Sliced Wasserstein Kernel for Persistence Diagrams

Mathieu Carriere, Marco Cuturi, Steve Oudot

Xiao Zha

1. Motivation and Related Work

- Persistence diagrams (PDs) play a key role in topological data analysis
- PDs enjoy strong stability properties and are widely used
- 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.
- To in corporate PDs in a convex learning pipeline, several kernels have been proposed with a strong emphasis on the stability of the resulting RKHS (Reproducing Kernel Hilbert Space) distance
- In this article, the authors use the sliced Wasserstein distance to define a new kernel for PDs
- Stable and discriminative

Related Work

- A series of recent contributions have proposed kernels for PDs, falling into two classes
- The first class of methods builds explicit feature maps
- One can compute and sample functions extracted from PDS (Bubenik, 2015; Adams et al., 2017; Robins & Turner, 2016)
- The second class of methods defines implicitly features maps by focusing instead 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

2. Background on TDA and Kernels

2. 1 Persistent Homology

- Persistent Homology is a technique inherited from algebraic topology for computing stable signature on real-valued functions
- Given $f : X \rightarrow \mathbb{R}$ as input, persistent homology outputs a planar point set with multiplicities, called the persistence diagram of f denoted by $Dg\,f$.
- It records the topological events (e.g. creation or merge of a connected component, creation or filling of a loop, void, etc)
- Each point in the persistence diagram represents the lifespan of a particular topological feature, with its creation and destruction times as coordinates

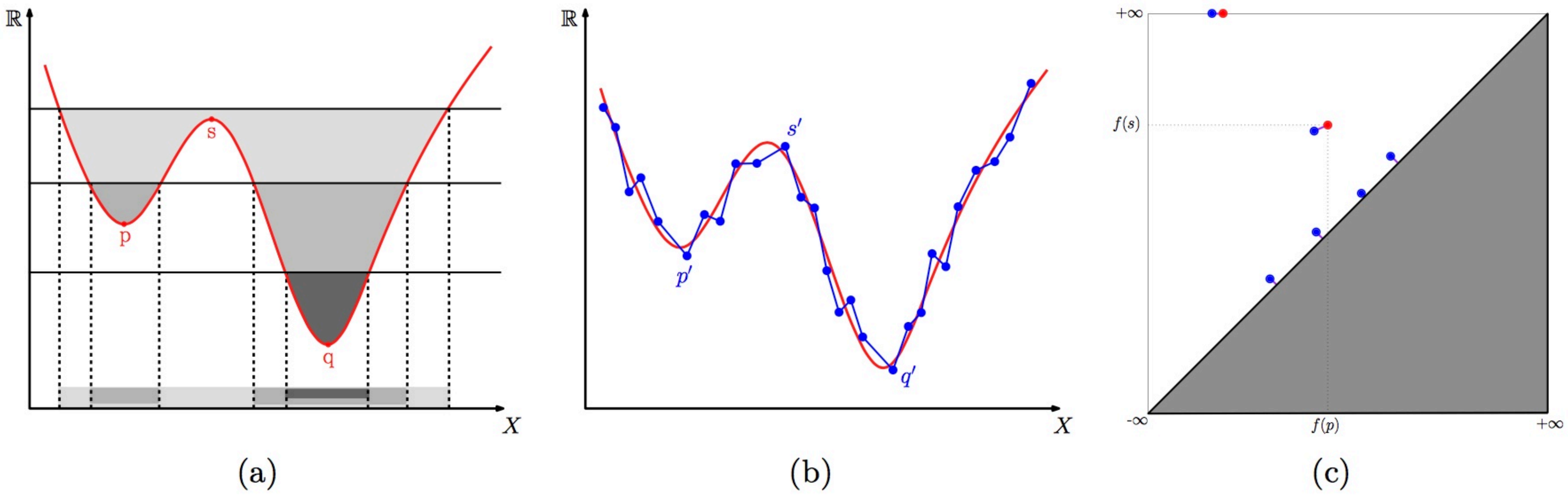

Figure 1: Sketch of persistent homology: (a) the horizontal lines are the boundaries of sublevel sets $f((-\infty,t])$, which are colored in decreasing shades of grey. The vertical dotted lines are the boundaries of their different connected components. For instance, a new connected component is created in the sublevel set $f^{-1}((-\infty,t])$ when $t = f(p)$, and it is merged (destroyed) when $t = f(s)$; its lifespan is represented by a copy of the point with coordinates $(f(p), f(s))$ in the persistence diagram of f (Figure (c)); (b) a piecewise-linear approximation g (blue) of the function f (red) from sampled values; (c) superposition of $\mathrm{Dg}(f)$ (red) and $\mathrm{Dg}(g)$ (blue), showing the partial matching of minimum cost (magenta) between the two persistence diagrams.

Distance between PDs

Let's define the pth diagram distance between PDs. Let $p \in \mathbb{N}$ and D_{g_1}, D_{g_2} be two PDs. Let $\Gamma : D_{g_1} \supseteq A \to B \subseteq D_{g_2}$ be a partial bijection between D_{g_1} and D_{g_1}. 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 $y \in (D_{g_1} \sqcup D_{g_2}) \setminus (A \sqcup B)$, the p-cost of y is defined as $c'_p(y) := \|y - \pi_\Delta(y)\|_\infty^p$, where π_Δ is the projection onto to the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}\}$. The cost $c_p(\Gamma)$ is defined as: $c_p(\Gamma) := \left(\sum_x c_p(x) + \sum_y c'_p(y)\right)^{1/p}$.

We then define the pth diagram distance d_p as the cost of the best partial bijection between the PDs:

$$d_p(\mathrm{Dg}_1, \mathrm{Dg}_2) = \inf_\Gamma c_p(\Gamma).$$

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.

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 $x_1, \dots, x_n$ of points in X, the matrix $\left[k(x_i, x_j)\right]_{i,j}$ is itself positive semi-definite. For brevity, positive definite kernels will be just called kernels in the rest of the paper.

It is known that kernels generalize scalar products, in the sense that, given a kernel k, there exists a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ 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:

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

Negative Definite and RBF Kernels

- A standard way to construct a kernel is to exponentiate the negative of a Euclidean distance.
- Gaussian kernel: $k_\sigma(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$, where $\sigma > 0$.
- Theorem of Berg et al. (1984) (Theorem 3.2.2, p.74) states that such an approach to build kernels, namely setting $k_\sigma(x,y) \stackrel{\text{def}}{=} \exp(-\frac{f(x,y)}{2\sigma^2})$, for an arbitrary function f can only yield a valid positive definite kernel for all $\sigma > 0$ if and only if f is a negative semi-definite function, namely that, for all integers n, $\forall x_1, \dots, x_n \in X$, $\forall a_1, \dots, a_n \in \mathbb{R}^n$ such that $\sum_i a_i = 0$, $\sum_{i,j} a_i a_j f(x_i, x_j) \le 0$.
- In this article, the authors use an approximation of d_1 with the Sliced Wasserstein distance and use it to define a RBF kernel

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

- 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:

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

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

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

where $\prod(\mu,\nu)$ 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

Proposition 2.1

- $\mathcal{W} = \mathcal{Q}_r = \mathcal{L}$. Additionally (i) $\mathcal{Q}_r$ is 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 $\mathcal{L}(\mu + \gamma, \nu + \gamma) = \mathcal{L}(\mu, \nu)$.

The equality between (2) and (3) is only valid for probability measures on the real line. Because the cost function $|\cdot|$ is homogeneous, we see that the scaling factor r can be removed when considering the quantile function and multiplied back. 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.

The definition of $\mathcal{Q}_r$ shows that the Wasserstein distance is the l_1 norm of $rM^{-1} - rN^{-1}$, and is therefore a negative definite kernel (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 statement is immediate.

- An important practical remark:

For two unnormalized uniform empirical measures $\mu = \sum_{i=1}^{n} \delta_{x_i}$ and $\nu = \sum_{i=1}^{n} \delta_{y_i}$ of the same size, with ordered $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$, one has: $\mathcal{W}(\mu, \nu) = \sum_{i=1}^{n} |x_i - y_i| = \|X - Y\|_1$, where $X = (x_1, \dots, x_n) \in \mathbb{R}^n$ and $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$

3. The Sliced Wasserstein Kernel

- 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 distances over all possible lines.

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

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

Since Q_r is negative semi-definite, we can conclude that SW itself is negative semi-definite.

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

Proof. Let $n \in \mathbb{N}^*$, $a_1, \ldots, a_n \in \mathbb{R}$ such that $\sum_i a_i = 0$ and $\mathrm{Dg}_1, \ldots, \mathrm{Dg}_n \in X$. Given $1 \le i \le n$, we let $\tilde{\mu}_i^\theta := \mu_i^\theta + \sum_{q \in \mathrm{Dg}_k, k \ne i} \delta_{\pi_\theta \circ \pi_\Delta(q)}$, $\tilde{\mu}_{ij\Delta}^\theta := \sum_{p \in \mathrm{Dg}_k, k \ne i,j} \delta_{\pi_\theta \circ \pi_\Delta(p)}$ and $d = \sum_i |\mathrm{Dg}_i|$. Then:

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

The result follows by linearity of integration. □

- Hence, the theorem of Berg et al. (1984) allows us to define a valid kernel with:

$$k_{\mathrm{SW}}(\mathrm{Dg}_1, \mathrm{Dg}_2) \stackrel{\text{def.}}{=} \exp\left(-\frac{\mathrm{SW}(\mathrm{Dg}_1, \mathrm{Dg}_2)}{2\sigma^2}\right)$$

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:

$$\frac{d_1(Dg_1, Dg_2)}{2M} \leq SW(Dg_1, Dg_2) \leq 2\sqrt{2} d_1(Dg_1, Dg_2)$$

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

Proof. Let $s^\theta : \mathrm{Dg}_1 \cup \pi_\Delta(\mathrm{Dg}_2) \to \mathrm{Dg}_2 \cup \pi_\Delta(\mathrm{Dg}_1)$ be the one-to-one bijection between $\mathrm{Dg}_1 \cup \pi_\Delta(\mathrm{Dg}_2)$ and $\mathrm{Dg}_2 \cup \pi_\Delta(\mathrm{Dg}_1)$ 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 $\mathrm{Dg}_1 \cup \pi_\Delta(\mathrm{Dg}_2)$ and $\mathrm{Dg}_2 \cup \pi_\Delta(\mathrm{Dg}_1)$ induced by the partial bijection achieving $d_1(\mathrm{Dg}_1, \mathrm{Dg}_2)$.

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

$$\begin{aligned}
&\mathcal{W}(\mu_1^\theta + \mu_{2\Delta}^\theta, \mu_2^\theta + \mu_{1\Delta}^\theta) = \sum |\langle p - s^\theta(p), \theta\rangle| \\
&\leq \sum |\langle p - s(p), \theta\rangle| \leq \sqrt{2} \sum \|p - s(p)\|_\infty \\
&\leq 2\sqrt{2} d_1(\mathrm{Dg}_1, \mathrm{Dg}_2),
\end{aligned}$$

where the sum is taken over all $p \in \mathrm{Dg}_1 \cup \pi_\Delta(\mathrm{Dg}_2)$. 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 $\mathrm{Dg}_1 \cup \pi_\Delta(\mathrm{Dg}_2)$ or $\mathrm{Dg}_2 \cup \pi_\Delta(\mathrm{Dg}_1)$. Then:

$$\begin{aligned}
&\int_{\Theta_i}^{\Theta_{i+1}} \sum |\langle p - s^\theta(p), \theta \rangle| \mathrm{d}\theta \\
&= \sum \|p - s^{\Theta_i}(p)\|_2 \int_{\Theta_i}^{\Theta_{i+1}} |\cos(\angle(p - s^{\Theta_i}(p), \theta))| \mathrm{d}\theta \\
&\geq \sum \|p - s^{\Theta_i}(p)\|_2 (\Theta_{i+1} - \Theta_i)^2 / 2\pi \\
&\geq (\Theta_{i+1} - \Theta_i)^2 d_1(\mathrm{Dg}_1, \mathrm{Dg}_2)/2\pi,
\end{aligned}$$

where the sum is again taken over all $p \in \mathrm{Dg}_1 \cup \pi_\Delta(\mathrm{Dg}_2)$, and where the inequality used to lower bound the integral of the cosine is obtained by concavity. The lower bound follows then from the Cauchy-Schwarz inequality. $\square$

Computation

In practice, the authors propose to approximate k_{SW} in $O(Nlog(N))$ time using Algorithm 1.

Algorithm 1 Computation of SW_M

Input: $\mathrm{Dg}_1 = \{p_1^1 \dots p_{N_1}^1\}$, $\mathrm{Dg}_2 = \{p_1^2 \dots p_{N_2}^2\}$, M.
Add $\pi_\Delta(\mathrm{Dg}_1)$ to Dg_2 and vice-versa.
Let $\mathrm{SW}_M = 0$; $\theta = -\pi/2$; $s = \pi/M$;
for $i = 1 \dots M$ **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;
  $\mathrm{SW}_M = \mathrm{SW}_M + s\|V_1 - V_2\|_1$;
  $\theta = \theta + s$;
end for
Output: $(1/\pi)\mathrm{SW}_M$;

---

4 Experiments

- PSS. The Persistence Scale Space kernel k_{PSS} (Reininghaus et al., 2015)
- PWG. The Persistence Weighted Gaussian kernel k_{PWG} (Kusano et al., 2016; 2017)
- Experiment: 3D shape segmentation. The goal is to produce point classifiers for 3D shapes.
- Use some categories of the mesh segmentation benchmark of Chen et al . (Chen et al., 2009), which contains 3D shapes classified in several categories ("airplane", "human", "ant", …). For each category, the goal is to design a classifier that can assign, to each point in the shape, a label that describes the relative location of that point in the shape. To train classifiers, we compute a PD per point using the geodesic distance function to this point.

Results

Task	k_{PSS}	k_{PWG}	k_{SW}
Human	68.5 ± 2.0	64.2 ± 1.2	$\mathbf{74.0} \pm 0.2$
Airplane	65.4 ± 2.4	61.3 ± 2.9	$\mathbf{72.6} \pm 0.2$
Ant	86.3 ± 1.0	87.4 ± 0.5	$\mathbf{92.3} \pm 0.2$
Bird	67.7 ± 1.8	$\mathbf{72.0} \pm 1.2$	67.0 ± 0.5
FourLeg	67.0 ± 2.5	64.0 ± 0.6	$\mathbf{73.0} \pm 0.4$
Octopus	77.6 ± 1.0	78.6 ± 1.3	$\mathbf{85.2} \pm 0.5$
Fish	76.1 ± 1.6	$\mathbf{79.8} \pm 0.5$	75.0 ± 0.4