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 incorporate 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 $f\left((-\infty, t]\right)$, 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 $f^{-1}\left((-\infty, t]\right)$ 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 $Dg(f)$ (red) and $Dg(g)$ (blue), showing the partial matching of minimum cost (magenta) between the two persistence diagrams.
Distance between PDs

Let’s define the $p$th 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 $c_p(x) := \|x - \Gamma(x)\|_p^p$, and for any point $y \in (D_{g_1} \cup D_{g_2}) \setminus (A \cup B)$, the $p$-cost of $y$ is defined as $c'_p(y) := \|y - \pi_\Delta(y)\|_\infty^p$, where $\pi_\Delta$ is the projection onto to the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}\}$. The cost $c_p(\Gamma)$ is defined as: $c_p(\Gamma) := (\sum_x c_p(x) + \sum_y c'_p(y))^{1/p}$.

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

$$d_p(D_{g_1}, D_{g_2}) = \inf_{\Gamma} c_p(\Gamma).$$

In the particular case $p = +\infty$, the cost of $\Gamma$ is defined as $c(\Gamma) := \max \{ \max_x c_1(x) + \max_y c'_1(y) \}$. The corresponding distance $d_\infty$ 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 $x_1, \ldots, x_n$ 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 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:

$$d_k^2(x_1, x_2) \overset{\text{def}}{=} k(x_1, x_1) + k(x_2, x_2) - 2k(x_1, x_2)$$
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 \( k_\sigma(x, y) \overset{\text{def}}{=} \exp(-\frac{f(x,y)}{2\sigma^2}) \), 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 \( n \), \( \forall x_1, ..., x_n \in X \), \( \forall a_1, ..., a_n \in \mathbb{R}^n \) such that \( \sum_i a_i = 0 \), \( \sum_{i,j} a_i a_j f(x_i, x_j) \leq 0 \).

• 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 $\mu$ and $\nu$ 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:

$$\mathcal{W}(\mu, \nu) = \inf_{P \in \Pi(\mu, \nu)} \iint_{\mathbb{R} \times \mathbb{R}} |x - y| P(dx, dy)$$  \hspace{1cm} (2)

$$Q_r(\mu, \nu) = r \int_{\mathbb{R}} |M^{-1}(x) - N^{-1}(x)| dx$$  \hspace{1cm} (3)

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

where $\Pi(\mu, \nu)$ is the set of measures on $\mathbb{R}^2$ with marginals $\mu$ and $\nu$, and $M^{-1}$ and $N^{-1}$ the generalized quantile functions of the probability measures $\mu/r$ and $\nu/r$ respectively.
Proposition 2.1

• \( \mathcal{W} = Q_r = \mathcal{L} \). Additionally (i) \( Q_r \) is negative definite on the space of measures of mass \( r \); (ii) for any three positive measures \( \mu, \nu, \gamma \) 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 \( Q_r \) 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 \( \mu \) and \( \nu \) 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: \( W(\mu, \nu) = \sum_{i=1}^{n} |x_i - y_i| = \|X - Y\|_1 \), where \( X = (x_1, \ldots, x_n) \in \mathbb{R}^n \) and \( Y = (y_1, \ldots, y_n) \in \mathbb{R}^n \).
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 $\mu^\theta_1 := \sum_{p \in Dg_1} \delta_{\pi_\theta(p)}$ and $\mu^\theta_{1\Delta} := \sum_{p \in Dg_1} \delta_{\pi_\theta \circ \pi_\Delta(p)}$, and similarly for $\mu^\theta_2$, where $\pi_\Delta$ is the orthogonal projection onto the diagonal. Then, the Sliced Wasserstein distance is defined as:

$$SW(Dg_1, Dg_2) \overset{\text{def}}{=} \frac{1}{2\pi} \int_{S^1} \mathcal{W}(\mu^\theta_1 + \mu^\theta_{2\Delta}, \mu^\theta_2 + \mu^\theta_{1\Delta}) d\theta$$

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 $n \in \mathbb{N}^*$, $a_1, ..., a_n \in \mathbb{R}$ such that $\sum_i a_i = 0$ and $Dg_1, ..., Dg_n \in X$. Given $1 \leq i \leq n$, we let $\mu_i^\theta := \mu_i^\theta + \sum_{q \in Dg, k \neq i} \delta_{\pi^\theta \circ \pi^\Delta}(q)$, $\tilde{\mu}_i^\theta \Delta := \sum_{p \in Dg, k \neq i, j} \delta_{\pi^\theta \circ \pi^\Delta}(p)$ and $d = \sum_i |Dg_i|$. Then:

$$
\sum_{i,j} a_i a_j \mathcal{W}(\mu_i^\theta + \mu_j^\theta, \mu_j^\theta + \mu_i^\theta)
= \sum_{i,j} a_i a_j \mathcal{L}(\mu_i^\theta + \mu_j^\theta, \mu_j^\theta + \mu_i^\theta)
= \sum_{i,j} a_i a_j \mathcal{L}(\mu_i^\theta + \mu_j^\theta \Delta, \mu_j^\theta + \mu_i^\theta \Delta)
= \sum_{i,j} a_i a_j \mathcal{L}(\tilde{\mu}_i^\theta \Delta, \tilde{\mu}_j^\theta \Delta) = \sum_{i,j} a_i a_j Q_d(\tilde{\mu}_i^\theta, \tilde{\mu}_j^\theta) \leq 0
$$

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

\[ k_{SW}(Dg_1, Dg_2) \overset{\text{def.}}{=} \exp\left(-\frac{SW(Dg_1, Dg_2)}{2\sigma^2}\right) \]

**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 $s^\theta : Dg_1 \cup \pi_\Delta(Dg_2) \rightarrow Dg_2 \cup \pi_\Delta(Dg_1)$ be the one-to-one bijection between $Dg_1 \cup \pi_\Delta(Dg_2)$ and $Dg_2 \cup \pi_\Delta(Dg_1)$ induced by $\mathcal{W}(\mu_1^\theta + \mu_2^\theta, \mu_2^\theta + \mu_1^\theta)$, 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:

$$\mathcal{W}(\mu_1^\theta + \mu_2^\theta, \mu_2^\theta + \mu_1^\theta) = \sum |\langle p - s^\theta(p), \theta \rangle|$$

$$\leq \sum |\langle p - s(p), \theta \rangle| \leq \sqrt{2} \sum \|p - s(p)\|_\infty$$

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

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 \( s^\theta \) is a piecewise-constant function of \( \theta \), and that it has at most \( 2 + 2N(2N - 1) \) critical values \( \Theta_0, \ldots, \Theta_M \) in \( [-\frac{\pi}{2}, \frac{\pi}{2}] \). Indeed, it suffices to look at all \( \theta \) 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:

\[
\begin{align*}
\int_{\Theta_i}^{\Theta_{i+1}} \sum_{\Theta_i} |\langle p - s^\theta(p), \theta \rangle| \, d\theta \\
= \sum \|p - s^{\Theta_i}(p)\|_2 \int_{\Theta_i}^{\Theta_{i+1}} \left| \cos(\angle(p - s^{\Theta_i}(p), \theta)) \right| \, d\theta \\
\geq \sum \|p - s^{\Theta_i}(p)\|_2 (\Theta_{i+1} - \Theta_i)^2 / 2\pi \\
\geq (\Theta_{i+1} - \Theta_i)^2 d_1(Dg_1, Dg_2) / 2\pi,
\end{align*}
\]

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 Cauchy-Schwarz inequality. \( \Box \)
Computation

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

### Algorithm 1 Computation of $SW_M$

**Input:** $Dg_1 = \{p_1^1 \ldots p_{N_1}^1\}$, $Dg_2 = \{p_1^2 \ldots p_{N_2}^2\}$, $M$.

Add $\pi_\Delta(Dg_1)$ to $Dg_2$ and vice-versa.

Let $SW_M = 0$; $\theta = -\pi/2$; $s = \pi/M$;

for $i = 1 \ldots M$ 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;
- $SW_M = SW_M + s \| V_1 - V_2 \|_1$;
- $\theta = \theta + s$;

end for

**Output:** $(1/\pi)SW_M$;
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.
<table>
<thead>
<tr>
<th>TASK</th>
<th>$k_{PSS}$</th>
<th>$k_{PWG}$</th>
<th>$k_{SW}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMAN</td>
<td>68.5 ± 2.0</td>
<td>64.2 ± 1.2</td>
<td>74.0 ± 0.2</td>
</tr>
<tr>
<td>AIRPLANE</td>
<td>65.4 ± 2.4</td>
<td>61.3 ± 2.9</td>
<td>72.6 ± 0.2</td>
</tr>
<tr>
<td>ANT</td>
<td>86.3 ± 1.0</td>
<td>87.4 ± 0.5</td>
<td>92.3 ± 0.2</td>
</tr>
<tr>
<td>BIRD</td>
<td>67.7 ± 1.8</td>
<td>72.0 ± 1.2</td>
<td>67.0 ± 0.5</td>
</tr>
<tr>
<td>FOURLEG</td>
<td>67.0 ± 2.5</td>
<td>64.0 ± 0.6</td>
<td>73.0 ± 0.4</td>
</tr>
<tr>
<td>OCTOPUS</td>
<td>77.6 ± 1.0</td>
<td>78.6 ± 1.3</td>
<td>85.2 ± 0.5</td>
</tr>
<tr>
<td>FISH</td>
<td>76.1 ± 1.6</td>
<td>79.8 ± 0.5</td>
<td>75.0 ± 0.4</td>
</tr>
</tbody>
</table>