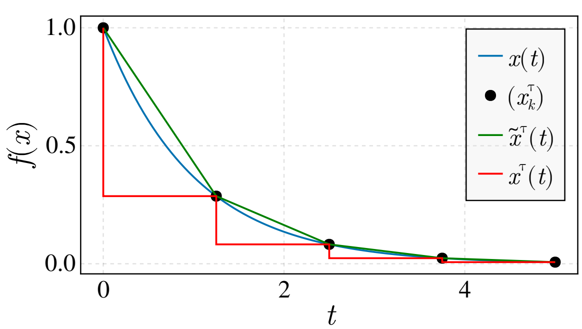


Figure 3.3: Exemplo de aproximação da solução $x(t)$ por meio de $x^{\tau}(t)$ usando interpolação linear dos pontos da sequência (x_k^{τ}).

Agora apresentamos o seguinte resultado que caracteriza o *Esquema de Movimento de Minimização* (3.11) com a solução do fluxo de gradiente.

Theorem 3.1.3. *Sejam $\tilde{x}^\tau, x^\tau$ construídas como apresentado acima. Suponha que $F(x_0) < +\infty$ e $F(x) > -\infty$ para todo x. Então, existe uma subsequência $\tau_i \to 0$, tal que ambas aproximações $\tilde{x}^\tau$ e x^τ convergem uniformemente para $x(t)$ em $[0, T]$, onde $x(t)$ é Absolutamente Contínua em $[0, T]$ e é L^2. Além disso, defina*

$$\boldsymbol{v}^\tau(t) = \frac{x_{k+1}^\tau - x_k^\tau}{\tau}, \text{ para } t \in (k\tau, (k+1)\tau]. \tag{3.15}$$

Assim, $\boldsymbol{v}^\tau$ converge fracamente em L^2 para $\boldsymbol{v} = x'$ e

1. *Se F é λ-convexo, então $\boldsymbol{v}(t) \in -\partial F(x(t))$ em quase todo t;*

2. *Se $F \in C^1$, então $\boldsymbol{v}(t) = -\nabla F(x(t))$ para quase todo t.*

3.2 Fluxo de Gradiente em Espaços Métricos

Começamos nossa exposição focando em $\mathbb{R}^n$, porém, espaços de Wasserstein são espaços métricos geodésicos onde não está claro, por exemplo, o que seria $x'(t)$ onde $x(t) = \mu_t \in \mathcal{P}(\mathbb{R}^n)$. Assim, vamos estender nossa análise subsequente para o caso de espaços métricos geodésicos.

3.2.1 Revisando Geodésicas

O conceito de geodésicas já foi apresentado no Capítulo anterior. Porém, vamos fazer uma breve revisão[5].

Definition 3.2.1 (Derivada Métrica)**.** Seja $\omega : [0, 1] \to X$ uma curva no espaço métrico (X, d). Definimos a sua derivada métrica como

$$|\omega'|(t) := \lim_{h \to 0} \frac{d(\omega(t+h), \omega(t))}{|h|}, \tag{3.16}$$

caso esse limite exista.

[5]Retirar essa seção na versão final do livro. Está aqui somente para que o capítulo seja mais auto-contido.

Definition 3.2.2 (Curva Absolutamente Contínua)**.** Uma curva $\omega : [0,1] \to \mathbb{X}$ é dita absolutamente contínua se existir $g \in L^1([0,1])$ tal que $d(\omega(t_0), \omega(t_1)) \leq \int_{t_0}^{t_1} g(s)ds$ para todo $t_0 < t_1$. O conjunto de todas as curvas absolutamente contínuas em (X, d) é denotado por $\mathrm{AC}(X)$, onde omitimos o d quando a métrica está clara.

Definition 3.2.3 (Comprimento)**.** Seja $\omega : [0,1] \to X$ um curva, defina o comprimento dessa curva como

$$\mathrm{Len}(\omega) := \sup \left\{ \sum_{k=0}^{n-1} d(\omega(t_k), \omega(t_{k+1})) : \; n \geq 1, 0 = t_0 < t_1, ..., t_n = 1 \right\}. \tag{3.17}$$

Definition 3.2.4 (Geodésica)**.** Uma curva $\omega : [0,1] \to X$ é uma geodésica entre x_0 e $x_1 \in X$ se $\omega(0) = x_0$, $\omega(1) = x_1$ e $\mathrm{Len}(\omega) = \min\{\mathrm{Len}(\tilde{\omega}) : \omega \in \mathrm{AC}(X), \tilde{\omega}(1) = x_1\}$. Ou seja, um curva ω é uma geodésica entre dois pontos caso ela seja a curva de menor distância.

Assim, fica claro que a idéia de geodésica é uma extensão da idéia de retas em espaços Euclidianos para espaços mais gerais, como superfícies curvas.

Definition 3.2.5 (Espaço Geodésico)**.** Um espaço métrico (X, d) é um espaço geodésico se para todo $x, y \in X$,

$$d(x, y) = \min\{\mathrm{Len}(\omega) : \omega \in \mathrm{AC}(X), \omega(0) = x, \omega(1) = y\}. \tag{3.18}$$

Ou seja, a distância entre dois pontos é dada pela geodésica de menor comprimento. Note que num espaço geodésico, essa geodésica sempre existe.

Definition 3.2.6 (Geodésica de Velocidade Constante)**.** Uma geodésica $\omega : [0,1] \to X$ é dita ter velocidade constante se

$$d(\omega(t), \omega(s)) = \frac{|t-s|}{1-0} d(\omega(0), \omega(1)), \forall t, s \in [0,1]. \tag{3.19}$$

O Teorema [2.1.2] mostra que para curvas $\omega \in \mathrm{AC}(X)$, temos que

$$\mathrm{Len}(\omega) = \int_0^1 |\omega'|(t)dt, \tag{3.20}$$

ou seja, se ω for de velocidade constante, temos que $\omega'(t)$ é a velocidade de ω e é uma constante.

Vamos mostrar agora que na verdade toda geodésica pode ser reparametrizada para possuir velocidade constante.

Proposition 3.2.1. *Seja (X, d) um espaço geodésico. Então, para todo $x_1, x_2 \in X$, existe uma geodésica de velocidade constante $\tilde{\omega}(t)$ ligando x_1 e x_2, i.e. $\tilde{\omega}(0) = x_1$ e $\tilde{\omega}(1) = x_2$.*

Proof. Seja $\omega(t)$ uma geodésica entre x_1 e $x_2 \in X$, teremos então

$$d(x_1, x_2) = \mathrm{Len}(\omega) = \int_0^1 |\omega'|(t)dt.$$

Faça agora $\tilde{\omega}(s) = \omega(g(s))$, onde

$$g(s) := \bar{r}_s \cdot d(x_1, x_2),$$

onde

$$\bar{r}_s := \inf \left\{ r \in [0,1] : \int_0^r |\omega'|(t)dt \geq s \right\}$$

Assim, $\mathrm{Len}(\tilde{\omega}) = d(x_1, x_2)$ e tem velocidade constante.

Exercício. Conclua demonstrando que de fato essa curva $\tilde{\omega}$ tem comprimento igual à ω e tem velocidade constante. Pensei nessa solução, mas não cheguei a verificar se de fato é verdade. □

Concluímos essa revisão de geodésicas com uma definição menos comum, chamada de convexidade geodésica.

Definition 3.2.7 (Convexidade Geodésica)**.** Seja (X, d) um espaço geodésico. Assim, a função $F : X \to \mathbb{R} \cup \{+\infty\}$ é *geodesicamente convexa* se para todo $x_0, x_1 \in X$, existe uma geodésica de velocidade constante $x(t)$ entre x_0 e x_1, onde a função F é convexa, ou seja,

$$F(x(t)) \leq (1-t)F(x(0)) + tF(x(1)). \tag{3.21}$$

E, similarmente, F é dita *geodesicamente λ-convexa* se

$$F(x(t)) \leq (1-t)F(x(0)) + tF(x(1)) - \lambda \frac{t(1-t)}{2} d^2(x_0, x_1). \tag{3.22}$$

Note que essa definição na verdade é mais fraca que convexidade ao longo de toda geodésica, pois, é possível que existam duas geodésicas entre x_0 e x_1, onde F é convexa em uma, mas não é na outra.

3.2.2 Formulação EDE e EVI

Considere um espaço métrico (X, d), com $F : X \to \mathbb{R} \cup +\infty$ inferiormente semi-continua. Assim, o *Esquema de Minimização de Movimento* pode ser reescrito como

$$x_{k+1}^{\tau} \in \underset{x}{\operatorname{argmin}}\, F(x) + \frac{d(x, x_k^{\tau})^2}{2\tau}. \tag{3.23}$$

Sendo nosso espaço geodésico, podemos usar novamente a idéia de obter uma sequência de pontos (x_k^{τ}), e obter $x^{\tau}(t)$ interpolando com geodésicas. Ou seja, para $t \in (k\tau, (k+1)\tau]$, faça $x^{\tau}(t)$ igual a geodésica de velocidade constante $d(x_k^{\tau}, x_{k+1}^{\tau})/\tau$, que sempre existirá se nosso espaço for geodésico.

Queremos assim que x^{τ} convirja para uma solução $x(t)$ onde $x'(t) = -\nabla F(x(t))$. Porém, para um espaço métrico qualquer não está claro que nem $x'(t)$ nem $-\nabla F(x(t))$ existem! Invés disso podemos usar a noção de derivada métrica que apresentamos e que sabemos que de fato existirá.

Bibliography

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[2] Luigi Ambrosio, Elia Brué, and Daniele Semola. Lectures on optimal transport, 2021.

[3] David JH Garling. *Analysis on Polish spaces and an introduction to optimal transportation*, volume 89. Cambridge University Press, 2018.

[4] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[5] Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted boltzmann machines. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3718–3726. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf.

[6] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[7] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[8] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[9] user125646 (https://math.stackexchange.com/users/125646/user125646). How to show that the set of all lipschitz functions on a compact set x is dense in c(x)? Mathematics Stack Exchange. URL https://math.stackexchange.com/q/665686. URL:https://math.stackexchange.com/q/665686 (version: 2014-02-07).

[10] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Chapter 4

Appendix

4.1 Auxiliary - Probability and Analysis

This section contains definitions and results in Probability and Analysis that are used throughout the text. These results are listed here mostly without proofs.

Definition 4.1.1. Let $d : X \times X \to \mathbb{R}_+$. We say that d is a metric on the set X if for all $x, y, z \in X$, the following three assertions are true:

i) $d(x, y) = 0 \iff x = y$

ii) $d(x, y) = d(y, x)$

iii) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

Definition 4.1.2. (Weak convergence) We say that $\mu_n \rightharpoonup \mu$ if and only if $\forall f$ continuous and bounded, we have $\int f \, d\mu_n \to \int f \, d\mu$.

Note that this is equivalent to the notion of convergence in distribution, which is more commonly known in probability.

Theorem 4.1.1. *(Portmanteau) Given $\mu \in \mathcal{P}(X)$, where X is a metric space. Then, the following statements are equivalent:*

i) $\mu_n \rightharpoonup \mu$*;*

ii) $\forall f$ *bounded and uniformly continuous, we have* $\int f \, d\mu_n \to \int f \, d\mu$*;*

iii) $\forall F \subset X$ *closed,* $\mu(F) \geq \limsup_n \mu_n(F)$*;*

iv) $\forall F \subset X$ *open,* $\mu(A) \leq \liminf_n \mu_n(A)$*;*

v) $\forall B$ *such that* $\mu(\partial B) = 0$*, then* $\mu_n(B) \to \mu(B)$*.*

Note that every set B *with* $\mu(\partial B) = 0$*is called a continuity set. And* ∂B *is the boundary set of* B*, hence* $\partial B := \hat{B} \setminus \mathring{B}$*.*

Theorem 4.1.2. *Let* X, Y *be metric spaces and* $\mu_n \rightharpoonup \mu$*. Given a continuous map* $h : X \to Y$*, then* $h_{\#}\mu_n = \mu_n \circ h^{-1} \rightharpoonup h_{\#}\mu$*.*

Corollary 4.1.1. *If* $\mu_n \rightharpoonup \mu$ *with* $h : X \to Y$ *such that* $\mu(D_h) = 0$ *where* D_h *is the set of points of discontinuity. Then,* $\mu_n \circ h^{-1} \rightharpoonup \mu \circ h^{-1}$*.*

Proposition 4.1.1. *If* X *is Polish, and* d *is a lower semi-continuous metric on* X*. For* $p \in [1, +\infty)$ *and* $x_0 \in X$*,* $\mu_n \rightharpoonup \mu$ *and* $\int_X d(x, x_0)^p d\mu_n \to \int_X d(x, x_0)^p d\mu$*, if, and only if,* $\mu_n \rightharpoonup \mu$ *and* $\lim_{R\to\infty} \int_{d(x,x_0)\geq R} d(x, x_0) d\mu_n \to 0$ *(uniformly integrable).*

Definition 4.1.3. (Tight) A family of probability measures $\mathcal{A}$ is tight if for $\epsilon > 0$, $\exists K \subset X$ compact, such that for any $\mu_\alpha \in \mathcal{A}$, $\mu_\alpha(X \setminus K) < \epsilon$

Theorem 4.1.3. *(Prokhorov) This theorem consists in two separate results.*

i) If the family $\mathcal{G} = \{\mu_\alpha\}_{\alpha\in\Lambda}$ *is tight, then* $\mathcal{G}$ *is sequentially pre-compact, i.e. for any* $(\mu_n) \subset \mathcal{G}$*,* $\exists \mu_{n_k} \rightharpoonup \mu$*, where* $\mu \in \overline{\mathcal{G}}$*;*

ii) If X *is Polish and* $\mathcal{G} = \{\mu_\alpha\}_{\alpha\in\Lambda} \subset \mathcal{P}(X)$ *is pre-compact. Then* $\mathcal{G}$ *is tight. In other words, for* X *polish, and* $\mu_n \in \mathcal{P}(X)$ *with* $\mu_n \rightharpoonup \mu$*, then the sequence* (μ_n) *is tight.*

Definition 4.1.4. (Disintegration)

For a Borel measurable space X with a measure μ. Given a function $f : X \to Y$. We say that the family $(\mu_y)_{y\in Y}$ is a Disintegration of μ according to f if every measure μ_y is concentrated on $f^{-1}(\{y\})$, and for every $\phi \in C(X)$, the map $\phi \mapsto \int_X \phi d\mu_y$ is Borel measurable with

$$\int_X \phi \, d\mu = \int_Y \int_X \phi \, d\mu_y(x) \, d\nu(y), \quad \text{where } \nu = f_{\#}\mu \tag{4.1}$$

Note that the existence and uniqueness of disintegration families depend on the spaces where the probabilities are defined, to which we introduce the next theorem.

Theorem 4.1.4. *(Garling [3] 16.10.1) Suppose that X and Y are Polish spaces, that $\mu \in \mathcal{P}(X)$ and that f is a Borel measurable map from X to Y. Then, the f-disintegration of μ exists, and is essentially unique (i.e. $\mu(f^{-1}(B)) = 0$, with $B := \{y \in f(X) : \mu_y \neq \mu'_y\}$ where μ_y and μ'_y are two disintegrations).*

Theorem 4.1.5. *$f : X \to \mathbb{R}$ is uniformly continuous $\iff$ $\exists\ \omega : \mathbb{R}_+ \to \mathbb{R}_+$, such that ω is increasing and $\lim_{x\to 0} w(x) = 0$ with $|f(x) - f(y)| \leq \omega(d(x, y))$, $\forall x, y \in X$. We call ω the modulus of continuity.*

Definition 4.1.5. (Equicontinuous) For a metric space X, the sequence of functions $f_n : X \to \mathbb{R}$ is equicontinuous if $\forall \epsilon > 0,\ \exists \delta > 0 :\ d(x, y) < \delta \implies d(f_n(x), f_n(y)) < \epsilon$ for every $n \in \mathbb{N}$.

Definition 4.1.6. (Equibounded) We say that a sequence (or family) of functions (f_n) is equibounded, if $\exists M > 0\ :\ |f_n(x)| < M < +\infty\ \forall n \in \mathbb{N}$. In words, there is a value M that bounds all functions in the sequence.

Theorem 4.1.6. *(Arzelà-Ascoli) If X is a compact metric space with f_n equicontinuous and equibounded, then $\exists f_{n_k} \to_{unif.} f$, where f is continuous.*

Theorem 4.1.7. *Let (X, d) be metric space. Thus, if X is compact, then $\mathrm{Lip}(X)$ is dense in $C(X)$.*

Proof. (Proof from user125646 [9]) Let $g : X \to \mathbb{R}$ be a continuous function, then since X is compact, g is uniformly continuous. Therefore, for any $\varepsilon > 0$, one can take a $\delta > 0$ such that $d(x, y) < \delta$ implies $|g(x)-g(y)| < \varepsilon$. Now, let $M = \sup_x |g(x)|$ and define

$$f(x) := \sup_y g(y) - \frac{2Md(x, y)}{\delta}$$

Now, note that f is Lipschitz, since

$$\begin{aligned} f(x_1) - f(x_2) &= \sup_y \left(g(y) - \frac{2Md(x_1, y)}{\delta} \right) - \sup_y \left(g(y) - \frac{2Md(x_2, y)}{\delta} \right) \\ &\leq \sup_y \frac{2M(d(x_1, y) - d(x_2, y))}{\delta}) \end{aligned}$$

By the triangle inequality, $d(x_1, y) - d(x_2, y) \leq d(x_1, x_2)$, then

$$\sup_y \frac{2M(d(x_1, y) - d(x_2, y))}{\delta} \leq \sup_y \frac{2Md(x_1, x_2)}{\delta} = \frac{2Md(x_1, x_2)}{\delta}$$

The same argument is valid by exchanging x_1 and x_2, so f has Lipschitz constant $\frac{2M}{\delta}$. Next, let's prove that $\sup_x |g(x) - f(x)| < \varepsilon$.

A first point to notice is that $f(x) \geq g(x)$, since for $y = x$, we have $f(x) = g(x)$. For $d(x, y) \geq \delta$,

$$f(x) = \sup_y g(y) - \frac{2Md(x,y)}{\delta} \leq \sup_y -2M \leq -M \leq g(x)$$

Hence $f(x) \geq g(x) \geq f(x)$, so we obtain an equality.

For $d(x, y) < \delta$,

$$f(x) - g(x) = \sup_y g(y) - g(x) - \frac{2Md(x,y)}{\delta} \leq \varepsilon - \frac{2Md(x,y)}{\delta} < \varepsilon$$

We conclude that $0 < f(x) - g(x) < \varepsilon$, so $\sup_x |f(x) - g(x)| < \varepsilon$.

□

4.2 Auxiliary - Inequalities

Lemma 4.2.1. *(Inf-Sup Inequality)*

$$|\inf_{x\in A} f(x) - \inf_{x\in A} g(x)| \leq \sup_{x\in A} |f(x) - g(x)| \tag{4.2}$$

Proof. Let's write $\sup_{x\in A} f(x)$ as $\sup_A f$ for simplicity. Note that $f = f - g + g$, hence,

$$\sup_A f = \sup_A f - g + g \leq \sup_A (f - g) + \sup_A g \implies$$
$$\sup_A f - \sup_A g \leq \sup_A f - g \leq \sup_A |f - g|$$

Using the same argument for g, we obtain that

$$|\sup_A f - \sup_A g| \leq \sup_A |f - g| \tag{4.3}$$

Finally, note that

$$|\sup_A f - \sup_A g| = |\inf_A (-f) - \inf_A (-g)| = | - \inf_A f + \inf_A g| =$$
$$= |\inf_A f - \inf_A g| \leq \sup_A |f - g|$$

□

Lemma 4.2.2. *(Minkowski's Inequality) Let X be a measurable space, for $p \in [1, +\infty)$ and $f, g \in L^p(X)$. Therefore,*

$$||f + g||_{L^p(X)} \leq ||f||_{L^p(X)} + ||g||_{L^p(X)} \tag{4.4}$$

Where $||f||^p_{L^p(X)} = \int_X |f|^p d\mu$.

Lemma 4.2.3 (Lemma de Gronwall)**.** *Seja $f : I \to \mathbb{R}$ uma função contínua e não negativa em um intervalo $I \subset \mathbb{R}$, tal que*

$$f(t) \leq a + \left| \int_u^t bf(s)ds \right| \tag{4.5}$$

para todo $t \in I$, com $a, b \geq 0$, e $u \in I$ constantes. Então

$$f(t) \leq ae^{b|t-u|}, \forall t \in I. \tag{4.6}$$