Polynomial Representation for Persistence Diagram - Supplementary Material

Zhichao Wang[1*] Qian Li[2*] Gang Li[3] Guandong Xu[2]
[1]School of Electrical Engineering and Telecommunications, University of New South Wales, Australia
[2]Advanced Analytics Institute, University of Technology Sydney, Australia
[3]School of Information Technology, Deakin University, Geelong, VIC 3216, Australia
zhichao.wang2@unsw.edu.au {qian.li, guandong.xu}@uts.edu.au gang.li@deakin.edu.au

In this supplementary material, we summarize the intuitive ideals and theory of topological data analysis, and also prove Lemma 1 given in the paper.

1. Topological data analysis theory

This section discusses the mathematical theory of topological data analysis. It is not possible to give a thorough introduction of the theory given space constraints. Here, we give some basic notations that are helpful for understanding the intuitive ideals of topological data analysis.

1.1. Basic topological objects

We first introduce the definition of simplices and simplicial complexes that are often considered as basic objects to understand topology theory [1, 2].

Definition 1 (Simplex). *A k-dimensional simplex σ refers to the convex hull of $k+1$ affinely independent vertices $\{x_0, x_2, \cdots, x_k\} \in \mathbb{R}^{k+1}$.*

With the definition of simplex, a higher dimensional generalization graphs can be built, called simplicial complexes. Simplicial complexes contain both topological and combinatorial properties that are very useful for TDA.

Definition 2 (Simplical complex). *A simplicial complex $\mathcal{K}$ is a finite collection of simplices σ such that the face τ of σ also in K, and the intersection $\sigma_1 \cap \sigma_2$ of any two simplices $\sigma_1, \sigma_2 \in K$ is either empty or a face of both σ_1 and σ_2.*

Given a data set or a topological space, there exist many ways to build simplicial complexes. Here, we present the widely used *Rips* complex.

Definition 3 (*Rips* complex). *Given a finite collection of data is $X = \{x_1, \cdots, x_k\}$ in Euclidean space and $r > 0$. The* Rips complex *denoted by $\mathcal{R}_r(X)$ is a simplicial complex. A k-simplex $[x_{i_0}, \cdots, x_{i_k}]$ is in $\mathcal{R}_r(X)$ if $\|x_{i_j} - x_{i_l}\| \leq 2r$ for all $0 \leq j,\ l \leq k$.*

1.2. Homology inference

To infer the homology, we begin with the concept of filtration.

Definition 4 (*Filtration*). *A filtration $\{\mathcal{K}\}_{k=0,\cdots,m}$ of a finite simplicial complex $\mathcal{K}$ is an nested sequence of subcomplexes satisfying*

- $\emptyset = \mathcal{K}_0 \subset \mathcal{K}_1 \cdots \subset \mathcal{K}_m = \mathcal{K}$
- $\mathcal{K}_{k+1} = \mathcal{K}_k \cup s^{k+1}$ *and* s^{k+1} *is a simplex of* $\mathcal{K}_{k+1}$.

Proposition 1. *Let $H_k(\mathcal{K}_i)$ be the k-th dimensional homology group for $\mathcal{K}_i$, the inclusion $\iota_{i,j} : \mathcal{K}_i \hookrightarrow \mathcal{K}_j$ induces a homomorphism on homology groups*

$$\iota_{i,j}^k : H_k(\mathcal{K}_i) \to H_k(\mathcal{K}_j) \tag{1}$$

Applying Proposition 1 to the filtration $\{\mathcal{K}\}_{i=1,\cdots,n}$, we have the following sequence of k-th dimensional chain complex:

$$H_k(\mathcal{K}_0) \to H_k(\mathcal{K}_1) \to \cdots \to H_k(\mathcal{K}_n) \tag{2}$$

Persistence of the k-th homology. The k-th homology group $H_k(\mathcal{K}_i)$ contains a set of homology classes that capture the k-dimensional cycles in complex $\mathcal{K}_i$. A homology class $[\alpha]$ is then said to be born at i, if $[\alpha] \in H_k(\mathcal{K}_i)$ and $[\alpha] \notin H_k(\mathcal{K}_{i-1})$. $[\alpha]$ is born at i dies at j, if j is the smallest index such that the class $[\alpha]$ is supported in the image of $\iota_{i-1,j}^k$. A homology class $[\alpha]$ is then said to be born at $\mathcal{K}_i$, if i is the smallest index such that $[\alpha]$ is nontrivial in $H_k(\mathcal{K}_i)$. The class $[\alpha]$ dies at $\mathcal{K}_j$, if j is the smallest index such that the class $[\alpha]$ is supported in the image of $\iota_{i,j}^*$. The interval between the birth time i and the death time j reflects the persistence of the topology feature represented by $[\alpha]$. The topological features with long lasting through $\{\mathcal{K}\}_{i=1,\cdots,n}$ can be regarded as reliable structures, while ones with small persistence are likely to be noise.

Homology and topological features. Homology characterizes sets based on connected components and holes. Note that 0-dimensional homology represents a connected component and 1-dimensional homology represents a "cycle".

[*] Equal contribution

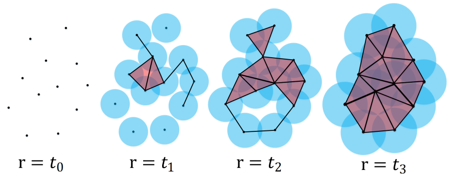

Figure 1. A filtration of *Rips* complex and the 1-skeleton of the complex. This filtration summarizes the births and deaths of one hole as the radius r increase.

2. A toy example of TDA

As stated in the introduction, the most popular topological descriptor in TDA is persistence diagram (PD) . Section 2 provides an example of PD construction using the height function. In addition to a height function, the Rips complex (Definition 3) can be also used to construct PD from data points as shown in Fig. 1. Given a union of balls centered on the set of points in Fig. 1, a *Rips* complex can be constructed by connecting points at scale r. Computing PD consists of two sequential steps: filtration (Definition 4) construction of Rips complex (Definition 3) and computing the birth-death pairs of topological features on the created filtration. By increasing r, we can produce a sequence of *Rips* complexes, defined as a filtration in Definition 4. As the radius grows r, features—such as connected components and holes—appear and disappear. For instance, when $r = t_1$, there are three connected components representing three 0-dimensional topological feature. These three connected components merge at $r = t_2$. Moreover, a 1-dimensional topological feature (the "cycle") appears at $r = t_2$ and disappears when it merges with an connected component at t_3. The hole is born at $r = t_2$ and dies at $r = t_3$, the lifespan of this hole is represented by a point (t_2, t_3) in the PD. The resulting information is encoded by PD where the coordinate of each point is the starting and the end point of the corresponding interval.

PD heavily relies on their stability with respect to perturbations of the data, which promotes varieties of famous TDA methods as reviewed in the related work. In addition to the stability property, another principle motivation of TDA is how to derive more effective topological features from PDs so as to fit standard machine learning methods. Aiming at above two motivations, we proposed a task-adapted polynomial representation for PDs and prove two attractive properties of the proposed method, i.e., stability and linear separability.

3. Proof of Lemma 1

Lemma 1. Let $u = (u_x, u_y) \in \mathcal{D}$ be the point in PD, exponential function $g_{u_x}(z) = e^{-\frac{(u_x - z)^2}{\sigma^2}}$ and weighted function $\omega(u) = \arctan(C(u_y - u_x)^2)$. Let $K = \sqrt{2}\left(1 + \frac{\sqrt{\pi}}{\sigma}\right)$, then we have

$$\int_{-\infty}^{\infty} |\omega(u) g_{u_x}(z) - \omega(v) g_{v_x}(z)|\, dz \le K\|u - v\|_\infty \quad (3)$$

Proof. Note that for technical reasons, the points on the diagonal $L = \{(u_x, u_y) : u_x = u_y\}$ are considered as part of every PD $\mathcal{D}$. Let $\omega(u) g_{u_x}(z) - \omega(v) g_{v_x}(z) = 0$, we have a unique real solution

$$z^* = \frac{v_x^2 - u_x^2 + 2\sigma^2 \ln(\omega(u)/\omega(v))}{2c}$$

Then, we have

$$|\omega(u) \cdot g_{u_x}(z) - \omega(v) \cdot g_{v_x}(z)| =$$
$$\left| \int_{-\infty}^{z^*} \omega(u) g_{u_x}(z) - \omega(v) g_{v_x}(z) + \int_{z^*}^{\infty} \omega(v) g_{v_x}(z) - \omega(u) g_{u_x}(z) dz \right|$$

Consequently, we have the following equation

$$\begin{aligned} &\int_{-\infty}^{\infty} |\omega(u) g_{u_x}(z) - \omega(v) g_{v_x}(z)|\, dz = \\ &2 \int_{-\infty}^{z^*} |\omega(u) g_{u_x}(z) - \omega(v) g_{v_x}(z)| dz \end{aligned} \quad (4)$$

Let $t = \frac{z - u_x}{\sqrt{2}\sigma}$, we then compute

$$\begin{aligned} \int_{-\infty}^{z^*} \omega(u) g_{u_x}(z) dz &= \frac{\omega(u)}{\sqrt{\pi}} \int_{-\infty}^{f(s,u,v)} e^{-t^2} dt \\ &= \frac{\omega(u)}{\sqrt{\pi}} \left(\int_{-\infty}^{0} e^{-t^2} dt + \int_{0}^{f(s,u,v)} e^{-t^2} dt \right) \\ &= \frac{\omega(u)}{2} (1 + \text{erf}(f(s,u,v))) \end{aligned} \quad (5)$$

where

$$\begin{cases} \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \\ f(s,u,v) = \frac{z^* - u_x}{\sqrt{2}\sigma} = \frac{s^2 + 2\sigma^2 \ln(\omega(u)/\omega(v))}{2\sqrt{2}\sigma s}, \quad s = v_x - u_x \end{cases} \quad (6)$$

Let $h(s,u,v) = \frac{-s^2 + 2\sigma^2 \ln(\omega(u)/\omega(v))}{2\sqrt{2}\sigma s}$ and $s = v_x - u_x$, we similarly obtain

$$\int_{-\infty}^{z^*} \omega(v) g_{v_x}(z) dz = \frac{\omega(v)}{2} (1 + \text{erf}(h(s,u,v))) \quad (7)$$

Based on Eq. (5) and Eq. (6), we have

$$\begin{aligned} &\int_{-\infty}^{\infty} |\omega(u) g_{u_x}(z) - \omega(v) g_{v_x}(z)|\, dz \\ &= |\omega(u) \cdot \text{erf}(f(s,u,v)) - \omega(v) \cdot \text{erf}(h(s,u,v))| \end{aligned} \quad (8)$$

Let $H(s) = |\omega(u) \cdot \text{erf}(f(s,u,v)) - \omega(v) \cdot \text{erf}(h(s,u,v))|$, we have $H(0) = |\omega(u) - \omega(v)|$ according to Eq. (6). By computing the roots of the second-order derivative $H(s)''$, we have $\sup H(s)' = \sqrt{\frac{2}{\pi}} \frac{\min\{w(u), w(v)\}}{\delta}$ The primitive function of $H(s)'$ satisfies

$$H(s) \leq |\omega(u) - \omega(v)| + \sqrt{\frac{2}{\pi}} \frac{\min\{w(u), w(v)\}}{\delta} |s|$$

Based on the definition of s and $\omega(\cdot)$, we have Based on the definition of s and $\omega(\cdot)$, we have

$$\begin{aligned}
&|\omega(u) \cdot \text{erf}(f(s,u,v)) - \omega(v) \cdot \text{erf}(h(s,u,v))| \\
&\leq |\omega(u) - \omega(v)| + \sqrt{\frac{2}{\pi}} \frac{\min\{\omega(u), \omega(v)\}}{\sigma} |u_x - v_x| \\
&\leq |\omega(u) - \omega(v)| + \sqrt{\frac{\pi}{2\sigma^2}} |u_x - v_x| \\
&\leq |\omega(u) - \omega(v)| + \sqrt{\frac{\pi}{2\sigma^2}} \|u - v\|_2 \\
&\leq \left(|\nabla\omega| + \sqrt{\frac{\pi}{2\sigma^2}} \right) \|u - v\|_2
\end{aligned} \tag{9}$$

$$\begin{aligned}
&\text{since } \|\cdot\|_2 \leq \sqrt{2}\|\cdot\|_\infty \in \mathbb{R}^2, \text{ then} \\
&\leq \left(\sqrt{2}|\nabla\omega| + \sqrt{\frac{2\pi}{\sigma^2}} \right) \|u - v\|_\infty \\
&\leq \left(\sqrt{2} + \sqrt{\frac{2\pi}{\sigma^2}} \right) \|u - v\|_\infty \qquad \text{since } |\nabla\omega| < 1 \\
&= \sqrt{2} \left(1 + \frac{\sqrt{\pi}}{\sigma} \right) \|u - v\|_\infty
\end{aligned}$$

Finally, Lemma 3.1 follows

$$\begin{aligned}
&\int_{-\infty}^{\infty} |\omega(u) g_{u_x}(z) - \omega(v) g_{v_x}(z)| \, dz \\
&\leq \sqrt{2} \left(1 + \frac{\sqrt{\pi}}{\sigma} \right) \|u - v\|_\infty
\end{aligned} \tag{10}$$

With Lemma 1 in mind, we turn to prove the persistence vector is stable w.r.t. *1-Wasserstein distance* between PDs. □

References

[1] H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[2] M. Ferri. Persistent topology for natural data analysis—a survey. In *Towards Integrative Machine Learning and Knowledge Extraction*, pages 117–133. Springer, 2017.