Report for CSE 5339 2018 — (OTMLSA) Optimal Transport in Machine Learning and Shape Analysis

Report on "Sliced Wasserstein Distance for Persistence Diagram"

Xiao Zha

1 Motivation and Related Work

The tools developed in Topological data analysis are built upon persistence homology theory, and their main output is a descriptor called persistence diagram (PD), which encodes the topology of a space at all scales in the form of a point cloud with multiplicities in the plane $\mathbb{R}^2$. PDs enjoy strong stability properties. However, they do not live in a space naturally endowed with a Hilbert structure and are usually compared with non-Hilbertian distances, such as the bottleneck distance. One approach to tackle this issue is to make use of the "kernel trick" by using a positive-definite kernel in order to map the persistence diagrams into a Hilbert space. A series of recent contributions have proposed kernels for PDs. One class of methods defines implicitly feature maps by focusing on building kernels for PDs. For instance, Reininghaus et al. (2015) use solutions of the heat differential equation in the plane and compare them with the usual $L^2(\mathbb{R}^2)$ dot product. Kusano et al. (2016) handle a persistence diagram as a discrete measure on the plane, and follow by using kernel mean embeddings with Gaussian kernels. Both kernels are provably stable, in the sense that the metric they induce in their respective reproducing kernel Hilbert space (RKHS) is bounded above by the distance between persistence diagrams. More generally, one of the reasons why the derivation of kernels for persistence diagrams is not straightforward is that the natural metrics between persistence diagrams, the *diagram distances* are not negative semi-definite. Indeed, these diagram distances are very similar to the *Wasserstein distance* between probability measures, which is not negative semi-definite. However, a relaxation of this metric called the *Sliced Wasserstein distance* has been shown to be negative semi-definite and was used to derive kernels for probability distributions in Yasuaki. H et al. (2016). In the article "Sliced Wasserstein distance for persistence diagrams", obviously, the authors used the Sliced Wasserstein distance to define a new kernel for persistence diagrams, which is proved to be both stable and discriminative. Specifically, they provide distortion bounds on the Sliced Wasserstein distance that quantify its ability to mimic the diagram distances between persistence diagrams.

2 Background

2.1 Persistent Homology

Persistent homology is an algebraic method for measuring topological features of shapes and functions. Given $f : X \to \mathbb{R}$ as input, persistent homology outputs a planar point set with multiplicities, called the *persistence diagram* of f and denoted by $Dg(f)$. To understand the meaning of each point in this diagram, it suffices to know that, to compute $Dg(f)$, persistent homology considers the family of *sublevel sets* of f, i.e. the sets of the form $f^{-1}((-\infty, t])$ for $t \in \mathbb{R}$, and it records the *topological events* (e.g. creation or merge of a connected component,

creation or filling of a loop, void, etc.) that occur in $f^{-1}((-\infty, t])$ as t ranges from $-\infty$ to $+\infty$. Then, each point $p \in Dg(f)$ represents the lifespan of a particular *topological feature* (connected component, loop, void, etc.), with its creation and destruction times as coordinates.

Distance between persistence diagrams. We now define the *pth diagram distance* between persistence diagrams. Let $p \in \mathbb{N}$ and Dg_1, Dg_2 be two persistence diagrams. Let Γ: $Dg_1 \supseteq A \to B \subseteq Dg_2$ be a *partial bijection* between Dg_1 and Dg_2. Then, for any point $x \in A$, the *p-cost* of x is defined as $c_p(x) = \|x - \Gamma(x)\|_\infty^p$, and for any point $y \in (Dg_1 \sqcup Dg_2) \backslash (A \sqcup B)$, the *p-cost* of y is defined as $c_p^{'}(y) = \|y - \pi_\Delta(y)\|_\infty^p$, where π_Δ is the projection onto the diagonal $\Delta = \{(x, x) : x \in \mathbb{R}\}$. The cost $c_p(\Gamma)$ is defined as : $c_p(\Gamma) = (\sum_x c_p(x) + \sum_y c_p^{'}(y))^{1/p}$. We then define the *pth diagram distance* d_p as the cost of the best partial bijection:

$$d_p(Dg_1, Dg_2) = \inf_\Gamma c_p(\Gamma).$$

In the particular case $p = +\infty$, the cost of Γ is defined as $c(\Gamma) = \max\{\max_x c_1(x) + \max_y c_1^{'}(y)\}$. The corresponding distance d_∞ is often called the bottleneck distance. One can show that $d_p \to d_\infty$ when $p \to +\infty$. A fundamental property of persistence diagrams is their stability with respect to small perturbations of their originating functions. Indeed, the stability theorem asserts that for any $f, g : X \to \mathbb{R}$, we have

$$d_\infty(Dg(f), Dg(g)) \leqslant \|f - g\|_\infty . \tag{1}$$

In practice, persistence diagrams can be used as descriptors for data via the choice of appropriate filtering function f, e.g. distance to the data in the ambient space, eccentricity, curvature, etc.

2.2 Kernel Methods

Positive Definite Kernels. Given a set X, a function $k : X \times X \to \mathbb{R}$ is called a *positive definite kernel* if for all integers n, for all families $x_1, ..., x_n$ of points in X, the matrix $[k(x_i, x_j)]_{i,j}$ is itself positive semi-definite. For brevity, positive definite kernels will be referred as kernels. It is known that kernels generalize scalar products, in the sense that, given a kernel k, there exists a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ and a *feature map* $\phi : X \to \mathcal{H}_k$ such that $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}_k}$. A kernel k also induces a distance d_k on X that can be computed as the Hilbert norm of the difference between two embeddings:

$$d_k^2(x_1, x_2) \stackrel{def.}{=} k(x_1, x_1) + k(x_2, x_2) - 2k(x_1, x_2).$$

Negative Definite and RBF Kernels. A standard way to construct a kernel is to exponentiate the negative of a Euclidean distance. Indeed, the Gaussian kernel for vectors with parameter $\sigma > 0$ does follow that template approach : $k_\sigma(x, y) = exp(-\frac{\|x-y\|^2}{2\sigma^2})$. An important theorem of Berg et al. (1984) (Theorem 3.2.2, p.74) states that such an approach to build kernels, namely setting

$$k_\sigma(x, y) \stackrel{def.}{=} exp(-\frac{f(x, y)}{2\sigma^2})$$

for an arbitrary function f can only yield a valid positive definite kernel for all $\sigma > 0$ if and only if f is a *conditionally negative definite* function, namely that, for all integers n, for all $x_1, ..., x_n \in X$, and for all $a_1, ..., a_n \in \mathbb{R}$ such that $\sum_i a_i = 0$, one has $\sum_{i,j} a_i a_j f(x_i, x_j) \leqslant 0$.

2.3 Wasserstein distance for unnormalized measures on $\mathbb{R}$

The Wasserstein distance is a distance between probability measures. In the article, the authors focus on a variant of that distance: the 1-Wasserstein distance for nonnegative, not necessarily normalized, measures on the real line. Let μ and ν be two nonnegative measures on the real

line such that $|\mu| = \mu(\mathbb{R})$ and $|\nu| = \nu(\mathbb{R})$ are equal to the same number r. Let's define the three following objects:

$$\mathcal{W}(\mu,\, \nu) = \inf_{P \in \Pi(\mu,\, \nu)} \iint_{\mathbb{R}\times\mathbb{R}} |x-y| P(dx, dy) \tag{2}$$

$$\mathcal{Q}_r(\mu,\, \nu) = r \int_{\mathbb{R}} |M^{-1}(x) - N^{-1}(x)| dx \tag{3}$$

$$\mathcal{L}(\mu,\, \nu) = \inf_{f \in 1-Lipschitz} \int_{\mathbb{R}} f(x)[\mu(dx) - \nu(dx)] \tag{4}$$

where $\Pi(\mu, \nu)$ is the set of measures on $\mathbb{R}^2$ with marginals μ and ν, and M^{-1} and N^{-1} the generalized quantile functions of the probability measures μ/r and ν/r respectively.

Prposition 2.1. $\mathcal{W} = \mathcal{Q}_r = \mathcal{L}$. Additionally (i) $\mathcal{Q}_r$ is conditionally negative definite on the space of measures of mass r; (ii) for any three positive measures μ, ν, γ such that $|\mu| = |\nu|$, we have $\mathcal{L}(\mu + \gamma, \nu + \gamma) = \mathcal{L}(\mu, \nu)$. *Proof.* The equality between (2) and (3) is known for probability measures on the real line - see Proposition 2.17 in [36] for instance, and can be trivially generalized to unnormalized measures. The equality between (2) and (4) is due to the well known Kantorovich duality for a distance cost which can also be trivially generalized to unnormalized measures, which proves the main statement of the proposition. The definition of $\mathcal{Q}_r$ shows that the Wasserstein distance is the l_1 norm of $rM^{-1} - rN^{-1}$, and is therefore conditionally negative definite (as the l_1) distance between two direct representations of μ and ν as functions rM^{-1} and rN^{-1}, proving point (i). The second statements is immediate.

3 The Sliced Wasserstein Kernel

3.1 The Sliced Wasserstein Kernel

A new kernel between persistence diagrams, called the *Sliced Wasserstein kernel*, can be defined based in the Sliced Wasserstein metric. The idea underlying this metric is to slice the plane with lines passing through the origin, to project the measures onto these lines where $\mathcal{W}$ is computed, and to integrate those distance over all possible lines.

Definition 3.1. Given $\theta \in \mathbb{R}^2$ with $\|\theta\|_2 = 1$, let $L(\theta)$ denote the line $\{\lambda\theta : \lambda \in \mathbb{R}\}$, and let $\pi_\theta : \mathbb{R}^2 \to L(\theta)$ be the orthogonal projection onto $L(\theta)$. Let Dg_1, Dg_2 be two persistence diagrams, and let $\mu_1^\theta = \sum_{p \in Dg_1} \delta_{\pi_{\theta(p)}}$ and $\mu_{1\Delta}^\theta = \sum_{p \in Dg_1} \delta_{\pi_\theta \circ \pi_\Delta(p)}$, and similarly for μ_2^θ, where π_Θ is the orthogonal projection onto the diagonal. Then, the Sliced Wasserstein distance is defined as:

$$SW(Dg_1, Dg_2) \overset{def.}{=} \frac{1}{2\pi} \int_{\mathcal{S}_1} \mathcal{W}(\mu_1^\theta + \mu_{2\Delta}^\theta,\, \mu_2^\theta + \mu_{1\Delta}^\theta)\, d\theta$$

Note that, by symmetry, once can restrict on the half-circle $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and normalize by π instead of 2π.

Lemma 3.2. Let X be the set of bounded and finite PDs. Then, SW is negative semi-definite on X.

Proof. Let $n \in \mathbb{N}^*, a_1, ..., a_n \in \mathbb{R}$ such that $\sum_i a_i = 0$ and $Dg_1, ..., Dg_n \in X$. Given $1 \leqslant i \leqslant n$, we

let $\tilde{\mu}_i^\theta := \mu_i^\theta + \sum_{q \in Dg_k, k \neq i} \delta_{\pi_\theta \circ \pi_\Delta(q)}$, $\tilde{\mu}_{ij\Delta}^\theta := \sum_{p \in Dg_k, k \neq i,j} \delta_{\pi_\theta \circ \pi_\Delta(p)}$ and $d = \sum_i |Dg_i|$. Then,

$$\begin{aligned}\sum_{i,j} a_i a_j \mathcal{W}(\mu_i^\theta + \mu_{j\Delta}^\theta, \mu_j^\theta + \mu_{i\Delta}^\theta) &= \sum_{i,j} a_i a_j \mathcal{L}(\mu_i^\theta + \mu_{j\Delta}^\theta, \mu_j^\theta + \mu_{i\Delta}^\theta) \\ &= \sum_{i,j} a_i a_j \mathcal{L}(\mu_i^\theta + \mu_{j\Delta}^\theta + \tilde{\mu}_{ij\Delta}^\theta, \mu_j^\theta + \mu_{i\Delta}^\theta + \tilde{\mu}_{ij\Delta}^\theta) \\ &= \sum_{i,j} a_i a_j \mathcal{L}(\tilde{\mu}_i^\theta, \tilde{\mu}_j^\theta) = \sum_{i,j} a_i a_j \mathcal{Q}_d(\tilde{\mu}_i^\theta, \tilde{\mu}_j^\theta) \leqslant 0\end{aligned}$$

The result follows by the linearity of integration.
Hence, the theorem of Berg et al.(1984) allows us to define a valid kernel with:

$$k_{SW}(Dg_1, Dg_2) \overset{def.}{=} exp\,(-\tfrac{SW(Dg_1, Dg_2)}{2\sigma^2})$$

Then, we have the main theoretical result of the article, which states that SW is *equivalent* to d_1.

Theorem 3.3. Let X be the set of bounded PDs with cardinalities bounded by $N \in \mathbb{N}^*$. Let $Dg_1, Dg_2 \in X$. Then, one has:

$$\tfrac{d_1(Dg_1, Dg_2)}{2M} \leqslant SW(Dg_1, Dg_2) \leqslant 2\sqrt{2} d_1(Dg_1, Dg_2)$$

where $M = 1 + 2N(2N-1)$

Proof. Let $s^\theta : Dg_1 \cup \pi_\Delta(Dg_2) \to Dg_2 \cup \pi_\Delta(Dg_1)$ be the one-to-one bijection between $Dg_1 \cup \pi_\Delta(Dg_2)$ and $Dg_2 \cup \pi_\Delta(Dg_1)$ induced by $\mathcal{W}(\mu_1^\theta + \mu_{2\Delta}^\theta, \mu_2^\theta + \mu_{1\Delta}^\theta)$, and let s be the one-to-one bijection between $Dg_1 \cup \pi_\Delta(Dg_2)$ and $Dg_2 \cup \pi_\Delta(Dg_1)$ induced by the partial bijection achieving $d_1(Dg_1, Dg_2)$

Upper bound. Recall that $\|\theta\|_2 = 1$. We have:

$$\begin{aligned}\mathcal{W}(\mu_1^\theta + \mu_{2\Delta}^\theta, \mu_2^\theta + \mu_{1\Delta}^\theta) &= \sum |\langle p - s^\theta(p), \theta \rangle| \\ &\leqslant \sum |\langle p - s(p), \theta \rangle| \leqslant \sum \|p - s(p)\|_2 \\ &\leqslant \sqrt{2} \sum \|p - s(p)\|_\infty \leqslant 2\sqrt{2} d_1(Dg_1, Dg_2),\end{aligned}$$

where the sum is taken over all $p \in Dg_1 \cup \pi_\Delta(Dg_2)$. The upper bound follows by linearity.

Lower bound. The idea is to use the fact that s^θ is a piecewise-constant function of θ, and that it has at most $2 + 2N(2N-1)$ critical values $\Theta_0, ..., \Theta_M$ in $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Indeed, it suffices to look at all θ such that $\langle p_1 - p_2, \theta \rangle = 0$ for some p_1, p_2 in $Dg_1 \cup \pi_\Delta(Dg_2)$ or $Dg_2 \cup \pi_\Delta(Dg_1)$. Then:

$$\begin{aligned}\int_{\Theta_i}^{\Theta_{i+1}} \sum |\langle p - s^\theta(p), \theta \rangle| d\theta &= \sum \left\|p - s^{\Theta_i}(p)\right\|_2 \int_{\Theta_i}^{\Theta_{i+1}} |\cos(\angle(p - s^{\Theta_i}(p), \theta))| d\theta \\ &\geqslant \sum \left\|p - s^{\Theta_i}(p)\right\|_2 \frac{(\Theta_{i+1} - \Theta_i)^2}{2\pi} \\ &\geqslant (\Theta_{i+1} - \Theta_i)^2 \, \frac{d_1(Dg_1, Dg_2)}{2\pi}\end{aligned}$$

where the sum is again taken over all $p \in Dg_1 \cup \pi_\Delta(Dg_2)$ and where the inequality used to lower bound the integral of the cosine is obtained by concavity. The lower bound follows then from the linearity and the Cauchy-Schwarz inequality.

3.2 Computation

Approximate computation. In practice, An algorithm is proposed to approximate k_{SW} in $O(Nlog(N))$ time.

Algorithm 1: Approximate computation of SW

Input: $\mathrm{Dg}_1 = \{p_1^1, \cdots, p_{N_1}^1\}$, $\mathrm{Dg}_2 = \{p_1^2, \cdots, p_{N_2}^2\}$, M.
Add $\pi_\Delta(\mathrm{Dg}_1)$ to Dg_2 and vice-versa.
Let $\mathrm{SW} = 0$; $\theta = -\pi/2$; $s = \pi/M$;
for $i = 1, \cdots, M$ **do**
  Store the products $\langle p_k^1, \theta \rangle$ in an array V_1;
  Store the products $\langle p_k^2, \theta \rangle$ in an array V_2;
  Sort V_1 and V_2 in ascending order;
  $\mathrm{SW} = \mathrm{SW} + s\|V_1 - V_2\|_1$;
  $\theta = \theta + s$;
end for
Output: $(1/\pi)\mathrm{SW}$;

References

[1] Edelsbrunner, H and Harer, J. Persistent homology - a survey. *Contemporary mathematics*, 453: 257-282, 2008

[2] Edelsbrunner, H. and Harer, J. *Computational Topology: an introduction.* AMS Bookstore, 2010.

[3] Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. A Stable Multi-Scale Kernel for Topological Machine Learning. In *Proceedings Conference Computer Vision and Pattern Recognition*, 2015.

[4] Kusano, G., Fukumizu, K., and Hiraoka, Y. Persistence Weighted Gaussian Kernel for Topological Data Analysis. In *Proceedings 33rd International Conference on Machine Learning*, pp. 2004-2013, 2016.

[5] Kusano, G., Fukumizu, K., and Hiraoka, Y. Kernel method for persistence diagrams via kernel embedding and weight factor. *CoRR*, abs/1706.03472, 2017.

[6] Berg, C., Christensen, J., and Ressel, P. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.* Springer, 1984

[7] Carlsson, G. Topology and Data. *Bulletin American Mathematical Society*, 46:255-308, 2009

[8] Yasuaki, H., Takenobu, N., Akihiki, H., Emerson, E., Kaname, M., and Yasumasa, N. Hierarchical structures of amorphous solids characterized by persistent homology. In *Proceedings of the National Academy of Science,* volume 26, 2016