Universidad Autónoma de Madrid

Mathematic Analysis Fundamentals

Optimal transport for Topological Data Analysis

End of Course Thesis.
2024-2025.

Author:
Gonzalo Ortega Carpintero

January 2025

1 Introduction

Transport maps were introduced in 1781 by Gaspard Monge to represent the idea of moving earth from one place into an other [1][1.1 Historical overview]. In this original formulation of the optimal transport problem, it was enough to consider $\mathbb{R}^3$ as the ambient space, using the Euclidean distance as the cost function of moving mass between two points.

In the 30's, Leonid Kantorovich reformulated the problem to describe the optimization process of supply and demand distributions of diverse problems. The mass could be divided between different origin and destinations, making it possible to interpret the problem as the way to measure the cost of transforming one probability distribution into an other. In this thesis, will introduce the p-Wasserstein distance as a metric on the probability measures with finite p-moment space. When $p = 1$, the distance will represent the metric introduced in the Kantorovich optimal transport problem, also used named Earth Mover's distance, used for machine learning algorithms and computer vision problems [2]. When $p = \infty$ it is named the bottleneck distance, and will be the main them of study of this thesis.

In topological data analysis, diagrams arise to represent the persistence of the homology groups of a data set through time. Those diagrams are named persistence diagrams, and those homology groups, persistence homology groups. We will introduce an analogous p-Wasserstein distance in the space of persistence diagrams and prove that there exists an isometric embedding from a separable metric space into the space of persistence diagrams with the Wasserstein distance.

2 Optimal transport

The main result of optimal transport theory is the solution of Kantorovich's problem for general costs: the existence of an optimal transport plan. Lets start by introducing Monge's and Kantorovich's problems, observing its main key difference. For that, we shall fist define a way to compare probability measures from two different spaces. We will denote the set of probability measures over a space X by $\mathcal{P}(X)$, and the class of Borel-measurable sets by $\mathcal{B}(X)$.

Definition 2.1 (Push-forward measure)**.** Let $T : X \to Y$ be a Borel map, and $\mu \in \mathcal{P}(X)$. Let $A \in \mathcal{B}$. The *push-forward measure* $T_{\#}\mu \in \mathcal{P}(Y)$ is defined as

$$T_{\#}\mu(A) := \mu(T^{-1}(A)).$$

Now we can introduce transport maps, as functions witch transform one probability measure into an other.

Definition 2.2 (Transport map)**.** Given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, a *transport map from μ to ν* is a Borel map $T : X \to Y$ that satisfies $T_{\#}\mu = \nu$.

Definition 2.3 (Transport plan)**.** Let $\pi_X : (X \times Y) \to X$ and $\pi_Y : (X \times Y) \to Y$ such that for every $(x, y) \in (X, Y)$, $\pi_X(x, y) = x$ and $\pi_Y(x, y) = y$. A *transport plan between μ and ν* is a probability measure $\gamma \in \mathcal{P}(X \times Y)$ where

$$(\pi_X)_{\#}\gamma = \mu \text{ and } (\pi_Y)_{\#}\gamma = \nu.$$

The set of all couplings between μ and ν is denoted $\Gamma(\mu, \nu)$.

While the set of transport maps between two given probability measures might be empty, transport plans are a more flexible generalization of them allowing to modulate one measure into the other. In probability theory, transport plans are named *couplings*, and $\Gamma(\mu, \nu)$ is the collection of all probability measures in $X \times Y$ with *marginals* μ and ν [3].

Given this definitions, we can introduce Monge and Kantorovich problems, $C_M(\mu, \nu)$ and $C_K(\mu, \nu)$ respectively, as follows.

Definition 2.4 (Transport problems)**.** Fix $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and consider a lower semicontinous map $c : X \times Y \to [0, \infty]$. Then

$$C_M(\mu, \nu) := \inf \left\{ \int_X c(x, T(x)) d\mu(x) : T_{\#}\mu = \nu \right\},$$
$$C_K(\mu, \nu) := \inf \left\{ \int_{X \times Y} c(x, y) d\gamma(x, y) : \gamma \in \Gamma(\mu, \nu) \right\}.$$

Next theorem asserts that it actually exists a minimizing transport plan that minimizes Kantorovich problem. This will prove useful to verify that Wasserstein distance exists and it is a well defined metric.

Theorem 2.5. *Let $c : X \times Y \to [0, \infty]$ be lower semicontinous, and let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then there exists a coupling $\bar{\gamma} \in \Gamma(\mu, \nu)$ that verifies*

$$C_K(\mu, \nu) = \int_{X \times Y} c(x, y) d\bar{\gamma}(x, y).$$

Proof. (to do) □

Example 2.6 (Mean and variance in $\mathbb{R}$)**.**

Definition 2.7 (Probability measures with finite p-moment)**.** Let (X, d) be a locally compact and separable, metric space. Let $1 \leq p < \infty$. The *set of probability measures with finite p-moment* is defined As

$$\mathcal{P}_p(X) := \left\{ \sigma \in \mathcal{P}(X) : \int_X d(x, x_0)^p d\mu(x) < \infty \text{ for some } x_0 \in X \right\}.$$

Proposition 2.8. *The definition of $\mathcal{P}_p(X)$ is independent of the base point x_0*

Proof. (to do) □

Definition 2.9 (p-Wasserstein distance)**.** Given $u, v \in \mathcal{P}_p(X)$, the *p-Wasserstein distance* is defined as

$$W_p(u, v) := \left(\inf_{\gamma \in \Gamma(u,v)} \int_{X \times X} d(x, y)^p d\gamma(x, y) \right)^{\frac{1}{p}}.$$

Proposition 2.10. *W_p is a distance on the space $\mathcal{P}_p(X)$.*

Proof. We will follow the steps made in [1][Theorem 3.1.5]. To prove the triangle inequality, let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(X)$ and

(to do) □

3 Wasserstein distance in persistence diagrams

In last section we have exposed the original optimal transport problem

We will denote the strict upper triangular region of the Euclidean plane as $\mathbb{R}^2_< := \{(x, y) \in \mathbb{R}^2 : x < y\}$, and the diagonal of the plane as $\Delta := \{(x, y) \in \mathbb{R}^2 : x = y\}$.

Definition 3.1 (Persistence diagram)**.** Let I be a countable set. A *persistence diagram* is a function $D : I \to \mathbb{R}^2_<$.

Definition 3.2 (Chebyshev distance)**.** (To do) $d_\infty := \max\{|a_x - b_x|, |a_y - b_y|\}$

Proposition 3.3. *If $a \in \mathbb{R}^2_<$, then $d_\infty(a, \Delta) = \inf_{t\in\Delta} d_\infty(a, t) = \frac{a_y - a_x}{2}$.*

Proof. (to do) □

Proposition 3.4. *The upper triangular region of the Euclidean plane with the Chebyshev distance $(\mathbb{R}^2_<, d_\infty)$ is a metric space.*

Proof. (To do) □

Definition 3.5 (Partial matching)**.** Let $D_1 : I_1 \to \mathbb{R}^2_<$ and $D_2 : I_2 \to \mathbb{R}^2_<$ be persistence diagrams. A *partial matching* between D_1 and D_2 is the triple (I'_1, I'_2, f) such that $f : I'_1 \to I'_2$ is a bijection with $I'_1 \subseteq I_1$ and $I'_2 \subseteq I_2$.

Definition 3.6. Let $D_1 : I_1 \to \mathbb{R}^2_<$ and $D_2 : I_2 \to \mathbb{R}^2_<$ be persistence diagrams. Let (I'_1, I'_2, f) be a partial matching between them. If $p < \infty$, the *p-cost of f* is defined as

$$\begin{aligned}\mathrm{cost}_p(f) := \Bigg(& \sum_{i\in I'_1} d_\infty(D_1(i), D_2(f(i)))^p \\ & + \sum_{i\in I_1\setminus I'_1} d_\infty(D_1(i), \Delta)^p \\ & + \sum_{i\in I_2\setminus I'_2} d_\infty(D_2(i), \Delta)^p \Bigg)^{\frac{1}{p}}.\end{aligned}$$

For $p = \infty$, the ∞-*cost of* f is defined as

$$\mathrm{cost}_\infty(f) := \max\left\{ \sup_{i\in I_1'} d_\infty(D_1(i), D_2(f_i)), \right.$$
$$\sup_{i\in I_1\setminus I_1'} d_\infty(D_1(i), \Delta),$$
$$\left. \sup_{i\in I_2\setminus I_2'} d_\infty(D_2(i), \Delta) \right\}.$$

Definition 3.7 (p-Wasserstein distance)**.** Let D_1, D_2 be persistence diagrams. Let $1 \leq p \leq \infty$. Define

$$\tilde{\omega_p}(D_1, D_2) = \inf\{\mathrm{cost}_p(f) : f \text{ is a partial matching between } D_1 \text{ and } D_2\}.$$

Let $\emptyset$ denote the unique persistence diagram with empty indexing set. Let $(\mathrm{Dgm}_p, \omega_p)$ be the space of persistence diagrams D that satisfy $\tilde{\omega_p}(D, \emptyset) < \infty$ modulo the equivalence relation $D_1 \sim D_2$ if $\tilde{\omega_p}(D_1, D_2) = 0$. The metric ω_p is called the *p-Wasserstein distance.*

Definition 3.8 (Bottleneck distance)**.** In the conditions of Definition 3.7, if $p = \infty$, the metric ω_∞ is called the *bottleneck distance.*

Proposition 3.9. *There is only one matching between* $D : I \to \mathbb{R}^2_<$ *and* $\emptyset$*. Hence,*

$$\tilde{\omega_p}(D, \emptyset) = \left(\sum_{i\in I} d_\infty(D(i), \Delta)^p \right)^{\frac{1}{p}}.$$

Proof. (To do) □

Next proposition will prove that, in indeed, the space of persistence diagrams with the p-Wasserstein distance $(\mathrm{Dgm}_p, \omega_p)$ is a metric space. Its proof is usually omitted in literature, as it based on the simple fact that d_∞ is a distance. We will give, however, an step by step version here.

Proposition 3.10. ω_p *is a distance on the space* $(\mathrm{Dgm}_p, \omega_p)$*.*

Proof. Let $D_1, D_2, D_3 \in \mathrm{Dgm}_p$, with $1 \leq p \leq \infty$.

First of all, $\omega_p(D_1, D_2) \geq 0$ because $d_\infty \geq 0$. $\omega_p(D_1, D_2) = 0$ if and only if $\tilde{\omega_p}(D_1, D_2) \geq 0$. Thus, because of the equivalence relationship used to define ω_p, it has to be $D_1 \sim D_2$.

To check symmetry, note that every partial matching f is bijective, therefore f^{-1} is a partial matching. But, for all $i \in I'_1$, exists $j \in I'_2$ such that $f(i) = j$ and

$$d_\infty(D_1(i), D_2(f(i))) = d_\infty(D_2(f(i)), D_1(i)) = d_\infty(D_2(j), D_1(f^{-1}(j))).$$

Then, $\text{cost}_p(f) = \text{cost}_p(f^{-1})$ and we have

$$\begin{aligned}
\omega_p(D_1, D_2) &= \inf\{\text{cost}_p(f) : f \text{ is a partial matching between } D_1 \text{ and } D_2\} \\
&= \inf\{\text{cost}_p(f^{-1}) : f^{-1} \text{ is a partial matching between } D_2 \text{ and } D_1\} \\
&= \omega_p(D_2, D_1).
\end{aligned}$$

Finally, lets prove the triangle inequality. If $f : I'_1 \to I'_2$ is a partial matching between D_1 and D_2 and $g : I'_2 \to I'_3$ is a partial matching between D_2 and D_3, $g \circ f : I'_1 \to I'_3$ is a partial matching between D_1 and D_3 as both f and g are bijective. Computing the cost of the matchings for $p < \infty$, we notice that

$$\begin{aligned}
&\sum_{i \in I'_1} d_\infty(D_1(i), D_2(f(i))) + \sum_{i \in I_1 \setminus I'_1} d_\infty(D_1(i), \Delta) + \sum_{i \in I_2 \setminus I'_2} d_\infty(D_2(i), \Delta) \\
&+ \sum_{i \in I'_2} d_\infty(D_2(i), D_3(g(i))) + \sum_{i \in I_2 \setminus I'_2} d_\infty(D_2(i), \Delta) + \sum_{i \in I_3 \setminus I'_3} d_\infty(D_3(i), \Delta) \\
&\geq \sum_{i \in I'_1} d_\infty(D_1(i), D_3(g \circ f(i))) + \sum_{i \in I_1 \setminus I'_1} d_\infty(D_1(i), \Delta) + \sum_{i \in I_3 \setminus I'_3} d_\infty(D_3(i), \Delta)
\end{aligned}$$

as $d_\infty(D_1(i), D_2(f(i))) + d_\infty(D_2(f(i)), D_2(g(f(i)))) \geq d_\infty(D_1(i), D_3(g \circ f(i)))$ using the triangle inequality of d_∞. Therefore, for all partial matchings f and g as described, we have $\text{cost}_p(f) + \text{cost}_p(g) \geq \text{cost}_p(g \circ f)$. Using the same reasoning, por $p = \infty$ we also obtain $\text{cost}_\infty(f) + \text{cost}_\infty(g) \geq \text{cost}_\infty(g \circ f)$. Hence, we have verified that

$$\omega_p(D_1, D_2) + \omega_p(D_2, D_3) \geq \omega_p(D_1, D_3).$$

$\square$

Definition 3.11 (Isometric embedding)**.** Let $(X, d_X), (Y, d_Y)$ be metric spaces. An *isometric embedding* $\eta : (X, d_X) \to (Y, d_Y)$ is a mapping that satisfies

$$d_X(x_1, x_2) = d_Y(\eta(x_1), \eta(x_2))$$

for all $x_1, x_2 \in X$.

Definition 3.12 (Ball in persistence diagrams)**.** Let $1 \leq p \leq \infty$. Let $D_0 \in \mathrm{Dgm}_p$. The *ball* at the space of persistence diagrams is defined as $B_p(D_0, r) := \{D \in \mathrm{Dgm}_p : w_p(D, D_0) < r\}$.

Theorem 3.13 (Isometric embeeding of metric spaces into persistance diagrams)**.** *Let (X, d) be a separable, bounded metric space. Then there exists an isometric embedding to the space of persistence diagrams $\eta : (X, d) \to (\mathrm{Dgm}_\infty, \omega_\infty)$ such that $\eta(X) \subseteq B(\emptyset, \frac{3c}{c}) \backslash B(\emptyset, c)$.*

Proof. We will follow the procedure followed in [4][Theorem 19]. As (X, d) is bounded, we can let $c > \sup\{d(x, y) : \ x, y \in X\}$. As (X, d) is separable, we can take $\{x_k\}_{k=1}^{\infty}$, a countable, dense subset of (X, d). Consider

$$\begin{aligned} \eta : (X, d) &\to (\mathrm{Dgm}_\infty, \omega_\infty) \\ x &\mapsto \{(2c(k-1), 2ck + d(x, x_k))\}_{k=1}^{\infty} \end{aligned}$$

For any $x \in X$ and $k \in \mathbb{N}$,

$$\begin{aligned} d_\infty((2c(k-1), 2ck + d(x, x_k)), \Delta) &= \frac{2ck + d(x, x_k) - 2c(k-1)}{2} \\ &= c + \frac{d(x, x_k)}{2} \\ &< c + \frac{c}{2} = \frac{3c}{2}. \end{aligned}$$

Because of Proposition 3.9, for every $x \in X$, $\tilde{\omega_\infty}(\eta(x), \emptyset) < \infty$ and η is well defined. Note that

$$\omega_\infty(\eta(x), \emptyset) = \sup_{1 \leq k < \infty} d_\infty((2c(k-1), 2ck + d(x, x_k)), \Delta),$$

so $\eta(x) \in B(\emptyset, \frac{3c}{c}) \backslash B(\emptyset, c)$.

Let $\eta(x)$ and $\eta(y)$ two equivalence classes of $(\mathrm{Dgm}_\infty, \omega_\infty)$. Choose the representative diagrams $D_x : \mathbb{N} \to \mathbb{R}^2_<$ and $D_y : \mathbb{N} \to \mathbb{R}^2_<$ and consider the partial matching $(\mathbb{N}, \mathbb{N}, \mathrm{id}_\mathbb{N})$. With it, for every $k \in \mathbb{N}$, $(2c(k-1), 2ck + d(x, x_k))$ is matched with $(2c(k-1), 2ck + d(y, x_k))$. The Chebyshev distance between those points is

$$\begin{aligned} d_\infty(D_x(k), D_y(k)) &= \max\big\{|2c(k-1) - 2c(k-1)|, \\ &\qquad |2ck + d(x, x_k) - b_y - (2ck + d(y, x_k))|\big\} \\ &= \max\{0, |d(x, x_k) - d(y, x_k)|\} \\ &= |d(x, x_k) - d(y, x_k)|. \end{aligned}$$

Hence, because of the triangle inequality, the cost of this partial matching is

$$\mathrm{cost}_\infty(\mathrm{id}_\mathbb{N}) = \sup_k |d(x, x_k - d(y, x_k)) \leq d(x, y).$$

Since $\{x_k\}_{k=1}^\infty$ is dense, for every $\epsilon > 0$, there exist a $k \in \mathbb{N}$ such that $d(x, x_k) \leq \epsilon$, so

$$\begin{aligned}
|d(x, x_k) - d(y, x_k)| &\geq d(y, x_k) - d(x, x_k) \\
&= d(y, x_k) + d(x, x_k) - d(x, x_k) - d(x, x_k) \\
&\geq d(x, y) - 2d(x, x_k) \\
&> d(x, y) - 2\epsilon.
\end{aligned}$$

Therefore, $\sup_k |d(x, x_k - d(y, x_k)) \geq d(x, y)$ and

$$\mathrm{cost}_\infty(\mathrm{id}_\mathbb{N}) = \sup_k |d(x, x_k - d(y, x_k)) = d(x, y).$$

Suppose $I, J \subseteq \mathbb{N}$ and (I, J, f) is a different partial matching between D_x and D_y. Then there exist a $k \in \mathbb{N}$ such that either $k \notin I$ or $k \in I$ and $f(k) = k \neq k$. If $k \notin I$, then

$$\mathrm{cost}_\infty(f) \geq d_\infty((2c(k-1), 2ck + d(x, x_k)), \Delta) \geq c.$$

If $k \in I$ and $f(k) = k \neq k$, then

$$\mathrm{cost}_\infty(f) \geq ||(2c(k-1), 2ck + d(x, x_k)) - (2c(k'-1), 2ck' + d(x, x_{k'}))||_\infty \geq 2\epsilon.$$

Hence, $\mathrm{cost}_\infty(f) \geq c > d(x, y)$ and $d(x, y) = \omega_\infty(\eta(x), \eta(y))$, proving that η is an isometric embedding of a metric space into the space of persistence diagrams. $\square$

References

[1] A. Figalli and F. Glaudo, *An invitation to Optimal Transport, Wasserstein distances, and gradient flows.* EMS Press, 2020.

[2] Y. Rubner, C. Tomasi, and L. J. Guibas, “The Earth Mover’s Distance as a metric for image retrieval,” *International Journal of Computer Vision*, 2000.

[3] C. R. Givens and R. M. Shortt, “A class of Wasserstein metrics for probability distributions,” *Michigan Technological University*, 1984.

[4] P. Bubenik and A. Wagner, “Embeddings of persistence diagrams into Hilbert spaces,” *Journal of Applied and Computational Topology*, 2020.