**Mathematic Analysis Fundamentals. End of Course Thesis. 2024-2025.**
**Title**: Optimal transport for Topological Data Analysis
**Name**: Gonzalo Ortega Carpintero

# 1 Optimal transport

The contents of this thesis are based on [1] and [2].

Along this text, 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 1.1** (Persistence diagram)**.** Let $I$ be a countable set. A *persistence diagram* is a function $D : I \to \mathbb{R}^2_<$.

**Definition 1.2** (Chebyshev distance)**.** (To do) $d_\infty := \max\{|a_x - b_x|, |a_y - b_y|\}$

**Proposition 1.1.** *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 1.2.** *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 1.3** (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 1.4.** 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

$$\operatorname{cost}_p(f) := \left( \sum_{i\in I_1'} d_\infty(D_1(i), D_2(f(i)))^p + \sum_{i\in I_1\setminus I_1'} d_infty(D_1(i),\Delta)^p + \sum_{i\in I_2\setminus I_2'} d_infty(D_2(i),\Delta)^p \right)^{\frac{1}{p}}.$$

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

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

**Definition 1.5** (p-Wasserstein distance)**.** Let $D_1, D_2$ be persistence diagrams. Let $1 \le p \le \infty$. Define

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

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 $\omega_p$ is called the *p-Wasserstein distance*.

**Definition 1.6** (Bottleneck distance)**.** In the conditions of Definition 1.5, if $p = \infty$, the metric $\omega_\infty$ is called the *bottleneck distance*.

**Proposition 1.3.** *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) □

**Proposition 1.4.** *The space of persistence diagrams with the p-Wasserstein distance* $(\mathrm{Dgm}_p, \omega_p)$ *is indeed a metric space.*

*Proof.* (To do) $\square$

**Definition 1.7** (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 1.8** (Ball)**.** 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 1.1** (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.* 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 $k \in \mathbb{N}$,

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

Because of Proposition 1.3, for every $x \in X$, $\tilde{\omega_\infty}(\eta(x), \emptyset) < \infty$ and $\eta$ 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

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

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 $\eta$ 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] P. Bubenik and A. Wagner, "Embeddings of persistence diagrams into hilbert spaces," 2020.