0.1 Optimal Transport Theory

The field of Optimal Transport has grown quite substantially in recent years [1], and going through the theory in order understand how it applies to Machine Learning can be a challenging task for ML researchers not acquainted with the field. Hence, we have filtered the main theoretical results necessary for understanding the applications of Optimal Transport to Machine Learning presented in this dissertation.

This section is mainly based on the book "Optimal Transport for Applied Mathematicians" by Santambrogio [1]. 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.

0.1.1 A Brief Introduction to Optimal Transport

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 Machine Learning 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?" [2][2]. 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), 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

[1] Villani [2] is roughly a thousand pages of theoretical results on OT.

[2] This is not a quote from Monge.

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 2).

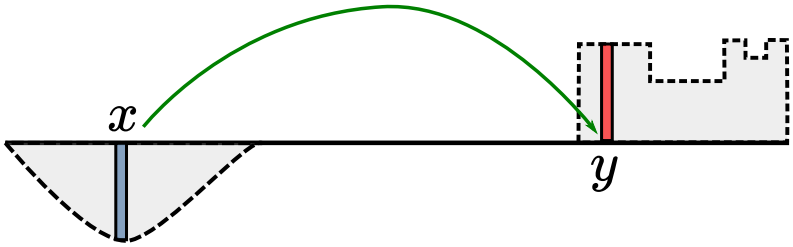

Figure 1: The figure illustrates the original Monge Problem, where all the mass is excavated from location x is transported to a deterministic location y. The transport assignment map is represented by the arrow in green.

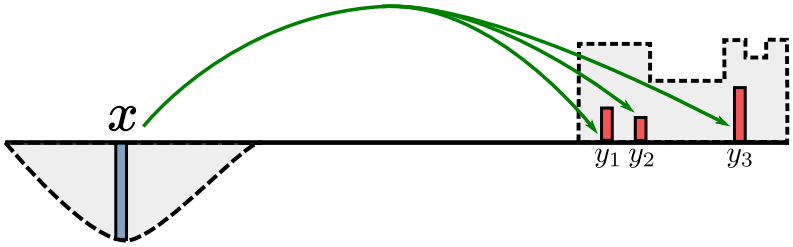

Figure 2: The figure illustrates the Optimal Transport Problem with the Kantorovich relaxation. The transportation assignment now can split the mass in blue, transporting it to many positions.

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 sections:

(i) **Monge Problem** - We formally define the Monge Problem;

(ii) **Kantorovich Problem** - We formally define the Kantorovich Problem 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;

(iii) **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;

(iv) **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;

(v) **Wasserstein Distance** - We define the Wasserstein and show that it is formally a metric (0.2.1). 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.

0.1.2 Monge Problem

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

Definition 0.1.1. 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 0.1.2. 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 0.1.3. (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:

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

Theorem 0.1.1. *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:*

$$\int_Y f dT_{\#}\mu = \int_X f \circ T d\mu \tag{2}$$

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,

$$\int_Y f \; dT_{\#}\mu = \int_Y \lim_{n\to+\infty} f_n \; dT_{\#}\mu = \lim_{n\to+\infty} \int_Y f_n \; dT_{\#}\mu =$$

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

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 0.1.4. (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:

$$(MP) \qquad \inf \left\{ \int_X c(x, T(x)) d\mu \quad : \quad T_{\#}\mu = \nu \right\} \tag{3}$$

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 3. A viable solution T to MP is called a **Transport Map**.

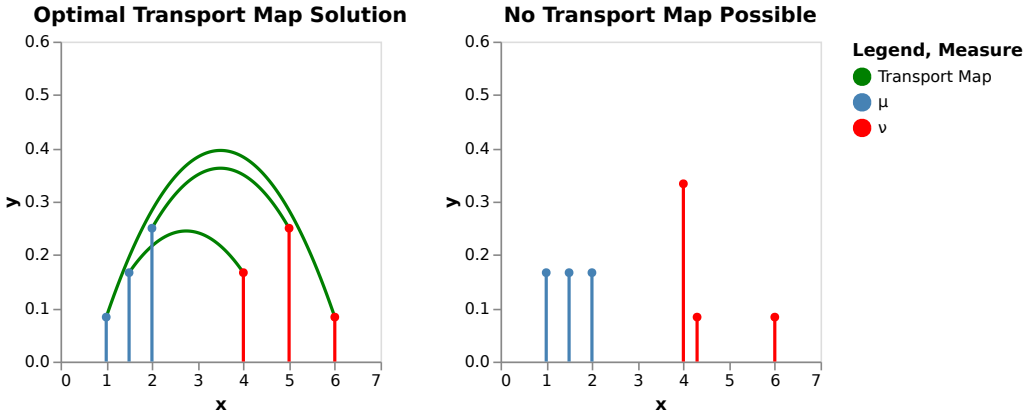

Figure 3: Example of two Optimal Transport Problems. On the left, there exists an optimal transport plan, while on the right there is no possible solution.

0.1.3 Kantorovich Problem

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 0.1.5. (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 0.1.1. *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,

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

$\impliedby$) 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,

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

□

Definition 0.1.6. (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**:

$$\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) \; : \; (\pi_x)_\# \gamma = \mu \quad \text{and} \quad (\pi_y)_\# \gamma = \nu\} \tag{4}$$

Finally, we can state the Kantorovich Problem.

Definition 0.1.7. (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:

$$(KP) \qquad \inf \left\{ \int_{X\times Y} c(x, y) d\gamma \; : \; \gamma \in \Pi(\mu, \nu) \right\} \tag{5}$$

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

$$\forall x \in X, \ f(x) \leq \liminf_{n \to +\infty} f(x_n) \tag{6}$$

Definition 0.1.9. (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:

$$\bar{F}(x) := \inf \left\{ \liminf_{n} F(x_n) \ : \ x_n \to x \right\} \tag{7}$$

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

Below in Figure 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)

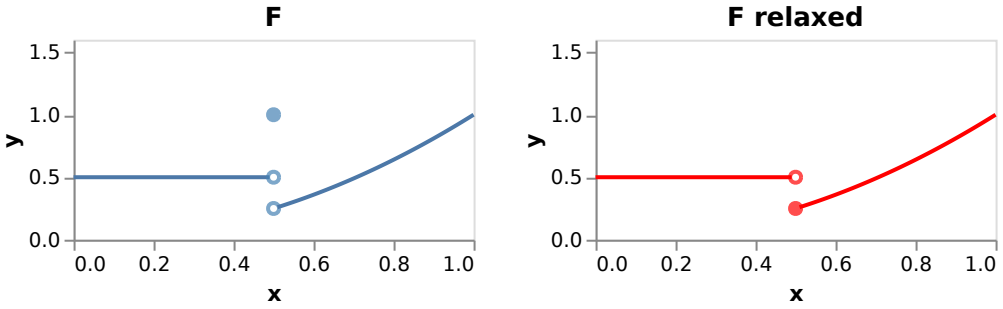

Figure 4: Example of a function F and it's relaxation.

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 [1].

Theorem 0.1.2. *(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 0.1.3. *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 0.2.2), 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 $K(\gamma) \leq J(\gamma)$.

By Theorem 0.1.2, 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,

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

We can conclude that

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

□

0.1.4 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 0.1.4. *(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 (0.2.3), because $(X \times Y)$ is compact. Then, by Prokhorov Theorem 0.2.3, $\exists \gamma_{n_k} \rightharpoonup \gamma$.

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

$$\int \phi(x) \, d\mu \underset{Cor.0.1.1}{=} \int \phi(x) \, d\gamma_{n_k} \to \int \phi(x) \, d\gamma$$
$$\int \psi(y) \, d\nu \underset{Cor.0.1.1}{=} \int \psi(y) \, d\gamma_{n_k} \to \int \psi(y) \, d\gamma$$

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 0.1.1. *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 $x_n \to x$:

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

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

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

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

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

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

□

Theorem 0.1.5. *(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 0.1.4. 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 0.1.6. *(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 (0.2.3), and therefore, by Prokhorov's Theorem (i) 0.2.3, it is pre-compact. Once this is done, the proof follows in the same manner as Theorem 0.1.4.

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) 0.2.3, 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.

□

0.1.5 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 0.1.2. *The Kantorovich Problem (0.1.7) is equivalent to:*

$$\inf_{\gamma\in\mathcal{M}_+(X\times Y)}\int_{X\times Y}c(x,y)d\gamma+\sup_{(\phi,\psi)\in B}\int_X\phi(x)\ d\mu + \int_Y\psi(y)\ d\nu-\int_{X\times Y}\phi(x)+\psi(y)\ d\gamma \quad (8)$$

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 (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 (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** [3]:

$$\inf_{\gamma\in\mathcal{M}_+(X\times Y)}\sup_{(\phi,\psi)\in B}\int_{X\times Y}c\ d\gamma+\int_X\phi\ d\mu+\int_Y\psi\ d\nu-\int_{X\times Y}\phi\oplus\psi\ d\gamma \quad (9)$$

- **Dual**:

$$\sup_{(\phi,\psi)\in B}\inf_{\gamma\in\mathcal{M}_+(X\times Y)}\int_{X\times Y}c\ d\gamma+\int_X\phi\ d\mu+\int_Y\psi\ d\nu-\int_{X\times Y}\phi\oplus\psi\ d\gamma \quad (10)$$

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

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

$$\sup_{(\phi,\psi)\in B} \int_X \phi \, d\mu + \int_Y \psi \, d\nu - \inf_{\gamma\in\mathcal{M}_+(X\times Y)} \int_{X\times Y} c - (\phi \oplus \psi) \, d\gamma \tag{11}$$

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 0.1.10. 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

$$\text{(DP)} \qquad \sup \left\{ \int_X \phi \, d\mu + \int_Y \psi \, d\nu \ : \phi \in C_b(X) \ , \psi \in C_b(y) \ , \ \phi \oplus \psi \leq c \right\} \tag{12}$$

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 0.1.3. *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,

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

□

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

Definition 0.1.11. (c-Transform) Given $f : X \to \overline{\mathbb{R}}$, and $c : X \times Y \to \overline{\mathbb{R}}$, the c-transform of f is:

$$f^c(y) := \inf_x c(x,y) - f(x) \tag{13}$$

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:

$$f^*(y) := \sup_x x \cdot y - f(x) \tag{14}$$

Lemma 0.1.4. *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* [4] *as c.*

Proof. By Theorem 0.2.5, 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 0.2.1

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

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 0.1.7. *(Santambrogio 1.11)*

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

$$\max(\text{DP}) = \max_{\phi \in c-conc.(X)} \int_X \phi \, d\mu + \int_Y \phi^c \, d\nu \tag{15}$$

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

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

[4] Check Theorem 0.2.5 for the definition of modulus of continuity

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 0.1.4, 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 (0.2.6), 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 0.2.5) and Equibounded (see definition 0.2.6).

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

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

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,

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

Also, for any $x \in X$:

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

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 0.2.6. 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.:

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

□

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(DP) = min(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 0.1.12. (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 $i \in \{1, ..., n\}$

$$\sum_{i=1}^n c(x_i, y_i) \leq \sum_{i=1}^n c(x_i, y_{\sigma(i)}) \tag{16}$$

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:

$$\langle x_1, y_1\rangle + \langle x_2, y_2\rangle \leq \langle x_1, y_2, \rangle + \langle x_2, y_1\rangle \tag{17}$$

Definition 0.1.13. For X a separable metric space, we define the support of a a measure μ as

$$\text{spt } \mu := \bigcap\{A \ : \ A \text{ is closed and } \mu(X \setminus A) = 0\} \tag{18}$$

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 $\text{spt}(\gamma)$ 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 $(x, y) \in \text{spt}(\gamma)$. Hence, this would prove that

$$\int_{X\times Y} c(x, y)\ d\gamma = \int_X \phi(x)\ d\mu + \int_Y \phi^c(y) d\nu \tag{19}$$

Theorem 0.1.8. *(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*

$$\Gamma \subset \{(x, y) \ : \ \phi(x) + \phi^c(y) = c(x, y)\} \tag{20}$$

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:

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

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 0.1.7). Since $\phi = \psi^c$, then, ϕ is continuous, therefore, it is measurable if we consider the Borel σ-algebra. □

Theorem 0.1.9. *(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 [1] for the complete proof. □

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

Theorem 0.1.10. *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 0.1.4, we obtain that $\exists\gamma \in \Pi(\mu, \nu)$ such that it minimizes the Kantorovich Problem, therefore, by Theorem 0.1.9, sptγ is c-CM.

By Proposition 0.1.7, we know that a solution to the Dual Problem can be found in the set of c-concave functions. Using 0.1.8, we can assert that there is a set of c-concave functions such that $\phi(x)+\phi^c(y) = c(x, y)$ for every $(x, y) \in$ spt γ. 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 $\max(\text{DP}) = \min(\text{KP})$. □

Theorem 0.1.11. *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 0.1.6 and affirm that exists an optimal solution γ to (KP).

By the same arguments used on the proof of Theorem 0.1.10, 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 0.1.8, 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 0.1.9), 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 $\sup \psi(y) > L > -\infty$:

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

Finally, using the same arguments as Theorem 0.1.10, we conclude that $\max(\text{DP}) = \min(\text{KP})$ and that (ϕ, ϕ^c) are a solution for the Dual Problem. □

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 0.1.12. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, with $c(x,y) = \frac{1}{2}|x-y|^2$. Suppose that $\int |x|^2 d\mu, \int |y|^2 d\nu < +\infty$ [5] . Instead of the original Dual Problem, consider*

[5] This is Theorem 1.40 in Santambrogio [1], 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 $\int |x|^2 d\mu, \int |y|^2 d\nu < +\infty$.

the following formulation:

$$(\mathrm{DP}') \quad \sup\left\{\int_{\mathbb{R}^d}\phi\ d\mu+\int_{\mathbb{R}^d}\psi\ d\nu\ :\phi\in L^1(\mu)\ ,\psi\in L^1(\nu)\ ,\ \phi\oplus\psi\leq c\right\} \tag{21}$$

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

Proof. First, in the same way as the proof of Theorem 0.1.11, (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 $h(x):=\frac{|x|^2}{2}-\phi(x)$

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

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

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

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 0.1.13. *(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,* $\sup(\text{DP}) = \min(\text{KP})$. *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. 0.2.1), then we can prove the following result:

Theorem 0.1.14. *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.

$\implies$) 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:

$$\begin{gathered} c(x, y) \leq c(x, z) + c(z, y) \implies \sup_y c(x, y) - c(y, z) \leq c(x, z) \\ c(y, z) \leq c(y, x) + c(x, z) \implies \sup_y c(y, z) - c(x, y) \leq c(x, z) \\ \therefore \\ \sup_y |c(y, z) - c(x, y)| \leq c(x, z) \end{gathered}$$

Therefore,

$$\begin{gathered} |f(x) - f(z)| = |\inf_y \{c(x, y) - g(y)\} - \inf_y \{c(z, y) - g(y)\}| \leq \\ \underset{0.2.1}{\leq} \sup_y |c(x, y) - c(z, y)| \leq c(x, z) \end{gathered}$$

$\Longleftarrow$) Let $f \in \text{Lip}_1^{(c)}$. 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 0.1.13 and 0.1.14, one obtains the famous Kantorovich-Rubinstein Duality:

Theorem 0.1.15. *(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

$$\sup \left\{ \int_X \phi \, d\mu - \int_X \phi \, d\nu \ : \phi \in Lip_1(X) \right\} \tag{22}$$

0.1.6 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 0.1.14. (Probability space with p-Moments)

$$\mathcal{P}_p(X) := \{\mu \in \mathcal{P}(X) : \int_{X \times X} d(x,y)^p \, d\mu(x) d\mu(y) < +\infty\} \tag{23}$$

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 **?**] Proposition 21.1.1.

Definition 0.1.15. (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:

$$W_p(\mu,\nu) := \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times X} d(x,y)^p \, d\gamma \right)^{1/p} \tag{24}$$

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 0.1.15 to obtain:

$$W_1(\mu, \nu) = \sup_{\phi \in Lip_1} \int_X f d(\mu - \nu) \tag{25}$$

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

Lemma 0.1.5. *(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. 0.2.4) 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 0.2.4). 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 0.1.1, 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 0.1.1. *$W_p(\cdot, \cdot)$ 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 0.1.5, 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,

$$\begin{aligned}
W_p(\mu,\nu) &\leq \left(\int_{X\times X} d(x,z)^p \, d\gamma\right)^{1/p} = \left(\int_{X\times X} d(x,z)^p \, d(\pi_{x,z})_{\#}\sigma\right)^{1/p} \\
&\underset{Thm.0.1.1}{=} \left(\int_{X\times X\times X} d(x,z)^p \, d\sigma\right)^{1/p} \\
&\leq \int_{X^3} (d(x,y)+d(y,z))^p \, d\sigma \\
&= ||d\circ(\pi_{x,y})(x,y,z) - d\circ(\pi_{y,z})(x,y,z)||_{L^p(\sigma)} \\
&\underset{0.2.2}{\leq} ||d\circ(\pi_{x,y})(x,y,z)||_{L^p(\sigma)} + ||d\circ(\pi_{y,z})(x,y,z)||_{L^p(\sigma)} \\
&= \left(\int_{X^3} d(x,y)^p \, d\sigma\right)^{1/p} + \left(\int_{X^3} d(y,z)^p \, d\sigma\right)^{1/p} \\
&= \left(\int_{X^2} d(x,y)^p \, d\gamma^{+}\right)^{1/p} + \left(\int_{X^2} d(y,z)^p \, d\gamma^{-}\right)^{1/p} \\
&= W_p(\mu,\rho) + W_p(\rho,\nu)
\end{aligned}$$

Which proves the triangle inequality for the Wasserstein distance.

□

Definition 0.1.16. (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 0.1.2. *For a bounded Polish space X, $p \in [1,+\infty)$, $\mu,\nu \in \mathcal{P}_p(X)$ and $C \in \mathbb{R}_+$, then*

$$W_1(\mu,\nu) \leq W_p(\mu,\nu) \leq C W_1(\mu,\nu)^{1/p} \tag{26}$$

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:

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

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 0.1.6. *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.

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

$\Longleftarrow$) 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,

$$\begin{aligned}\left|\int g d\mu_n - \int g d\mu\right| &= \left|\int g' d\mu_n - \int g' d\mu\right| \\ &\leq \left|\int g' d\mu_n - \int f_k d\mu_n\right| + \left|\int f_k d\mu_n - \int f_k d\mu\right| + \left|\int f_k d\mu - \int g' d\mu\right|\end{aligned}$$

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,

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

□

Theorem 0.1.16. *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 0.1.4 the Kantorovich Problem has a solution. Also, by Theorem 0.1.10, 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 $f \in \text{Lip}_1$, $\int f d\mu_n \to \int f d\mu$. 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 0.2.7), which leads us to conclude that $\mu_n \rightharpoonup \mu$ (by Portmanteau 0.2.1). Now, by Proposition 0.1.2, 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 $\phi_{n_k} \in \text{Lip}_1$ 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 (0.2.6) to obtain a sub-subsequence that converges uniformly to a $\phi \in \mathrm{Lip}_1(X)$. 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 0.1.2:

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

□

Theorem 0.1.17. *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*

$$W_p(\mu_n, \mu) \to 0 \iff \int_X d(x, x_0)^p d\mu_n \to \int_X d(x, x_0)^p d\mu \text{ and } \mu_n \rightharpoonup \mu \quad (27)$$

Proof.

$\implies$) Let $W_p(\mu_n, \mu) \to 0$. Since X is Polish, and c is a continuous function, by Theorem 0.1.6 the Kantorovich Problem has a solution. Also, by Theorem 0.1.13, we obtain that $\sup(\mathrm{DP}) = \min(\mathrm{KP})$. 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 0.2.7), we can use Lemma 0.1.6 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

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

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:

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

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,

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

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

Lastly, note that since $\overline{B(R)}$ is compact, then we can use Theorem 0.1.16 to stablish that

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

We can then conclude that

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

□

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 [2] under Theorem 6.9.

Theorem 0.1.18. *For (X, d) a Polish metric space, $\mu_n, \mu \in \mathcal{P}_p(X)$ and $x_0 \in X$. Then*

$$W_p(\mu_n, \mu) \to 0 \iff \int_X d(x, x_0)^p d\mu_n \to \int_X d(x, x_0)^p d\mu \text{ and } \mu_n \rightharpoonup \mu \quad (28)$$

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 [2] 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 5 highlights how W_p takes into account the underlying geometry compared to the Kullback-Leibler divergence, which does not.

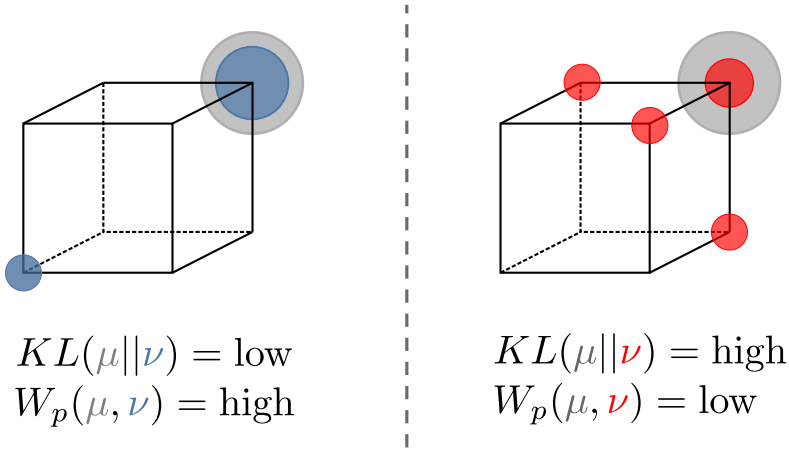

Figure 5: Comparison between Wasserstein distance and KL Divergence, based on **?**]. On the left, there is a large overlap between the two distributions, but a large geometrical distance for a portion. On the right, there is much less overlap, but the whole distribution is geometrically closer. These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

Bibliography

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

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

0.2 Appendix

0.2.1 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 0.2.1. 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 0.2.2. (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 0.2.1. *(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 0.2.2. *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 0.2.1. *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 0.2.1. *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 0.2.3. (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 0.2.3. *(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 0.2.4. (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

$$\int_X \phi \, d\mu = \int_Y \int_X \phi \, d\mu_y(x) \, d\nu(y), \qquad \text{where } \nu = f_{\#}\mu \tag{29}$$

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 0.2.4. *(?] 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 0.2.5. *$f : X \to \mathbb{R}$ 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 0.2.5. (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 0.2.6. (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 0.2.6. *(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 0.2.7. *Let (X, d) be metric space. Thus, if X is compact, then* $\mathrm{Lip}(X)$ *is dense in $C(X)$.*

Proof. (Proof from **?**]) 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 $\sup_x |f(x) - g(x)| < \varepsilon$.

□

0.2.2 Auxiliary - Inequalities

Lemma 0.2.1. *(Inf-Sup Inequality)*

$$|\inf_{x\in A} f(x) - \inf_{x\in A} g(x)| \leq \sup_{x\in A} |f(x) - g(x)| \tag{30}$$

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

$$|\sup_A f - \sup_A g| \leq \sup_A |f - g| \tag{31}$$

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

$$||f + g||_{L^p(X)} \leq ||f||_{L^p(X)} + ||g||_{L^p(X)} \tag{32}$$

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