Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

Topological Properties of the Wasserstein metrics

Tesi di Laurea in Probabilità e Statistica

Relatore:
Professor
Andrea Pascucci

Presentata da:
Riccardo Roda

Anno Accademico 2024/2025

Alle persone a me care

Introduction

When studying convergence of probability measures, an important issue is the choice of a probability metric; it might be useful to quantify that convergence in terms of some probability distance such that one may study convergence in metric terms. Under certain conditions one may choose a metric over another one.

It will be shown that if $\mathcal{X}$ is a separable metric space, then the choice of the Levy-Prokhorov metric seems to be theoretically the more appropriate, despite being difficult to compute. This metric, metrizes weak convergence if $\mathcal{P}(\mathcal{X})$ is equipped with the topology of weak convergence. Moreover, if $\mathcal{X}$ is Polish, then $\mathcal{P}(\mathcal{X})$ is Polish. It follows that any measure μ might be approximated by a discrete measure with finite support; another important property is that any Cauchy sequence in the Levy-Prokhorov distance converges weakly to some measure.

In the main chapter, Wasserstein metrics of order p are introduced and they are denoted by W_p, where $p \in [1, \infty)$. They are defined over a complete and separable metric space, or a Polish space, $(\mathcal{X}, d)$ and their construction comes from the theory of optimal transportation, when one introduces a cost function. In our case, the cost function is defined in terms of the distance of the space. We denote by $\mathcal{P}_p(\mathcal{X})$ the Wasserstein space of order p, in which any measure has finite p-th moment. Since $\mathcal{X}$ is Polish, then $\mathcal{P}_p(\mathcal{X})$ is Polish and the Wasserstein distance metrizes weak convergence. Moreover there is a rich duality that comes with it, that may be useful in many occasions, for instance when $p = 1$. Many techniques are used in order to prove those powerful results, like coupling techniques and few theorems or lemmas correlated, from which follow fundamental properties, like the existence of an optimal coupling, lower semicontinuity of the cost functional with respect to *weak topology* and tightness of transference plans.

Quando si studia la convergenza di misure di probabilità, risulta essere utile quantificare questa convergenza in termini di una distanza probabilistica, così da poterne studiare la convergenza in termini metrici. La scelta di una di queste metriche può rivelarsi fondamentale in quanto, sotto certe condizioni, potrebbe risultarne più conveniente una rispetto a un'altra.

Verrà affrontata la distanza di Levy-Prokhorov, che teoricamente risulta essere fondamentale su uno spazio metrico separabile $\mathcal{X}$, nonostante la difficoltà nel calcolarla. Questo perché tale metrica, metrizza la convergenza debole su $\mathcal{P}(\mathcal{X})$ se dotata della topologia indotta dalla convergenza debole, in gergo *weak topology*. Inoltre, se $\mathcal{X}$ è spazio Polacco, allora anche $\mathcal{P}(\mathcal{X})$ lo è; segue che ogni misura di probabilità è approssimabile a una misura discreta con supporto finito e che ogni successione di Cauchy rispetto alla distanza di Levy-Prokhorov è convergente in $\mathcal{P}(\mathcal{X})$.

Infine, nel capitolo principale, affronteremo le metriche di Wasserstein di ordine p, W_p, definite su uno spazio Polacco $(\mathcal{X}, d)$. Si parte dalla teoria del trasporto ottimale, in cui viene introdotta una funzione costo; se tale funzione è scritta in termini di d, si può costruire una classe di metriche al variare di $p \geq 1$. Considereremo il sottospazio di $\mathcal{P}(\mathcal{X})$ in cui le misure hanno momento p-esimo finito e lo chiameremo $\mathcal{P}_p(\mathcal{X})$. Si mostrerà essere anch'esso uno spazio Polacco e che W_p ne metrizza la convergenza debole. W_p possiede una definizione duale, detta *Kantorovich duality*, che risulterà essere utile soprattutto nel caso in cui $p = 1$. Per dimostrare questi risultati, si utilizzeranno tecniche di *coupling* e teoremi o lemmi da cui ne conseguono proprietà fondamentali, tra cui l'esistenza di un "trasporto" (o coupling) ottimale, semicontinuità dal basso del funzionale rispetto alla *weak topology* e la *tightness* dei "piani di trasporto".

Contents

Chapter 1

Preliminaries and notations

1.1 Conventions

Sets, structures and function spaces

Id is the identity mapping, regardless of the space.

if A is a set, then the function 1_A is the indicator function of A: $1_A(x) = 1$ if $x \in A$ and 0 otherwise. If f and g are two functions, then (f, g) is the function $x \longmapsto (f(x), g(x))$; the map $f \times g$ is the function $(x, y) \longmapsto (f(x), g(y))$.

The Euclidian scalar product between two vectors $a, b \in \mathbb{R}^n$, where n is a positive integer, is denoted by $a \cdot b$ or $\langle a, b \rangle$; for instance, the first notation will be seen in the *second step* of the proof of Proposition 3.5.

A Polish space is, by definition, a separable and completely metrizable space, in other words, it can be equipped with a metric that makes this space a complete and separable metric space. For this reason we are going to use "Polish space" and "complete and separable metric space" as synonyms; the notation used will be $\mathcal{X}$, $(\mathcal{X}, d)$ or (E, d).

If (E, d) is a metric space, the open ball of radius r centered in x is denoted by $B_r(x)$; the closed ball will be denoted by $\overline{B_r(x)}$. In general, if $A \in E$ is a set, its closure will be $\overline{A}$, which is also the set of all limits of sequences with value in A. Its intern part will be denoted by $\mathring{A}$ and is the the biggest open set in A. With A^ε we indicate the open ε-neighborhood of A.

A map f between two metric spaces (E, d) and (E', d') is said to be C-Lipschitz if $d'(f(x), f(y)) \leq Cd(x, y)$ for all $x, y \in E$; the lowest admissible constant C is denoted by $\|f\|_{\mathrm{L}}$.

$C(E)$ is the space of continuous function $E \longrightarrow \mathbb{R}$ and $bC(E) \subset C(E)$ is the space of all bounded continuous function.

The notation $a \wedge b$ stands for $\min\{a, b\}$. This notion can be extended in the usual way

to functions and also signed measures.

Probability measures

δ_x is the Dirac mass at point x.

If E is a Polish space, or more in general a metric space, it will be equipped with its natural Borel σ-algebra $\mathcal{B}$, so any probability measure on E is a Borel measure. Only when introducing Total Variation distance in Chapter 2 we have a generic measurable space.

The space of all probability measures on E is denoted by $\mathcal{P}(E)$ and the weak topology on it is induced by convergence against $bC(E)$, i.e. *bounded continuous* test functions. If E is Polish, even $\mathcal{P}(E)$ is Polish (see Theorem 2.5) and becomes a complete and separable metric space once being equipped with Levy-Prokhorov distance.

The integral of a function f with respect to a probability measure μ is denoted by $\int_E f(x)\ \mathrm{d}\mu(x)$ or in a more simple way $\int f\ \mathrm{d}\mu$.

If μ is a Borel measure on a topological space $\mathcal{X}$, a set S is said to be μ-negligible if it is included in a Borel set of zero μ-measure. Then μ is said to be concentrated on a set C if $\mu(\mathcal{X} \setminus C) = 0$, and its support ($\mathrm{Spt}\mu$) is the smallest closed set in which μ is concentrated.

If μ is a Borel measure and $T : \mathcal{X} \longrightarrow \mathcal{Y}$ a Borel map, then the *push-formard* measure of μ, denoted by $T_{\#}\mu$, is a Borel measure on $\mathcal{Y}$ and is defined as $T_{\#}\mu(A) = \mu(T^{-1}(A))$ for any Borel set $A \in \mathcal{Y}$.

The law of a random variable X defined on a probability space $(\Omega, \mathcal{G}, \mathbb{P})$ is denoted by $\mathcal{L}(X)$. The expected value is denoted by $\mathbb{E}$.

Let $(\mathcal{X}, \mathcal{G})$ be a measurable space. A measure ν is absolutely continuous with respect to μ, and is written $\nu << \mu$, if $\mu(A) = 0$ implies $\nu(A) = 0$. By Radon-Nikodym theorem, there is a measurable function $f : \mathcal{X} \longrightarrow [0, \infty]$ such that $\nu(A) = \int_A f(x)\ \mathrm{d}\mu(x)$ for any $A \in \mathcal{G}$ and $f = \mathrm{d}\nu/\mathrm{d}\mu$ is called Radon-Nikodym derivative. Notice that if $\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) < 1$, then ν is a sub-probability measure.

A sequence of probability measures $\{\mu_n\}_n$ is said to converge weakly to a measure μ, and is written $\mu_n \xrightarrow{\mathrm{w}} \mu$, if $\int f\ \mathrm{d}\mu_n \to \int f\ \mathrm{d}\mu$ for all bounded continuous function f.

Notations specific to optimal transport

If $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, then $\Pi(\mu, \nu)$ is the set of all joint distributions on $\mathcal{X} \times \mathcal{Y}$ whose marginals are μ and ν. Given a cost function $c(x, y)$, the optimal total cost between μ and ν is $C(\mu, \nu) = \inf \mathbb{E}\left[c(X, Y)\right]$, where $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$. Let

introduce the projection to the first and second factor:

$$\begin{aligned} p_1 : \mathcal{X} \times \mathcal{Y} &\longrightarrow \mathcal{X} & \qquad p_2 : \mathcal{X} \times \mathcal{Y} &\longrightarrow \mathcal{Y} \\ (x, y) &\longmapsto x & (x, y) &\longmapsto y \ . \end{aligned}$$

Then if $\pi \in \Pi(\mu, \nu)$, we have $(p_1)_{\#}(\pi) = \mu$ and $(p_2)_{\#}(\pi) = \nu$.

1.2 Couplings and its Properties

Coupling techniques and properties will turn out to be essential in order to demonstrate those important results of the main chapter, so it is necessary to introduce them. We start by giving a definition.

Definition 1.1 (Coupling)**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces. Coupling μ and ν means constructing two random variables X and Y on some probability space $(\Omega, \mathbb{P})$ such that $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$. The couple (X, Y) is called a coupling of (μ, ν), by abuse of language, the law of (X, Y) is also called a coupling.*

If μ and ν are the only laws in the problem, without loss of generality, one may assume Ω to be $\mathcal{X} \times \mathcal{Y}$. In other terms, coupling μ and ν means constructing a measure π on $\mathcal{X} \times \mathcal{Y}$ such that admits μ and ν as marginals on $\mathcal{X}$ and $\mathcal{Y}$ respectively. The following three statements are equivalent to each other and they are all ways to rephrase what have just been explained:

- $(p_1)_{\#}(\pi) = \mu$ and $(p_2)_{\#}(\pi) = \nu$, where p_i are the projection to the first or second factor;
- For all measurable sets $A \in \mathcal{X}$, $B \in \mathcal{Y}$, one has $\pi(A \times \mathcal{Y}) = \mu(A)$ and $\pi(\mathcal{X} \times B) = \nu(B)$;
- For all integrable (or nonnegative) measurable function $\phi,\ \psi$ on $\mathcal{X}, \mathcal{Y}$

$$\int_{\mathcal{X} \times \mathcal{Y}} (\phi(x) + \psi(y)) \ \mathrm{d}\pi(x, y) = \int_{\mathcal{X}} \phi(x) \ \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \ \mathrm{d}\nu(y).$$

Remark 1.2. A coupling always exists, because we have always the trivial case in which the random variables X, Y are independent, thus the joint distribution is the product measure between μ, ν. However, since the value of X does not give any information about Y, this can hardly be called a coupling.

Another extreme case is when all the information about the value of Y is contained in the value of X, in other words, Y is a deterministic function of X. Notice that this is not a symmetric property in general.

Definition 1.3 (Deterministic Couplings)**.** *With the same notation of Definition 1.1, a coupling (X, Y) is said to be a deterministic one if there exists a measurable function $T : \mathcal{X} \longrightarrow \mathcal{Y}$ such that $Y = T(X)$.*

Saying that (X, Y) is a deterministic coupling of μ and ν, is strictly equivalent to one of the four statements below:

- (X, Y) is a coupling of μ and ν whose law π is concentrated on the *graph* of a measurable function $T : \mathcal{X} \longrightarrow \mathcal{Y}$;
- X has law μ and $Y = T(X)$, where $T_{\#}\mu = \nu$;
- X has law μ and $Y = T(X)$, where T is a **change of variables** form μ to ν: for all ν-integrable (or nonnegative) function ϕ,
$$\int_{\mathcal{Y}} \phi(y)\ \mathrm{d}\nu(y) = \int_{\mathcal{X}} \phi(T(x))\ \mathrm{d}\mu(x);$$
- $\pi = (Id, T)_{\#}\mu$.

The map T is the same in all these statements and is uniquely defined μ-almost surely, once the law of (X, Y) has been fixed. The converse is also true: if T and $\tilde{T}$ coincide μ-almost surely, then $T_{\#}\mu = \tilde{T}_{\#}\mu$. However deterministic couplings do not always exist such as common couplings; if μ is a Dirac mass and ν is not, we cannot find a measurable function T. In other cases one may find infinitely many deterministic couplings between two given probability measures.

We give an example of an important coupling, that we will se to be coherent with the main chapter.

Example 1.4 (Optimal coupling or Optimal transport). Here one introduces a **cost function** $c(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ that can be interpreted as the work needed to move one unit of mass from location x to location y. Then one considers the *Monge-Kantorovich minimization problem*

$$\inf \left\{ \mathbb{E}[c(X, Y)] : \mathcal{L}(X) = \mu,\ \mathcal{L}(Y) = \nu \right\};$$

In terms of measures, it is equivalent to find

$$C(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\ \mathrm{d}\pi(x, y).$$

Those joint measures are called transference plan and those achieving the infimum are called **optimal transference plan**.

Once we let c to be lower semicontinuous and $\mathcal{X}, \mathcal{Y}$ to be Polish spaces, nontrivial results can be obtained.

Definition 1.5 (Lower semicontinuity)**.** *Let $\mathcal{X}$ be a topological space and $f : \mathcal{X} \longrightarrow \overline{\mathbb{R}}$. f is said to be **lower semicontinuous** in $x_0 \in \mathcal{X}$ if $f(x_0) \in \mathbb{R}$ and for all $\varepsilon > 0$, there is a neighborhood U of x_0 such that $f(x) > f(x_0) - \varepsilon$, for all $x \in U$. Another way to say that is* $\liminf_{x \to x_0} f(x) \geq f(x_0)$.

Remark 1.6. With the same hypothesis, except for $f(x) < f(x_0)+\varepsilon$, we get the definition of **upper semicontinuous function**. Like in previous definition, this is equivalent to say $\limsup_{x \to x_0} f(x) \leq f(x_0)$.

The following lemma will be used to prove that the Wasserstein Metrics, defined in Chapter 3 Section 3.1, satisfy the triangle inequality.

Lemma 1.7 (Gluing lemma)**.** *Let $(\mathcal{X}_i, \mu_i)$, $i = 1, 2, 3$, be Polish spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and (Y_2, Y_3) is a coupling of (μ_2, μ_3), then one can construct a triple of random variables (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same law of (X_1, X_2) and (Z_2, Z_3) has the same law of (Y_2, Y_3).*

The idea behind it is that if π_{12} is the law of (X_1, X_2) on $\mathcal{X}_1 \times \mathcal{X}_2$ and π_{23} is the law of (Y_2, Y_3) on $\mathcal{X}_2 \times \mathcal{X}_3$, then to construct the law π_{123} of (Z_1, Z_2, Z_3), one just has to *glue* π_{12} and π_{23} along their common marginal. In other terms disintegrate π_{12} and π_{23} as follows:

$$\pi_{12}(\mathrm{d}x_1\ \mathrm{d}x_2) = \pi_{12}(\mathrm{d}x_1|x_2)\mu_2(\mathrm{d}x_2),$$
$$\pi_{23}(\mathrm{d}x_2\ \mathrm{d}x_3) = \pi_{23}(\mathrm{d}x_3|x_2)\mu_2(\mathrm{d}x_2);$$

then reconstruct π_{123} as

$$\pi_{123}(\mathrm{d}x_1\ \mathrm{d}x_2\ \mathrm{d}x_3) = \pi_{12}(\mathrm{d}x_1|x_2)\mu_2(\mathrm{d}x_2)\pi_{23}(\mathrm{d}x_3|x_2).$$

Correlated to Example 1.4 there is a theorem and two lemmas that are really important in sight of Chapter 3, but first let introduce **Prokhorov's Theorem**.

Theorem 1.8 (Prokhorov's theorem)**.** *If $\mathcal{X}$ is a Polish space, then a set $\mathcal{K} \subset \mathcal{P}(\mathcal{X})$ is precompact for the weak topology if and only if it is tight.*

Let clarify first some of those terms.

Definition 1.9 (Precompact with respect to weak topology)**.** *$\mathcal{K} \in \mathcal{P}(\mathcal{X})$ is said to be precompact if any sequence in $\mathcal{K}$ has a subsequence that converges weakly to some probability meausure on $\mathcal{X}$. Then its closure with respect to weak topology is said to be weakly compact.*

Definition 1.10 (Tightness)**.** *Let $\mathcal{X}$ be a Polish space, a subset $\mathcal{K}$ of $\mathcal{P}(\mathcal{X})$ is tight if for all $\varepsilon > 0$ there exists a compact set $K_\varepsilon \in \mathcal{X}$ such that $\mu(\mathcal{X} \setminus K_\varepsilon) \leq \varepsilon$ for all $\mu \in \mathcal{K}$.*

Corollary 1.11. *Let $(\mathcal{X}, d)$ be a Polish space and let $\mathcal{K} \subset \mathcal{P}(\mathcal{X})$ be a finite subset, then $\mathcal{K}$ is tight.*

Proof. Let $\varepsilon > 0$ and $\mu \in \mathcal{P}(\mathcal{X})$. Using the separability of $\mathcal{X}$, for any $1 \leq p \in \mathbb{N}$, there exists an $m(p) \in \mathbb{N}$ such that

$$\mu\left(\bigcup_{i=1}^{m(p)} B_{2^{-p}\varepsilon}(x_i)\right) > 1 - 2^{-p}\varepsilon,$$

with each x_i is in the dense of $\mathcal{X}$. Let consider

$$K = \bigcap_{p\geq 1} \bigcup_{i=1}^{m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)}.$$

It is closed as an intersection of finite unions of closed sets, thus it is complete since $\mathcal{X}$ is complete. For any $\delta > 0$, by simply choosing a p large enough so that $2^{-p}\varepsilon < \delta$,

$$K \subset \bigcup_{i=1}^{m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)} \subset \bigcup_{i=1}^{m(p)} B_\delta(x_i),$$

and K is totally bounded, proving that K is compact.
Last thing to check is the following:

$$\begin{aligned}\mu(\mathcal{X} \setminus K) &= \mu\left(\mathcal{X} \setminus \bigcap_{p\geq 1} \bigcup_{i=1}^{m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)}\right) = \mu\left(\bigcup_{p\geq 1} \mathcal{X} \setminus \bigcup_{i=1}^{m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)}\right) \\ &\leq \sum_{p\geq 1} \mu\left(E \setminus \bigcup_{i=1}^{m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)}\right) \leq \varepsilon \sum_{p\geq 1} 2^{-p} = \varepsilon.\end{aligned}$$

Therefore if $\mathcal{K} = \{\mu_1, \cdots, \mu_l\}$ we find l compact sets $\left\{K^1, \cdots, K^l\right\}$; then by taking

$$K := \bigcup_{j=1}^{l} K^j,$$

$\mu(K) > 1 - \varepsilon$, for all $\mu \in \mathcal{K}$, concluding the proof.

□

The first good thing about optimal couplings is that they exist.

Theorem 1.12 (Existence of an optimal coupling)**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces; let $a : \mathcal{X} \longrightarrow \mathbb{R} \cup \{-\infty\}$ and $b : \mathcal{Y} \longrightarrow \mathbb{R} \cup \{-\infty\}$ be two upper semicontinuous functions such that $a \in L^1(\mathcal{X}, \mu)$, $b \in L^1(\mathcal{Y}, \nu)$.*
Let $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous cost function such that $c(x, y) \geq a(x) + b(y)$ for all x, y. Then there is a coupling of (μ, ν) that minimize the total cost $\mathbb{E}\,[c(X, Y)]$ among all possible couplings (X, Y).

Remark 1.13. The lower bound assumption on c guarantees that the expected cost $\mathbb{E}\,[c(X, Y)]$ is well-defined on $\mathbb{R} \cup \{+\infty\}$. For instance, if c is a distance, one may choose a and b to be zero.

The proof relies on two properties: lower semicontinuity and compactness.

Lemma 1.14 (Lower semicontinuity of the cost functional)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces and $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous cost function. Let $h : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R} \cup \{-\infty\}$ be an upper semicontinuous cost function such that $c \geq h$. Let $\{\pi_k\}_{k \in \mathbb{N}}$ be a sequence of probability measures on $\mathcal{X} \times \mathcal{Y}$, converging weakly to some $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ in such a way that $h \in L^1(\mathcal{X} \times \mathcal{Y}, \pi_k)$, $h \in L^1(\mathcal{X} \times \mathcal{Y}, \pi)$ and*

$$\int_{\mathcal{X} \times \mathcal{Y}} h \, \mathrm{d}\pi_k \underset{k \to \infty}{\longrightarrow} \int_{\mathcal{X} \times \mathcal{Y}} h \, \mathrm{d}\pi.$$

Then

$$\int_{\mathcal{X} \times \mathcal{Y}} c \, \mathrm{d}\pi \leq \liminf_{k \to \infty} \int_{\mathcal{X} \times \mathcal{Y}} c \, \mathrm{d}\pi_k.$$

In particular, if c is nonnegative, then $F : \pi \to \int c \, \mathrm{d}\pi$ is a lower semicontinuous function on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, equipped with the topology of weak convergence.

Remark 1.15. In case $c = d$ is a distance, it is always nonnegative, therefore, the Wasserstein distances introduced in Chapter 3 are lower semicontinuous functions.

Proof of Lemma 1.14. Replacing c with $c - h$, one may assume that c is a nonnegative lower semicontinuous function. Then c can be written as the pointwise limit of a non-decrescending bounded continuous real-valued function $(c_l)_{l \in \mathbb{N}}$: $c_l \leq c_{l+1} \nearrow c$ and by monotone convergence,

$$\int c \, \mathrm{d}\pi = \lim_{l \to \infty} \int c_l \, \mathrm{d}\pi = \lim_{l \to \infty} \lim_{k \to \infty} \int c_l \, \mathrm{d}\pi_k \leq \liminf_{k \to \infty} \int c \, \mathrm{d}\pi_k.$$

□

Lemma 1.16 (Tightness of transference plan)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. Let $\mathcal{K} \subset \mathcal{P}(\mathcal{X})$ and $\mathcal{H} \subset \mathcal{P}(\mathcal{Y})$ be two tight subsets. Then the set $\Pi(\mathcal{K}, \mathcal{H})$ of all transference plans whose marginals lie in $\mathcal{K}$ and $\mathcal{H}$ respectively, is itself tight in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$.*

Proof of Lemma 1.16. Let $\mu \in \mathcal{K}$, $\nu \in \mathcal{H}$ and $\pi \in \Pi(\mu, \nu)$. By assumption, for any $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset \mathcal{X}$, independent of the choice of μ in $\mathcal{K}$, such that $\mu(\mathcal{X} \setminus K_\varepsilon) \leq \varepsilon$. Similarly there is a compact set $H_\varepsilon \subset \mathcal{Y}$, independent of the choice of ν in $\mathcal{H}$, such that $\nu(\mathcal{Y} \setminus H_\varepsilon) \leq \varepsilon$. Then for any coupling (X, Y) of (μ, ν),

$$\mathbb{P}((X, Y) \notin K_\varepsilon \times H_\varepsilon) \leq \mathbb{P}(X \notin K_\varepsilon) + \mathbb{P}(Y \notin H_\varepsilon) \leq 2\varepsilon.$$

The result follows since this bound is independent of the choice of the coupling and $K_\varepsilon \times H_\varepsilon$ is compact in $\mathcal{X} \times \mathcal{Y}$.

□

Proof of Theorem 1.12. Since $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces, $\{\mu\}$ is tight in $\mathcal{P}(\mathcal{X})$, similarly $\{\nu\}$ is tight in $\mathcal{P}(\mathcal{Y})$. By Lemma 1.16, $\Pi(\mu, \nu)$ is tight in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and by Prokhorov's Theorem, this set has a compact closure. Let $\{\pi_n\}_{n\in\mathbb{N}}$ be a sequence in $\Pi(\mu, \nu)$; extracting a subsequence if necessary, always denoted by $\{\pi_n\}$, we know it converges weakly to some $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Therefore even $(p_1)_\#(\pi_n)$ converges weakly, to some measure on $\mathcal{X}$, but the marginals are fixed; we can apply the same reasoning with $(p_2)_\#$. That proves that even the marginals of π are μ and ν, so the limit measure lies in $\Pi(\mu, \nu)$. Thus $\Pi(\mu, \nu)$ is closed. Moreover it is compact by sequences, so it is in fact compact.
Now let $\{\pi_k\}_{k\in\mathbb{N}}$ be a sequence of probability measures on $\mathcal{X} \times \mathcal{Y}$ such that $\int c \, \mathrm{d}\pi_k$ converges to the infimum transport cost. Extracting a subsequence if necessary, we may assume that $\{\pi_k\}$ converges weakly to some $\pi \in \Pi(\mu, \nu)$. Consider the function $h : (x, y) \mapsto a(x) + b(y)$; we notice that h lies in $L^1(\pi_k)$ and in $L^1(\pi)$, moreover $\int h \, \mathrm{d}\pi_k = \int h \, \mathrm{d}\pi = \int a \, \mathrm{d}\mu + \int b \, \mathrm{d}\nu$. By assumption $c \geq h$, so Lemma 1.14 implies

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\pi \leq \liminf_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\pi_k.$$

Thus π is minimizing, concluding the proof.

□

Another important result is the following:

Theorem 1.17. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces and let $c \in C(\mathcal{X} \times \mathcal{Y})$ be a real-valued continuous cost function such that $\inf c > -\infty$. Let $(c_k)_{k\in\mathbb{N}}$ be a sequence of continuous cost function converging uniformly to c on $\mathcal{X} \times \mathcal{Y}$. Let $\{\mu_k\}_{k\in\mathbb{N}}$ and $\{\nu_k\}_{k\in\mathbb{N}}$ be two sequences of probability measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Assume that μ_k converges weakly to μ and ν_k converges weakly to ν, then for each k, let π_k be an optimal transference plan between μ_k and ν_k. If for each $k \in \mathbb{N}$,*

$$\int_{\mathcal{X}\times\mathcal{Y}} c_k(x, y) \, \mathrm{d}\pi_k(x, y) < +\infty,$$

then, up to an extraction of a subsequence of π_k, still denoted the same for simplicity, it converges weakly to some c-cyclically monotone transference plan $\pi \in \Pi(\mu,\nu)$. *If moreover*

$$\liminf_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c_k(x,y)\ \mathrm{d}\pi_k(x,y) < +\infty,$$

then the optimal transport cost $C(\mu,\nu)$ is finite and π is optimal.

Definition 1.18 (Cyclical monotonicity)**.** *Let $\mathcal{X}$, $\mathcal{Y}$ be to arbitrary sets and $c : \mathcal{X}\times\mathcal{Y} \longrightarrow (-\infty,\infty]$ be a function. A subset $\Gamma \subset \mathcal{X}\times\mathcal{Y}$ is said to be c-cyclically monotone if, for any $N \in \mathbb{N}$ and any family $(x_1,y_1),\cdots,(x_N,y_N)$ of points in Γ, holds the inequality*

$$\sum_{i=1}^{N} c(x_i,y_i) \le \sum_{i=1}^{N} c(x_i,y_{i+1}), \qquad y_{N+1} = y_1. \tag{1.1}$$

A transference plan is said to be c-cyclically monotone if it is concentrated on a c-cyclically monotone set.

Informally, a c-cyclically monotone plan is a plan that cannot be improved, it is impossible to perturb it and get something more "economical".

Remark 1.19 (*Kantorovich duality theorem*; Villani - Optimal Transport, pag. 58). Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $c : \mathcal{X}\times\mathcal{Y} \longrightarrow \mathbb{R}\cup\{+\infty\}$ a continuous cost function such that $c(x,y) \ge a(x) + b(y)$ for some continuous function $a \in L^1(\mathcal{X},\mu)$ and $b \in L^1(\mathcal{Y},\nu)$. Then if $C(\mu,\nu) = \int c\ \mathrm{d}\pi < +\infty$, there is a measurable c-cyclically monotone closed set $\Gamma \in \mathcal{X}\times\mathcal{Y}$ such that for any $\pi \in \Pi(\mu,\nu)$, the following two statements are equivalent:

a) $\pi \in \Pi(\mu,\nu)$ is optimal;

b) π is c-cyclically monotone.

Define then

$$\mathcal{C}(N) = \{(x_1,y_1),\cdots,(x_N,y_N) : \text{holds (1.1)}\}.$$

Proof of Theorem 1.17. Since μ_k and ν_k are convergent sequences, by Prokhorov's theorem, they constitute tight subsets, and by Lemma 1.16, $\Pi(\{\mu_k\}_k,\{\nu_k\}_k)$ is a tight set of $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$. Therefore, let π_k be an optimal transference plan between μ_k and ν_k. Up to extracting a subsequence, still denoted by π_k, we know that π_k converges weakly to some $\pi \in \Pi(\mu,\nu)$. The idea now is to prove that π is c-cyclically monotone. Since each π_k is optimal, it is concentrated on a c_k-cyclically monotone set, so $\pi_k^{\otimes N}$ is concentrated on the set

$$\mathcal{C}_k(N) = \{(x_1,y_1),\cdots,(x_N,y_N) : \text{holds (1.1) for } c_k\}.$$

So if $\varepsilon > 0$ and N are given, for k large enough $\pi_k^{\otimes N}$ is concentrated on the set $\mathcal{C}_\varepsilon(N)$ defined by

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}) + \varepsilon.$$

Since this is a closed set, the same is true for $\pi^{\otimes N}$, and then by letting $\varepsilon \searrow 0$, we see that $\pi^{\otimes N}$ is concentrated on $\mathcal{C}(N)$, so the support of π is c-cyclically monotone, as desired. By the same argument of Theorem 1.12

$$\int c \, \mathrm{d}\pi \leq \liminf_{k\to\infty} \int c_k \, \mathrm{d}\pi_k < +\infty.$$

In particular $C(\mu, \nu) < +\infty$, then Remark 1.19 guarantees the optimality of π.

□

1.3 Other important results

In this section we see few more results that are quite interesting by themselves but have useful applications as well. We first see Urysohn's Lemma; it is necessary that the space $\mathcal{X}$ is a $T4$ topological space: for all $A, B \in \mathcal{X}$ disjoint closed sets, there exist $U, V \in \mathcal{X}$ disjoint open sets such that $A \subset U$ and $B \subset V$. Since our focus is on Polish spaces or metric spaces in general, they are indeed $T4$.

Lemma 1.20 (Urysohn's Lemma)**.** *Let $\mathcal{X}$ be a $T4$ topological space and let $A, B \in \mathcal{X}$ disjoint closed sets. Then there is a continuous function $f : \mathcal{X} \longrightarrow [0, 1]$ such that*

- $f(x) = 0$ *for all* $x \in A$;
- $f(x) = 1$ *for all* $x \in B$;
- $f(x) \in (0, 1)$ *for all* $x \in \mathcal{X} \setminus (A \cup B)$.

Proof. Let start by having in mind that $\mathcal{D} := \mathbb{Q} \cap [0, 1]$ is dense in $[0, 1]$. Thus we can write $\mathcal{D} = \{q_j\}_{j\in\mathbb{N}}$, $q_0 = 0$ and taken in ascending order. For all $q \in \mathcal{D}$, we can construct an open set $U_q \subset \mathcal{X}$ so that:

- $A \subset U_0$;
- $\overline{U_q} \subset U_r$ for all $q < r$;
- $B \cap \bigcup_{q<1} U_q = \emptyset$.

For all $r > q$ we use the property $T4$ of $\mathcal{X}$ by separating the closed set $\overline{U_q}$ with B. We now achieved an increasing sequence of open sets $\{U_q : q \in \mathcal{D}\}$ such that

$$A \subset U_0 \subset \bigcup_{0 \le q \le 1} U_q \subseteq \mathcal{X} \setminus B.$$

We are ready to define f: Let $f(x) = \inf\{q \in \mathcal{D} : x \in U_q\}$ and let $f(x) = 1$ if $x \notin U_q$ for any $q \in \mathcal{D}$, which is equivalent to define $f(x) = 1$ for all $x \in B$. Last thing to check is that f is continuous. It is enough to prove that the pre-image of an open set in $[0, 1]$ is open. Notice that if $0 \le a < b \le 1$, $(a, b) = [0, b) \cap (a, 1]$.

$$\begin{aligned} f^{-1}([0, b)) &= \{x \in \mathcal{X} : f(x) < b\} = \bigcup_{q \in \mathcal{D},\ q < b} U_q; \\ f^{-1}((a, 1]) &= \{x \in \mathcal{X} : f(x) > a\} \\ &= \mathcal{X} \setminus \{x \in \mathcal{D} : f(x) \le a\} = \mathcal{X} \setminus \overline{\bigcup_{q \in \mathcal{D},\ q \le a} U_q}. \end{aligned}$$

They are both open sets in $\mathcal{X}$, therefore f is continuous.

□

Using this Lemma, it is possible to show a very important theorem, which gives us a much way easier method to prove if a sequence of probability measures converges weakly to some $\mu \in \mathcal{P}(\mathcal{X})$:

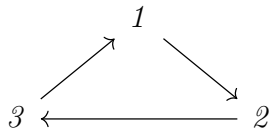
Theorem 1.21 (Portmanteau Theorem)**.** *Let $(\mathcal{X}, d)$ be a separable metric space, then the following statements are equivalent:*

1) $\mu_k \xrightarrow{\text{w}} \mu$ *;*

2) $\limsup_{k \to \infty} \mu_k(F) \le \mu(F)$*, for all $F \in \mathcal{X}$ closed;*

3) $\liminf_{k \to \infty} \mu_k(U) \ge \mu(U)$*, for all $U \in \mathcal{X}$ open.*

Proof.



First step: 1) ⇒ 2). Let $F \in \mathcal{X}$ closed, $B^{1/n} = \{y \in \mathcal{X} : \exists x \in F \text{ s.t. } d(x, y) < 1/n\}$ and $G = \mathcal{X} \setminus B^{1/n}$. Notice that F and G are disjoint closed sets, so, by Urysohn's Lemma 1.20, there is a continuous function $f : \mathcal{X} \longrightarrow [0, 1]$ such that $f(F) = 1$ and $f(G) = 0$. Of course we have

$$1_F(x) \le f(x) \le 1_{B^{1/n}}(x) \qquad \text{for all } x \in \mathcal{X}.$$

Then

$$\mu_k(F) = \int_{\mathcal{X}} 1_F(x)\ \mathrm{d}\mu_k(x) \leq \int_{\mathcal{X}} f(x)\ \mathrm{d}\mu_k(x).$$

Applying the supremum limit and knowing that $f \in bC(\mathcal{X})$ we get

$$\begin{aligned}\limsup_{k\to\infty} \mu_k(F) &\leq \int_{\mathcal{X}} f(x)\ \mathrm{d}\mu(x) \\ &\leq \int_{\mathcal{X}} 1_{B^{1/n}}\ \mathrm{d}\mu(x) \\ &= \mu(B^{1/n}),\end{aligned}$$

On the other hand, $B^{1/n} \searrow F$, in fact

$$\bigcap_{n>0} B^{1/n} = F,$$

therefore

$$\lim_{n\to\infty} \mu(B^{1/n}) = \mu(F).$$

At this point we proved that for all $n \in \mathbb{N}$, $\limsup_{k\to\infty} \mu_k(F) \leq \mu(B^{1/n})$, then

$$\limsup_{k\to\infty} \mu_k(F) \leq \lim_{n\to\infty} \mu(B^{1/n}) = \mu(F).$$

Second step: 2) ⇒ 3). Passing to the complementary, we get

$$\limsup_{k\to\infty} [1 - \mu_k(\mathcal{X} \setminus F)] = 1 - \liminf_{k\to\infty} \mu_k(\mathcal{X} \setminus F) \leq 1 - \mu(\mathcal{X} \setminus F).$$

Notice that $\mathcal{X} \setminus F$ is an open set. Therefore for any open set $U \in \mathcal{X}$ holds

$$\mu(U) \leq \liminf_{k\to\infty} \mu_k(U).$$

Third step: 3) ⇒ 1). Let $f \in bC(\mathcal{X})$ and $D = \|f\|_\infty < +\infty$. Without loss of generality one may suppose $0 \leq f \leq 1$, otherwise let $\tilde{f} = \frac{D+f}{2D}$. Let $\varepsilon > 0$; we can approximate f from below to a function $g = \sum_{i=1}^m a_i 1_{A_i}$, in which each A_i is an open set, $0 \leq a_i \leq 1$ and $0 \leq g \leq f$, such that

$$\int_{\mathcal{X}} f\ \mathrm{d}\mu - \int_{\mathcal{X}} g\ \mathrm{d}\mu \leq \varepsilon.$$

Then

$$\begin{aligned}\int_{\mathcal{X}} f\ \mathrm{d}\mu \leq \int_{\mathcal{X}} g\ \mathrm{d}\mu + \varepsilon &= \sum_{i=1}^m a_i \mu(A_i) + \varepsilon \\ &\leq \liminf_{k\to\infty} \sum_{i=1}^m a_i \mu_k(A_i) + \varepsilon = \liminf_{k\to\infty} \int_{\mathcal{X}} g\ \mathrm{d}\mu_k + \varepsilon \\ &\leq \liminf_{k\to\infty} \int_{\mathcal{X}} f\ \mathrm{d}\mu_k + \varepsilon.\end{aligned}$$

Then letting $\varepsilon \searrow 0$ we have $\int_{\mathcal{X}} f \, \mathrm{d}\mu \leq \liminf_{k\to\infty} \int_{\mathcal{X}} f \, \mathrm{d}\mu_k$. Now, reasoning with $-f$, we will get

$$\begin{aligned} -\int_{\mathcal{X}} f \, \mathrm{d}\mu &= \int_{\mathcal{X}} -f \, \mathrm{d}\mu \\ &\leq \liminf_{k\to\infty} \int_{\mathcal{X}} -f \, \mathrm{d}\mu_k = \liminf_{k\to\infty} \left(-\int_{\mathcal{X}} f \, \mathrm{d}\mu_k \right) \\ &= -\limsup_{k\to\infty} \int_{\mathcal{X}} f \, \mathrm{d}\mu_k. \end{aligned}$$

Therefore

$$\int_{\mathcal{X}} f \, \mathrm{d}\mu \leq \liminf_{k\to\infty} \int_{\mathcal{X}} f \, \mathrm{d}\mu_k \leq \limsup_{k\to\infty} \int_{\mathcal{X}} f \, \mathrm{d}\mu_k \leq \int_{\mathcal{X}} f \, \mathrm{d}\mu.$$

Thus it is actually an equality and

$$\int_{\mathcal{X}} f \, \mathrm{d}\mu_k \longrightarrow \int_{\mathcal{X}} f \, \mathrm{d}\mu.$$

□

Remark 1.22. We used the previous theorem in the proof after Definition 3.8, to prove that *i* implies *iii*, but why is it legit? If X_k is a random variable on a probability space $(\Omega, \mathcal{G}, \mathbb{P})$ with law μ_k converging weakly to a random variable X with law μ, then

$$\begin{aligned} \limsup_{k\to\infty} \int_{\mathcal{X}} 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu_k(x) &= \limsup_{k\to\infty} \mathbb{P}(d(x_0, X_k) \geq R) \\ &= \limsup_{k\to\infty} \mathbb{P}(X_k \in H) \leq \mathbb{P}(X \in H) \\ &= \int_{\mathcal{X}} 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu(x). \end{aligned}$$

Where $H := \{x \in \mathcal{X} : d(x_0, x) \geq R\} \subset \mathcal{X}$ is a closed set.

To prove that the Wasserstein metric, in Chapter 3, satisfies the triangle inequality we used the Minkovsky inequality, in addition to the Gluing Lemma 1.7.
Let $(\mathcal{X}, \mathcal{G}, \lambda)$ be a measurable space and let $f, g : \mathcal{X} \longrightarrow \overline{\mathbb{R}}$ two measurable functions such that

$$f, g \in L^p(\mathcal{X}) := \left\{ h : \mathcal{X} \longrightarrow \overline{\mathbb{R}} \text{ measurable} : \int_{\mathcal{X}} |h|^p \, \mathrm{d}\lambda < +\infty \right\} / \sim, \qquad p \geq 1.$$

where the equivalence relationship is given by:
$f \sim g$ if and only if the set $A = \{x \in \mathcal{X} : f(x) \neq g(x)\}$ is λ-negligible.
Define the L^p-norm as follows

$$\|h\|_{L^p} := \left(\int_{\mathcal{X}} |h|^p \, \mathrm{d}\lambda \right)^{1/p}.$$

Thanks to Holder's inequality it is easy to show that $\|\cdot\|_{L^p}$ satisfies the triangle inequality:

$$\begin{aligned}\|f+g\|_{L^p}^p &= \int_E |f+g|^p \,\mathrm{d}\lambda \\ &\leq \int_E |f+g|^{p-1}|f| \,\mathrm{d}\lambda + \int_E |f+g|^{p-1}|g| \,\mathrm{d}\lambda \\ &\leq \left(\int_E |f+g|^p \,\mathrm{d}\lambda\right)^{\frac{p-1}{p}} \|f\|_{L^p} + \left(\int_E |f+g|^p \,\mathrm{d}\lambda\right)^{\frac{p-1}{p}} \|g\|_{L^p} \\ &= \|f+g\|_{L^p}^{p-1} \|f\|_{L^p} + \|f+g\|_{L^p}^{p-1} \|g\|_{L^p},\end{aligned}$$

Concluding once we divide both sides by $\|f+g\|_{L^p}^{p-1}$, supposing it is not zero, otherwise it would be obvious.

Another important thing that deserves some attention is Corollary 3.10; it will be proved that W_p is continuous even though it will be demonstrated by using the definition of sequentially continuity. Those two definitions are not necessarily equivalent in general, however, if $(\mathcal{X}, d)$ is a metric space, they are.

Proposition 1.23. *Let $f : (\mathcal{X}, d) \longrightarrow \mathbb{R}$ and $x_0 \in (\mathcal{X}, d)$, then f is continuous in x_0 if and only if it is sequentially continuous: for any sequence $(x_n)_n$ such that $x_n \to x_0$, then $f(x_n) \to f(x_0)$.*

Proof. First step. Suppose f is continuous in x_0. Then for any $\varepsilon > 0$, there is a $\delta > 0$ such that $f(B_\delta(x_0)) \subset B_\varepsilon(f(x_0))$. Since $x_n \to x_0$, there exists an $N \in \mathbb{N}$ such that $x_n \in B_\delta(x_0)$ for all $n \geq N$. Therefore $f(x_n) \in B_\varepsilon(f(x_0))$ and letting $\varepsilon \searrow 0$, we conlude.

Second step. Suppose that f is not continuous in x_0, then there exists an open neighborhood $V \ni f(x_0)$ such that for all $U \ni x_0$, $f(U) \not\subset V$. In particular for any $\delta > 0$ there in an $N \in \mathbb{N}$ such that $x_n \in B_\delta(x_0)$ for all $n \geq N$. Therefore $f(x_n) \notin V$ and $f(x_n) \not\to f(x_0)$, which is a contradiction.

□

Chapter 2

Distances between measures

In this chapter we are going to introduce three probability metrics. Total Variation distance has an interesting coupling characterization and is important because it is strictly correlated to the L^1-norm of measures. Levy-Prokhorov distance is theoretically important because it gives a strong topological structure to $\mathcal{P}(\mathcal{X})$ if $\mathcal{X}$ is a complete and separable metric space. The Kantorovich-Rubenstein distance has a rich duality, that will be be proved in the next chapter.

2.1 Total Variation distance

Let $(\mathcal{X}, \mathcal{F})$ be any measurable space. Total Variation distance between two probability measures μ and ν is defined as follows:

$$d_{TV}(\mu, \nu) := \sup_{A \subset \mathcal{X}} |\mu(A) - \nu(A)| \tag{2.1}$$

where A is a measurable subset of $\mathcal{X}$.

Proposition 2.1. *Total variation distance is half the L^1-norm.*

Proof. By definition, $\|\mu - \nu\|_{TV} = |\mu - \nu|(\mathcal{X})$. Let $\sigma = \mu - \nu$, then by Hahn-Jordan decomposition, since σ is finite, there exist two positive finite measure such that $\sigma = \sigma_+ - \sigma_-$. Notice that $\sigma(\mathcal{X}) = 0$, that implies that $\sigma_+(\mathcal{X}) = \sigma_-(\mathcal{X})$, therefore

$$|\sigma|(\mathcal{X}) = \sigma_+(\mathcal{X}) + \sigma_-(\mathcal{X}) = 2\sigma_+(\mathcal{X}).$$

Then by definition

$$\sup_{A \in \mathcal{X}} |\sigma(A)| \triangleq \sigma_+(\mathcal{X}).$$

That concludes the proof, in fact

$$\begin{aligned} d_{TV}(\mu,\nu) &= \sup_{A\in\mathcal{X}} |\mu(A)-\nu(A)| = \sup_{A\in\mathcal{X}} |\sigma(A)| \\ &= \sigma_+(\mathcal{X}) = \frac{1}{2}\left[2\sigma_+(\mathcal{X})\right] = \frac{1}{2}\left|\sigma\right|(\mathcal{X}) \\ &= \frac{1}{2}\left|\mu-\nu\right|(\mathcal{X}) = \frac{1}{2}\|\mu-\nu\|_{TV}. \end{aligned}$$

□

For a countable state space $\mathcal{X}$, the definition above becomes

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sum_{x\in\mathcal{X}} \left|\mu(x)-\nu(x)\right|,$$

which is of course half the L^1-norm between the two measures, as we just proved. In that case we can explicitly find the one subset of $\mathcal{X}$ that verifies $\diamondsuit$.

Proof. We see $\mathcal{X}$ as a set of indices, therefore we are going to write $\mu(i) := \mu(\{x_i\})$. Our target is the biggest set $I \subset \mathcal{X}$ such that $\mu(i) > \nu(i)$ for all $i \in I$. It is quite simple to understand why: if we remove one of those indices or add one in I^c, in the first case we left behind a positive number to add, in the second case we are adding a non-positive number. We then notice that

$$\begin{aligned} \sum_{i\in I} \mu(i)-\nu(i) &= \sum_{i\in\mathcal{X}} \mu(i)-\nu(i) - \sum_{i\in I^c} \mu(i)-\nu(i) \\ &= \sum_{i\in I^c} \nu(i)-\mu(i) \end{aligned}$$

Therefore

$$\begin{aligned} d_{TV}(\mu,\nu) &= \sum_{i\in I} \mu(i)-\nu(i) \\ &= \frac{1}{2}\left(\sum_{i\in I} \mu(i)-\nu(i) + \sum_{i\in I^c} \nu(i)-\mu(i)\right) \\ &= \frac{1}{2}\sum_{i\in\mathcal{X}} \left|\mu(i)-\nu(i)\right|. \end{aligned}$$

That means that $(\mu-\nu)_+(\,\cdot\,) = (\mu-\nu)(\,\cdot\cap I)$ and $(\mu-\nu)_-(\,\cdot\,) = -(\mu-\nu)(\,\cdot\cap I^c)$.

□

In all cases it assumes values in $[0,1]$.

We now give its coupling characterization:

$$\begin{aligned} d_{TV}(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} 1_{\{x\neq y\}}\, \mathrm{d}\pi(x,y) \\ &= \inf\left\{\mathbb{P}(X\neq Y)\ :\ \mathcal{L}(X)=\mu,\ \mathcal{L}(Y)=\nu\right\}. \end{aligned} \tag{2.2}$$

It is now important to understand why (2.2) is equivalent to (2.1). In general, let (X, Y) be a coupling of μ, ν on a probability space $(\Omega, \mathcal{G}, \mathbb{P})$. For all $A \in \mathcal{X}$ measurable,

$$\begin{aligned}
|\mu(A) - \nu(A)| &= |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\
&= |\mathbb{P}(X \in A \cap X \neq Y) + \mathbb{P}(X \in A \cap X = Y) + \\
&\quad - \mathbb{P}(Y \in A \cap X \neq Y) - \mathbb{P}(Y \in A \cap X = Y)| \\
&= |\mathbb{P}(X \in A \cap X \neq Y) - \mathbb{P}(Y \in A \cap X \neq Y)| \\
&\leq \max\{\mathbb{P}(X \in A \cap X \neq Y), \mathbb{P}(Y \in A \cap X \neq Y)\} \\
&\leq \mathbb{P}(X \neq Y).
\end{aligned}$$

Since the result does not depend on A or the coupling chosen, applying the supremum over all measurable subset of $\mathcal{X}$ and taking the infimum over all couplings of μ and ν, the inequality remains true. The idea now is to find a coupling that gives the equality.

Let $\Delta = \{(x, x) : x \in \mathcal{X}\}$ be the diagonal of $\mathcal{X} \times \mathcal{X}$. Let $\psi : (\mathcal{X}, \mathcal{F}) \longrightarrow (\mathcal{X} \times \mathcal{X}, \mathcal{F} \times \mathcal{F})$ defined as $x \longmapsto (x, x)$. We observe it is a measurable function, in fact, given $A, B \in \mathcal{F}$,

$$\begin{aligned}
\psi^{-1}(A \times B) &= \psi^{-1}(A \times B \cap \Delta) \\
&= \psi^{-1}(\{(x, y) \in A \times B : x = y\}) \\
&= A \cap B \in \mathcal{F}.
\end{aligned}$$

Define $\lambda := \mu + \nu$; since $\mu, \nu \leq \lambda$, both μ and ν are absolutely continuous with respect to λ, so we can define

$$g = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \qquad g' = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}.$$

Now define η on $(\mathcal{X}, \mathcal{F})$ as

$$\frac{\mathrm{d}\eta}{\mathrm{d}\lambda} = g \wedge g'$$

and the push-forward measure $\psi_{\#}\eta = \eta \circ \psi^{-1}$. Notice they are both sub-probability measures, otherwise we would conclude that $g = g'$ λ-almost surely and $\mu = \nu$; thus it would be obvious that both the definitions of the Total Variation distance are zero. Then $\psi_{\#}\eta$ gives all its mass on Δ, let call $\gamma = \psi_{\#}\eta(\Delta) = \eta(\mathcal{X})$. Let

$$\zeta = \mu - \eta, \qquad \zeta' = \nu - \eta, \qquad \mathbb{P} = \frac{\zeta \otimes \zeta'}{1 - \gamma} + \psi_{\#}\eta.$$

Then

$$\begin{aligned}
\mathbb{P}(A \times \mathcal{X}) &= \frac{\zeta(A) \otimes \zeta'(\mathcal{X})}{1 - \gamma} + \psi_{\#}\eta(A \times \mathcal{X}) \\
&= \frac{(\mu(A) - \eta(A))\ (1 - \gamma)}{1 - \gamma} + \eta(A) \\
&= \mu(A).
\end{aligned}$$

With the same reasoning, we get $\mathbb{P}(\mathcal{X} \times A) = \nu(A)$, therefore $\mathbb{P}$ is a coupling. We remind that $d_{TV}(\mu, \nu) = \frac{1}{2}\|\mu - \nu\|_{L^1}$ and if μ and ν are absolutely continuous with respect to a common measure λ, then $\|\mu - \nu\|_{L^1} = \int_{\mathcal{X}} |g - g'| \, \mathrm{d}\lambda$, where g and g' are the density functions. Then, notice that $|g - g'| = g + g' - 2(g \wedge g')$; so we get

$$\begin{aligned} d_{TV}(\mu, \nu) = \frac{1}{2}\|\mu - \nu\|_{L^1} &= \frac{1}{2}\int_{\mathcal{X}} |g - g'| \, \mathrm{d}\lambda \\ &= 1 - \int_{\mathcal{X}} (g \wedge g') \mathrm{d}\lambda \overset{\spadesuit}{=} 1 - \eta(\mathcal{X}) \\ &= 1 - \gamma \overset{\diamondsuit}{=} \mathbb{P}(\Delta^c) = \mathbb{P}(X \neq Y), \end{aligned}$$

where X, Y are random variables with respective laws μ, ν. The equality signed by $\spadesuit$ is by definition of η and the one signed by $\diamondsuit$ is because of $\psi_{\#}\eta(\Delta^c) = 0$, and $(\zeta \otimes \zeta')(\Delta^c) = \zeta(\mathcal{X})\ \zeta'(\mathcal{X}) = (1 - \gamma)^2$.

2.2 Levy-Prokhorov metric

Let $(\mathcal{X}, d)$ be any metric space;

$$d_{LP}(\mu, \nu) := \inf \{\varepsilon > 0 \ : \ \mu(B) \leq \nu(B^\varepsilon) + \varepsilon, \text{ for all Borel sets B}\} \tag{2.3}$$

where $B^\varepsilon = \{x \in \mathcal{X} \ : \ \exists \, y \in B \text{ such that } d(x, y) < \varepsilon\}$. It assumes value in $[0, 1]$.
It is easy to show that this metric is symmetric in μ and ν, in fact

$$\mu(B) - \nu(B^\varepsilon) \leq \varepsilon \qquad \Longrightarrow \qquad \nu((B^\varepsilon)^c) - \mu(B^c) \leq \varepsilon.$$

Therefore, letting $C := (B^\varepsilon)^c$, it is quite simple to understand that $B^c = C^\varepsilon$.
It also satisfies the triangle inequality. Let $\epsilon_1 = d_{LP}(\mu, \eta)$ and $\epsilon_2 = d_{LP}(\eta, \nu)$. Then

$$\begin{aligned} \mu(A) - \nu(A^{\epsilon_1 + \epsilon_2}) &= \mu(A) - \eta(A^{\epsilon_1}) \\ &\quad + \eta(A^{\epsilon_1}) - \nu((A^{\epsilon_1})^{\epsilon_2}) \\ &\leq \epsilon_1 + \epsilon_2. \end{aligned}$$

Since $\epsilon_1 + \epsilon_2$ verifies the condition of Levy-Prokhorov between μ and ν, it follows that $d_{LP}(\mu, \nu) \leq \epsilon_1 + \epsilon_2 = d_{LP}(\mu, \eta) + d_{LP}(\eta, \nu)$.
While not easy to compute, this metric is theoretically important because it metrizes weak convergence on any separable metric space; in other terms, $d_{LP}(\mu_n, \mu) \to 0$ as n approaches infinity, implies that the sequence of the $\{\mu_n\}_n$ converges weakly to μ, and viceversa. Moreover, it is precisely the minimum distance *in probability* between two random variables with respect distribution fixed.

Proposition 2.2 (Levy-Prokhorov distance is bounded by Total variation distance)**.** *For any metric space we have the following bound:*

$$d_{LP} \leq d_{TV}. \tag{2.4}$$

Proof. Let $\epsilon := d_{TV}(\mu, \nu)$, then for any Borel set B, we have $\mu(B) - \nu(B) \leq \epsilon$. Since we know, by definition, that B^ϵ contains B, $\nu(B) \leq \nu(B^\epsilon)$, therefore $\mu(B) - \nu(B^\epsilon) \leq \epsilon$. In other words, the chosen ϵ verify the condition of Levy-Prokhorov's metric, so it's obvious that $\mathrm{d}_{LP}(\mu, \nu) \leq \epsilon$.

□

Theorem 2.3. *If $(\mathcal{X}, d)$ is a separable metric space, then Levy-Prokhorov distance metrize weak convergence. In particular, convergence under the Levy-Prokhorov distance generates the same topology induced by convergence against bounded continuous test functions, i.e. the weak topology.*

Proof. First step. $d_{LP}(\mu_n, \mu) \longrightarrow 0 \implies \mu_n \xrightarrow{\mathrm{w}} \mu$.
By definition, since $\mathcal{X}$ is separable, for all $k \in \mathbb{N}$, there is an $N \in \mathbb{N}$ such that for all $n \geq N$, $\mu_n(A) - \mu(A^{1/k}) \leq 1/k$ and $\mu(A) - \mu_n(A^{1/k}) \leq 1/k$, for all $A \in \mathcal{B}$. The idea is to prove that the condition for closed sets of Portmanteau Theorem 1.21 holds. Let F be a closed set. Then $F^{1/k} \searrow F$ as $k \to \infty$, in fact, $F = \bigcap_{k \geq 1} F^{1/k}$. Consider $\mu_n(F) \leq \mu(F^{1/k}) + 1/k$; since the right-hand side does not depend on n, the inequality remains true even after applying the supremum limit on the left-hand side. Therefore, for any k, holds

$$\limsup_{n \to \infty} \mu_n(F) \leq \mu(F^{1/k}) + 1/k,$$

concluding the first step once we let $k \to \infty$.

Second step. $\mu_n \xrightarrow{\mathrm{w}} \mu \implies d_{LP}(\mu_n, \mu) \longrightarrow 0$.
We want to prove that for any $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that $d(\mu, \mu_n) \leq \varepsilon$ for all $n \geq N$. For any measurable set $B \in \mathcal{X}$, let $F = \overline{B}$ and $U = B^\varepsilon$. Notice that F is closed, U is open and $F \subset U$, regardless the choice of B. By Portmanteau Theorem 1.21, we know that for any closed set H and open set V

$$\limsup_{n \to \infty} \mu_n(H) \leq \mu(H) \quad \text{and} \quad \liminf_{n \to \infty} \mu_n(V) \geq \mu(V).$$

That means that for any $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that for any $n \geq \mathbb{N}$,

$$\mu_n(H) \leq \mu(H) + \varepsilon \quad \text{and} \quad \mu_n(V) \geq \mu(V) - \varepsilon.$$

In particular, this holds for F and U chosen earlier, therefore

$$\mu_n(B) \leq \mu_n(F) \leq \mu(F) + \varepsilon \leq \mu(B^\varepsilon) + \varepsilon$$
$$\mu(B) \leq \mu(B^\varepsilon) \leq \mu_n(B^\varepsilon) + \varepsilon,$$

which is exactly $d_{LP}(\mu, \mu_n) \leq \varepsilon$, for all $n \geq N$.

□

Lemma 2.4. *Let $\mathcal{X}$ be a Polish space. Then Cauchy sequences in the Levy-Prokhorov distance are tight.*

Proof. Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence, then $d_{LP}(\mu_j, \mu_k) \underset{j,k\to\infty}{\longrightarrow} 0$. Now, let $\varepsilon > 0$; by definition there is an $N \in \mathbb{N}$ such that $d_{LP}(\mu_k, \mu_N) < \varepsilon$ for any $k > N$. Notice that since $\{\mu_1, \cdots, \mu_N\}$ is finite, it is tight, so there exists a compact set $K \in \mathcal{X}$ such that $\mu_j(K) > 1 - \varepsilon,\ j \in \{1, \cdots, N\}$. Then

$$K \subset \bigcup_{i=1}^{m} B_\varepsilon(x_i) \subset K^\varepsilon \subset \bigcup_{i=1}^{m} B_{2\varepsilon}(x_i) \subset \bigcup_{i=1}^{m} \overline{B_{2\varepsilon}(x_i)}.$$

Next, for all $k > N$, we get the following:

$$\begin{aligned} \mu_k(K^\varepsilon) &= \mu_N(K) + [\mu_k(K^\varepsilon) - \mu_N(K)] \\ &\overset{\spadesuit}{>} 1 - \varepsilon - \varepsilon = 1 - 2\varepsilon. \end{aligned}$$

where the inequality signed by ♠ follows from the choice of N and by definition, in fact $\mu_N(K) - \mu_k(K^\varepsilon) \leq \varepsilon$. Therefore

$$\mu_k\left(\bigcup_{i=1}^{m} \overline{B_{2\varepsilon}(x_i)}\right) > 1 - 2\varepsilon, \qquad \text{for all } k \in \mathbb{N}.$$

With the same reasoning, replacing ε with $2^{-(p+1)}\varepsilon$, we get

$$\mu_k\left(\bigcup_{i=1}^{m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)}\right) > 1 - 2^{-p}\varepsilon, \qquad \text{for all } k \in \mathbb{N}.$$

Thus, the set

$$S = \bigcap_{p\geq 1} \bigcup_{i=1}^{m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)},$$

satisfies $\mu_k(\mathcal{X} \setminus S) \leq \varepsilon$ for all k. S is closed as an intersection of finite unions of closed balls. Since it is closed, it is complete. Then for any $\delta > 0$ one could choose p large enough such that $2^{-p}\varepsilon < \delta$, so

$$S \subset \bigcup_{i=1}^{m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)} \subset \bigcup_{i=1}^{m(p)} B_\delta(x_i),$$

and S is totally bounded. Therefore S is compact.

□

At this point, if $\mathcal{X}$ is separable, we are able to say that $\mathcal{P}(\mathcal{X})$, equipped with the topology of weak convergence, is in fact metrizable with respect to Levy-Prokhorov metric. In other words, $(\mathcal{P}(\mathcal{X}), d_{LP})$ is a metric space, in fact d_{LP} is symmetric, satisfies the triangle inequality and $d_{LP}(\mu, \nu) = 0$ implies that $\mu = \nu$.

Theorem 2.5. *Let $\mathcal{X}$ be a Polish space, then $\mathcal{P}(\mathcal{X})$ is itself Polish. In an equivalent way, one could say that $\mathcal{P}(\mathcal{X})$ equipped with d_{LP} is a complete and separable metric space.*

Proof. The fact that $(\mathcal{P}(\mathcal{X}), d_{LP})$ is a metric space was already explained.
Let prove the *separability* first. Let denote $\mathcal{D}$ the dense in $\mathcal{X}$ and let $\mu \in \mathcal{P}(\mathcal{X})$. Since $\{\mu\}$ is tight, for any $\varepsilon > 0$ there exists a compact set K such that $\mu(\mathcal{X} \setminus K) \leq \varepsilon$. The goal is to find a discrete measure $\tilde{\mu}$ that approximate μ with an error at most of ε. By compactness of K and separability of $\mathcal{X}$, there are $\{x_1, \cdots, x_m\} \in \mathcal{D}$ such that

$$K \subseteq \bigcup_{i=1}^{m} B_{\varepsilon/2}(x_i).$$

Then, by taking $D_1 = B_{\varepsilon/2}(x_1) \cap K$ and $D_i = (B_{\varepsilon/2}(x_i) \cap K) \setminus \bigcup_{j<i} B_{\varepsilon/2}(x_j)$, we notice that this is a finite partition and each D_i has diameter smaller then ε. Therefore

$$K = \biguplus_{i=1}^{m} D_i.$$

Let $x_0 \in (\mathcal{X} \setminus K) \cap \mathcal{D}$. Consider now the following measurable function

$$\begin{aligned} f \ &: \mathcal{X} \longrightarrow \mathcal{X} \\ &x \longmapsto x_i \qquad \text{if } \ x \in D_i \\ &x \longmapsto x_0 \qquad \text{if } \ x \in \mathcal{X} \setminus K, \end{aligned}$$

and define $\mu' = f_{\#}\mu$; of course μ' can be written as $\sum_{i=0}^{m} q_i \delta_{x_i}$. Notice that $\mu'(\{x_0\}) = \mu(\mathcal{X} \setminus K) \leq \varepsilon$. For any $A \in \mathcal{B}$, if we take $A^\varepsilon = \{x \in \mathcal{X} \ : \ \exists \ y \in A \text{ such that } d(x, y) < \varepsilon\}$, we notice that

$$\begin{aligned} \mu'(A \cap K) = \mu\left(f^{-1}(A \cap K)\right) &= \mu\left(f^{-1}\left(\bigcup_{x_i \in A \cap K} \{x_i\}\right)\right) \\ &= \mu\left(\bigcup_{x_i \in A \cap K} D_i\right) \leq \mu(A^\varepsilon); \\ \mu(A \cap K) \leq \mu\left(\bigcup_{x_i \in A^\varepsilon \cap K} D_i\right) &= \mu'\left(\bigcup_{x_i \in A^\varepsilon \cap K} \{x_i\}\right) \\ &\leq \mu'\left(\bigcup_{x_i \in A^\varepsilon} \{x_i\}\right) = \mu'(A^\varepsilon). \end{aligned}$$

Therefore

$$\mu(A) = \mu(A \cap K) + \mu(A \setminus K) \leq \mu(A \cap K) + \mu(E \setminus K) \leq \mu'(A^\varepsilon) + \varepsilon;$$
$$\mu'(A) = \mu'(A \cap K) + \mu'(A \setminus K) \leq \mu'(A \cap K) + \mu'(\{x_0\}) \leq \mu(A^\varepsilon) + \varepsilon.$$

Denote now $\tilde{\mu} = \sum_{i=0}^{m} p_i \delta_{x_i}$, with $p_i \in \mathbb{Q} \cap [0,1]$; then by Proposition 2.2,

$$d_{LP}(\tilde{\mu}, \mu') \leq d_{TV}(\tilde{\mu}, \mu') = \frac{1}{2} \sum_{i=0}^{m} |q_i - p_i| < \varepsilon,$$

for some well chosen rational coefficients. Then

$$d_{LP}(\mu, \tilde{\mu}) \leq d_{LP}(\mu, \mu') + d_{LP}(\mu', \tilde{\mu}) \leq 2\varepsilon.$$

Therefore the set of all discrete measures with finite support and rational coefficients is dense in $\mathcal{P}(\mathcal{X})$ because $\mathcal{D}$ is countable, as well as $\mathbb{Q} \cap [0,1]$.
Let prove the *completeness*. Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence. Thanks to the previous Lemma, we know it is tight, therefore, Prokhorov's Theorem guarantees the existence of a sub-sequence $\{\mu_{j'}\}$ converging weakly to some measure μ. Let $\varepsilon > 0$, then there exists an $N \in \mathbb{N}$ such that for all $j', n > N$, we get

$$d_{LP}(\mu_n, \mu) \leq d_{LP}(\mu_n, \mu_{j'}) + d_{LP}(\mu_{j'}, \mu) \leq \varepsilon + \varepsilon = 2\varepsilon.$$

So $\mu_n \xrightarrow{\text{w}} \mu$ and that ends the argument.

□

2.3 Kantorovich-Rubenstein metric

For any complete and separable metric space $(\mathcal{X}, d)$, the Kantorovich-Rubenstein distance is defined as follows:

$$W_1(\mu, \nu) := \sup \left\{ \int_{\mathcal{X}} h \, \mathrm{d}\mu - \int_{\mathcal{X}} h \, \mathrm{d}\nu \ : \ \|h\|_L \leq 1 \right\}, \tag{2.5}$$

where d is the metric of $\mathcal{X}$ and the supremum is taken over all h satisfying the Lipschitz condition $|h(x) - h(y)| \leq d(x,y)$.
It assumes value in $[0, \mathrm{diam}(\mathcal{X})]$, where $\mathrm{diam}(\mathcal{X}) = \sup\{d(x,y) : x, y \in \mathcal{X}\}$. If $\mathcal{X}$ is compact, W_1 metrizes weak convergence; it will be rather easy to prove (see above Theorem 3.9). If $\mathrm{diam}(\mathcal{X})$ is not bounded, then it is not guaranteed that W_1 is always bounded, therefore it might not be a metric in the strict sense. However, this problem can be overcome by reasoning on a sub-space of $\mathcal{P}(\mathcal{X})$, in which will take finite value.

Remark 2.6. The notation chosen for this metric is due to *Kantorovich-Rubenstein theorem* (Proposition 3.5), that affirms that the Kantorovich metric is equal to the *Wasserstein metric of order one*, i.e.

$$W_1(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)\ \mathrm{d}\pi(x,y), \tag{2.6}$$

where the infimum is taken over all joint distributions with fixed marginals μ, ν.
The proof to that is seen in Corollary 3.6.

Chapter 3

Wasserstein Metric

In this chapter, we are going to define several properties of the Wasserstein metric, for instance, his relationship with weak convergence in the Wasserstein space. Last but not list we are going to describe the topological properties that the Wasserstein space inherits from the state space.

We denote by (E, d) a complete and separable metric space, but for abuse of notation we refer to it as a *Polish space*; we implicitly assume that d is the one distance on E making E a complete and separable metric space. The σ-algebra $\mathcal{E}$ equipping E is always assumed to be the Borel σ-algebra $\mathcal{B}(E)$, which does not depend on the choice of the particular compatible distance.
Before we actually define the *Wasserstein metric*, it comes natural to consider first the **optimal transport cost** between two measure

$$C(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{E \times E} c(x, y) \, \mathrm{d}\pi(x, y), \tag{3.1}$$

with respect to a cost function $c(x, y)$. We can think of c as the function representing the cost to transport one unit of mass from point x to y. It is natural to see (3.1) as a kind of distance between μ and ν, but in general it does not satisfy the axioms of distance. However, when the cost function is defined in terms of a distance, we can easily construct a class of metrics.

3.1 Wasserstein Metrics and Wasserstein Spaces

Definition 3.1 (Wasserstein metric)**.** *Let (E, d) be a Polish space, and let $p \in [1, \infty)$. The Wasserstein metric of order p between two measures μ and ν is defined by*

$$W_p(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{E\times E} d(x,y)^p \, \mathrm{d}\pi(x,y)\right)^{\frac{1}{p}} \tag{3.2}$$
$$= \inf\left\{\mathbb{E}\left[d(X,Y)^p\right]^{\frac{1}{p}} \ : \ \mathcal{L}(X)=\mu, \mathcal{L}(Y)=\nu\right\}.$$

Remark 3.2. By putting $p = 1$, we have (2.6).

At the moment, W_p is still defined on the space of all probability measures over E, denoted by $\mathcal{P}(E)$, thus it might take the value $+\infty$. That means it is still not a distance in the strict sense, however it satisfies the axioms of distance.

Proof that W_p satisfies the axioms of a distance. Since $d(x, y)^p$ is non negative, Wasserstein metric is non negative as well, and by Fubini's theorem, you can change the order of integration, that means it's also symmetric in μ and ν. We now prove that satisfies triangle inequality: thanks to Lemma 1.16 we know that $\Pi(\mu, \nu)$ is a nonempty compact set, which means that the infimum in the Definition 3.1 is always attained by at least one coupling. To be more clear, the subset of $\Pi(\mu, \nu)$ satisfying the infimum is denoted by

$$\Pi_p^{\mathrm{opt}}(\mu,\nu) = \left\{\pi\in\Pi(\mu,\nu) \ : \ W_p(\mu,\nu)^p = \int_{E\times E} d(x,y)^p \, \mathrm{d}\pi(x,y)\right\}.$$

Next, let μ_1, μ_2 and μ_3 be three probability measures on E, and let (X_1, X_2) be an optimal coupling of (μ_1, μ_2) and (Z_2, Z_3) be an optimal coupling of (μ_2, μ_3). By the Gluing Lemma 1.7, there exist random variables (Y_1, Y_2, Y_3) with $\mathcal{L}(Y_1, Y_2) = \mathcal{L}(X_1, X_2)$ and $\mathcal{L}(Y_2, Y_3) = \mathcal{L}(Z_2, Z_3)$. In particular (Y_1, Y_3) is a coupling of (μ_1, μ_3), so

$$\begin{aligned} W_p(\mu_1,\mu_3) \le \mathbb{E}\left[d(Y_1,Y_3)^p\right]^{\frac{1}{p}} &\le \mathbb{E}\left[(d(Y_1,Y_2)+d(Y_2,Y_3))^p\right]^{\frac{1}{p}} \\ &\overset{\spadesuit}{\le} \mathbb{E}\left[d(Y_1,Y_2)^p\right]^{\frac{1}{p}} + \mathbb{E}\left[d(Y_2,Y_3)^p\right]^{\frac{1}{p}} \\ &= W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3) \end{aligned}$$

where the last equality follows from (Y_1, Y_2) and (Y_2, Y_3) being optimal couplings and the inequality signed by ♠ follows from Minkovsky inequality. So W_p satisfies the triangle inequality. Finally we assume that $W_p(\mu, \nu) = 0$; then there exist a transference plan which is entirely concentrated on the diagonal $(y = x)$ in $E \times E$, so $\nu = Id_{\#}(\mu) = \mu$.

□

To complete the construction, it is natural to restrict W_p to a subset of $\mathcal{P}(E) \times \mathcal{P}(E)$ on which it takes finite value.

Definition 3.3 (Wasserstein space)**.** *With the same hypothesis of Definition 3.1, the Wasserstein space of order p is defined as*

$$\mathcal{P}_p(E) := \left\{ \mu \in \mathcal{P}(E) \ : \ \int_E d(x_0, x)^p \ \mathrm{d}\mu(x) < \infty \right\},$$

where $x_0 \in E$ is arbitrary. This space does not depend on the chosen point x_0. Then W_p defines a finite measure on $\mathcal{P}_p(E)$.

Proof that the definition above does not depend on x_0. Let $x_0, y_0 \in E$. The following inequalities

$$d(x_0, x)^p \leq 2^{p-1}(d(x_0, y_0)^p + d(y_0, x)^p)$$
$$d(y_0, x)^p \leq 2^{p-1}(d(x_0, y_0)^p + d(x_0, x)^p),$$

tell us that $d(x_0, \ \cdot \ )^p$ is μ-integrable if and only if $d(y_0, \ \cdot \ )^p$ is μ-integrable.

□

Proof that W_p is finite on $\mathcal{P}_p(E)$. Let $\pi \in \Pi(\mu, \nu)$, with $\mu, \nu \in \mathcal{P}_p(E)$. The inequality $d(x, y)^p \leq 2^{p-1}(d(x, x_0)^p + d(x_0, y)^p)$ shows that if $d(\ \cdot\ , x_0)^p$ is μ-integrable and $d(x_0, \ \cdot \ )^p$ is ν-integrable, then $d(x, y)^p$ is $\pi(\mathrm{d}x \ \mathrm{d}y)$-integrable, which means that $W_p(\mu, \nu) < \infty$.

□

Since W_p is finite on $\mathcal{P}_p(E)$ and satisfies the axioms of a distance, it is actually a distance on the Wasserstein space of order p, therefore $(\mathcal{P}_p(E), W_p)$ is a metric space.

Remark 3.4. if $p \leq q$, then $W_p \leq W_q$.

Proof. We recall the Jensen inequality first: if $f : E \to \mathbb{R}$ is a continuous concave function, and Z is a random variable on $(\Omega, \mathcal{G}, \mathbb{P})$ which takes values in E, then

$$\mathbb{E}\left[f(Z)\right] \leq f(\mathbb{E}\left[Z\right]). \tag{3.3}$$

Let us get into the proof. The function $f(x) = x^{p/q}$ is concave if restricted to the non-negative values of x, since $p \leq q$. Let (X, Y) be an optimal coupling for the Wasserstein distance of order q; the existence is guaranteed by Theorem 1.12. Then

$$\begin{aligned} W_p(\mu, \nu)^p &\leq \mathbb{E}[d(X, Y)^p] \\ &= \mathbb{E}[(d(X, Y)^q)^{p/q}] \\ &\overset{(3.3)}{\leq} \mathbb{E}[(d(X, Y)^q)]^{p/q} \\ &= W_q(\mu, \nu)^p. \end{aligned}$$

Last equality is due to the choice of the coupling, concluding the proof.

□

The next result is known as the *Kantorovich duality* theorem, which is central to the theory of optimal transportation.

Proposition 3.5. *Let (E,d) a Polish space, $p \geq 1$ and $\mu,\nu \in \mathcal{P}_p(E)$, then*

$$W_p(\mu,\nu)^p = \sup_{\phi,\psi:\ \phi(x)+\psi(y)\leq d(x,y)^p} \left[\int_E \phi(x)\ \mathrm{d}\mu(x) + \int_E \psi(y)\ \mathrm{d}\nu(y)\right], \tag{3.4}$$

where the supremum is taken over all real valued bounded continuous functions ϕ and ψ on E. Moreover, if $\pi \in \Pi_p^{\mathrm{opt}}(\mu,\nu)$ is an optimal transport plan between μ and ν, then there exists $\phi \in L^1(E,\mu)$ and $\psi \in L^1(E,\nu)$ such that for π-almost every $(x,y) \in E \times E$,

$$\phi(x) + \psi(y) = d(x,y)^p.$$

Proof. First step. Let us denote by $\tilde{W}_p(\mu,\nu)^p$ the right-end side of (3.4). We want to prove that satisfies the triangle inequality in the sense that given μ,ν and η three probability measures on $\mathcal{P}_p(E)$, we have:

$$\tilde{W}_p(\mu,\nu) \leq \tilde{W}_p(\mu,\eta) + \tilde{W}_p(\eta,\nu).$$

The idea is borrowed from classical analysis proof that the norm of L^p spaces satisfies the triangle inequality. Let $(0,\infty)^2 \ni (x,y) \longmapsto c_i(x,y) \in (0,\infty),\ i=1,2$, be deterministic function such that

$$(a+b)^p = \inf_{s,t>0} \left[c_1(s,t)a^p + c_2(s,t)b^p\right]. \tag{3.5}$$

Now let ϕ and ψ be two real valued bounded continuous functions on E satisfying $\phi(x)+\psi(y) \leq d(x,y)^p$ for all $x,y \in E$. Notice that (3.5) implies that for any $s,t>0$ and $x,y,z \in E$ we have:

$$\phi(x)+\psi(y) \leq d(x,y)^p \leq (d(x,z)+d(z,y))^p \leq c_1(s,t)d(x,z)^p + c_2(s,t)d(z,y)^p. \tag{3.6}$$

Now, having s and t fixed, we define for all $z \in E$,

$$\xi(z) = \inf_{x\in E} \left[c_1(s,t)d(x,z)^p - \phi(x)\right].$$

By construction for all $x,z \in E$,

$$\phi(x) + \xi(z) \leq c_1(s,t)d(x,z)^p. \tag{3.7}$$

Moreover, using (3.6), we get

$$\begin{aligned}
\psi(y) - \xi(z) &= \phi(x) + \psi(y) - \xi(z) - \phi(x) \\
&\leq c_1(s,t)d(x,z)^p + c_2(s,t)d(z,y)^p - \xi(z) - \phi(x) \\
&= \left[c_1(s,t)d(x,z)^p - \phi(x)\right] + c_2(s,t)d(z,y)^p - \xi(z).
\end{aligned}$$

Since the left-hand side does not depend upon $x \in E$, the inequality remains true after taking the infimum on the right-hand side. Thus,

$$\begin{aligned}\psi(y) - \xi(z) &\leq \inf_{x\in E} [c_1(s,t)d(x,z)^p - \phi(x)] + c_2(s,t)d(z,y)^p - \xi(z) \\ &= c_2(s,t)d(z,y)^p.\end{aligned} \tag{3.8}$$

Putting together (3.7) and (3.8) we get:

$$\begin{aligned}\int_E \phi(x)\ \mathrm{d}\mu(x) + \int_E \psi(y)\ \mathrm{d}\nu(y) &\leq \left[\int_E \phi(x)\ \mathrm{d}\mu(x) + \int_E \xi(z)\ \mathrm{d}\eta(z)\right] \\ &\quad + \left[\int_E \psi(y)\ \mathrm{d}\nu(y) - \int_E \xi(z)\ \mathrm{d}\eta(z)\ \right] \\ &\leq c_1(s,t)\tilde{W}_p(\mu,\eta)^p + c_2(s,t)\tilde{W}_p(\eta,\nu)^p,\end{aligned}$$

where we used the definition of $\tilde{W}_p$. Taking the supremum on the left-hand side over all function ϕ and ψ satisfying $\phi(x) + \psi(y) \leq d(x,y)^p$, we proved

$$\tilde{W}_p(\mu,\nu)^p \leq c_1(s,t)\tilde{W}_p(\mu,\eta)^p + c_2(s,t)\tilde{W}_p(\eta,\nu)^p.$$

Therefore, applying the infimum on the right-hand side over $s,t>0$ we get

$$\begin{aligned}\tilde{W}_p(\mu,\nu)^p &\leq \inf_{s,t>0}\left[c_1(s,t)\tilde{W}_p(\mu,\eta)^p + c_2(s,t)\tilde{W}_p(\eta,\nu)^p\right] \\ &= \left[\tilde{W}_p(\mu,\eta) + \tilde{W}_p(\eta,\nu)\right]^p\end{aligned}$$

concluding the first step.

Second step. We now prove that (3.4) holds when E is finite. If $E = \{e_1, \cdots, e_n\}$, we use the notation $\mu(i) = \mu(\{e_i\})$ and $\nu(i) = \nu(\{e_i\})$ for $i = 1, \cdots, n$. By definition

$$W_p(\mu,\nu)^p = \inf\left\{\sum_{1\leq i,j\leq n} d(e_i,e_j)^p\pi(i,j) : \pi(i,j)\geq 0,\ \sum_{j=1}^n \pi(i,j) = \mu(i),\ \sum_{i=1}^n \pi(i,j) = \nu(j)\right\}.$$

If we treat the $n\times n$ matrix $(d(e_i,e_j)^p)_{1\leq i,j\leq n}$ as an n^2 vector b, then the Wasserstein metric of order p between μ and ν is given by the value of a *plain linear program.* We consider $b = (b_1, \cdots, b_n)$, in which each b_k is an n vector and $b_k = (d(e_k,e_1)^p, \cdots, d(e_k,e_n)^p)$; with the same strategy we define the n^2 vector $\pi = (\pi(i,j))_{1\leq i,j\leq n}$. Then we take $A = (A_{l,(i,j)})_{1\leq l\leq 2n, 1\leq i,j\leq n}$ to be a $2n\times n^2$ matrix with $A_{l,(i,j)} = 1_{i=l}$, if $l \leq n$ and $A_{l,(i,j)} = 1_{j=l-n}$ if $l > n$. Let c be a $2n$ vector in which $c(l) = \mu(l)$ if $l \leq n$ and $c(l) = \nu(l-n)$ if $l > n$. Therefore we can think of this problem as a *primal* problem:

$$\inf_{\pi(i,j)\geq 0,\ A\pi = c} b\cdot\pi,$$

and then write that its value is given by the value of the corresponding *dual* problem. For finite dimensional linear programming, we recall the classical duality theory that gives the following equality:

$$\sup_{A^T x \leq b} c \cdot x = \inf_{y \geq 0,\ Ay = c} b \cdot y.$$

If we denote $x = (\phi(1), \cdots, \phi(n), \psi(1), \cdots, \psi(n))$, then

$$\begin{aligned} W_p(\mu,\nu)^p &= \inf_{y\geq 0,\ Ay=c} b\cdot y = \sup_{A^T x \leq b} c \cdot x \\ &\overset{\heartsuit}{=} \sup_{\phi,\psi,\ \phi(i)-\psi(j)\leq d(e_i,e_j)^p} \left[\sum_{i=1}^n \phi(i)\mu(i) + \sum_{j=1}^n \psi(j)\mu(j)\right] \\ &= \tilde{W}_p(\mu,\nu)^p, \end{aligned}$$

where the equality signed by $\heartsuit$ is because given the vector $A^T x$, if $k = h + (t-1)n$ where $t = 1, \cdots, n$ and $h = 1, \cdots, n$, then $A^T x(k) = \phi(t) + \psi(h) \leq d(e_t, e_h)^p$.

Third step. Let now prove the inequality $W_p(\mu,\nu)^p \geq \tilde{W}_p(\mu,\nu)^p$ in full generality. If ϕ and ψ are real valued bounded continuous function on E satisfying $\phi(x)+\psi(y) \leq d(x,y)^p$, then for any coupling $\pi \in \Pi(\mu,\nu)$ we have:

$$\begin{aligned} \int_E \phi(x)\ \mathrm{d}\mu(x) + \int_E \psi(y)\ \mathrm{d}\nu(y) &= \int_{E\times E} \phi(x)\ \mathrm{d}\pi(x,y) + \int_{E\times E} \psi(y)\ \mathrm{d}\pi(x,y) \\ &= \int_{E\times E} \phi(x) + \psi(y)\ \mathrm{d}\pi(x,y) \\ &\leq \int_{E\times E} d(x,y)^p\ \mathrm{d}\pi(x,y). \end{aligned}$$

Since the right-hand side does not depend on the coupling chosen, the inequality holds even after applying the infimum over all couplings and we get $W_p(\mu,\nu)^p$, therefore the right-hand side is still an upper bound for the left-hand side. We can now take the supremum on the left-hand side over all the couples (ϕ,ψ), to obtain

$$\tilde{W}_p(\mu,\nu)^p \leq W_p(\mu,\nu)^p.$$

Fourth step. Finally, we prove the remaining equality by an approximation procedure. Let $x_0 \in E$ be fixed. Since $\{\mu,\nu\}$ is a tight subset of $\mathcal{P}_p(E)$, for all $\varepsilon > 0$ there exists a compact set K_ε such that

$$\int_{K_\varepsilon^c} d(x_0,x)^p\,[\mathrm{d}\mu(x) + \mathrm{d}\nu(x)] < \varepsilon^p.$$

Since K_ε is compact, we can construct a finite partition $(D_k)_{1\leq k\leq n}$ and each D_k has diameter at most ε. For each $k \in \{1, \cdots, n\}$, we can pick an element $x_k \in D_k$, and

construct a Borel map

$$\begin{aligned} \psi \ &: E \longrightarrow E \\ &x \longmapsto x_k \qquad \text{if } x \in D_k \\ &x \longmapsto x_0 \qquad \text{if } x \in K_\varepsilon^c. \end{aligned}$$

Clearly ψ is a coupling map between μ and $\tilde{\mu} := \psi_{\#}\mu$ and ν and $\tilde{\nu} := \psi_{\#}\nu$. Then we notice that

$$\begin{aligned} W_p(\mu, \tilde{\mu})^p &= \sum_{k=1}^{n} \int_{D_k} d(x_k, x)^p \, \mathrm{d}\mu(x) \\ &\quad + \int_{K_\varepsilon^c} d(x_0, x)^p \, \mathrm{d}\mu(x) \\ &\leq \varepsilon^p \sum_{k=1}^{n} \mu(D_k) + \varepsilon^p \\ &= \varepsilon^p \ \mu(K_\varepsilon) + \varepsilon^p \\ &\leq 2\varepsilon^p. \end{aligned}$$

With the same reasoning, $W_p(\nu, \tilde{\nu})^p \leq 2\varepsilon^p$. Using the triangle inequality for W_p, we get:

$$W_p(\mu, \nu) \leq W_p(\mu, \tilde{\mu}) + W_p(\tilde{\mu}, \tilde{\nu}) + W_p(\nu, \tilde{\nu}) \leq W_p(\tilde{\mu}, \tilde{\nu}) + 2^{1+1/p}\varepsilon. \tag{3.9}$$

Using the result proven for probability measures on finite spaces in the *second step*, we know that $\tilde{W}_p(\tilde{\mu}, \tilde{\nu}) = W_p(\tilde{\mu}, \tilde{\nu})$. Then, thanks to the triangle inequality proved in the *first step*, we obtain:

$$\begin{aligned} \tilde{W}_p(\tilde{\mu}, \tilde{\nu}) &\leq \tilde{W}_p(\mu, \tilde{\mu}) + \tilde{W}_p(\mu, \nu) + \tilde{W}_p(\nu, \tilde{\nu}) \\ &\leq W_p(\mu, \tilde{\mu}) + \tilde{W}_p(\mu, \nu) + W_p(\nu, \tilde{\nu}) \\ &\leq \tilde{W}_p(\mu, \nu) + 2^{1+1/p}\varepsilon, \end{aligned} \tag{3.10}$$

where in the second inequality we used what we proved in the *third step*. Combining (3.9) with (3.10) we get:

$$W_p(\mu, \nu) \leq \tilde{W}_p(\mu, \nu) + 2^{2+1/p}\varepsilon,$$

concluding the proof once we let $\varepsilon \searrow 0$.

□

Corollary 3.6. *If (E, d) is a Polish space and $\mu, \nu \in \mathcal{P}_1(E)$, then*

$$W_1(\mu, \nu) = \sup_{\phi : |\phi(x) - \phi(y)| \leq d(x,y)} \int_E \phi(x) \ \mathrm{d}(\mu - \nu)(x).$$

Remark 3.7. That proves that the Wasserstein distance of order one and the Kantorovich-Rubenstein distance introduced in Chapter 2 Section 2.3, are equivalent.

Proof. Let ϕ and ψ be two bounded continuous function such that $\phi(x)+\psi(y) \leq d(x,y)$. The previous inequality can be replaced by:

$$\phi(x) = \inf_{y \in E} \left[d(x,y) - \psi(y) \right], \qquad x \in E. \tag{3.11}$$

We immediately observe that ϕ is 1-Lipschitz, in fact

$$\begin{aligned} \phi(x) - \phi(z) &= \inf_{y \in E} \left[d(x,y) - \psi(y) \right] - \inf_{y \in E} \left[d(z,y) - \psi(y) \right] \\ &\leq d(x,\tilde{y}) - \psi(\tilde{y}) - d(z,\tilde{y}) + \psi(\tilde{y}) \\ &= d(x,\tilde{y}) - d(z,\tilde{y}) \\ &\leq d(x,z). \end{aligned}$$

where $\tilde{y}$ is the one $y \in E$ that satisfies the infimum for $\phi(z)$. With the same reasoning we get $\phi(z) - \phi(x) \leq d(x,z)$, therefore

$$|\phi(x) - \phi(z)| \leq d(x,z).$$

So, limiting ourself to functions ϕ that are 1-Lipschitz, the inequality $\phi(x) + \psi(y) \leq d(x,y)$ can be replaced by:

$$\psi(y) = \inf_{x \in E} \left[d(x,y) - \phi(x) \right] = -\phi(y), \qquad y \in E.$$

Therefore, in the Kantorovich duality, it is enough to maximize over pairs of 1-Lipschitz functions $(\phi, -\phi)$, which completes the proof.

□

3.2 Weak Convergence in the Wasserstein Spaces

We first define a characterization of convergence in the Wasserstein space.

Definition 3.8 (Weak convergence in $\mathcal{P}_p$)**.** *Let (E,d) be a Polish space and $p \in [1,\infty)$. Let $\{\mu_k\}_{k\in\mathbb{N}}$ be a sequence of probability measures in $\mathcal{P}_p(E)$ and μ be another element of $\mathcal{P}_p(E)$. Then $\{\mu_k\}$ is said to converge weakly in $\mathcal{P}_p(E)$ if any of the following equivalent properties is satisfied for some (and then any) $x_0 \in E$:*

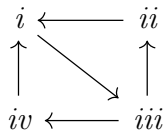
i) $\mu_k \xrightarrow{\text{w}} \mu$ *and* $\int_E d(x_0,x)^p \, \mathrm{d}\mu_k(x) \longrightarrow \int_E d(x_0,x)^p \, \mathrm{d}\mu(x)$;

ii) $\mu_k \xrightarrow{\text{w}} \mu$ *and* $\limsup_{k\to\infty} \int_E d(x_0,x)^p \, \mathrm{d}\mu_k(x) \leq \int_E d(x_0,x)^p \, \mathrm{d}\mu(x)$;

iii) $\mu_k \xrightarrow{\text{w}} \mu$ *and* $\lim_{R\to\infty} \limsup_{k\to\infty} \int_{d(x_0,x)\geq R} d(x_0,x)^p \, \mathrm{d}\mu_k(x) = 0$;

iv) For all continuous function ϕ *such that* $|\phi(x)| \leq C(1 + d(x_0,x)^p), C > 0$, $\int_E \phi(x) \, \mathrm{d}\mu_k(x) \to \int_E \phi(x) \, \mathrm{d}\mu(x)$.

Proof that the four statements in Definition 3.8 are equivalent.



$ii \Rightarrow i$) Lemma 1.14 guarantees that

$$\int_E d(x_0,x)^p \, \mathrm{d}\mu(x) \leq \liminf_{k\to\infty} \int_E d(x_0,x)^p \, \mathrm{d}\mu(x).$$

Therefore we have

$$\limsup_{k\to\infty} \int_E d(x_0,x)^p \, \mathrm{d}\mu_k(x) \leq \int_E d(x_0,x)^p \, \mathrm{d}\mu(x) \leq \liminf_{k\to\infty} \int_E d(x_0,x)^p \, \mathrm{d}\mu(x),$$

meaning it is instead an equality, therefore the limit exists and

$$\int_E d(x_0,x)^p \, \mathrm{d}\mu_k(x) \longrightarrow \int_E d(x_0,x)^p \, \mathrm{d}\mu(x).$$

$i \Rightarrow iii$) We can write $d(x_0,x)^p 1_{d(x_0,x)\geq R}$ as following:

$$d(x_0,x)^p 1_{d(x_0,x)\geq R} = d(x_0,x)^p - [d(x_0,x) \wedge R]^p + R^p 1_{d(x_0,x)\geq R}.$$

Integrating both side of this equality with respect to μ_k, and applying the supremum limit, we get

$$\begin{aligned}\limsup_{k\to\infty} \int_E d(x_0,x)^p 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu_k(x) = &\limsup_{k\to\infty} \int_E d(x_0,x)^p \, \mathrm{d}\mu_k(x) \\ &- \limsup_{k\to\infty} \int_E [d(x_0,x) \wedge R]^p \, \mathrm{d}\mu_k(x) \\ &+ R^p \limsup_{k\to\infty} \int_E 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu_k(x).\end{aligned}$$

The first term of the right-hand side converges by hypothesis to $\int_E d(x_0,x)^p \, \mathrm{d}\mu(x)$. The second one is converging to $\int_E [d(x_0,x) \wedge R]^p \, \mathrm{d}\mu(x)$ because the integrand is bounded continuous and μ_k converges weakly to μ. The last term is bounded from above by $R^p \int_E 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu(x)$ thanks to Portmanteau Theorem 1.21. Therefore

$$\limsup_{k\to\infty} \int_E d(x_0,x)^p 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu_k(x) \leq \int_E d(x_0,x)^p 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu(x)$$

Since $0 \leq d(x_0,x)^p 1_{d(x_0,x)\geq R} \leq d(x_0,x)^p \in L^1(E,\mu)$ and $\lim_{R\to\infty} d(x_0,x)^p 1_{d(x_0,x)\geq R} = 0$ pointwise, thanks to Dominant Convergence Theorem,

$$\begin{aligned}\lim_{R\to\infty} \limsup_{k\to\infty} \int_E d(x_0,x)^p 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu_k(x) &\leq \lim_{R\to\infty} \int_E d(x_0,x)^p 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu(x) \\ &= \int_E \lim_{R\to\infty} d(x_0,x)^p 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu(x) \\ &= 0.\end{aligned}$$

iii $\Rightarrow$ *ii*) Let start by writing $d(x_0,x)^p$ as following:

$$d(x_0,x)^p = [d(x_0,x) \wedge R]^p + [d(x_0,x)^p - R^p]_+ \ .$$

integrating both sides and applying the supremum limit, we get

$$\begin{aligned}\limsup_{k\to\infty} \int_E d(x_0,x)^p \, \mathrm{d}\mu_k(x) &= \limsup_{k\to\infty} \int_E [d(x_0,x) \wedge R]^p \, \mathrm{d}\mu_k(x) \\ &\quad + \limsup_{k\to\infty} \int_E [d(x_0,x)^p - R^p]_+ \, \mathrm{d}\mu_k(x) \\ &\leq \int_E [d(x_0,x) \wedge R]^p \, \mathrm{d}\mu(x) + \limsup_{k\to\infty} \int_E d(x_0,x)^p 1_{d(x_0,x)\geq R} \, \mathrm{d}\mu_k(x).\end{aligned}$$

Applying the limit as R approaches infinity, the second term on the right-hand side goes to 0 by hypothesis. In the first term the limit can go inside the integral sign thanks to Monotone Convergence Theorem, therefore

$$\limsup_{k\to\infty} \int_E d(x_0,x)^p \, \mathrm{d}\mu_k(x) \leq \int_E d(x_0,x)^p \, \mathrm{d}\mu(x).$$

At this point we proved that *i*, *ii* and *iii* are equivalent.

iv $\Rightarrow$ *i*) Just take $\phi(x) = d(x_0,x)^p$ to prove the second part. Then we notice that any bounded continuous function φ satisfies $|\varphi(x)| \leq C(1 + d(x_0,x)^p)$, therefore

$$\int_E \varphi(x) \, \mathrm{d}\mu_k(x) \longrightarrow \int_E \varphi(x) \, \mathrm{d}\mu(x)$$

and by definition, that is equivalent to μ_k converging weakly to μ.

iii $\Rightarrow$ *iv*) For any continuous function ϕ satisfying $|\phi(x)| \leq C(1 + d(x_0,x)^p)$, one may suppose ϕ to be nonnegative. That is because $\phi(x) = \phi_+(x) - \phi_-(x)$, where $\phi_+(x) = \max\{0, \phi(x)\}$ and $\phi_-(x) = \max\{0, -\phi(x)\}$ are by construction nonnegative continuous function. Therefore if we prove that, for any nonnegative function satisfying the hypothesis, we get

$$\int_E \phi \, \mathrm{d}\mu_k = \int_E \phi_+ \, \mathrm{d}\mu_k - \int_E \phi_- \, \mathrm{d}\mu_k \longrightarrow \int_E \phi_+ \, \mathrm{d}\mu - \int_E \phi_- \, \mathrm{d}\mu = \int_E \phi \, \mathrm{d}\mu,$$

for any ϕ, concluding the proof because it is obvious that ϕ satisfies the hypothesis as well.
if $0 \leq \phi(x) \leq C(1 + d(x_0, x)^p)$, then

$$\phi(x) = [\phi(x) \wedge C(1 + R^p)] + [\phi(x) - C(1 + R^p)]\, 1_{\phi(x) \geq C(1+R^p)},$$

and we observe that

$$\begin{aligned}
&[\phi(x) - C(1 + R^p)]\, 1_{\phi(x) \geq C(1+R^p)} \leq [C(1 + d(x_0, x)^p) - C(1 + R^p)]\, 1_{C(1+d(x_0,x)^p) \geq C(1+R^p)} \\
&= [C(d(x_0, x)^p - R^p)]\, 1_{d(x_0,x) \geq R} \leq C d(x_0, x)^p 1_{d(x_0,x) \geq R}.
\end{aligned}$$

Therefore,

$$\int_E \phi(x)\ \mathrm{d}\mu_k(x) \leq \int_E [\phi(x) \wedge C(1 + R^p)]\ \mathrm{d}\mu_k(x) + \int_E C d(x_0, x)^p 1_{d(x_0,x) \geq R}\ \mathrm{d}\mu_k(x).$$

Let $\phi_R(x) = [\phi(x) \wedge C(1 + R^p)]$. Notice it is a bounded continuous function, thus

$$\begin{aligned}
\limsup_{k\to\infty} \int_E \phi(x)\ \mathrm{d}\mu_k(x) &\leq \limsup_{k\to\infty} \int_E [\phi(x) \wedge C(1 + R^p)]\ \mathrm{d}\mu_k(x) \\
&\quad + C \limsup_{k\to\infty} \int_E d(x_0, x)^p 1_{d(x_0,x) \geq R}\ \mathrm{d}\mu_k(x) \\
&= \int_E [\phi(x) \wedge C(1 + R^p)]\ \mathrm{d}\mu(x) + C \limsup_{k\to\infty} \int_E d(x_0, x)^p 1_{d(x_0,x) \geq R}\ \mathrm{d}\mu_k(x).
\end{aligned}$$

$\phi_R(x)$ satisfies the hypothesis of Monotone Convergence Theorem, thus

$$\begin{aligned}
\limsup_{k\to\infty} \int_E \phi(x)\ \mathrm{d}\mu_k(x) &\leq \lim_{R\to\infty} \int_E [\phi(x) \wedge C(1 + R^p)]\ \mathrm{d}\mu(x) \\
&\quad + C \lim_{R\to\infty} \limsup_{k\to\infty} \int_E d(x_0, x)^p 1_{d(x_0,x) \geq R}\ \mathrm{d}\mu_k(x) \\
&= \int_E \lim_{R\to\infty} [\phi(x) \wedge C(1 + R^p)]\ \mathrm{d}\mu(x) = \int_E \phi(x)\ \mathrm{d}\mu(x).
\end{aligned}$$

By Lemma 1.14,

$$\int_E \phi(x)\ \mathrm{d}\mu(x) \leq \liminf_{k\to\infty} \int_E \phi(x)\ \mathrm{d}\mu_k(x),$$

therefore it is in fact an equality and

$$\int_E \phi(x)\ \mathrm{d}\mu_k(x) \longrightarrow \int_E \phi(x)\ \mathrm{d}\mu(x).$$

□

It comes natural to wonder how convergence of measures in $\mathcal{P}_p(E)$ is related to weak convergence of measures. When E is compact, it is quite simple to prove.

Proof when E is compact. First, we remember that a metric space is compact if and only if is complete and totally bounded, in particular it is bounded, which means that for all $x, y \in E$, there exists a constant $D < \infty$, such that $d(x, y) \leq D$. Therefore $d(x_0, x)^p$ is a bounded continuous function on E and, by definition of weak convergence,

$$\int_E d(x_0, x)^p \, \mathrm{d}\mu_k(x) \longrightarrow \int_E d(x_0, x)^p \, \mathrm{d}\mu(x),$$

proving the Property (i) of Definition 3.8.

□

Theorem 3.9 (W_p metrizes $\mathcal{P}_p$)**.** *Let (E, d) be a Polish space and $p \in [1, \infty)$. Then the Wasserstein distance metrize the weak convergence in $\mathcal{P}_p(E)$. In other words, if $\{\mu_k\}_{k\in\mathbb{N}}$ is a sequence of measures in $\mathcal{P}_p(E)$ and μ another measure in $\mathcal{P}_p(E)$, then the following statements are equivalent:*

1. *μ_k converges weakly to μ in $\mathcal{P}_p(E)$;*
2. *$\lim_{k\to\infty} W_p(\mu_k, \mu) = 0$.*

From Theorem 3.9 follows this corollary:

Corollary 3.10 (Continuity of W_p)**.** *If (E, d) is a Polish space and $p \in [1, \infty)$, then W_p is continuous in $\mathcal{P}_p(E)$. Which means that, if μ_k converges weakly to μ in $\mathcal{P}_p(E)$, respectively ν_k with ν, then*

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

Proof of Corollary 3.10. Thanks to the triangle inequality, we get two bounds:

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

Remembering that $W_p(\mu_k, \mu) \to 0$ and $W_p(\nu, \nu_k) \to 0$ as $k \to \infty$, applying the supremum limit on the first inequality, and the infimum limit on the second one, we get

$$\limsup_{k\to\infty} W_p(\mu_k, \nu_k) \leq W_p(\mu, \nu) \leq \liminf_{k\to\infty} W_p(\mu_k, \nu_k).$$

That proves what we claimed, which is $\lim_{k\to\infty} W_p(\mu_k, \nu_k) = W_p(\mu, \nu)$.

□

The definition of continuity is well defined: since $(\mathcal{P}_p(E), W_p)$ is a metric space, $\mathcal{P}_p(E) \times \mathcal{P}_p(E)$ is metrizable as well, for instance just take the maximum distance

$D_M((\mu,\nu),(\mu',\nu')) = \max\{W_p(\mu,\mu'), W_p(\nu,\nu')\}$ that generates the product topology. Therefore $W_p : \mathcal{P}_p(E) \times \mathcal{P}_p(E) \longrightarrow [0,\infty)$ is sequentially continuous on a metric space if and only if it is topologically continuous, see Proposition 1.23.

Before getting to the proof of Theorem 3.9 it will be good making some more comments. The short version of the theorem is that *Wasserstein metric metrize weak convergence*, however there are many ways to metrize weak convergence; for instance, we already know that the Levy-Prokhorov distance does it. So why bother with Wasserstein distances?

- The definitions of Wasserstein distances make them convenient in problems where optimal transport is naturally involved, such as many problems coming from partial differential equation;

- The Wasserstein distances have a rich duality. This is especially useful for p=1, which is that for any $\mu, \nu \in \mathcal{P}_1(E)$,

$$W_1(\mu,\nu) = \sup_{\|\psi\|_{Lip}\leq 1} \left\{ \int_E \psi \, \mathrm{d}\mu - \int_E \psi \, \mathrm{d}\nu \right\},$$

  see Corollary 3.6. Going back and forth from the original definition to the dual one might be convenient.

- Being defined by an infimum Wasserstein distances are often relatively easy to bound form above, since any *coupling* of μ and ν yields a bound between their distance.

  Example 3.11. Any C-Lipschitz mapping $f : (E,d) \to (E',d')$ induces a C-Lipschitz mapping $f_\# : (\mathcal{P}_p(E), W_p) \to (\mathcal{P}_p(E'), W_p')$ defined by $\mu \longmapsto f_\#\mu$.

  Proof. First we remember that since f is C-Lipschitz, $d'(f(x),f(y)) \leq Cd(x,y)$. Let

$$T : E \times E \longrightarrow E' \times E' \quad \text{defined as} \quad (x,y) \longmapsto (f(x),f(y)),$$

  which induces the function

$$T_\# : P_p(E \times E) \longrightarrow P_p(E' \times E') \quad \text{defined as} \quad \pi \longmapsto T_\#\pi.$$

  Notice that if π is a coupling of μ and ν, then $T_\#\pi$ is a coupling of $f_\#\mu$ and $f_\#\nu$: let introduce first the two projection $p_1, p_2 : E \times E \longrightarrow E$ defined as $(x_1,x_2) \mapsto x_i$ for $i = 1,2$. If π is a coupling of μ,ν then $(p_1)_\#(\pi) = \mu$ and $(p_2)_\#(\pi) = \nu$.

$$\begin{array}{ccc} \mathcal{P}_p(E \times E) & \xrightarrow{T_\#} & \mathcal{P}_p(E' \times E') \\ \downarrow{\scriptstyle (p_1)_\#} & & \downarrow{\scriptstyle (p_1)_\#} \\ \mathcal{P}_p(E) & \xrightarrow{f_\#} & \mathcal{P}_p(E') \end{array}$$

The previous diagram commutes, that means that

$$\begin{aligned}(p_1)_\#(T_\#\pi) &= (p_1)_\#(T_\#)(\pi)\\ &= (f_\#)(p_1)_\#(\pi)\\ &= f_\#(\mu) = f_\#\mu.\end{aligned}$$

We get the same result if we used $(p_2)_\#$, therefore $T_\#\pi$ is a coupling. We are now ready to prove that $f_\#$ is C-Lipschitz. Let π be an optimal coupling of μ and ν.

$$\begin{aligned}W_p'(f_\#\mu, f_\#\nu)^p &\leq \int_{E'\times E'} d'(u,v)^p \ \mathrm{d}T_\#\pi(u,v)\\ &= \int_{E\times E} d'(f(x),f(y))^p \ \mathrm{d}\pi(x,y)\\ &\leq \int_{E\times E} C^p d(x,y)^p \ \mathrm{d}\pi(x,y)\\ &= C^p W_p(\mu,\nu)^p.\end{aligned}$$

That proves the wanted result: $W_p'(f_\#\mu, f_\#\nu) \leq C W_p(\mu,\nu)$.

□

- Wasserstein distance incorporate a lot of geometry of the space. For instance, the map $E \to \mathcal{P}_p(E)$ defined as $x \longmapsto \delta_x$ is an isometric embedding, i.e. it preserves distances: for all $x, y \in E,\ d(x,y) = W_p(\delta_x, \delta_y)$.

We are almost ready to prove Theorem 3.9. First, it is necessary to introduce a lemma, which is quite interesting by itself.

Lemma 3.12 (Cauchy sequences in W_p are tight)**.** *Let (E,d) be a Polish space, $p \in [1,\infty)$ and $\{\mu_k\}_{k\in\mathbb{N}}$ be a Cauchy sequence in $(\mathcal{P}_p(E), W_p)$. Then $\{\mu_k\}_k$ is tight.*

Proof of Lemma 3.12. Let $\{\mu_k\}_{k\in\mathbb{N}}$ be a Cauchy sequence in $(\mathcal{P}_p(E), W_p)$, this means that for all $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that $W_p(\mu_k, \mu_h) < \varepsilon$ for all $k, h \geq N$. In particular,

$$\int_E d(x_0,x)^p \ \mathrm{d}\mu_k(x) = W_p(\delta_{x_0}, \mu_k)^p \leq [W_p(\delta_{x_0}, \mu_1) + W_p(\mu_1, \mu_k)]^p$$

remains bounded as $k \to \infty$.
Thanks to Remark 3.4 we know that the sequence $\{\mu_k\}$ is also Cauchy in the W_1 sense. Let $\varepsilon > 0$ be given and let $N \in \mathbb{N}$ be such that

$$k \geq N \implies W_1(\mu_N, \mu_k) < \varepsilon^2 \tag{3.12}$$

Then for any $k \in \mathbb{N}$ there is $j \in \{1, ..., N\}$ such that $W_1(\mu_j, \mu_k) < \varepsilon^2$ (if $k < N$, choose $j = k$, if $k \geq N$ it is just (3.12) by choosing $j = N$).
Since the finite set $\{\mu_1, ..., \mu_N\}$ is tight (see Corollary 1.11), there exists a compact set $K \subset E$ such that $\mu_j(E \setminus K) < \varepsilon$ for all $j \in \{1, ..., N\}$. By compactness, K can be covered by a finite number of small open ball:

$$K \subset U := B_\varepsilon(x_1) \cup ... \cup B_\varepsilon(x_m);$$

We define

$$U_\varepsilon := \{x \in E \ : \ d(x, U) < \varepsilon\} \subset B_{2\varepsilon}(x_1) \cup ... \cup B_{2\varepsilon}(x_m);$$

$$\phi(x) := \left(1 - \frac{d(x, U)}{\varepsilon}\right)_+.$$

We note that $1_U \leq \phi \leq 1_{U_\varepsilon}$ and ϕ is $1/\varepsilon$-Lipschitz. By using these bounds and the Kantorovic-Rubenstein duality (2.5), we find that for $j \leq N$ and k arbitrary,

$$\begin{aligned}
\mu_k(U_\varepsilon) &\geq \int_E \phi \ \mathrm{d}\mu_k \\
&= \int_E \phi \ \mathrm{d}\mu_j + \left(\int_E \phi \ \mathrm{d}\mu_j - \int_E \phi \ \mathrm{d}\mu_k\right) \\
&\geq \int_E \phi \ \mathrm{d}\mu_j - \frac{W_1(\mu_k, \mu_j)}{\varepsilon} \\
&\geq \mu_j(U) - \frac{W_1(\mu_k, \mu_j)}{\varepsilon}.
\end{aligned}$$

On the other end, $\mu_j(U) \geq \mu_j(K) \geq 1 - \varepsilon$ if $j \leq N$; for each k we can find $j = j(k)$ such that $W_1(\mu_k, \mu_j) \leq \varepsilon^2$. So we get

$$\mu_k(U_\varepsilon) \geq 1 - \varepsilon - \frac{\varepsilon^2}{\varepsilon} = 1 - 2\varepsilon.$$

At this point we have shown the following: for each $\varepsilon > 0$ there is a finite family $(x_i)_{1 \leq i \leq m}$ such that all measures μ_k give mass at least $1 - 2\varepsilon$ to the set $Z := \bigcup \overline{B_{2\varepsilon}(x_i)}$. The point is that Z might not be compact. To avoid that we repeat the same reasoning with ε replaced by $2^{-(p+1)}\varepsilon, \ p \in \mathbb{N}$; so there will be $(x_i)_{1 \leq i \leq m(p)}$ such that

$$\mu_k\left(E \setminus \bigcup_{1 \leq i \leq m(p)} B_{2^{-p}\varepsilon}(x_i)\right) \leq 2^{-p}\varepsilon.$$

Thus, if

$$S := \bigcap_{1 \leq p \leq \infty} \bigcup_{1 \leq i \leq m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)},$$

then

$$\begin{aligned}
\mu_k(E \setminus S) &= \mu_k \left(E \setminus \bigcap_{1\le p\le\infty} \bigcup_{1\le i\le m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)} \right) \\
&= \mu_k \left(\bigcup_{1\le p\le\infty} E \setminus \bigcup_{1\le i\le m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)} \right) \\
&\le \sum_{p\ge 1} \mu_k \left(E \setminus \bigcup_{1\le i\le m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)} \right) \\
&\le \varepsilon \sum_{p\ge 1} 2^{-p} \\
&= \varepsilon.
\end{aligned}$$

By construction, S can be covered by finitely many balls of radius δ where it could be arbitrarily small: for instance just take p large enough so that $2^{-p}\varepsilon < \delta$, then

$$S \subset \bigcup_{1\le i\le m(p)} \overline{B_{2^{-p}\varepsilon}(x_i)} \subset \bigcup_{1\le i\le m(p)} B_\delta(x_i).$$

That proves that S is totally bounded and it is closed, as an intersection of finite unions of closed balls. Since E is a complete metric space, we know that a subset of E is complete if and only if is closed. Then S is complete and totally bounded, which is equivalent to S being compact.

□

We are now ready to prove Theorem 3.9.

Proof of Theorem 3.9. Let $\{\mu_k\}_{k\in\mathbb{N}}$ be such that $\mu_k \to \mu$ in distance W_p, so that is obvious that $\{\mu_k\}$ is a Cauchy sequence in $(\mathcal{P}_p(E), W_p)$. The goal is to show that μ_k converges to μ in $P_p(E)$. First, by Lemma 3.12 the sequence $\{\mu_k\}_{k\in\mathbb{N}}$ is tight, so there is a subsequence $\{\mu_{k'}\}$ such that $\mu_{k'}$ converges weakly to some probability measure $\tilde{\mu}$. Then by Lemma 1.14

$$W_p(\tilde{\mu}, \mu) \le \liminf_{k'\to\infty} W_p(\mu_{k'}, \mu) = 0.$$

So $\tilde{\mu} = \mu$ and the whole sequence $\{\mu_k\}$ has to converge to μ. This only shows the weak convergence in the usual sense, not yet in $\mathcal{P}_p(E)$.
For any $\varepsilon > 0$ there exists a constant C_ε such that for all nonnegative real numbers a, b, we have

$$(a+b)^p \le (1+\varepsilon)\ a^p + C_\varepsilon\ b^p.$$

Combining the latter inequality and the usual triangle inequality, we see that whenever x_0, x and y are points in E, one has

$$d(x, x_0)^p \leq (1+\varepsilon)\ d(x_0, y)^p + C_\varepsilon\ d(x, y)^p. \tag{3.13}$$

Now let $\{\mu_k\}$ be a sequence in $\mathcal{P}_p(E)$ such that $W_p(\mu_k, \mu) \to 0$, and for each k, let π_k be an optimal transport plan between μ_k and μ. Integrating the inequality (3.13) in $\mathrm{d}\pi_k$ and using the marginal property, we get

$$\int_E d(x, x_0)^p\ \mathrm{d}\mu_k(x) \leq (1+\varepsilon) \int_E d(x_0, y)^p\ \mathrm{d}\mu(y) + C_\varepsilon \int_{E\times E} d(x, y)^p\ \mathrm{d}\pi_k(x, y).$$

But of course we observe that

$$\int_{E\times E} d(x, y)^p\ \mathrm{d}\pi_k(x, y) = W_p(\mu_k, \mu)^p \longrightarrow 0, \text{ as } k \to \infty;$$

Therefore, applying the lim sup on both sides,

$$\limsup_{k\to\infty} \int_E d(x_0, x)^p\ \mathrm{d}\mu_k(x) \leq (1+\varepsilon) \int_E d(x_0, y)^p\ \mathrm{d}\mu(y).$$

Letting $\varepsilon \searrow 0$, we see that Property (ii) of Definition 3.8 holds true, so μ_k does converge weakly to μ in $\mathcal{P}_p(E)$.

Conversely, assume μ_k converges weakly to μ in $\mathcal{P}_p(E)$, and for each k let π_k be an optimal transport plan between μ_k and μ.

By Prokhorov's Theorem, $\{\mu_k\}$ forms a tight sequence and also μ is tight. By Lemma 1.16 the sequence $\{\pi_k\}$ is also tight in $\mathcal{P}(E \times E)$. So, up to extraction of a subsequence, still denoted by $\{\pi_k\}$, one may assume

$$\pi_k \longrightarrow \pi \qquad \text{weakly in } \mathcal{P}(E \times E).$$

Since each π_k is optimal, Theorem 1.17 guarantees that π is an optimal coupling of μ and μ, so this is the trivial coupling, where $\pi = (Id, Id)_\# \mu$ and in terms of random variables $Y = X$. Since this is independent of the extracted subsequence, actually π is the limit to the whole sequence $\{\pi_k\}$.

Now let x_0 in E and $R > 0$. If $d(x, y) > R$, then we have that one of the numbers $d(x, x_0)$ and $d(x_0, y)$ has to be larger than $R/2$ and both of them would be larger than $d(x, y)/2$. We denote

$$U_1 := \{d(x, x_0) \geq R/2 \text{ and } d(x, x_0) \geq d(x, y)/2\};$$
$$U_2 := \{d(x_0, y) \geq R/2 \text{ and } d(x_0, y) \geq d(x, y)/2\}.$$

Then we definitely have that $\{d(x, y) \geq R\} \subset U_1 \cup U_2$, therefore

$$1_{d(x,y)\geq R} \leq 1_{U_1} + 1_{U_2}.$$

So, obviously

$$\begin{aligned}[d(x,y)^p - R^p]\,1_{d(x,y)\geq R} &\leq d(x,y)^p 1_{U_1} + d(x,y)^p 1_{U_2} \\ &\leq 2^p d(x,x_0)^p 1_{d(x,x_0)\geq R/2} + 2^p d(x_0,y)^p 1_{d(x,x_0)\geq R/2}.\end{aligned}$$

It follows that

$$\begin{aligned}W_p(\mu_k,\mu)^p &= \int_{E\times E} d(x,y)^p \,\mathrm{d}\pi_k(x,y) \\ &= \int_{E\times E} [d(x,y)\wedge R]^p \,\mathrm{d}\pi_k(x,y) + \int_{E\times E} [d(x,y)^p - R^p]\,1_{d(x,y)\geq R}\,\mathrm{d}\pi_k(x,y) \\ &\leq \int_{E\times E} [d(x,y)\wedge R]^p \,\mathrm{d}\pi_k(x,y) + 2^p \int_{d(x,x_0)\geq R/2} d(x,x_0)^p \,\mathrm{d}\pi_k(x,y) \\ &\qquad + 2^p \int_{d(x_0,y)>R/2} d(x_0,y)^p \,\mathrm{d}\pi_k(x,y) \\ &= \int_{E\times E} [d(x,y)\wedge R]^p \,\mathrm{d}\pi_k(x,y) + 2^p \int_{d(x,x_0)\geq R/2} d(x,x_0)^p \,\mathrm{d}\mu_k(x) \\ &\qquad + 2^p \int_{d(x_0,y)\geq R/2} d(x_0,y)^p \,\mathrm{d}\mu(y).\end{aligned}$$

Since π_k converges weakly to π,

$$\int_{E\times E} [d(x,y)\wedge R]^p \,\mathrm{d}\pi_k(x,y) \longrightarrow \int_{E\times E} [d(x,y)\wedge R]^p \,\mathrm{d}\pi(x,y) = 0,$$

because we are integrating over the diagonal of $E\times E$, in which $[d(x,y)\wedge R]^p$ is equal to zero. So

$$\begin{aligned}\limsup_{k\to\infty} W_p(\mu_k,\mu)^p &\leq \lim_{R\to\infty} 2^p \left[\limsup_{k\to\infty} \int_{d(x,x_0)\geq R/2} d(x,x_0)^p \,\mathrm{d}\mu_k(x)\right] \\ &\quad + \lim_{R\to\infty} 2^p \left[\int_{d(x_0,y)\geq R/2} d(x_0,y)^p \,\mathrm{d}\mu(y)\right] \\ &= 0.\end{aligned}$$

This concludes the argument.

□

Remark 3.13. The notion of weak convergence in $\mathcal{P}_p(E)$ is stronger than the usual one. Simply saying that $\mu_k \xrightarrow{\mathrm{w}} \mu$ does not necessarily imply that we have convergence of the moments of order p, for instance, it is not guaranteed if a sequence of measure is not tight and $W_p(\mu_k,\mu)$ might not converge to zero. Only the converse would be true, in fact if μ_k converges to μ in W_p, then it is tight, therefore μ_k converges weakly to μ and we

have convergence of the moments of order p. Thus the hypothesis of convergence of the p-th moment cannot be removed.
Having said that, we notice that topology induced by convergence in W_p is finer that the usual *weak topology*, because the first one implies the second one but the only weak convergence is not enough to have convergence in W_p.

3.3 Topological properties of the Wasserstein Spaces

Theorem 3.14 (Wasserstein distances are controlled by weighted Total Variation)**.** *Let μ and ν be two probability measures on a Polish space (E, d). Let $p \in [1, \infty)$ and $x_0 \in E$. Then*

$$W_p(\mu, \nu) \leq 2^{1/p'} \left(\int_E d(x_0, x)^p \, \mathrm{d} \, |\mu - \nu| \, (x) \right)^{1/p}, \qquad \frac{1}{p} + \frac{1}{p'} = 1, \tag{3.14}$$

where $|\mu - \nu| = (\mu - \nu)_+ + (\mu - \nu)_-$.

Proof. Let π be the transference plan obtained by keeping fixed all the mass shared by μ and ν and distributing the rest uniformly:

$$\pi = (Id, Id)_{\#}(\mu \wedge \nu) + \frac{1}{a}(\mu - \nu)_+ \otimes (\mu - \nu)_-,$$

where $(\mu \wedge \nu) = \mu - (\mu - \nu)_+$ and $a = (\mu - \nu)_+(E) = (\mu - \nu)_-(E)$. Notice that a is well define because $\mu - \nu = (\mu - \nu)_+ - (\mu - \nu)_-$ and

$$0 = \mu(E) - \nu(E) = (\mu - \nu)_+(E) - (\mu - \nu)_-(E).$$

π is a coupling of μ and ν, in fact $\pi(E \times E) = 1$ and:

1. if $\mu(A) < \nu(A)$, we have $\pi(A \times E) = \mu(A) + \frac{1}{a}(0 \cdot a) = \mu(A)$
   $\pi(E \times A) = \mu(A) + \frac{1}{a}\left[a \cdot (\nu(A) - \mu(A))\right] = \nu(A)$;

2. if $\mu(A) > \nu(A)$, we have $\pi(E \times A) = \nu(A) + \frac{1}{a}(a \cdot 0) = \nu(A)$
   $\pi(A \times E) = \nu(A) + \frac{1}{a}\left[(\mu(A) - \nu(A)) \cdot a\right] = \mu(A)$.

Now, using the definition of W_p, the definition of π, the triangle inequality for d, the

elementary inequality $(C+D)^p \leq 2^{p-1}(C^p + D^p)$ and the definition of a, we get:

$$\begin{aligned}
W_p(\mu,\nu)^p &\leq \int_{E\times E} d(x,y)^p \, \mathrm{d}\pi(x,y) \\
&= \frac{1}{a}\int_{E\times E} d(x,y)^p \, \mathrm{d}(\mu-\nu)_+(x)\mathrm{d}(\mu-\nu)_-(y) \\
&\leq \frac{2^{p-1}}{a}\int_{E\times E} [d(x,x_0)^p + d(x_0,y)^p] \, \mathrm{d}(\mu-\nu)_+(x)\mathrm{d}(\mu-\nu)_-(y) \\
&\leq 2^{p-1}\left[\int_E d(x,x_0)^p \, \mathrm{d}(\mu-\nu)_+(x) + \int_E d(x_0,y)^p \, \mathrm{d}(\mu-\nu)_-(y)\right] \\
&= 2^{p-1}\int_E d(x_0,x)^p \, \mathrm{d}\left[(\mu-\nu)_+ + (\mu-\nu)_-\right](x) \\
&= 2^{p-1}\int_E d(x_0,x)^p \, \mathrm{d}\left|\mu-\nu\right|(x).
\end{aligned}$$

□

With the next theorem we are going to show that $\mathcal{P}_p(E)$ inherits several properties of the space E.

Theorem 3.15. *Let E be a complete and separable metric space and $p \in [1,\infty)$. Then the Wasserstein space $\mathcal{P}_p(E)$, metrized by the Wasserstein distance W_p, is also a complete and separable metric space. Therefore, if E is Polish, $\mathcal{P}_p(E)$ is itself Polish. Moreover, any probability measure on $\mathcal{P}_p(E)$ can be approximated by a sequence of probability measures with finite support.*

Proof. The fact that $\mathcal{P}_p(E)$ equipped with W_p is a metric space was already explained. It remains to check the *separability* and *completeness*.
Let us prove the separability. Let $\mathcal{D}$ be a dense set in E and let $\mathcal{R}$ be the space of all probability measures that can be written as $\sum b_j\delta_{x_j}$ where $b_j \in \mathbb{Q}$ and x_j are finitely many elements in $\mathcal{D}$. It will turn out that $\mathcal{R}$ is dense in $\mathcal{P}_p(E)$. To prove this, let $\varepsilon > 0$ be given and let x_0 be an arbitrary element of $\mathcal{D}$. If $\mu \in \mathcal{P}_p(E)$, since $\{\mu\}$ is tight, there exists a compact set $K \subset E$ such that

$$\int_{E\setminus K} d(x_0,x)^p \, \mathrm{d}\mu(x) \leq \varepsilon^p.$$

K is totally bounded then we can cover it by a finite family of balls of radius $\varepsilon/2$ centred in $x_j \in \mathcal{D}$, in other words

$$K \subseteq \bigcup_{1\leq j\leq N} B_{\varepsilon/2}(x_j).$$

We define

$$D_1 = B_{\varepsilon/2}(x_1) \qquad \text{and} \qquad D_k = B_{\varepsilon/2}(x_k) \setminus \bigcup_{j<k} B_{\varepsilon/2}(x_j),$$

then all D_k are disjoints and still cover K. Define f on E by

$$f(D_k \cap K) = \{x_k\}, \qquad f(E \setminus K) = \{x_0\}.$$

Then, for any $x \in K,\ d(x, f(x)) \leq \varepsilon$. So

$$\begin{aligned}
\int_E d(x, f(x))^p \, \mathrm{d}\mu(x) &= \int_K d(x, f(x))^p \, \mathrm{d}\mu(x) + \int_{E\setminus K} d(x, f(x))^p \, \mathrm{d}\mu(x) \\
&\leq \varepsilon^p \int_K \mathrm{d}\mu(x) + \int_{E\setminus K} d(x, x_0)^p \, \mathrm{d}\mu(x) \\
&\leq \varepsilon^p \ \mu(K) + \varepsilon^p \\
&\leq 2\varepsilon^p.
\end{aligned}$$

Since (Id, f) is a coupling of μ and $f_{\#}\mu,\ W_p(\mu, f_{\#}\mu) \leq 2^{1/p}\varepsilon \leq 2\varepsilon$, and of course $f_{\#}\mu$ can be written as $\sum a_j \delta_{x_j},\ 0 \leq j \leq N$. This shows that μ might be approximated, with arbitrary precision, by a finite combination of Dirac masses. To conclude, it is sufficient to show that the coefficients a_j might be replaced by rational coefficients.
By Theorem 3.14

$$\begin{aligned}
W_p\left(\sum_{j\leq N} a_j \delta_{x_j}, \sum_{j\leq N} b_j \delta_{x_j}\right) &\leq 2^{1/p'} \left[\max_{k,h} d(x_k, x_h)\right] \left(\sum_{j\leq N} |a_j - b_j|\right)^{1/p} \\
&\overset{\diamondsuit}{\leq} 2^{1/p'} \left[\max_{k,h} d(x_k, x_h)\right] \sum_{j\leq N} |a_j - b_j|^{1/p} \\
&\leq \varepsilon,
\end{aligned}$$

as long as we choose some rational coefficient b_j close enough to a_j. The inequality signed by $\diamondsuit$, follows from $(\sum_{j\leq N} a_j)^\alpha \leq \sum_{j\leq N} a_j^\alpha$, if $\alpha \in (0, 1]$ and $a_j \geq 0$. Therefore we have

$$\begin{aligned}
W_p\left(\mu, \sum_{j\leq N} b_j \delta_{x_j}\right) &\leq W_p\left(\sum_{j\leq N} a_j \delta_{x_j}, \sum_{j\leq N} b_j \delta_{x_j}\right) + W_p\left(\mu, \sum_{j\leq N} a_j \delta_{x_j}\right) \\
&< \varepsilon + 2\varepsilon \\
&= 3\varepsilon.
\end{aligned}$$

Finally, let us prove the completeness. Let $\{\mu_k\}_{k\in\mathbb{N}}$ be a Cauchy sequence in $\mathcal{P}_p(E)$. By Lemma 3.12, it admits a subsequence $\{\mu_{k'}\}$ which converges weakly (in the usual sense) to a measure μ. Then,

$$\int_E d(x_0, x)^p \, \mathrm{d}\mu(x) \leq \liminf_{k'\to\infty} \int_E d(x_0, x)^p \, \mathrm{d}\mu_{k'}(x) < \infty,$$

so μ belongs to $\mathcal{P}_p(E)$. Moreover, by lower semicontinuity of W_p,

$$W_p(\mu, \mu_{j'}) \leq \liminf_{k' \to \infty} W_p(\mu_{k'}, \mu_{j'}).$$

In particular

$$\lim_{j' \to \infty} W_p(\mu, \mu_{j'}) \leq \limsup_{k', j' \to \infty} W_p(\mu_{k'}, \mu_{j'}) = 0,$$

because a subsequence of a Cauchy sequence is still a Cauchy sequence. Which means that $\mu_{j'}$ converges weakly to μ in the W_p sense, see Theorem 3.9. Since $\{\mu_k\}$ is a Cauchy sequence with a converging subsequence, it follows that the whole sequence is converging: for all $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that for all $k, j' \geq N$, one has $W_p(\mu, \mu_{j'}) < \varepsilon$ and $W_p(\mu_{j'}, \mu_k) < \varepsilon$, therefore

$$W_p(\mu, \mu_k) \leq W_p(\mu, \mu_{j'}) + W_p(\mu_{j'}, \mu_k) \leq 2\varepsilon.$$

That proves that $\mu_k \xrightarrow{\text{w}} \mu$ in $\mathcal{P}_p(E)$, concluding the proof.

□

Corollary 3.16. *If E is compact, then $\mathcal{P}_p(E)$ is also compact.*

Proof. We use the same notation of the proof of Theorem 3.15. Since E is compact, it is totally bounded, which means that for all $\varepsilon > 0$ there exists $\{x_1, ..., x_N\} \in E$ such that

$$E = \bigcup_{i=1}^{N} B_\varepsilon(x_i).$$

Let now denote

$$V_i := \left\{ \mu \in P_p(E) \ : \ \int_E d(x, x_i)^p \, \mathrm{d}\mu(x) < \varepsilon^p \right\}.$$

Claim: for all i, V_i is an open set of $\mathcal{P}_p(E)$ and the union of all V_i, that we remind to be a finite union, covers $\mathcal{P}_p(E)$. That would conclude the argument, in fact we would be able to say that the Wasserstein Space over a complete and totally bounded metric space, is complete (we already know it from Theorem 3.15) and totally bounded itself, therefore compact.

First step. V_i is open in $\mathcal{P}_p(E)$. Of course it is a non-empty set because there exists a $y \in \mathcal{D}$ such that $d(x_i, y) < \varepsilon$, therefore δ_y is in V_i. For any $\epsilon_1 > 0$ and for all $\mu \in V_i$ there exist a probability measure $\nu \in \mathcal{P}_p(E)$ such that $W_p(\mu, \nu) < \epsilon_1$. If we take ϵ_1 to be smaller than the quantity $\varepsilon - W_p(\mu, \delta_{x_i})$, then ν is a probability measure on V_i. In fact

$$\begin{aligned} W_p(\nu, \delta_{x_i}) &\leq W_p(\mu, \delta_{x_i}) + W_p(\mu, \nu) \leq \\ &\leq W_p(\mu, \delta_{x_i}) + \epsilon_1 \\ &< \varepsilon, \end{aligned}$$

To prove the first step.

Second step. For all $\mu \in \mathcal{P}_p(E)$ there exists a collection of probability measures on V_i such that

$$\mu = \sum_{i=1}^{N} p_i \, \mu_i,$$

where $\mu_i \in V_i$ for all i. Let

$$D_1 := B_\varepsilon(x_1) \qquad\qquad D_k := B_\varepsilon(x_k) \setminus \bigcup_{j=1}^{k-1} B_\varepsilon(x_j).$$

Of course

$$E = \bigcup_{i=1}^{N} B_\varepsilon(x_i) = \biguplus_{i=1}^{N} D_i.$$

where the last one is a disjoint union, so it comes natural to use the Total Probability formula:

$$\begin{aligned} \mu(\,\cdot\,) &= \sum_{i=1}^{N} \mu(D_i) \; \mu(\,\cdot\,|D_i) \\ &=: \sum_{i=1}^{N} p_i \; \mu_i(\,\cdot\,). \end{aligned}$$

Last thing to check is proving that $\mu_i \in V_i$.

$$\begin{aligned} W_p(\mu_i, \delta_{x_i})^p &= \int_E d(x, x_i)^p \; \mathrm{d}\mu_i(x) \\ &= \frac{1}{\mu(D_i)} \int_{D_i} d(x, x_i)^p \; \mathrm{d}\mu(x) \\ &\overset{(\bigstar)}{<} \frac{1}{\mu(D_i)} \int_{D_i} \varepsilon^p \; \mathrm{d}\mu(x) \\ &= \frac{1}{\mu(D_i)} \varepsilon^p \mu(D_i) \\ &= \varepsilon^p. \end{aligned}$$

The inequality signed by $\bigstar$ follows from D_i being contained in $B_\varepsilon(x_i)$, where $d(x, x_i) < \varepsilon$, by definition. Therefore we proved that

$$\mathcal{P}_p(E) = \bigcup_{i=1}^{N} V_i.$$

□

Bibliography

[CD18] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I*. vol. 83. Springer, 2018, pp. 350–360. ISBN 978-3-319-56437-1. https://doi.org/10.1007/978-3-319-58920-6

[GS02] Alison L. Gibbs and Francis Edward Su. *On Choosing and Bounding Probability Metrics*. International Statistical Review, vol. 70, n. 3, pp. 423–425, 2002. DOI https://doi.org/10.1111/j.1751-5823.2002.tb00178.x. https://arxiv.org/abs/math/0209021

[Vil09] Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, vol 338. Springer, 2009. ISBN 978-3-540-71049-3. https://link.springer.com/book/10.1007/978-3-540-71050-9

Ringraziamenti

Per prima cosa vorrei ringraziare i miei genitori che mi hanno permesso di intraprendere questo percorso e che hanno sempre creduto nelle mie capacità. Ringrazio i miei fratelli Lorenzo, Francesca e Maria, i nonni e gli zii che mi hanno sostenuto, soprattutto in quei momenti in cui la persona che ci credeva di meno ero io.

Un ruolo importante in questo percorso l'hanno avuto i miei compagni di corso, con loro mi sono divertito e ho instaurato vere amicizie che mi auguro possano continuare.

Grazie a Mattia, Davide e Andrea, i miei amici di lunga data, con loro ho passato tanti bei momenti e per questo sarò sempre riconoscente.

Un ringraziamento speciale va alla mia Professoressa di matematica del liceo, Laura Piazzi. È merito suo se mi sono appassionato alla materia e se l'ho voluta approfondire all'Università. Non dimenticherò mai i suoi consigli e la sua gentilezza.

Infine, ma non per importanza, ringrazio te, Beatrice. Da quando sei entrata a far parte della mia vita è cambiato tutto, sono una persona migliore e con te al mio fianco penso sia tutto più bello. Da te ho imparato a non arrendermi, me lo insegni tutti i giorni. Non ci sono parole a sufficienza per descrivere quanto ti sono riconoscente per tutto (anche perché probabilmente non le conosco).