Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Kernel method for persistence diagrams via kernel embedding and weight factor

Genki Kusano GENKI.KUSANO.R5@DC.TOHOKU.AC.JP
Graduate School of Science
Tohoku University
Sendai, Miyagi 980-8578, Japan

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics
Tachikawa, Tokyo 190-8562, Japan

Yasuaki Hiraoka HIRAOKA@TOHOKU.AC.JP
Advanced Institute for Materials Research
Tohoku University
Sendai, Miyagi 980-0811, Japan

Editor:

Abstract

Topological data analysis (TDA) is an emerging mathematical concept for characterizing shapes in complicated data. In TDA, persistence diagrams are widely recognized as a useful descriptor of data, distinguishing robust and noisy topological properties. This paper introduces a kernel method for persistence diagrams to develop a statistical framework in TDA. The proposed kernel is stable under perturbation of data, enables one to explicitly control the effect of persistence by a weight function, and allows an efficient and accurate approximate computation. The method is applied into practical data on granular systems, oxide glasses and proteins, showing advantages of our method compared to other relevant methods for persistence diagrams.

Keywords: Topological data analysis, persistence diagrams, kernel method, kernel embedding, persistence weighted Gaussian kernel

1. Introduction

Recent years have witnessed an increasing interest in utilizing methods of algebraic topology for statistical data analysis. In terms of algebraic topology, conventional clustering methods are regarded as charactering 0-dimensional topological features which mean connected components of data. Furthermore, higher dimensional topological features also represent informative shape of data, such as rings (1-dimension) and cavities (2-dimension). The research analyzing these topological features in data is called *topological data analysis* (TDA) (Carlsson, 2009), which has been successfully applied to various areas including information science (Carlsson et al., 2008; de Silva and Ghrist, 2007), biology (Kasson et al., 2007; Xia and Wei, 2014), brain science (Lee et al., 2011; Petri et al., 2014; Singh et al., 2008), biochemistry (Gameiro et al., 2015), material science (Hiraoka et al., 2016; Nakamura et al., 2015; Saadatfar et al., 2017), and so on. In many of these applications, data have compli-

cated geometric structures, and thus it is important to extract informative topological features from the data.

A *persistent homology* (Edelsbrunner et al., 2002), which is a key mathematical tool in TDA, extracts robust topological information from data, and it has a compact expression called a *persistence diagram*. While it is applied to various problems such as the ones listed above, statistical or machine learning methods for analysis on persistence diagrams are still limited. In TDA, analysts often elaborate only single persistence diagram and, in particular, methods for handling many persistence diagrams, which can contain randomness from the data, are at the beginning stage (see the end of this section for related works). Hence, developing a framework of statistical data analysis on persistence diagrams is a significant issue for further success of TDA and, to this goal, this paper discusses kernel methods for persistence diagrams.

1.1 Topological descriptor

In order to provide some intuitions for the persistent homology, let us consider a typical way of constructing persistent homology from data points in a Euclidean space, assuming that the point set lies on a submanifold. The aim is to make inference on the topology of the underlying manifold from finite data points. We consider the r-balls (balls with radius r) to recover the topology of the manifold, as popularly employed in constructing an r-neighbor graph in many manifold learning algorithms. While it is expected that, with an appropriate choice of r, the r-ball model can represent the underlying topological structures of the manifold, it is also known that the result is sensitive to the choice of r. If r is too small (resp. large), the union of r-balls consists simply of the disjoint r-balls (resp. a contractible space). Then, by considering not one specific r but all r, the persistent homology gives robust topological features of the point set.

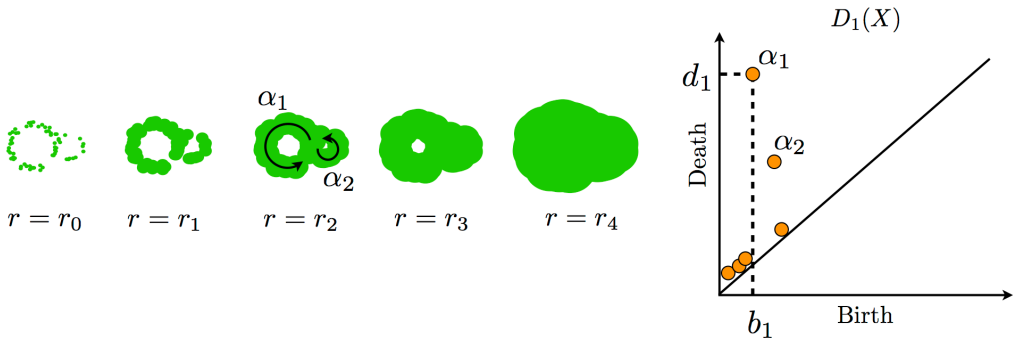

Figure 1: Unions of r-balls at data points (left) and its 1-st persistence diagram (right). The point (b_1, d_1) in the persistence diagram represents the ring α_1, which appears at $r = b_1$ and disappears at $r = d_1$. The noisy rings are plotted as the points close to the diagonal.

As a useful representation of persistent homology, a persistence diagram is often used in topological data analysis. The persistence diagram is given in the form of a multiset $D = \{(b_i, d_i) \in \mathbb{R}^2 \mid i \in I,\ b_i < d_i\}$ (Figure 1). Every point $(b_i, d_i) \in D$, called a *generator* of the persistent homology, represents a topological property (e.g., connected components, rings, cavities, etc.) which appears at $r = b_i$ and disappears at $r = d_i$ in the ball model. Then, the *persistence* $d_i - b_i$ of the generator shows the robustness of the topological property under the radius parameter. A generator with large persistence can be regarded as a reliable structure, while that with small persistence (points close to the diagonal) is likely to be a structure caused by noise. In this way, persistence diagrams encode topological and geometric information of data points.

1.2 Contribution

Since a persistence diagram is a point set of variable size, it is not straightforward to apply standard methods of statistical data analysis, which typically assume vectorial data. To vectorize persistence diagrams, we employ the framework of kernel embedding of (probability and more general) measures into reproducing kernel Hilbert spaces (RKHS). This framework has recently been developed and leading various new methods for nonparametric inference (Muandet et al., 2017; Smola et al., 2007; Song et al., 2013; Lopez-Paz et al., 2015; Szabó et al., 2016). It is known (Sriperumbudur et al., 2011) that, with an appropriate choice of kernels, a signed Radon measure can be uniquely represented by the Bochner integral of the feature vectors with respect to the measure. Since a persistence diagram can be regarded as a sum of Dirac delta measures, it can be embedded into an RKHS by the Bochner integral. Once such a vector representation is obtained, we can introduce any kernel methods for persistence diagrams systematically (see Figure 2).

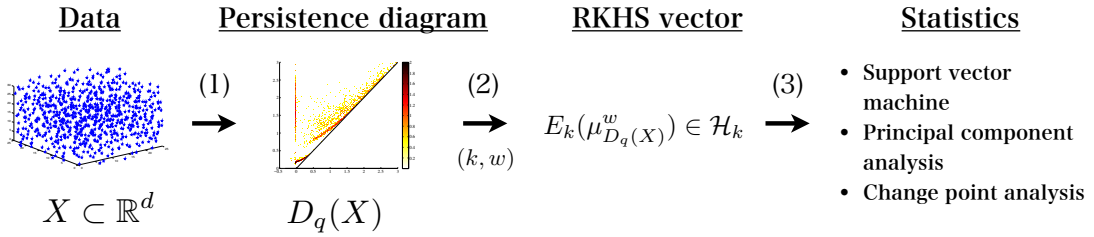

Figure 2: (1) A data set X is transformed into a persistence diagram $D_q(X)$ (Section 2.1). (2) The persistence diagram $D_q(X)$ is mapped to an RKHS vector $E_k(\mu^w_{D_q(X)})$, where k is a positive definite kernel and w is a weight function controlling the effect of persistence (Section 3.1). (3) Statistical methods are applied to those vector representations of persistence diagrams (Section 4).

Furthermore, since each generator in a persistence diagram is equipped with a persistence which indicates a robustness of the topological features, we will utilize it as a weight on the generator. For embedding persistence diagrams in an RKHS, we propose a useful class of positive definite kernels, called *persistence weighted Gaussian kernel* (PWGK). The

advantages of this kernel are as follows: (i) We can explicitly control the effect of persistence by a weight function, and hence discount the noisy generators appropriately in statistical analysis. (ii) As a theoretical contribution, the distance defined by the RKHS norm for the PWGK satisfies the stability property, which ensures the continuity from data to the vector representation of the persistence diagram. (iii) The PWGK allows efficient computation by using the random Fourier features (Rahimi and Recht, 2007), and thus it is applicable to persistence diagrams with a large number of generators.

We demonstrate the performance of the proposed kernel method with synthesized and real-world data, including granular systems (taken by X-ray Computed Tomography on granular experiments), oxide glasses (taken by molecular dynamics simulations) and protein datasets (taken by NMR and X-ray crystallography experiments). We remark that these real-world problems have physical and biochemical significance in their own right, as detailed in Section 4.

1.3 Related works

There are already some relevant works on statistical approaches to persistence diagrams. Some studies discuss how to transform a persistence diagram to a vector (Adams et al., 2017; Bubenik, 2015; Cang et al., 2015; Carrière et al., 2015; Donatini et al., 1998; Reininghaus et al., 2015; Robins and Turner, 2016). In these methods, a transformed vector is typically expressed in a Euclidean space $\mathbb{R}^k$ or a function space L^p, and simple and ad-hoc summary statistics like means and variances are used for data analysis such as principal component analysis (PCA) and support vector machines (SVMs). In this paper, we will compare the performance among the PWGK, the persistence scale-space kernel (Reininghaus et al., 2015), the persistence landscape (Bubenik, 2015), the persistence image (Adams et al., 2017), and the molecular topological fingerprint (Cang et al., 2015) in several machine learning tasks. Furthermore, we show that our vectorization is a generalization of the persistence scale-space kernel and the persistence image although the constructions are different. We also remark that there are some works discussing statistical properties of persistence diagrams for random data points: Chazal et al. (2015) show convergence rates of persistence diagram estimation, and Fasy et al. (2014) discuss confidence sets in a persistence diagram. These works consider a different but important direction to the statistical methods for persistence diagrams.

The remaining of this paper is organized as follows: In Section 2, we review some basics on persistence diagrams and kernel embedding methods. In Section 3, the PWGK is proposed, and some theoretical and computational issues are discussed. Section 4 shows experimental results and compares the proposed kernel method with other methods.

This paper is an extended version of our ICML paper (Kusano et al., 2016). The difference from this conference version is as follows: (i) Comparisons with other relevant methods, in particular, persistence landscapes and persistence images, have been added to this version. (ii) New experimental results in comparison with other relevant methods. (iii) The theoretical results have been generalized to a class of kernels that satisfy the assumption (K) and weight functions that satisfy assumption (W1) (Propositions 8 and 10). The proofs have been modified accordingly, while the basic line of ideas are the same.

2. Backgrounds

We review the concepts of persistence diagrams and kernel methods. For readers who are not familiar with algebraic topology, we give a brief summary in Appendix A. See also Hatcher (2002) as an accessible introduction to algebraic topology.

2.1 Persistence diagram

In order to define a persistence diagram, we transform a data set X into a filtration $\mathbb{F}\mathrm{ilt}(X)$ and compute its persistent homology $H_q(\mathbb{F}\mathrm{ilt}(X))$. In this section, we will first introduce this mathematical framework of persistence diagrams. Then, by using a ball model filtration, we will intuitively explain geometrical meanings of persistence diagrams. The ball model filtrations can be generalized toward two constructions using Čech complexes and sub-level sets. The former construction is useful for computations of persistence diagrams and the latter is useful to discuss theoretical properties.

2.1.1 Mathematical framework of persistence diagrams

Let K be a coefficient field of homology[1]. Let $\mathbb{F}\mathrm{ilt} = \{F_a \mid a \in \mathbb{R}\}$ be a (right continuous) *filtration* of simplicial complexes, i.e., F_a is a subcomplex of F_b for $a \leq b$ and $F_a = \bigcap_{a<b} F_b$. Alternatively, $\mathbb{F}\mathrm{ilt}$ may be a filtration of topological spaces: in this case F_a is a subspace of F_b with the same condition as above. For $a \leq b$, the K-linear map induced from the inclusion $F_a \hookrightarrow F_b$ is denoted by $\rho_a^b : H_q(F_a) \to H_q(F_b)$, where $H_q(F_a)$ is the q-th homology of F_a. The q-th *persistent homology* $H_q(\mathbb{F}\mathrm{ilt}) = (H_q(F_a), \rho_a^b)$ of $\mathbb{F}\mathrm{ilt}$ is defined by the family of homology $\{H_q(F_a) \mid a \in \mathbb{R}\}$ and the induced linear maps $\{\rho_a^b \mid a \leq b\}$.

A *homological critical value* of $H_q(\mathbb{F}\mathrm{ilt})$ is the number $a \in \mathbb{R}$ such that the linear map $\rho_{a-\varepsilon}^{a+\varepsilon} : H_q(F_{a-\varepsilon}) \to H_q(F_{a+\varepsilon})$ is not isomorphic for any sufficiently small $\varepsilon > 0$. The persistent homology $H_q(\mathbb{F}\mathrm{ilt})$ is called *tame* if $\dim_K H_q(F_a) < \infty$ for any $a \in \mathbb{R}$ and the number of homological critical values is finite. A tame persistent homology $H_q(\mathbb{F}\mathrm{ilt})$ has a nice decomposition property:

Theorem 1 (Zomorodian and Carlsson (2005)) *A tame persistent homology can be uniquely expressed*[2] *by*

$$H_q(\mathbb{F}\mathrm{ilt}) \cong \bigoplus_{i \in I} \mathbb{I}[b_i, d_i], \tag{1}$$

where $\mathbb{I}[b_i, d_i] = (U_a, \iota_a^b)$ *consists of a family of* K*-vector spaces*

$$U_a = \begin{cases} K, & b_i \leq a < d_i \\ 0, & \text{otherwise} \end{cases},$$

and $\iota_a^b = \mathrm{id}_K$ *for* $b_i \leq a \leq b < d_i$.

1. In this setting, all homology are K-vector spaces. You may simply consider the case $K = \mathbb{R}$, but the theory is built with an arbitrary field.
2. To be more precise, a persistent homology can be seen as an object of the functor category from the poset category defined by $(\mathbb{R}, \leq)$ to the category of finite dimensional vector spaces. The symbols $\cong$ and $\oplus$ represent the isomorphism and the direct sum in the functor category. It is far beyond the scope of this paper to provide precise definitions of these notions. Interested readers can see Bubenik and Scott (2014) for more details.

Each summand $\mathbb{I}[b_i, d_i]$ means a topological feature in $\mathbb{F}\mathrm{ilt}$ that appears at $a = b_i$ and disappears at $a = d_i$. The birth-death pair $x = (b_i, d_i)$ is called a *generator* of the persistent homology, and $\mathrm{pers}(x) := d_i - b_i$ a *persistence* of x. We note that, when $\dim_K H_q(F_a) \neq 0$ for any $a < 0$ (resp. for any $a > 0$), the decomposition (1) should be understood in the sense that some b_i takes the value $-\infty$ (resp. $d_i = \infty$), where $-\infty, \infty$ are the elements in the extended real $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Through the decomposition in Theorem 1, a persistent homology $H_q(\mathbb{F}\mathrm{ilt})$, which is an algebraic object and is not suitable to be analyzed by statistical methods, is transformed into a multi-set[3] of 2-dimensional vectors

$$D_q(\mathbb{F}\mathrm{ilt}) = \left\{ (b_i, d_i) \in \overline{\mathbb{R}}^2 \;\middle|\; i \in I \right\}$$

and we call it the *persistence diagram* of $\mathbb{F}\mathrm{ilt}$.

In this paper, we assume that all persistence diagrams have finite cardinality because a tame persistent homology defines a finite persistence diagram. Moreover, we also assume that all birth-death pairs are bounded[4], that is, all elements in a persistence diagram take neither ∞ nor $-\infty$. In the following, we also use abstract persistence diagrams (denoted by D or E) given by finite multi-sets above the diagonal $\mathbb{R}^2_{\mathrm{ad}} := \{(b, d) \in \mathbb{R}^2 \mid b < d\}$.

2.1.2 Ball model filtrations

The example used in Figure 1 can be expressed as follows. Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a finite subset in a metric space (M, d_M) and $X_a := \bigcup_{i=1}^n B(\boldsymbol{x}_i; a)$ be a union of balls $B(\boldsymbol{x}_i; a) = \{\boldsymbol{x} \in M \mid d_M(\boldsymbol{x}_i, \boldsymbol{x}) \leq a\}$ with radius $a \geq 0$. For convenience, let $X_a := \emptyset$ $(a < 0)$. Since $\mathbb{X} = \{X_a \mid a \in \mathbb{R}\}$ is a right-continuous filtration of topological spaces and X is a finite set, $H_q(\mathbb{X})$ is tame and the persistence diagram $D_q(\mathbb{X})$ is well-defined. For notational simplicity, the persistence diagram of this ball model filtration is denoted by $D_q(X)$.

We remark that, in this model, there is only one generator in $D_0(X)$ that does not disappear in the filtration; its lifetime is ∞. From now on, we deal with $D_0(X)$ by removing this infinite lifetime generator[5]. Let $\mathrm{diam}(X)$ be the diameter of X defined by $\max_{\boldsymbol{x}_i, \boldsymbol{x}_j \in X} d_M(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Then, all generators appear after $a = 0$ and disappear before $a = \mathrm{diam}(X)$ because $X_{\mathrm{diam}(X)}$ becomes a contractible space. Thus, for any dimension q, all birth-death pairs of $D_q(X)$ have finite values.

2.1.3 Geometric complexes

We review some standard methods of constructing a filtration from finite sets in a metric space. See also Chazal et al. (2014) for more details.

Let (M, d_M) be a metric space and $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a finite subset in M. For a fixed $a \geq 0$, we form a q-simplex $[\boldsymbol{x}_{i_0} \cdots \boldsymbol{x}_{i_q}]$ as a subset $\{\boldsymbol{x}_{i_0}, \ldots, \boldsymbol{x}_{i_q}\}$ of X whenever there exists $\bar{\boldsymbol{x}} \in M$ such that $d_M(\boldsymbol{x}_{i_j}, \bar{\boldsymbol{x}}) \leq a$ for all $j = 0, \ldots, q$, or equivalently, $\cap_{j=0}^q B(\boldsymbol{x}_{i_j}; a) \neq \emptyset$. The set of these simplices forms a simplicial complex, called the *Čech complex* of X with parameter a, denoted by $\check{\mathrm{C}}\mathrm{ech}(X; a)$. For $a < 0$, we define $\check{\mathrm{C}}\mathrm{ech}(X; a)$ as an empty set. Since there is a natural inclusion $\check{\mathrm{C}}\mathrm{ech}(X; a) \hookrightarrow \check{\mathrm{C}}\mathrm{ech}(X; b)$ whenever $a \leq b$, $\check{\mathbb{C}}\mathrm{ech}(X) =$

3. A *multi-set* is a set with multiplicity of each point. We regard a persistence diagram as a multi-set, since several generators can have the same birth-death pairs.
4. This assumption will be justified in Section 2.1.2.
5. This is called the *reduced persistence diagram*.

$\{\check{\mathrm{C}}\mathrm{ech}(X;a) \mid a \in \mathbb{R}\}$ is a filtration. When M is a subspace of $\mathbb{R}^d$, from the nerve lemma (Hatcher, 2002), it is known that the topology of $\check{\mathrm{C}}\mathrm{ech}(X;a)$ is the same[6] as X_a (Figure 3), and hence $D_q(\check{\mathbb{C}}\mathrm{ech}(X)) = D_q(X)$.

As a similar concept to the Čech complex, the Rips complex (or Vietoris-Rips complex) is also often used in TDA. While the Rips complex gives different topology from the Čech complex, it can be computed much more efficiently; the Rips complex needs only pairwise distances, while the Čech complex needs all the $(q+1)$-combinations among n points for q-th homology, which easily becomes infeasible for large n. For a fixed $a \geq 0$, we form a q-simplex $[\boldsymbol{x}_{i_0} \cdots \boldsymbol{x}_{i_q}]$ as a subset $\{\boldsymbol{x}_{i_0}, \ldots, \boldsymbol{x}_{i_q}\}$ of X that satisfies $d_M(\boldsymbol{x}_{i_j}, \boldsymbol{x}_{i_k}) \leq 2a$ for all $j, k = 0, \ldots, q$. The set of these simplices forms a simplicial complex, called the *Rips complex* of X with parameter a, denoted by $\mathrm{Rips}(X;a)$. Similarly, the Rips complex also forms a filtration $\mathbb{R}\mathrm{ips}(X)$. In general, $D_q(\mathbb{R}\mathrm{ips}(X))$ is not the same as $D_q(X)$ (see Figure 3). In experiments in this paper, all persistence diagrams are computed by a ball model filtration, which is equivalent to the Čech complex filtration, and we do not use the Rips complex filtration. We remark that, however, there are also applications of Rips complexes (e.g., sensor networks (de Silva and Ghrist, 2007)), and our kernel method and stability results shown in Proposition 8 and Proposition 10 can be applied not only the ball model filtration but also any filtrations including the Rips complex filtration.

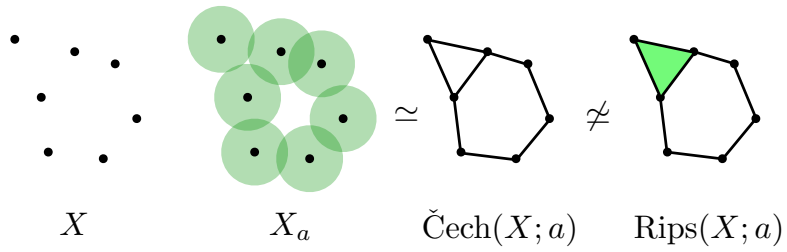

Figure 3: A point set X, the union of balls X_a, the Čech complex $\check{\mathrm{C}}\mathrm{ech}(X;a)$ and the Rips complex $\mathrm{Rips}(X;a)$. There are two rings in X_a and $\check{\mathrm{C}}\mathrm{ech}(X;a)$. However, $\mathrm{Rips}(X;a)$ has only one ring because there is a 2-simplex.

2.1.4 sub-level sets

Another popular way of constructing a filtration is to use sub-level sets. This is useful when data is given in the from of a function like a gray-scale image on a two dimensional region. Let M be a topological space and $f : M \to \mathbb{R}$ be a continuous map. Then, we define a *sub-level set* by $\mathrm{Sub}(f;a) := f^{-1}((-\infty, a])$ for $a \in \mathbb{R}$ and its filtration by $\mathbb{S}\mathrm{ub}(f) := \{\mathrm{Sub}(f;a) \mid a \in \mathbb{R}\}$. Here, $f : M \to \mathbb{R}$ is said to be *tame* if $H_q(\mathbb{S}\mathrm{ub}(f))$ is tame.

For a finite set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in a metric space (M, d_M), we define the distance function $\mathrm{dist}_X : M \to \mathbb{R}$ by

$$\mathrm{dist}_X(\boldsymbol{x}) := \min_{\boldsymbol{x}_i \in X} d_M(\boldsymbol{x}, \boldsymbol{x}_i).$$

6. Precisely, they are *homotopy equivalent*.

Then, we can see $\mathrm{Sub}(\mathrm{dist}_X; a) = \bigcup_{\boldsymbol{x}_i \in X} B(\boldsymbol{x}_i; a)$ and $D_q(\mathbb{S}\mathrm{ub}(\mathrm{dist}_X)) = D_q(X)$. This means that the ball model is a special case of the sub-level set, and the Čech complex and the sub-level set with the distance function dist_X give the same persistence diagram.

2.2 Stability of persistence diagrams

When we consider data analysis based on persistence diagrams, it is useful to introduce a distance measure among persistence diagrams for describing their relations. In introducing a distance measure, it is desirable that, as a representation of data, the mapping from data to a persistence diagram is continuous with respect to the distance. In many cases, data involve noise or stochasticity, and thus the persistence diagrams should be stable under perturbation of data.

The *bottleneck distance* d_{W_∞} between two persistence diagrams D and E is defined by

$$d_{\mathrm{W}_\infty}(D, E) := \inf_{\gamma} \sup_{x \in D \cup \Delta} \|x - \gamma(x)\|_\infty ,$$

where $\Delta := \{(a, a) \mid a \in \mathbb{R}\}$ is the diagonal set with infinite multiplicity and γ ranges over all multi-bijections[7] from $D \cup \Delta$ to $E \cup \Delta$. Here, for $z = (z_1, z_2) \in \mathbb{R}^2$, $\|z\|_\infty$ denotes $\max\{|z_1|, |z_2|\}$. We note that there always exists such a multi-bijection γ because the cardinalities of $D \cup \Delta$ and $E \cup \Delta$ are equal by considering the diagonal set Δ with infinite multiplicity. For sets X and Y in a metric space (M, d_M), let us recall the *Hausdorff distance* d_{H} given by

$$d_{\mathrm{H}}(X, Y) := \max \left\{ \sup_{\boldsymbol{x} \in X} \inf_{\boldsymbol{y} \in Y} d_M(\boldsymbol{x}, \boldsymbol{y}), \sup_{\boldsymbol{y} \in Y} \inf_{\boldsymbol{x} \in X} d_M(\boldsymbol{x}, \boldsymbol{y}) \right\}.$$

Then, the bottleneck distance satisfies the following stability property.

Proposition 2 (Chazal et al. (2014); Cohen-Steiner et al. (2007)) *Let X and Y be finite subsets in a metric space (M, d_M). Then the persistence diagrams satisfy*

$$d_{\mathrm{W}_\infty}(D_q(X), D_q(Y)) \leq d_{\mathrm{H}}(X, Y).$$

Proposition 2 provides a geometric intuition of the stability of persistence diagrams. Assume that two point sets X and Y are close to each other with $\varepsilon = d_{\mathrm{H}}(X, Y)$. If there is a generator $(b, d) \in D_q(Y)$, then we can find at least one generator in X which is born in $(b - \varepsilon, b + \varepsilon)$ and dies in $(d - \varepsilon, d + \varepsilon)$ (see Figure 4). Thus, the stability guarantees the similarity of two persistence diagrams, and hence we can infer the true topological features from the persistence diagrams given by noisy observation (see also Fasy et al. (2014)).

For $1 \leq p < \infty$, the *p-Wasserstein distance* d_{W_p}, which is also used as a distance between two persistence diagrams D and E, is defined by

$$d_{\mathrm{W}_p}(D, E) = \inf_{\gamma} \left(\sum_{x \in D \cup \Delta} \|x - \gamma(x)\|_\infty^p \right)^{\frac{1}{p}},$$

where γ ranges over all multi-bijections from $D \cup \Delta$ to $E \cup \Delta$. Here, we define the *degree-p total persistence* of D by $\mathrm{Pers}_p(D) := \sum_{x \in D} \mathrm{pers}(x)^p$ for $1 \leq p < \infty$.

7. A *multi-bijection* is a bijective map between two multi-sets counted with their multiplicity.

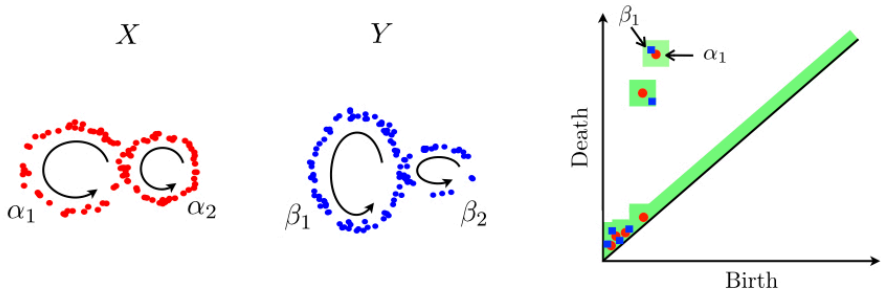

Figure 4: Two point sets X and Y (left) and their persistence diagrams (right). The green region is an ε-neighborhood of $D_q(Y)$ and all generators in $D_q(X)$ are in the ε-neighborhood.

Proposition 3 (Cohen-Steiner et al. (2010)) *Let $1 \leq p' \leq p < \infty$, and D and E be persistence diagrams whose degree-p' total persistences are bounded from above. Then,*

$$d_{\mathrm{W}_p}(D, E) \leq \left(\frac{\mathrm{Pers}_{p'}(D) + \mathrm{Pers}_{p'}(E)}{2} \right)^{\frac{1}{p}} d_{\mathrm{W}_\infty}(D, E)^{1-\frac{p'}{p}}.$$

For a persistence diagram D, its degree-p total persistence is bounded from above by $\mathrm{card}\,(D) \times \max_{x \in D} \mathrm{pers}(x)^p$, where $\mathrm{card}\,(D)$ denotes the number of generators in D. However, this bound may be weak because, in general, $\mathrm{card}\,(D)$ cannot be bounded from above. In particular, if data set has noise, the persistence diagram often has many generators close to the diagonal. Thus, it is desirable that the total persistence is bounded from above independently of $\mathrm{card}\,(D)$. In the case of persistence diagrams obtained from a ball model filtration, we have the following upper bound (see Appendix B for the proof):

Lemma 4 *Let M be a triangulable compact subspace in $\mathbb{R}^d$, X be a finite subset of M, and $p > d$. Then,*

$$\mathrm{Pers}_p(D_q(X)) \leq \frac{p}{p-d} C_M \mathrm{diam}(M)^{p-d},$$

where C_M is a constant depending only on M.

Hence, we have the following by combining Proposition 3 and Lemma 4.

Corollary 5 *Let M be a triangulable compact subspace in $\mathbb{R}^d$, X, Y be finite subsets of M, and $p \geq p' > d$. Then*

$$\begin{aligned} d_{\mathrm{W}_p}(D_q(X), D_q(Y)) &\leq \left(\frac{p'}{p'-d} C_M \mathrm{diam}(M)^{p'-d} \right)^{\frac{1}{p}} d_{\mathrm{W}_\infty}(D_q(X), D_q(Y))^{1-\frac{p'}{p}} \\ &\leq \left(\frac{p'}{p'-d} C_M \mathrm{diam}(M)^{p'-d} \right)^{\frac{1}{p}} d_{\mathrm{H}}(X, Y)^{1-\frac{p'}{p}} \end{aligned}$$

where C_M is a constant depending only on M.

2.3 Kernel methods for representing signed measures

As a preliminary to our proposal of vector representation for persistence diagrams, we briefly summarize a method for embedding signed measures with a positive definite kernel.

Let Ω be a set and $k : \Omega \times \Omega \to \mathbb{R}$ be a *positive definite kernel* on Ω, i.e., k is symmetric, and for any number of points $x_1, \ldots, x_n$ in Ω, the Gram matrix $(k(x_i, x_j))_{i,j=1,\ldots,n}$ is nonnegative definite. A popular example of positive definite kernel on $\mathbb{R}^d$ is the Gaussian kernel $k_{\mathrm{G}}(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ ($\sigma > 0$), where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. From the Moore-Aronszajn theorem, it is also known that every positive definite kernel k on Ω is uniquely associated with a reproducing kernel Hilbert space $\mathcal{H}_k$ (RKHS).

We use a positive definite kernel to represent persistence diagrams by following the idea of the kernel mean embedding of distributions (Muandet et al., 2017; Smola et al., 2007; Sriperumbudur et al., 2011). Let Ω be a locally compact Hausdorff space, $M_{\mathrm{b}}(\Omega)$ be the space of all finite signed Radon measures[8] on Ω, and k be a bounded measurable kernel on Ω. Since $\int \|k(\cdot, x)\|_{\mathcal{H}_k} d\mu(x)$ is finite, the integral $\int k(\cdot, x) d\mu(x)$ is well-defined as the Bochner integral (Diestel and Uhl Jr, 1977). Here, we define a mapping from $M_{\mathrm{b}}(\Omega)$ to $\mathcal{H}_k$ by

$$E_k : M_{\mathrm{b}}(\Omega) \to \mathcal{H}_k, \quad \mu \mapsto \int k(\cdot, x) d\mu(x). \tag{2}$$

For a locally compact Hausdorff space Ω, let $C_0(\Omega)$ denote the space of continuous functions vanishing at infinity[9]. A kernel k on Ω is said to be C_0-kernel if $k(\cdot, x) \in C_0(\Omega)$ for any $x \in \Omega$. If k is C_0-kernel, the associated RKHS $\mathcal{H}_k$ is a subspace of $C_0(\Omega)$. A C_0-kernel k is called *C_0-universal* if $\mathcal{H}_k$ is dense in $C_0(\Omega)$. It is known that the Gaussian kernel k_{G} is C_0-universal on $\mathbb{R}^d$ (Sriperumbudur et al., 2011). When k is C_0-universal, the vector $E_k(\mu)$ in the RKHS uniquely determines the finite signed measure μ, and thus serves as a representation of μ. We summarize the property as follows:

Proposition 6 (Sriperumbudur et al. (2011)) *Let Ω be a locally compact Hausdorff space. If k is C_0-universal on Ω, the mapping E_k is injective. Thus,*

$$d_k(\mu, \nu) = \|E_k(\mu) - E_k(\nu)\|_{\mathcal{H}_k}$$

defines a distance on $M_{\mathrm{b}}(\Omega)$.

3. Kernel methods for persistence diagrams

We propose a positive definite kernel for persistence diagrams, called the persistence weighted Gaussian kernel (PWGK), to embed the persistence diagrams into an RKHS. This vectorization of persistence diagrams enables us to apply any kernel methods to persistence diagrams and explicitly control the effect of persistence. We show the stability theorem with respect to the distance defined by the embedding and discuss the efficient and precise approximate computation of the PWGK.

8. A *Radon measure* μ on Ω is a Borel measure on Ω satisfying (i) $\mu(C) < \infty$ for any compact subset $C \subset \Omega$, and (ii) $\mu(B) = \sup\{\mu(C) \mid C \subset B,\ C\text{:compact}\}$ for any B in the Borel σ-algebra of Ω.
9. A function f is said to *vanish at infinity* if for any $\varepsilon > 0$ there is a compact set $K \subset \Omega$ such that $\sup_{x \in K^c} |f(x)| \leq \varepsilon$.

3.1 Vectorization of persistence diagrams

We propose a method for vectorizing persistence diagrams using the kernel embedding (2) by regarding a persistence diagram as a discrete measure. In vectorizing persistence diagrams, it is desirable to have flexibility to discount the effect of generators close to the diagonal, since they often tend to be caused by noise. To this goal, we explain slightly different two ways of embeddings, which turn out to give the same inner product for two persistence diagrams.

First, for a persistence diagram D, we introduce a measure $\mu_D^w := \sum_{x \in D} w(x)\delta_x$ with a weight $w(x) > 0$ for each generator $x \in D$ (Figure 5), where δ_x is the Dirac delta measure at x. By appropriately choosing $w(x)$, the measure μ_D^w can discount the effect of generators close to the diagonal. A concrete choice of $w(x)$ will be discussed later.

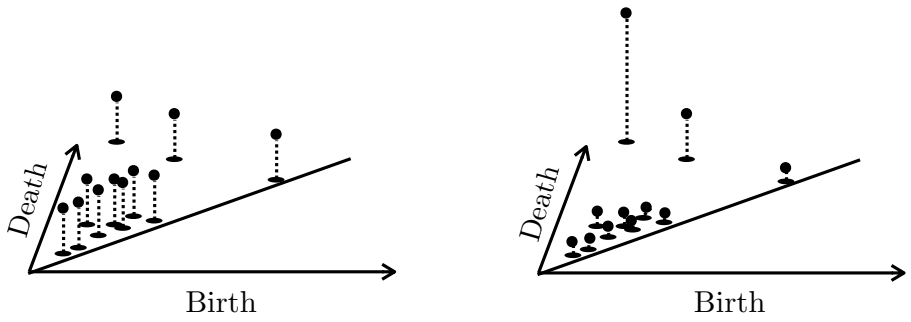

Figure 5: Unweighted (left) and weighted (right) measures.

As discussed in Section 2.3, given a C_0-universal kernel k above the diagonal $\mathbb{R}^2_{\mathrm{ad}} = \{(b, d) \in \mathbb{R}^2 \mid b < d\}$, the measure μ_D^w can be embedded as an element of the RKHS $\mathcal{H}_k$ via

$$\mu_D^w \mapsto E_k(\mu_D^w) := \sum_{x \in D} w(x)k(\cdot, x). \tag{3}$$

From the injectivity in Proposition 6, this mapping identifies a persistence diagram; in other words, it does not lose any information about persistence diagrams. Hence, $E_k(\mu_D^w) \in \mathcal{H}_k$ serves as a vector representation of the persistence diagram.

As the second construction, let

$$k^w(x, y) := w(x)w(y)k(x, y)$$

be the weighted kernel with the same weight function as above[10]. Then the mapping

$$E_{k^w} : \mu_D \mapsto \sum_{x \in D} w(x)w(\cdot)k(\cdot, x) \in \mathcal{H}_{k^w} \tag{4}$$

also defines a vectorization of persistence diagrams. The first construction may be more intuitive by directly weighting a measure, while the second one is also practically useful

10. From the facts that the product of positive definite kernels are also a positive definite kernel and $f(x, y) = w(x)w(y)$ is a positive definite kernel, k^w is actually a positive definite kernel.

since all the parameter tuning is reduced to kernel choice. We note that the inner products introduced by two RKHS vectors (3) and (4) are the same:

$$\langle E_k(\mu_D^w), E_k(\mu_E^w)\rangle_{\mathcal{H}_k} = \langle E_{k^w}(\mu_D), E_{k^w}(\mu_E)\rangle_{\mathcal{H}_{k^w}}.$$

In addition, these two RKHS vectors (3) and (4) are essentially equivalent, as seen from the next proposition:

Proposition 7 *Let k be C_0-universal on $\mathbb{R}^2_{\mathrm{ad}}$ and w be a positive function on $\mathbb{R}^2_{\mathrm{ad}}$. Then the following mapping*

$$\mathcal{H}_k \to \mathcal{H}_{k^w}, \quad f \mapsto wf$$

defines an isomorphism between the RKHSs. Under this isomorphism, $E_k(\mu_D^w)$ and $E_{k^w}(\mu_D)$ are identified.

Proof Let $\tilde{\mathcal{H}} := \{wf : \mathbb{R}^2_{\mathrm{ad}} \to \mathbb{R} \mid f \in \mathcal{H}_k\}$ and define its inner product by

$$\langle wf, wg\rangle_{\tilde{\mathcal{H}}} := \langle f, g\rangle_{\mathcal{H}_k}.$$

Then, it is easy to see that $\tilde{\mathcal{H}}$ is a Hilbert space and the mapping $f \mapsto wf$ gives an isomorphism between $\tilde{\mathcal{H}}$ and $\mathcal{H}_k$ of the Hilbert spaces. In fact, we can show that $\tilde{\mathcal{H}}$ is the same as $\mathcal{H}_{k^w}$. To see this, it is sufficient to check that k^w is a reproducing kernel of $\tilde{\mathcal{H}}$ from the uniqueness property of a reproducing kernel for an RKHS. The reproducing property is proven from

$$\langle wf, k^w(\cdot, x)\rangle_{\tilde{\mathcal{H}}} = \langle f, w(x)k(\cdot, x)\rangle_{\mathcal{H}_k} = w(x)f(x) = (wf)(x).$$

The second assertion is obvious. ■

3.2 Stability with respect to the kernel embedding

Given a data set X, we compute the persistence diagram $D_q(X)$ and vectorize it as an element $E_k(\mu^w_{D_q(X)})$ of the RKHS. Then, for practical applications, this map $X \mapsto E_k(\mu^w_{D_q(X)})$ should be stable with respect to perturbations to the data as discussed in Section 2.2.

Let D and E be persistence diagrams and $\gamma : D \cup \Delta \to E \cup \Delta$ be any multi-bijection. Here, we partition D (resp. Δ) into D_1 and D_2 (resp. Δ_1 and Δ_2) such as

$$\gamma(D_1) \subset \mathbb{R}^2_{\mathrm{ad}},\ \gamma(D_2) \subset \Delta,\ \gamma(\Delta_1) \subset \mathbb{R}^2_{\mathrm{ad}},\ \gamma(\Delta_2) \subset \Delta.$$

Then $D_1 \cup \Delta_1$ and E are bijective under γ. Now, let a weight function w be zero on the diagonal Δ. Then, the norm of the difference between RKHS vectors is calculated as follows:

$$
\begin{aligned}
&\|E_k(\mu_D^w) - E_k(\mu_E^w)\|_{\mathcal{H}_k} \\
&= \left\| \sum_{x \in D} w(x) k(\cdot, x) - \sum_{y \in E} w(y) k(\cdot, y) \right\|_{\mathcal{H}_k} \\
&= \left\| \sum_{x \in D} w(x) k(\cdot, x) - \sum_{x \in D_1 \cup \Delta_1} w(\gamma(x)) k(\cdot, \gamma(x)) \right\|_{\mathcal{H}_k} \\
&= \left\| \sum_{x \in D \cup \Delta_1} \Big(w(x) k(\cdot, x) - w(\gamma(x)) k(\cdot, \gamma(x)) \Big) + \sum_{x \in D_2} w(\gamma(x)) k(\cdot, \gamma(x)) \right\|_{\mathcal{H}_k} \\
&= \left\| \sum_{x \in D \cup \Delta_1} \Big(w(x) k(\cdot, x) - w(\gamma(x)) k(\cdot, \gamma(x)) \Big) \right\|_{\mathcal{H}_k} \\
&= \left\| \sum_{x \in D} w(x) \Big(k(\cdot, x) - k(\cdot, \gamma(x)) \Big) + \sum_{x \in D \cup \Delta_1} \Big(w(x) - w(\gamma(x)) \Big) k(\cdot, \gamma(x)) \right\|_{\mathcal{H}_k} \\
&\leq \sum_{x \in D} w(x) \left\| k(\cdot, x) - k(\cdot, \gamma(x)) \right\|_{\mathcal{H}_k} + \sum_{x \in D \cup \Delta_1} \left| w(x) - w(\gamma(x)) \right| \left\| k(\cdot, \gamma(x)) \right\|_{\mathcal{H}_k}.
\end{aligned}
$$

Here, let k be a C_0-universal kernel and satisfy the following:

(K) There exist constants $B_k, L_k > 0$ such that

$$
\|k(\cdot, x)\|_{\mathcal{H}_k} \leq B_k, \quad \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}_k} \leq L_k \|x - y\|_\infty \quad (x, y \in \mathbb{R}^2).
$$

Then, we have

$$
\|E_k(\mu_D^w) - E_k(\mu_E^w)\|_{\mathcal{H}_k} \leq L_k \sum_{x \in D} w(x) \|x - \gamma(x)\|_\infty + B_k \sum_{x \in D \cup \Delta_1} |w(x) - w(\gamma(x))|. \tag{5}
$$

In this sequel, we consider the Gaussian kernel $k_{\mathrm{G}}(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ $(\sigma > 0)$ for a C_0-universal kernel satisfying (K) by $B_{k_{\mathrm{G}}} = 1$ and $L_{k_{\mathrm{G}}} = \frac{\sqrt{2}}{\sigma}$ (Lemma 16 in Appendix C). Note that the Laplace kernel $k(x, y) = e^{-\alpha \sum_i |x_i - y_i|}$ $(\alpha > 0)$ also satisfies (K) by $B_k = 1$ and $L_k = 4\alpha$.

For a weight function, we consider the following assumption:

(W1) For any persistence diagrams D and E, and any multi-bijection $\gamma : D \cup \Delta \to E \cup \Delta$, there exist constants $B_1, L_1 > 0$ such that

$$
\sum_{x \in D} |w(x)| \leq B_1, \quad \sum_{x \in D \cup \Delta} |w(x) - w(\gamma(x))| \leq L_1 \sup_{x \in D \cup \Delta} \|x - \gamma(x)\|_\infty. \tag{6}
$$

If the weight function w satisfies (W1), from Equation (5), we have

$$\|E_k(\mu_D^w) - E_k(\mu_E^w)\|_{\mathcal{H}_k} \leq (L_k B_1 + B_k L_1) \sup_{x \in D \cup \Delta} \|x - \gamma(x)\|_\infty .$$

Since this inequality holds for any multi-bijection γ, we obtain the bottleneck stability.

Proposition 8 *Let D and E be persistence diagrams, a C_0-universal kernel k satisfy (K), and a weight function w satisfy (W1). Then,*

$$\|E_k(\mu_D^w) - E_k(\mu_E^w)\|_{\mathcal{H}_k} \leq (L_k B_1 + B_k L_1)\, d_{\mathrm{W}_\infty}(D, E).$$

In this paper, among many choices, we propose to use a weight function

$$w_{\mathrm{arc}}(x) = \arctan(C\mathrm{pers}(x)^p) \quad (C > 0,\ p \in \mathbb{Z}_{>0}).$$

This is a bounded and increasing function of $\mathrm{pers}(x)$. The corresponding positive definite kernel is

$$k_{\mathrm{PWG}}(x, y) = w_{\mathrm{arc}}(x) w_{\mathrm{arc}}(y) e^{-\frac{\|x-y\|^2}{2\sigma^2}}. \tag{7}$$

We call it *persistence weighted Gaussian kernel* (PWGK). This function w_{arc} gives a small (resp. large) weight on a noisy (resp. essential) generator. In addition, by appropriately adjusting the parameters C and p in w_{arc}, we can control the effect of the persistence. In order to check whether w_{arc} satisfies (W1), we first have

$$\sum_{x \in D} |w_{\mathrm{arc}}(x)| \leq C\mathrm{Pers}_p(D) \tag{8}$$

from the fact $w_{\mathrm{arc}}(x) \leq C\mathrm{pers}(x)^p$ $(x \in \mathbb{R}^2)$, and

$$\begin{aligned}
&\sum_{x \in D \cup \Delta} |w_{\mathrm{arc}}(x) - w_{\mathrm{arc}}(\gamma(x))| \\
&\leq 2pC \sum_{x \in D \cup \Delta_1} \max\{\mathrm{pers}(x)^{p-1}, \mathrm{pers}(\gamma(x))^{p-1}\} \|x - \gamma(x)\|_\infty \quad \text{(Lemma 18 in Appendix C)} \\
&\leq 2pC \left(\mathrm{Pers}_{p-1}(D \cup \Delta) + \mathrm{Pers}_{p-1}(\gamma(D \cup \Delta))\right) \sup_{x \in D \cup \Delta} \|x - \gamma(x)\|_\infty \\
&\leq 2pC \left(\mathrm{Pers}_{p-1}(D) + \mathrm{Pers}_{p-1}(E)\right) \sup_{x \in D \cup \Delta} \|x - \gamma(x)\|_\infty . \qquad (9)
\end{aligned}$$

Although total persistences in Equation (8) and Equation (9) are not constant, by restricting a class of persistence diagrams to that of a ball model filtration, w_{arc} satisfies (W1). Therefore, we obtain the bottleneck stability for PWGK:

Theorem 9 *Let M be a triangulable compact subspace in $\mathbb{R}^d$, $X, Y \subset M$ be finite subsets, $p > d + 1$, and a C_0-universal kernel k satisfy (K). Then,*

$$\left\| E_k(\mu_{D_q(X)}^{w_{\mathrm{arc}}}) - E_k(\mu_{D_q(Y)}^{w_{\mathrm{arc}}}) \right\|_{\mathcal{H}_k} \leq L_{k,\mathrm{arc}} d_{\mathrm{W}_\infty}(D_q(X), D_q(Y)),$$

where $L_{k,\mathrm{arc}}$ is a constant independent of X and Y.

Proof From Lemma 4, for $p-1>d$, there exists a constant $C_M>0$ such that

$$\mathrm{Pers}_p(D_q(X)) \leq \frac{p}{p-d} C_M \mathrm{diam}(M)^{p-d},$$
$$\mathrm{Pers}_{p-1}(D_q(X)),\ \mathrm{Pers}_{p-1}(D_q(Y)) \leq \frac{p-1}{p-1-d} C_M \mathrm{diam}(M)^{p-1-d}.$$

Thus, from Equation (8) and Equation (9), we obtain the constants in (W1) as

$$B_1 := \frac{p}{p-d} C C_M \mathrm{diam}(M)^{p-d}, \quad L_1 := \frac{4p(p-1)}{p-1-d} C C_M \mathrm{diam}(M)^{p-1-d},$$

and the statement is proven from Proposition 8. Note that

$$\begin{aligned} L_{k,\mathrm{arc}} &:= L_k B_1 + B_k L_1 \\ &= \left(\frac{\pi L_k}{2} \frac{p}{p-d} \mathrm{diam}(M) + B_k \frac{4p(p-1)}{p-1-d} \right) C C_M \mathrm{diam}(M)^{p-1-d}, \end{aligned}$$

is actually a constant independent of X and Y. ■

Let $\mathcal{P}_{\mathrm{finite}}(M)$ be the set of finite subsets in a triangulable compact subspace $M \subset \mathbb{R}^d$. Since the constant $L_{k_{\mathrm{G}},\mathrm{arc}}$ is independent of X and Y, Proposition 2 and Theorem 9 conclude that the map

$$\mathcal{P}_{\mathrm{finite}}(M) \to \mathcal{H}_{k_{\mathrm{G}}}, \quad X \mapsto E_{k_{\mathrm{G}}}(\mu^{w_{\mathrm{arc}}}_{D_q(X)})$$

is Lipschitz continuous. Note again that this implies a desirable stability property of the PWGK with the ball model: small perturbation of data points in terms of the Hausdorff distance causes only small perturbation of the persistence diagrams in terms of the RKHS distance with the PWGK. Note also that the RKHS of the PWGK is infinite dimensional. This can be seen from Proposition 7 and the fact that the Gaussian kernel defines an infinite dimensional RKHS on $\mathbb{R}^2_{\mathrm{ad}}$.

As the most relevant work to the PWGK, the persistence scale-space kernel (PSSK, Reininghaus et al. (2015))[11] provides another kernel method for vectorization of persistence diagrams and its stability result is shown with respect to 1-Wasserstein distance. However, to the best of our knowledge, 1-Wasserstein stability with respect to the Hausdorff distance is not shown, that is, for point sets X and Y, $d_{\mathrm{W}_1}(D_q(X), D_q(Y))$ is not estimated by $d_{\mathrm{H}}(X,Y)$ such as Proposition 2 or Corollary 5. Furthermore, it is shown (Reininghaus et al., 2015, Theorem 3) that the PSSK does not satisfy the stability with respect to p-Wasserstein distance for $p>1$, including the bottleneck distance d_{W_∞}, and hence it is not ensured that results obtained from the PSSK are stable under perturbation of data points in terms of the Hausdorff distance. On the other hand, since the PWGK has the desirable stability (Theorem 9), it is one of the advantages of our method over the previous research.

For completeness of theoretical discussions, we will show some mathematical results on the the stability with respect to 1-Wasserstein distance for PWGK along the line of Reininghaus et al. (2015). Now, we consider the following assumption (W2) which is weaker than (W1).

11. See Section 4.1.1.

(W2) For any $x, y \in \mathbb{R}^2$, there exist constants $B_2, L_2 > 0$ such that

$$|w(x)| \leq B_2, \quad |w(x) - w(y)| \leq L_2 \left\| x - y \right\|_\infty . \tag{10}$$

Proposition 10 *Let D and E be persistence diagrams, a C_0-universal kernel k satisfy (K), and a weight function w satisfy (W2). Then,*

$$\left\| E_k(\mu_D^w) - E_k(\mu_E^w) \right\|_{\mathcal{H}_k} \leq (L_k B_2 + B_k L_2)\, d_{\mathrm{W}_1}(D, E).$$

Proof From Equation (5), we have

$$\begin{aligned}
\left\| E_k(\mu_D^w) - E_k(\mu_E^w) \right\|_{\mathcal{H}_k} &\leq L_k \sum_{x \in D} w(x) \left\| x - \gamma(x) \right\|_\infty + B_k \sum_{x \in D \cup \Delta_1} |w(x) - w(\gamma(x))| \\
&\leq L_k B_2 \sum_{x \in D} \left\| x - \gamma(x) \right\|_\infty + B_k L_2 \sum_{x \in D \cup \Delta_1} \left\| x - \gamma(x) \right\|_\infty
\end{aligned}$$

Since this inequality holds for any multi-bijection γ, the statement is proven. ■

Here, we remark the relation between a weight function and stability. As a weight function, we also consider the following two natural weight functions

$$\begin{aligned}
w_{\mathrm{pers}}(x) &:= \begin{cases} 0 & (\mathrm{pers}(x) < 0) \\ \frac{1}{L}\mathrm{pers}(x) & (0 \leq \mathrm{pers}(x) \leq L) \\ 1 & (\mathrm{pers}(x) > L) \end{cases}, \\
w_{\mathrm{one}}(x) &\equiv 1,
\end{aligned} \tag{11}$$

where $L > 0$ is a parameter. Similar to w_{arc}, a piecewise linear weighting function w_{pers} gives a weight to a generator dependent on its persistence, but it does not satisfy satisfies (W1) since $\sum_{x \in D} w_{\mathrm{pers}}(x) = \frac{1}{L}\mathrm{Pers}_1(D)$, which is not a constant. For an unweighted function w_{one}, it also does not satisfy (W1) since $\sum_{x \in D} w_{\mathrm{one}}(x) = \mathrm{card}\,(D)$. Thus, it is still unknown whether the bottleneck distance stability holds for w_{pers} or w_{one}. On the other hand, since w_{pers} and w_{one} satisfy (W2)[12], the 1-Wasserstein stability holds for these weight functions. Regarding w_{arc}, we proposed it to satisfy (W1) with restriction to the class of persistence diagrams, and obtained the bottleneck stability. For $p = 1$, w_{arc} satisfies (W2) by $B_2 = \frac{\pi}{2}$ and $L_2 = 2C$ without any assumptions on persistence diagrams.

Corollary 11 *Let D and E be persistence diagrams and a C_0-universal kernel k satisfy (K). Then,*

$$\begin{aligned}
\left\| E_k(\mu_D^{w_{\mathrm{pers}}}) - E_k(\mu_E^{w_{\mathrm{pers}}}) \right\|_{\mathcal{H}_k} &\leq \left(L_k + \frac{2B_k}{L} \right) d_{\mathrm{W}_1}(D, E), \\
\left\| E_k(\mu_D^{w_{\mathrm{one}}}) - E_k(\mu_E^{w_{\mathrm{one}}}) \right\|_{\mathcal{H}_k} &\leq L_k d_{\mathrm{W}_1}(D, E), \\
\left\| E_k(\mu_D^{w_{\mathrm{arc}}}) - E_k(\mu_E^{w_{\mathrm{arc}}}) \right\|_{\mathcal{H}_k} &\leq \left(\frac{\pi L_k}{2} + 2B_k C \right) d_{\mathrm{W}_1}(D, E) \quad (p = 1 \textit{ in } w_{\mathrm{arc}}).
\end{aligned}$$

12. Regarding w_{pers}, L_2 in (W2) is given by $\frac{2}{L}$. See Lemma 17 in Appendix C

For $p > 1$, in general, w_{arc} does not satisfy (W2) since Ct^p is not Lipschitz continuous with respect to $t \in \mathbb{R}$. Similar to Theorem 9, by restricting to the class of persistence diagrams, we have the 1-Wasserstein stability:

Corollary 12 *Let M be a triangulable compact subspace in $\mathbb{R}^d$, $X, Y \subset M$ be finite subsets, $p > d+1$, and a C_0-universal kernel k satisfy (K). Then,*

$$\begin{aligned}&\left\|E_k(\mu^{w_{\mathrm{arc}}}_{D_q(X)}) - E_k(\mu^{w_{\mathrm{arc}}}_{D_q(Y)})\right\|_{\mathcal{H}_k}\\ &\leq \left(\frac{\pi L_k}{2} + B_k \frac{4p(p-1)}{p-1-d} C C_M \mathrm{diam}(M)^{p-1-d}\right) d_{\mathrm{W}_1}(D_q(X), D_q(Y)),\end{aligned}$$

for some constant $C_M > 0$.

Proof For any multi-bijection $\gamma : D_q(X) \cup \Delta \to D_q(Y) \cup \Delta$, we have

$$\begin{aligned}&\sum_{x \in D_q(X) \cup \Delta} |w_{\mathrm{arc}}(x) - w_{\mathrm{arc}}(\gamma(x))|\\ &\leq 2pC \left(\mathrm{Pers}_{p-1}(D_q(X)) + \mathrm{Pers}_{p-1}(D_q(Y))\right) \sup_{x \in D_q(X) \cup \Delta} \|x - \gamma(x)\|_\infty \quad \text{(from Equation (8))}\\ &\leq \frac{4p(p-1)}{p-1-d} C C_M \mathrm{diam}(M)^{p-1-d} \sum_{x \in D_q(X) \cup \Delta} \|x - \gamma(x)\|_\infty\end{aligned}$$

From Equation (5) and $\arctan(t) \leq \frac{\pi}{2}$ $(t \in \mathbb{R})$, the statement is proven. ■

3.3 Kernel methods on RKHS

Once persistence diagrams are represented as RKHS vectors, we can apply any kernel methods to those vectors by defining a kernel over the vector representation. In a similar way to the standard vectors, the simplest choice is to consider the inner product as a linear kernel

$$K_{\mathrm{L}}(D, E; k, w) := \langle E_k(\mu^w_D), E_k(\mu^w_E)\rangle_{\mathcal{H}_k} = \sum_{x \in D} \sum_{y \in E} w(x) w(y) k(x, y) \tag{12}$$

on the RKHS and we call it the (k, w)-*linear kernel*.

If k is a C_0-universal kernel and w is strictly positive on $\mathbb{R}^2_{\mathrm{ad}}$, from Proposition 6, $\|E_k(\mu^w_D) - E_k(\mu^w_E)\|_{\mathcal{H}_k}$ defines a distance on the persistence diagrams and it is computed as

$$\sqrt{K_{\mathrm{L}}(D, D; k, w) + K_{\mathrm{L}}(E, E; k, w) - 2K_{\mathrm{L}}(D, E; k, w)}.$$

Then, we can also consider a nonlinear kernel

$$K_{\mathrm{G}}(D, E; k, w) = \exp\left(-\frac{1}{2\tau^2} \|E_k(\mu^w_D) - E_k(\mu^w_E)\|^2_{\mathcal{H}_k}\right) \ (\tau > 0) \tag{13}$$

on the RKHS and we call it the (k, w)-*Gaussian kernel*.

In this paper, if there is no confusion, we also refer to the $(k_{\mathrm{G}}, w_{\mathrm{arc}})$-Gaussian kernel as the PWGK. Muandet et al. (2012) observed better performance with nonlinear kernels for some complex tasks and this is one of the reasons that we will use the Gaussian kernel on the RKHS.

3.4 Computation of Gram matrix

Let $\mathcal{D} = \{D_\ell \mid \ell = 1, \dots, n\}$ be a collection of persistence diagrams. In many practical applications, the number of generators in a persistence diagram can be large, while n is often relatively small; in Section 4.4, for example, the number of generators is about 30000, while $n = 80$.

If the persistence diagrams contain at most m points, each element of the Gram matrix $(K_{\mathrm{G}}(D_i, D_j; k_{\mathrm{G}}, w))_{i,j=1,\dots,n}$ involves $O(m^2)$ evaluations of $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$, resulting the complexity $O(m^2n^2)$ for obtaining the Gram matrix. Hence, reducing computational cost with respect to m is an important issue.

We solve this computational issue by using the random Fourier features (Rahimi and Recht, 2007). To be more precise, let $z_1, \dots, z_{M_{\mathrm{rff}}}$ be random variables from the 2-dimensional normal distribution $N((0,0), \sigma^{-2}I)$ where I is the identity matrix. This method approximates $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ by $\frac{1}{M_{\mathrm{rff}}} \sum_{a=1}^{M_{\mathrm{rff}}} e^{\sqrt{-1}z_a^T x}(e^{\sqrt{-1}z_a^T y})^*$, where $*$ denotes the complex conjugate. Then, $\sum_{x\in D_i}\sum_{y\in D_j} w(x)w(y)k_{\mathrm{G}}(x,y)$ is approximated by $\frac{1}{M_{\mathrm{rff}}}\sum_{a=1}^{M_{\mathrm{rff}}} B_i^a (B_j^a)^*$, where $B_\ell^a = \sum_{x\in D_\ell} w(x) e^{\sqrt{-1}z_a^T x}$. As a result, the computational complexity of the approximated Gram matrix is $O(mnM_{\mathrm{rff}} + n^2 M_{\mathrm{rff}})$, which is linear to m. In Section 4.3 and Section 4.4, we set $M_{\mathrm{rff}} = 10^5$. For the convergence rate of this approximation with respect to M_{rff}, please see Appendix D.

We note that the approximation by the random Fourier features can be sensitive to the choice of σ. If σ is much smaller than $\|x-y\|$, the relative error can be large. For example, in the case of $x = (1,2), y = (1,2.1)$ and $\sigma = 0.01$, $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ is about 10^{-22} while we observed the approximated value can be about 10^{-4} with $M_{\mathrm{rff}} = 10^5$. As a whole, these m^2 errors may cause a critical error to the statistical analysis. Moreover, if σ is largely deviated from the ensemble $\|x-y\|$ for $x \in D_i, y \in D_j$, then most values $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ become close to 0 or 1.

In order to obtain a good approximation and extract meaningful values, the choice of parameters is important. For unsupervised case, we follow the heuristics proposed in Gretton et al. (2007) and set

$$\sigma = \mathrm{median}\{\sigma(D_\ell) \mid \ell = 1, \dots, n\}, \text{ where } \sigma(D) = \mathrm{median}\{\|x_i - x_j\| \mid x_i, x_j \in D,\ i < j\},$$

so that σ takes close values to many $\|x-y\|$. For the parameter C, we also set

$$C = (\mathrm{median}\{\mathrm{pers}(D_\ell) \mid \ell = 1, \dots, n\})^{-p}, \text{ where } \mathrm{pers}(D) = \mathrm{median}\{\mathrm{pers}(x_i) \mid x_i \in D\}.$$

Similarly, the parameter τ in the (k,w)-Gaussian kernel is defined by

$$\mathrm{median}\left\{ \left\| E_k(\mu_{D_i}^w) - E_k(\mu_{D_j}^w) \right\|_{\mathcal{H}_k} \;\middle|\; 1 \le i < j \le n \right\}. \tag{14}$$

For supervised learning such as SVM, we use the cross-validation (CV) approach and do not use the random Fourier features in Section 4.2 and Section 4.5.

4. Experiments

In this section, we apply the kernel method of the PWGK to synthesized and real data, and compare the performance between the PWGK and other statistical methods of persistence diagrams. All persistence diagrams are obtained from the ball model filtrations and computed by CGAL (Da et al., 2015) and PHAT (Bauer et al., 2014). With respect to the dimension of persistence diagrams, we use 2-dimensional persistence diagrams in Section 4.3 and 1-dimensional ones in other parts.

4.1 Comparison to previous works

4.1.1 Persistence scale-space kernel

The most relevant work to our method is the one proposed by Reininghaus et al. (2015). Inspired by the heat equation, they propose a positive definite kernel called *persistence scale-space kernel* (PSSK) K_{PSS} on the persistence diagrams:

$$K_{\mathrm{PSS}}(D,E) = \langle \Phi_t(D), \Phi_t(E) \rangle_{L^2(\mathbb{R}^2)} = \frac{1}{8\pi t} \sum_{x\in D} \sum_{y\in E} e^{-\frac{\|x-y\|^2}{8t}} - e^{-\frac{\|x-\bar{y}\|^2}{8t}}, \tag{15}$$

where $\Phi_t(D)(x) = \frac{1}{4\pi t} \sum_{y\in D} e^{-\frac{\|x-y\|^2}{4t}} - e^{-\frac{\|x-\bar{y}\|^2}{4t}}$ and $\bar{y} := (y^2, y^1)$ for $y = (y^1, y^2)$. We note that $\Phi_t(D)$ also takes zero on the diagonal by subtracting the Gaussian kernels for y and $\bar{y}$.

In fact, we can verify that the (k, w)-linear kernel contains the PSSK. Let $\tilde{D} := D \cup D^*$ where $D^* = \{(d, b) \in \mathbb{R}^2 \mid (b, d) \in D\}$. Then, $\Phi_t(D)$ can also be expressed as

$$\Phi_t(D) = \frac{1}{4\pi t} \sum_{y\in\tilde{D}} w_{\mathrm{PSS}}(y) k_{\mathrm{G}}(\cdot, y) \quad \text{where} \quad w_{\mathrm{PSS}}(y) = \begin{cases} 1, & y^2 > y^1 \\ 0, & y \in \Delta \\ -1, & y^2 < y^1 \end{cases},$$

which is equal to $\frac{1}{4\pi t} E_{k_{\mathrm{G}}}(\mu_{\tilde{D}}^{w_{\mathrm{PSS}}})$. Furthermore, the inner product in $\mathcal{H}_{k_{\mathrm{G}}}$ is

$$K_{\mathrm{L}}(\tilde{D}, \tilde{E}; k_{\mathrm{G}}, w_{\mathrm{PSS}}) = \langle E_{k_{\mathrm{G}}}(\mu_{\tilde{D}}^{w_{\mathrm{PSS}}}), E_{k_{\mathrm{G}}}(\mu_{\tilde{E}}^{w_{\mathrm{PSS}}}) \rangle_{\mathcal{H}_{k_{\mathrm{G}}}} = 2 \sum_{x\in D} \sum_{y\in E} k_{\mathrm{G}}(x, y) - k_{\mathrm{G}}(x, \bar{y}). \tag{16}$$

By scaling the variance parameter σ in the Gaussian kernel k_{G} and multiplying by an appropriate scalar, Equation (15) is the same as Equation (16). Thus, the PSSK can also be approximated by the random Fourier features. When we apply the random Fourier features for the PSSK, we set $\tilde{\sigma} = \mathrm{median}\{\sigma(\tilde{D}_\ell) \mid \ell = 1, \cdots, n\}$ as before and $t = \frac{\tilde{\sigma}^2}{4}$.

While both methods discount noisy generators, the PWGK has the following advantages over the PSSK. (i) The PWGK can control the effect of the persistence by C and p in w_{arc} independently of the bandwidth parameter σ in the Gaussian factor, while in the PSSK only one parameter t cannot adjust the global bandwidth and the effect of persistence simultaneously. (ii) The PSSK does not satisfy the stability with respect to the bottleneck distance (see also remarks after Theorem 9).

4.1.2 Persistence landscape

The *persistence landscape* (Bubenik, 2015) is a well-known approach in TDA for vectorization of persistence diagrams. For a persistence diagram D, the persistence landscape λ_D is defined by

$$\lambda_D(k,t) = k\text{-th largest value of } \min\{t-b_i, d_i-t\}_+,$$

where c_+ denotes $\max\{c,0\}$, and it is a vector in the Hilbert space $L^2(\mathbb{N}\times\mathbb{R})$. Here, we define a positive definite kernel of persistence landscapes as a linear kernel on $L^2(\mathbb{N}\times\mathbb{R})$:

$$K_{\mathrm{PL}}(D,E) := \langle \lambda_D, \lambda_E \rangle_{L^2(\mathbb{N}\times\mathbb{R})} = \int_{\mathbb{R}} \sum_{k=1} \lambda_D(k,t)\lambda_E(k,t)dt. \tag{17}$$

Since a persistence landscape does not have any parameters, we do not need to consider the parameter tuning. However, the integral computation is required and it causes much computational time. Let $\mathcal{D} = \{D_\ell \mid \ell = 1,\dots,n\}$ be a collection of persistence diagrams which contain at most m points. Since $\lambda_{D_i}(k,t) \equiv 0$ for any $k > m$, $t \in \mathbb{R}$, $i = 1,\cdots,n$, calculating $\{\lambda_{D_i}(k,t) \mid k \in \mathbb{Z}_{\geq 0}\}$, which needs sorting, is in $O(m\log m)$ (see also Bubenik and Dłotko (2017)). For a fixed t, we can calculate $(\sum_{k=1}\lambda_{D_i}(k,t)\lambda_{D_j}(k,t))_{i,j=1,\cdots,n}$ in $O(nm\log m + n^2 m)$, and the Gram matrix $(K_{\mathrm{PL}}(D_i,D_j))_{i,j=1,\cdots,n}$ in $O(M_{\mathrm{int}}(nm\log m + n^2 m))$, where M_{int} is the number of partitions in the integral interval. Theoretically speaking, this implies that it takes more time to calculate the Gram matrix of K_{PL} than the PWGK and the PSSK by the random Fourier features.

4.1.3 Persistence image

As a finite dimensional vector representation of a persistence diagram, a *persistence image* is proposed in Adams et al. (2017). First, we prepare a differentiable probability density function $\phi_x : \mathbb{R}^2 \to \mathbb{R}$ with mean x and a weight function $w : \mathbb{R}^2_{\mathrm{ad}} \to \mathbb{R}$. For a persistence diagram D, the *corresponding persistence surface* is defined by

$$\rho_D(z) := \sum_{x\in D} w(x)\phi_x(z). \tag{18}$$

Then, for fixed points $a_0 < \cdots < a_M$ ($a_i \in \mathbb{R}$), the *persistence image*[13] $\mathrm{PI}(D)$ is defined by an $M \times M$ matrix whose (i,j)-element is assigned to the integral of ρ_D over the pixel $P_{i,j} := (a_{i-1}, a_i] \times (a_{j-1}, a_j]$, i.e.,

$$\mathrm{PI}(D)_{i,j} := \int_{P_{i,j}} \rho_D(z)dz.$$

Since the persistence image can be regarded as an M^2-dimensional vector, we define a vector $\mathrm{PIV}(D) \in \mathbb{R}^{M^2}$ by

$$\mathrm{PIV}(D)_{i+M(j-1)} := \mathrm{PI}(D)_{i,j}, \tag{19}$$

13. Adams et al. (2017) use a persistence diagram in birth-persistence coordinates. That is, by a linear transformation $T(b,d) = (b, d-b)$, a persistence diagram D is transformed into $T(D)$. In this paper, in order to compare with the persistence image and the PWGK, we use birth-death coordinates.

and, in this paper, call it the persistence image vector.

In Adams et al. (2017), they use the 2-dimensional Gaussian distribution $\frac{1}{2\pi\sigma^2}k_{\mathrm{G}}(x,z)$ as $\phi_x(z)$ and a piecewise linear weighting function $w_{\mathrm{pers}}(x)$. In this paper, for a collection of persistence diagrams $\mathcal{D} = \{D_\ell \mid \ell = 1, \ldots, n\}$, we set a parameter L in Equation (11) as

$$L = \max\{L(D_\ell) \mid \ell = 1, \cdots, n\}, \text{ where } L(D) = \max\{d_i \mid (b_i, d_i) \in D\}.$$

For points $a_0 < \cdots < a_M$ of a pixel $P_{i,j} = (a_{i-1}, a_i] \times (a_{j-1}, a_j]$, we set $a_M = L$ and $a_i = \frac{i}{M}a_M$ for $0 \leq i \leq M$[14].

Here, by choosing ϕ_x and w in the proposed way, we define a positive definite kernel of persistence image vector as a linear kernel on $\mathbb{R}^{M^2}$:

$$\begin{aligned} K_{\mathrm{PI}}(D,E) &:= \langle \mathrm{PIV}(D), \mathrm{PIV}(E) \rangle_{\mathbb{R}^{M^2}} \\ &= \sum_{i,j=1}^{M} \mathrm{PI}(D)_{i,j}\mathrm{PI}(E)_{i,j} \\ &= \frac{1}{(2\pi\sigma^2)^2} \sum_{x\in D}\sum_{y\in E} w_{\mathrm{pers}}(x)w_{\mathrm{pers}}(y) \sum_{i,j=1}^{M} \int_{P_{i,j}} k_{\mathrm{G}}(x,z)dz \int_{P_{i,j}} k_{\mathrm{G}}(y,z)dz. \quad (20) \end{aligned}$$

If we choose $\phi_x(z)$ as a (normalized) positive definite kernel $k(x,z)$, the corresponding persistence surface ρ_D (18) is the same as the RKHS vector $E_k(\mu_D^w)$. Thus, it may be expected that the persistence image and the PWGK show similar performance for data analysis. However, there are several differences between the persistence image and the PWGK. (i) Underlying vector spaces are different: the PWGK vector $E_k(\mu_D^w)$ is always in the RKHS and the corresponding persistence surface ρ_D is in $L^p(\mathbb{R}^2)$ with appropriate conditions. Hence, the inner product structures are also different[15]. (ii) Regarding the mapping from a persistence diagram to the corresponding persistence surface, the injectivity is not discussed in the original paper (Adams et al., 2017). On the other hand, from Proposition 6, we can easily check the injectivity of the RKHS vector $E_k(\mu_D^w)$ due to its construction based on kernel method. (iii) It is also shown that the persistence image has a stability result with respect to 1-Wasserstein distance, but it does not satisfy the bottleneck stability (Remark 1 in Adams et al. (2017)) or the Haussdorff stability as noted after Theorem 9. This instability is considered to be caused by the norm of the persistence image, which is different from the RKHS. (iv) The computational complexity of a persistence image does not depend on the number of generators in a persistence diagram, but instead, it depends on the number of pixels M^2. Precisely, the Gram matrix $(K_{\mathrm{PI}}(D_i, D_j))_{i,j=1,\cdots,n}$ is calculated

14. Here, we set $a_0 = 0$ because all generators in the ball model filtrations are born after $b = 0$.

15. Since the persistence image vector $\mathrm{PIV}(D)$ (19) is a discretization of ρ_D, the inner product (20) can be also seen as a discretization of L^2 inner product of the corresponding persistence surfaces

$$\langle \rho_D, \rho_E \rangle_{L^2(\mathbb{R}^2)} = \frac{1}{(2\pi\sigma^2)^2} \sum_{x\in D}\sum_{y\in E} w_{\mathrm{pers}}(x)w_{\mathrm{pers}}(y) \int_{\mathbb{R}^2} k_{\mathrm{G}}(x,z)k_{\mathrm{G}}(y,z)dz.$$

Furthermore, since $\int_{\mathbb{R}^2} e^{-\frac{\|x-z\|^2}{2\sigma^2}} e^{-\frac{\|y-z\|^2}{2\sigma^2}} dz \propto e^{-\frac{\|x-y\|^2}{4\sigma^2}}$, $K_{\mathrm{PI}}(D,E)$ is also a discretization of the inner product of the RKHS vectors $K_{\mathrm{L}}(D,E;k_{\mathrm{G}},w_{\mathrm{arc}})$ by scaling the variance parameter σ in k_{G}. However, this is a special case, and it is not always to be true for any positive definite kernel.

in $O(n^2M^2)$. We can reduce the computational time of the persistence image by choosing a small mesh size M. However, some situations need a fine mesh (i.e., a large mesh size), and thus, we have to be careful with the choice of mesh size. In Section 4.2.2, we will discuss the effect of the mesh size on the classification performance of the persistence image.

4.2 Classification with synthesized data

We compare the performance among the PWGK, the PSSK, the persistence landscape, and the persistence image for a simple binary classification task with SVMs.

4.2.1 Synthesized data

In this experiment, we design data sets so that important generators close to the diagonal must be taken into account to solve the classification task.

Let $S^1(x, y, r, N)$ be a set composed of N points sampled with equal distance from a circle in 2-dimensional Euclidean space with radius r centered at (x, y). When we compute the persistence diagram of $S^1(x, y, r, N)$ for $N > 3$, there always exists a generator whose birth time is approximately $\frac{\pi r}{N}$ (here we use $\sin\theta \approx \theta$ for small θ) and death time is r (Figure 6).

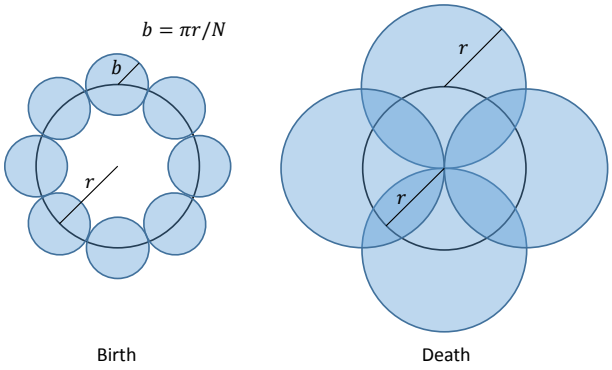

Figure 6: Birth and death of the generator for $S^1(x, y, r, N)$.

In order to add randomness on $S^1(x, y, r, N)$, we extend it into $\mathbb{R}^3$ and change $S^1(x, y, r, N)$ to $S^1_z(x, y, r, N)$ and $\tilde{S}^1_z(x, y, r, N)$ as follows:

$$S^1_z(x, y, r, N) := \{(z_1, z_2, z_3) \mid (z_1, z_2) \in S^1(x, y, r, N),\ z_3 \text{ is uniformly sampled from } [0, 0.01]\}$$
$$\tilde{S}^1_z(x, y, r, N) := S^1_z(x + W_x^2, y + W_y^2, r + W_r^2, \lceil N + 2W_N \rceil),$$

where $W_x, W_y \sim N(0, 2)$[16], $W_r, W_N \sim N(0, 1)$ and $\lceil c \rceil$ is the smallest integer greater than or equal to c. Then, we add $S_2 := S^1_z(x_2, y_2, r_2, N_2)$ to $S_1 := \tilde{S}^1_z(x_1, y_1, r_1, N_1)$ with probability 0.5 and use it as the synthesized data.

16. $N(\mu, \sigma^2)$ is the 1-dimensional normal distribution with mean μ and variance σ^2.

In this paper, we choose parameters by

$$\begin{aligned} r_1 &= 1 + 8W^2 \quad (W \sim N(0,1)), \\ x_1 = y_1 &= 1.5 r_1, \\ N_1 &\ : \text{ a random integer with equal probability in } (\lceil \frac{\pi r}{2} \rceil, 4\pi r), \end{aligned}$$

and set (x_2, y_2, r_2, N_2) as $(0, 0, 0.2, 10)$ (Figure 7).

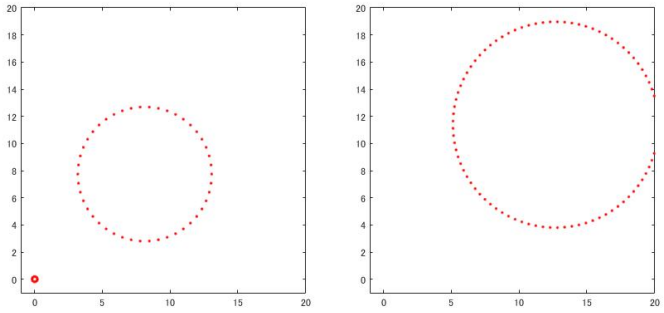

Figure 7: Examples of synthesized data. Left: S_2 exists. Right: S_2 does not exist.

For the binary classification, we introduce the following labels:

$z_0 = 1$ if there exists a generator (b, d) in the persistence diagram such that $b \leq 1$ and $d \geq 4$.

$z_1 = 1$ if S_2 exists.

The class label of the data set is then given by $\mathbf{XOR}(z_0, z_1)$. By this construction, identifying z_0 requires relatively smooth function in the area of long lifetimes, while classifying the existing of z_1 needs delicate control of the resolution around the diagonal.

4.2.2 SVM results

SVMs are trained from persistence diagrams given by 100 data sets, and evaluated with 100 independent test data sets. As a positive definite kernel k, we choose the Gaussian kernel k_{G} and the linear kernel $k_{\mathrm{L}}(x, y) := \langle x, y \rangle_{\mathbb{R}^2}$. For a weight function w, we use the proposed function $w_{\mathrm{arc}}(x) = \arctan(C\mathrm{pers}(x)^p)$, the piecewise linear weighting function $w_{\mathrm{pers}}(x)$, and an unweighted function $w_{\mathrm{one}}(x) \equiv 1$. The hyper-parameters (σ, C) in the PWGK and t in the PSSK are chosen by the 10-fold cross-validation, and the degree p in $w_{\mathrm{arc}}(x)$ is set as $1, 5, 10$. For K_{PSS} and K_{PL}, while they originally consider only the inner product, we also apply the Gaussian kernels on RKHS following Equation (13). Since K_{PI} can be seen as a discretization of the $(k_{\mathrm{G}}, w_{\mathrm{pers}})$-linear kernel, we also construct another kernel of persistence image by replacing w_{pers} with w_{arc}, which is considered as a discretization of the PWGK. In order to check whether the persistence image with w_{arc} is an appropriate discretization of the PWGK, we try several mesh size $M = 20, 50, 100$.

In Table 1, we can see that the PWGK $\triangle$ and the Gaussian kernel on the persistence image with w_{arc} and large mesh size $\square_{100}$ show higher classification rates (85% accuracy)

Table 1: Results of SVMs with the (k, w)-linear/Gaussian kernel, the PSSK, the persistence landscape, and the persistence image. Average classification rates (%) and standard deviations for 100 test data sets are shown.

		Linear	Gaussian
PWGK			
kernel	**weight**		
	w_{arc} $(p=1)$	75.7 ± 2.31	85.8 ± 5.19 (PWGK)
	w_{arc} $(p=5)$	75.8 ± 2.47 ($\triangle$)	85.6 ± 5.01 (PWGK, $\square$)
k_{G}	w_{arc} $(p=10)$	76.0 ± 2.39	86.0 ± 4.98 (PWGK)
	w_{pers}	49.3 ± 2.72	52.3 ± 6.60
	w_{one}	53.8 ± 4.76	55.1 ± 8.42
	w_{arc} $(p=5)$	49.3 ± 6.92	51.8 ± 3.52
k_{L}	w_{pers}	51.0 ± 6.84	55.7 ± 8.68
	w_{one}	50.5 ± 6.90	53.0 ± 4.89
PWGK with Persistence image			
$M=20$	w_{arc} $(p=5)$	48.8 ± 3.75 ($\triangle_{20}$)	52.0 ± 5.65 ($\square_{20}$)
$M=50$	w_{arc} $(p=5)$	49.2 ± 5.77 ($\triangle_{50}$)	51.8 ± 7.23 ($\square_{50}$)
$M=100$	w_{arc} $(p=5)$	75.0 ± 2.20 ($\triangle_{100}$)	85.8 ± 4.15 ($\square_{100}$)
PSSK		50.5 ± 5.60 (K_{PSS})	53.6 ± 6.69
Persistence landscape		50.6 ± 5.92 (K_{PL})	48.8 ± 4.25
Persistence image			
$M=20$	w_{pers}	51.1 ± 4.38 (K_{PI})	51.7 ± 6.86
$M=50$	w_{pers}	49.0 ± 6.14 (K_{PI})	52.3 ± 7.21
$M=100$	w_{pers}	54.5 ± 8.76 (K_{PI})	52.1 ± 6.70

than the other methods (K_{PSS} : 50%, K_{PL} : 50%, and K_{PI} : 55%). Although the $(k_{\mathrm{G}}, w_{\mathrm{pers}})$-Gaussian kernel and the persistence image with the original weight w_{pers} discount noisy generators, the classification rates are close the chance level. These unfavorable results must be caused by the difficulty in handling the local and global locations of generators simultaneously. While the result of the persistence image with a large mesh size is similar to that of the PWGK (e.g., $\square$ and $\square_{100}$), a small mesh size gives bad approximation results (e.g., $\square$ and $\square_{50}$). The reason is because a small mesh size makes rough pixels, and S_2 itself and noisy generators are treated in some rough pixel. On the other hand, we remark that a large mesh size M needs much computational time.

We observe that the classification accuracies are not sensitive to p. Thus, in the rest of this paper, we set $p = 5$ because the assumption $p > d + 1$ in Theorem 9 ensures the continuity in the kernel embedding of persistence diagrams and all data points are obtained from $\mathbb{R}^3$.

4.3 Analysis of granular system

We apply the PWGK, the PSSK, the persistence landscape, and the persistence image to persistence diagrams obtained by experimental data in a granular packing system (Francois et al., 2013). In this example, a partially crystallized packing with 150,000 monosized beads (diameter = 1mm, polydispersity = 0.025mm) in a container is obtained by experiments, where the configuration of the beads is imaged by means of X-ray Computed Tomography. One of the fundamental interests in the study of granular packings is to understand the transition from random packings to crystallized packings. In particular, the maximum packing density ϕ_* that random packings can attain is still a controversial issue (e.g., see Torquato et al. (2000)). Here, we apply the change point analysis to detect ϕ_*.

In oder to observe configurations of various densities, we divide the original full system into 35 cubical subsets containing approximately 4000 beads. The data are provided by the authors of the paper (Francois et al., 2013). The packing densities of the subsets range from $\phi = 0.590$ to $\phi = 0.730$. Saadatfar et al. (2017) computed a persistence diagram for each subset by taking the beads configuration as a finite subset in $\mathbb{R}^3$, and found that the persistence diagrams characterize different configurations in random packings (small ϕ) and crystallized packings (large ϕ). Hence, it is expected that the change point analysis applied to these persistence diagrams can detect the maximum packing density ϕ_* as a transition from the random to crystallized packings.

Our strategy is to regard the maximum packing density as the change point and detect it from a collection $\mathcal{D} = \{D_\ell \mid \ell = 1, \ldots, n\}$ $(n = 35)$ of persistence diagrams made by beads configurations of granular systems, where ℓ is the index of the packing densities listed in the increasing order. As a statistical quantity for the change point detection, we use the kernel Fisher discriminant ratio (Harchaoui et al., 2009) defined by

$$\mathrm{KFDR}_{n,\ell,\gamma}(\mathcal{D}) = \frac{\ell(n-\ell)}{n} \left\| \left(\frac{\ell}{n} \hat{\Sigma}_{1:\ell} + \frac{n-\ell}{n} \hat{\Sigma}_{\ell+1:n} + \gamma I \right)^{-\frac{1}{2}} (\hat{\mu}_{\ell+1:n} - \hat{\mu}_{1:\ell}) \right\|_{\mathcal{H}_K}, \qquad (21)$$

where the empirical mean element $\hat{\mu}_{i:j}$ and the empirical covariance operator $\hat{\Sigma}_{i:j}$ with data D_i through D_j $(i < j)$ are given by

$$\hat{\mu}_{i:j} = \frac{1}{j-i+1} \sum_{\ell=i}^{j} K(\cdot, D_\ell),$$

$$\hat{\Sigma}_{i:j} = \frac{1}{j-i+1} \sum_{\ell=i}^{j} (K(\cdot, D_\ell) - \hat{\mu}_{i:j}) \otimes (K(\cdot, D_\ell) - \hat{\mu}_{i:j})$$

respectively, and γ is a regularization parameter (in this paper we set $\gamma = 10^{-3}$). The index ℓ achieving the maximum of $\mathrm{KFDR}_{n,\ell,\gamma}(\mathcal{D})$ corresponds to the estimated change point. In Figure 8, all the four methods detect $\ell = 23$ as the sharp maximizer of the KFDR. This result indicates that the maximum packing density ϕ_* exists in the interval $[0.604, 0.653]$ and supports the traditional observation $\phi_* \approx 0.636$ (Anonymous, 1972).

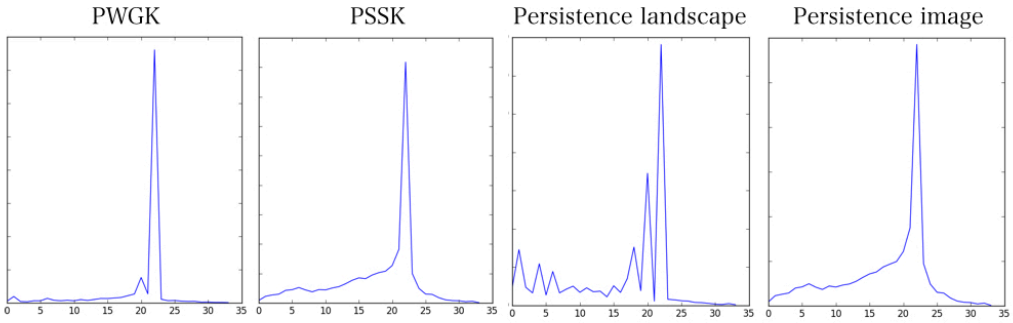

Figure 8: The KFDR graphs of the PWGK, the PSSK, the persistence landscape, and the persistence image.

We also apply kernel principal component analysis (KPCA) to the same collection of the 35 persistence diagrams. Figure 9 shows the 2-dimensional KPCA plots where each blue cross (resp. red circle) indicates the persistence diagram of random packing (resp. crystallized packing). We can see clear two-cluster structure corresponding to two physical states.

4.4 Analysis of SiO_2

When we rapidly cool down the liquid state of SiO_2, it avoids the usual crystallization and changes into a glass state. Understanding the liquid-glass transition is an important issue for the current physics and industrial applications (Greaves and Sen, 2007). Glass is an amorphous solid, which does not have a clear structure in the configuration of molecules, but it is also known that the medium distance structure such as rings have important influence on the physical properties of the material. It is thus promising to apply the persistent homology to express the topological and geometrical structure of the glass configuration. For estimating the glass transition temperature by simulations, a traditional physical method is to prepare atomic configurations of SiO_2 for a certain range of temperatures by molecular

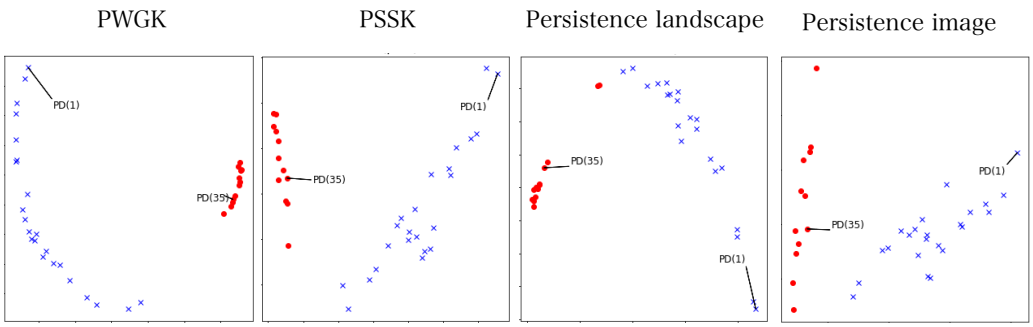

Figure 9: The KPCA plots of the PWGK (contribution rate: 92.9%), the PSSK (99.7%), the persistence landscape (83.8%), and the persistence image (98.7%).

dynamics simulations, and then draw the temperature-enthalpy graph. The graph consists of two lines in high and low temperatures with slightly different slopes which correspond to the liquid and the glass states, respectively, and the glass transition temperature is conventionally estimated as an interval of the transient region combining these two lines (e.g., see Elliott (1990)). However, since the slopes of two lines are close to each other, determining the interval is a subtle problem. Usually only the rough estimate of the interval is available. Hence, we apply our framework of topological data analysis with kernels to detect the glass transition temperature.

Let $\{D_\ell \mid \ell = 1, \ldots, 80\}$ be a collection of the persistence diagrams made by atomic configurations of SiO_2 and sorted by the decreasing order of the temperature. The same data was used in the previous works by Hiraoka et al. (2016); Nakamura et al. (2015). The interval of the glass transition temperature T estimated by the conventional method explained above is $2000K \leq T \leq 3500K$, which corresponds to $35 \leq \ell \leq 50$.

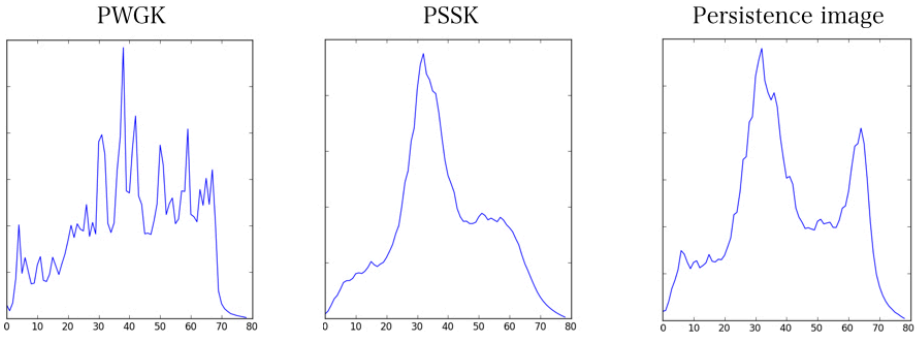

Figure 10: The KFDR graphs of the PWGK (left), the PSSK (center) and the persistence image (right).

In Figure 10, the KFDR plots show that the change point is estimated as $\ell = 39$ by the PWGK, $\ell = 33$ by the PSSK, and $\ell = 33$ by the persistence image. For the persistence

landscape, we cannot obtain the KFDR or the KPCA results with reasonable computational time.

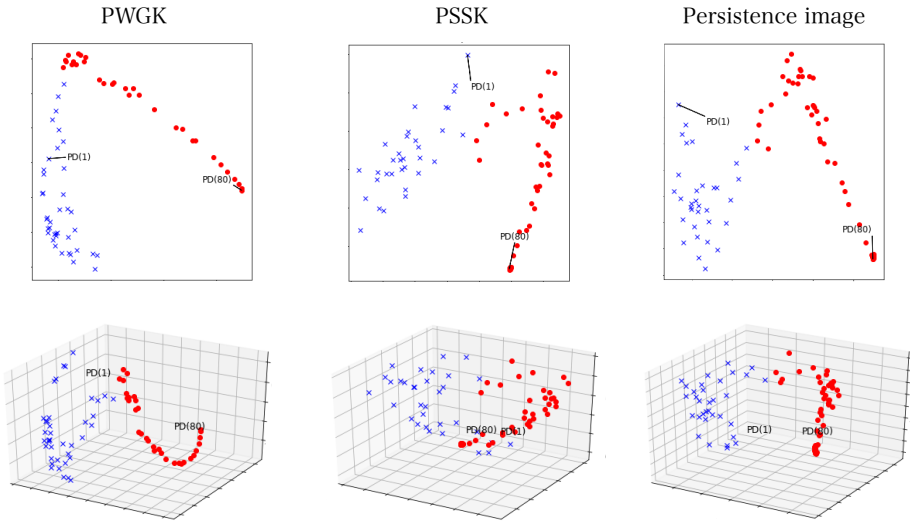

Figure 11: The 2-dimensional and 3-dimensional KPCA plots of the PWGK (contribution rates for 2-dimension: 81.7%, 3-dimension: 92.1%), the PSSK (97.2%, 99.3%) and the persistence image (99.9%, 99.9%).

As we see from the 2-dimensional plots given by KPCA (Figure 11), the PWGK presents sharp change of the gradients between before (blue cross) and after (red circle) the change point determined by the KFDR. This matches with the analysis in physics that expects a sharp change of slope in the temperature-enthalpy plane. This strongly suggests that the glass transition occurs at the detected change point. On the other hand, in the results of PSSK and persistence images we cannot observe a sharp change of the gradients at the boundary of the estimated two phases. We also remark that clearer structures are observed in the 3-dimensional KPCA plots of the PWGK.

4.5 Protein classification

We apply the PWGK to two classification tasks studied in Cang et al. (2015). They introduced the molecular topological fingerprint (MTF) as a feature vector constructed from the persistent homology, and used it for the input to the SVM. The MTF is given by the 13-dimensional vector whose elements consist of the persistences of some specific generators[17] in persistence diagrams. We compare the performance between the PWGK and the MTF method under the same setting of the SVM reported in Cang et al. (2015).

17. The MTF method is not a general method for persistence diagrams because some elements of the MTF vector are specialized for protein data, e.g., the ninth element of the MTF vector is defined by the number of Betti 1 bars that locate at $[4.5, 5.5]$Å, divided by the number of atoms. For the details, see Cang et al. (2015).

Table 2: CV classification rates (%) of SVMs with the PWGK and the MTF (cited from Cang et al. (2015)).

	Protein-Drug	Hemoglobin
PWGK	100	88.90
MTF	(nbd) 93.91 / (bd) 98.31	84.50

The first task is a protein-drug binding problem, where the binding and non-binding of drug to the M2 channel protein of the influenza A virus is to be classified. For each of the two forms, 15 data were obtained by NMR experiments, and 10 data are used for training and the remaining for testing. We randomly generate 100 ways of partitions and calculate the average classification rates.

In the second problem, the taut and relaxed forms of hemoglobin are to be classified. For each form, 9 data were collected by the X-ray crystallography. We select one data from each class for testing and use the remaining for training. All the 81 combinations are performed to calculate the CV classification rates.

The results of the two problems are shown in Table 2. We can see that the PWGK achieves better performance than the MTF in both problems.

5. Conclusion and Discussions

One of the contributions of this paper is to introduce a kernel framework to topological data analysis with persistence diagrams. We applied the kernel embedding approach to vectorize the persistence diagrams, which enables us to utilize any standard kernel methods for data analysis. Another contribution is to propose a kernel specific to persistence diagrams, that is called persistence weighted Gaussian kernel (PWGK). As a significant advantage, our kernel enables one to control the effect of persistence in data analysis. We have also proven the stability property with respect to the distance in the Hilbert space. Furthermore, we have analyzed the synthesized and real data by using the proposed kernel. The change point detection, the principal component analysis, and the support vector machine derived meaningful results for the tasks. From the viewpoint of computations, our kernel can utilize an efficient approximation to compute the Gram matrix.

One of the main theoretical results of this paper is the bottleneck stability of the PWGK (Theorem 9). It is obtained by restricting the class of persistence diagrams to that obtained from ball model filtrations. The reason of this restriction is because the total persistence can be bounded from above independent of the persistence diagram. Thus, one direction to extend this work is to examine the boundedness condition about the total persistence of other persistence diagrams, for example obtained from Rips complexes or sub-level sets.

Another direction to extend this work is to generalize the class of weight functions. The reason of the choice of w_{arc} is mainly for the stability property, but in principle, we can apply any weight function to data analysis. Even if we do not concern about stability properties, which weight function is practically good for data analysis? Suppose generators close to the diagonal are sometimes seen as important features. Then, our statistical framework can treat such small generators as significant ones by a weight function which has large weight

close to the diagonal, while other statistical methods for persistence diagrams always see small generators as noisy ones. In addition, the weight function becomes better when it is constructed to satisfy the assumption (W1) or (W2), which implies the stability property.

Acknowledgement

We thank Ulrich Bauer for giving us useful comments in Section 4.1.1, Mohammad Saadatfar and Takenobu Nakamura for providing experimental and simulation data used in Section 4.3 and 4.4, and the anonymous referees for their valuable comments and suggestions This work is partially supported by JST CREST Mathematics (15656429), JSPS KAKENHI Grant Number 26540016, Structural Materials for Innovation Strategic Innovation Promotion Program D72, Materials research by Information Integration Initiative (MI^2I) project of the Support Program for Starting Up, Innovation Hub from JST, and JSPS Research Fellow (17J02401).

Appendix A. Topological tools

This section summarizes some topological tools used in the paper. To study topological properties algebraically, simplicial complexes are often considered as basic objects. We start with a brief explanation of simplicial complexes, and gradually increase the generality from simplicial homology to singular and persistent homology. For more details, see Hatcher (2002).

A.1 Simplicial complex

We first introduce a combinatorial geometric model called simplicial complex to define homology. Let $P = \{1, \ldots, n\}$ be a finite set (not necessarily points in a metric space). A *simplicial complex* with the vertex set P is defined by a collection S of subsets in P satisfying the following properties:

1. $\{i\} \in S$ for $i = 1, \ldots, n$, and
2. if $\sigma \in S$ and $\tau \subset \sigma$, then $\tau \in S$.

Each subset σ with $q+1$ vertices is called a q-simplex. We denote the set of q-simplices by S_q. A subcollection $T \subset S$ which also becomes a simplicial complex (with possibly less vertices) is called a subcomplex of S.

We can visually deal with a simplicial complex S as a polyhedron by pasting simplices in S into a Euclidean space. The simplicial complex obtained in this way is called a geometric realization, and its polyhedron is denoted by $|S|$. In this context, the simplices with small q correspond to points ($q = 0$), edges ($q = 1$), triangles ($q = 2$), and tetrahedra ($q = 3$).

Example 1 Figure 12 shows two polyhedra of simplicial complexes

$$S = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\},$$
$$T = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}\}.$$

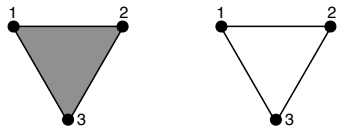

Figure 12: The polyhedra of the simplicial complexes S (left) and T (right).

A.2 Homology

A.2.1 Simplicial homology

The procedure to define homology is summarized as follows:

1. Given a simplicial complex S, build a chain complex $C_*(S)$. This is an algebraization of S characterizing the boundary.

2. Define homology by quotienting out certain subspaces in $C_*(S)$ characterized by the boundary.

We begin with the procedure 1 by assigning orderings on simplices. When we deal with a q-simplex $\sigma = \{i_0, \dots, i_q\}$ as an ordered set, there are $(q+1)!$ orderings on σ. For $q > 0$, we define an equivalence relation $i_{j_0}, \dots, i_{j_q} \sim i_{\ell_0}, \dots, i_{\ell_q}$ on two orderings of σ such that they are mapped to each other by even permutations. By definition, two equivalence classes exist, and each of them is called an oriented simplex. An oriented simplex is denoted by $\langle i_{j_0}, \dots, i_{j_q} \rangle$, and its opposite orientation is expressed by adding the minus $-\langle i_{j_0}, \dots, i_{j_q} \rangle$. We write $\langle \sigma \rangle = \langle i_{j_0}, \dots, i_{j_q} \rangle$ for the equivalence class including $i_{j_0} < \cdots < i_{j_q}$. For $q = 0$, we suppose that we have only one orientation for each vertex.

Let K be a field. We construct a K-vector space $C_q(S)$ as

$$C_q(S) = \mathrm{Span}_K\{\langle \sigma \rangle \mid \sigma \in S_q\}$$

for $S_q \neq \emptyset$ and $C_q(S) = 0$ for $S_q = \emptyset$. Here, $\mathrm{Span}_K(A)$ for a set A is a vector space over K such that the elements of A formally form a basis of the vector space. Furthermore, we define a linear map called the *boundary map* $\partial_q : C_q(S) \to C_{q-1}(S)$ by the linear extension of

$$\partial_q \langle i_0, \dots, i_q \rangle = \sum_{\ell=0}^{q} (-1)^\ell \langle i_0, \dots, \widehat{i_\ell}, \dots, i_q \rangle, \tag{22}$$

where $\widehat{i_\ell}$ means the removal of the vertex i_ℓ. We can regard the linear map ∂_q as algebraically capturing the $(q-1)$-dimensional boundary of a q-dimensional object. For example, the image of the linear map ∂_2 of a basis $\langle 1, 2, 3 \rangle$ in the vector space $C_2(S)$ is given by

$$\partial_2 \langle 1, 2, 3 \rangle = \langle 2, 3 \rangle - \langle 1, 3 \rangle + \langle 1, 2 \rangle = \langle 1, 2 \rangle + \langle 2, 3 \rangle + \langle 3, 1 \rangle \tag{23}$$

as linear combinations of bases in the vector space $C_1(S)$. The above sentence is written only in the language of linear algebra and there is no meaning of $+$ or $-$ in Equation (23) except for its vector space structure. On the other hand, the geometric realization of $\partial_2 \langle 1, 2, 3 \rangle$ is a boundary of $\langle 1, 2, 3 \rangle$ in a geometric sense (see Figure 12). In this way, we analyze geometric properties of a simplicial complex algebraically.

In practice, by arranging some orderings of the oriented q- and $(q-1)$- simplices, we can represent the boundary map as a matrix $M_q = (M_{\sigma,\tau})_{\sigma \in S_{q-1}, \tau \in S_q}$ with the entry $M_{\sigma,\tau} = 0, \pm 1$ given by the coefficient in Equation (23). For the simplicial complex S in Example 1, the matrix representations M_1 and M_2 of the boundary maps are given by

$$M_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \quad M_1 = \begin{bmatrix} -1 & 0 & -1 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \tag{24}$$

Here, the 1-simplices (resp. 0-simplices) are ordered by $\langle 1, 2 \rangle, \langle 2, 3 \rangle, \langle 1, 3 \rangle$ (resp. $\langle 1 \rangle$, $\langle 2 \rangle$, $\langle 3 \rangle$).

We call a sequence of the vector spaces and linear maps

$$\cdots \longrightarrow C_{q+1}(S) \xrightarrow{\partial_{q+1}} C_q(S) \xrightarrow{\partial_q} C_{q-1}(S) \longrightarrow \cdots$$

the *chain complex* of S. As an easy exercise, we can show $\partial_q \circ \partial_{q+1} = 0$. Hence, the subspaces $Z_q(S) = \ker\partial_q$ and $B_q(S) = \mathrm{im}\partial_{q+1}$ satisfy $B_q(S) \subset Z_q(S)$. Then, the q-th (simplicial) *homology* is defined by taking the quotient space

$$H_q(S) = Z_q(S)/B_q(S).$$

Note that $H_q(S)$ is a K-vector space, and the dimension can be considered in a standard way. Intuitively, the dimension of $H_q(S)$ counts the number of q-dimensional holes in S and each generator of the vector space $H_q(S)$ corresponds to these holes. We remark that the homology as a vector space is independent of the orientations of simplices. For $q = 0$, each generator of $H_0(S)$ corresponds to a path-connected component of S. This can be seen from the fact that any two vertices are in the same equivalence class modulo the boundary $B_0(S)$ if and only if they are connected by a path.

For a subcomplex T of S, the inclusion map $\rho : T \hookrightarrow S$ naturally induces a linear map in homology $\rho_q : H_q(T) \to H_q(S)$. Namely, an element $[c] \in H_q(T)$ is mapped to $[c] \in H_q(S)$, where the equivalence class $[c]$ is taken in each vector space.

For example, the simplicial complex S in Example 1 has

$$Z_1(S) = \mathrm{Span}_K[\ 1 \quad 1 \quad -1\ ]^T = B_1(S)$$

from (24). Hence $H_1(S) = 0$, meaning that there are no 1-dimensional hole (ring) in S. On the other hand, since $Z_1(T) = Z_1(S)$ and $B_1(T) = 0$, we have $H_1(T) \cong K$, meaning that T consists of one ring. Hence, the induced linear map $\rho_1 : H_1(T) \to H_1(S)$ means that the ring in T disappears in S under $T \hookrightarrow S$.

A topological space X is called *triangulable* if there exists a geometric realization of a simplicial complex S whose polyhedron is homeomorphic[18] to X. For such a triangulable topological space, the homology is defined by $H_q(X) = H_q(S)$. This is well-defined, since a different geometric realization provides an isomorphic homology.

A.2.2 Singular homology

We here extend the homology to general topological spaces. Let $e_0, \ldots, e_q$ be the standard basis of $\mathbb{R}^{q+1}$ (i.e., $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, 1 at $(i+1)$-th position, and 0 otherwise), and set

$$\Delta_q = \left\{ \sum_{i=0}^{q} \lambda_i e_i \;\middle|\; \sum_{i=0}^{q} \lambda_i = 1, \lambda_i \geq 0 \right\},$$
$$\Delta_q^\ell = \left\{ \sum_{i=0}^{q} \lambda_i e_i \;\middle|\; \sum_{i=0}^{q} \lambda_i = 1, \lambda_i \geq 0, \lambda_\ell = 0 \right\}.$$

We also denote the inclusion by $\iota_q^\ell : \Delta_q^\ell \hookrightarrow \Delta_q$.

For a topological space X, a continuous map $\sigma : \Delta_q \to X$ is called a singular q-simplex, and let X_q be the set of q-simplices. We construct a K-vector space $C_q(X)$ as

$$C_q(X) = \mathrm{Span}_K\{\sigma \mid \sigma \in X_q\}.$$

18. A continuous map $f : X \to Y$ is said to be *homeomorphic* if $f : X \to Y$ is bijective and the inverse $f^{-1} : Y \to X$ is also continuous.

The boundary map $\partial_q : C_q(X) \to C_{q-1}(X)$ is defined by the linear extension of

$$\partial_q \sigma = \sum_{\ell=0}^{q} (-1)^\ell \sigma \circ \iota_q^\ell.$$

Even in this setting, we can show that $\partial_q \circ \partial_{q+1} = 0$, and hence the subspaces $Z_q(X) = \ker\partial_q$ and $B_q(X) = \mathrm{im}\partial_{q+1}$ satisfy $B_q(X) \subset Z_q(X)$. Then, the q-th (singular) *homology* is similarly defined by

$$H_q(X) = Z_q(X)/B_q(X).$$

It is known that, for a triangulable topological space, the homology of this definition is isomorphic to that defined in A.2.1. From this reason, we hereafter identify simplicial and singular homology.

The induced linear map in homology for an inclusion pair of topological space $Y \subset X$ is similarly defined as in A.2.1.

Appendix B. Total persistence

Let (M, d_M) be a triangulable compact metric space. For a Lipschitz function $f : M \to \mathbb{R}$, we define the degree-p total persistence over t by

$$\mathrm{Pers}_p(D_q(\mathbb{S}\mathrm{ub}(f)), t) = \sum_{\substack{x \in D_q(\mathbb{S}\mathrm{ub}(f)) \\ \mathrm{pers}(x) > t}} \mathrm{pers}(x)^p$$

for $0 \le t \le \mathrm{Amp}(f)$, where $\mathrm{Amp}(f) := \max_{\boldsymbol{x} \in M} f(\boldsymbol{x}) - \min_{\boldsymbol{x} \in M} f(\boldsymbol{x})$ is the amplitude of f. Let S be a triangulated simplicial complex of M by a homeomorphism $\vartheta : |S| \to M$. The diameter of a simplex $\sigma \in S$ and the mesh of the triangulation S are defined by $\mathrm{diam}(\sigma) = \max_{\boldsymbol{x}, \boldsymbol{y} \in \sigma} d_M(\vartheta(\boldsymbol{x}), \vartheta(\boldsymbol{y}))$ and $\mathrm{mesh}(S) = \max_{\sigma \in S} \mathrm{diam}(\sigma)$, respectively. Furthermore, let us set $N(r) = \min_{\mathrm{mesh}(S) \le r} \mathrm{card}\,(S)$. Then, the degree-$p$ total persistence over t is bounded from above as follows:

Lemma 13 (Cohen-Steiner et al. (2010)) *Let M be a triangulable compact metric space and $f : M \to \mathbb{R}$ be a tame Lipschitz function. Then,* $\mathrm{Pers}_p(D_q(\mathbb{S}\mathrm{ub}(f)), t)$ *is bounded from above by*

$$t^p N\left(\frac{t}{\mathrm{Lip}(f)}\right) + p \int_{\varepsilon=t}^{\mathrm{Amp}(f)} N\left(\frac{\varepsilon}{\mathrm{Lip}(f)}\right) \varepsilon^{p-1} d\varepsilon,$$

where $\mathrm{Lip}(f)$ *is the Lipschitz constant of f.*

For a compact triangulable subspace M in $\mathbb{R}^d$, the number of d-cubes with length $r > 0$ covering M is bounded from above by $O(\frac{1}{r^d})$, and hence there exists some constant C_M depending only on M such that $N(r) \le \frac{C_M}{r^d}$.

For $p > d$, we can find the upper bounds for the both terms as follows:

$$t^p N\left(\frac{t}{\mathrm{Lip}(f)}\right) \le t^p C_M \frac{\mathrm{Lip}(f)^d}{t^d} \;\to\; 0 \quad (t \to 0)$$

and

$$p\int_{\varepsilon=t}^{\mathrm{Amp}(f)} N\left(\frac{\varepsilon}{\mathrm{Lip}(f)}\right)\varepsilon^{p-1}d\varepsilon \leq \frac{p}{p-d}C_M\mathrm{Lip}(f)^d\mathrm{Amp}(f)^{p-d}.$$

Then, the upper bound of the total persistence $\mathrm{Pers}_p(D_q(\mathbb{S}\mathrm{ub}(f))) = \mathrm{Pers}_p(D_q(\mathbb{S}\mathrm{ub}(f)), 0)$ is given as follows:

Lemma 14 *Let M be a triangulable compact subspace in $\mathbb{R}^d$ and $p > d$. For any Lipschitz function $f : M \to \mathbb{R}$,*

$$\mathrm{Pers}_p(D_q(\mathbb{S}\mathrm{ub}(f))) \leq \frac{p}{p-d}C_M\mathrm{Lip}(f)^d\mathrm{Amp}(f)^{p-d},$$

where C_M is a constant depending only on M.

In the case of a finite subset $X \subset \mathbb{R}^d$, there always exists an R-ball M containing X for some $R > 0$, which is a triangulable compact subspace in $\mathbb{R}^d$. Moreover, by estimating $\mathrm{Lip}(\mathrm{dist}_X)^d\mathrm{Amp}(\mathrm{dist}_X)^{p-d}$, we show Lemma 4 as a corollary of Lemma 14:

Proof [Lemma 4] The Lipschitz constant of dist_X is 1, because, for any $\boldsymbol{x}, \boldsymbol{y} \in M$,

$$\begin{aligned}\mathrm{dist}_X(\boldsymbol{x}) - \mathrm{dist}_X(\boldsymbol{y}) &= \min_{\boldsymbol{x}_i\in X} d_M(\boldsymbol{x}, \boldsymbol{x}_i) - \min_{\boldsymbol{x}_i\in X} d_M(\boldsymbol{y}, \boldsymbol{x}_i)\\ &\leq \min_{\boldsymbol{x}_i\in X}(d_M(\boldsymbol{x}, \boldsymbol{y}) + d_M(\boldsymbol{y}, \boldsymbol{x}_i)) - \min_{\boldsymbol{x}_i\in X} d_M(\boldsymbol{y}, \boldsymbol{x}_i)\\ &= d_M(\boldsymbol{x}, \boldsymbol{y}).\end{aligned}$$

Moreover,

$$\mathrm{Amp}(\mathrm{dist}_X) \leq \mathrm{diam}(M) := \max_{\boldsymbol{x}_i,\boldsymbol{x}_j\in M} d_M(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

because $\min_{\boldsymbol{x}\in M}\mathrm{dist}_X(\boldsymbol{x}) = 0$ and $\max_{\boldsymbol{x}\in M}\mathrm{dist}_X(\boldsymbol{x}) \leq \mathrm{diam}(M)$. Thus, for some constant C_M depending only on M, we have

$$\begin{aligned}\mathrm{Pers}_p(D_q(X)) &= \mathrm{Pers}_p(D_q(\mathbb{S}\mathrm{ub}(\mathrm{dist}_X)))\\ &\leq \frac{p}{p-d}C_M\mathrm{Lip}(\mathrm{dist}_X)^d\mathrm{Amp}(\mathrm{dist}_X)^{p-d}\\ &\leq \frac{p}{p-d}C_M\mathrm{diam}(M)^{p-d}.\end{aligned}$$

■

For a persistence diagram $D = \{x_1, \ldots, x_n\}$, we construct a n-dimensional vector

$$v(D) := (\mathrm{pers}(x_1), \ldots, \mathrm{pers}(x_n)).$$

Then, the degree-p total persistence is represented as

$$\mathrm{Pers}_p(D) = \|v(D)\|_p^p,$$

where $\|\cdot\|_p$ denotes the ℓ^p-norm of $\mathbb{R}^n$. Since $\|v\|_q \leq \|v\|_p$ $(v \in \mathbb{R}^n,\ 1 \leq p \leq q < \infty)$, we have

$$\mathrm{Pers}_q(D)^{\frac{1}{q}} = \|v(D)\|_q \leq \|v(D)\|_p = \mathrm{Pers}_p(D)^{\frac{1}{p}}.$$

Proposition 15 *If $1 \leq p \leq q < \infty$ and $\mathrm{Pers}_p(D)$ is bounded from above, $\mathrm{Pers}_q(D)$ is also bounded from above.*

Appendix C. Lemmata for Section 3.2

Lemma 16 *For any* $x, y \in \mathbb{R}^2$, $\|k_{\mathrm{G}}(\cdot, x) - k_{\mathrm{G}}(\cdot, y)\|_{\mathcal{H}_{k_{\mathrm{G}}}} \leq \frac{\sqrt{2}}{\sigma} \|x - y\|_\infty$.

Proof

$$\begin{aligned}
\|k_{\mathrm{G}}(\cdot, x) - k_{\mathrm{G}}(\cdot, y)\|^2_{\mathcal{H}_{k_{\mathrm{G}}}} &= k_{\mathrm{G}}(x, x) + k_{\mathrm{G}}(y, y) - 2k_{\mathrm{G}}(x, y) \\
&= 1 + 1 - 2e^{-\frac{\|x-y\|^2}{2\sigma^2}} \\
&= 2\left(1 - e^{-\frac{\|x-y\|^2}{2\sigma^2}}\right) \\
&\leq \frac{1}{\sigma^2} \|x - y\|^2 \qquad (25) \\
&\leq \frac{2}{\sigma^2} \|x - y\|^2_\infty . \qquad (26)
\end{aligned}$$

We have used the fact $1 - e^{-t} \leq t$ $(t \in \mathbb{R})$ in Equation (25) and $\|x\|^2 \leq 2\|x\|^2_\infty$ $(x \in \mathbb{R}^2)$ in Equation (26). ■

Lemma 17 *For any* $x, y \in \mathbb{R}^2$, *the difference of persistences* $|\mathrm{pers}(x) - \mathrm{pers}(y)|$ *is less than or equal to* $2\|x - y\|_\infty$.

Proof For $x = (x_1, x_2), y = (y_1, y_2)$, we have

$$\begin{aligned}
|\mathrm{pers}(x) - \mathrm{pers}(y)| &= |(x_2 - x_1) - (y_2 - y_1)| \\
&\leq |x_2 - y_2| + |x_1 - y_1| \\
&\leq 2\|x - y\|_\infty .
\end{aligned}$$

■

Lemma 18 *For any* $x, y \in \mathbb{R}^2$, *we have*

$$|w_{\mathrm{arc}}(x) - w_{\mathrm{arc}}(y)| \leq 2pC \max\{\mathrm{pers}(x)^{p-1}, \mathrm{pers}(y)^{p-1}\} \|x - y\|_\infty .$$

Proof

$$\begin{aligned}
&|w_{\mathrm{arc}}(x) - w_{\mathrm{arc}}(y)| \\
&= |\arctan(C\mathrm{pers}(x)^p) - \arctan(C\mathrm{pers}(y)^p)| \qquad (27) \\
&\leq C\,|\mathrm{pers}(x)^p - \mathrm{pers}(y)^p| \\
&\leq C\,|\mathrm{pers}(x) - \mathrm{pers}(y)|\, p \max\{\mathrm{pers}(x)^{p-1}, \mathrm{pers}(y)^{p-1}\} \qquad (28) \\
&\leq 2pC \max\{\mathrm{pers}(x)^{p-1}, \mathrm{pers}(y)^{p-1}\} \|x - y\|_\infty . \qquad (29)
\end{aligned}$$

We have used the fact that the Lipschitz constant of arctan is 1 in Equation (27),

$$\begin{aligned} s^p - t^p &= (s-t)(s^{p-1} + s^{p-2}t + \cdots + t^{p-1}) \\ &\leq (s-t)p\max\{s^{p-1}, t^{p-1}\} \end{aligned}$$

for any $s, t > 0$ in Equation (28), and Lemma 17 in Equation (29). ■

Appendix D. The number of the random Fourier features

Let $\mathcal{M}$ be a compact subset of $\mathbb{R}^2$ and $\delta > 0$ be a positive number, then it is known (Rahimi and Recht, 2007) that

$$\sup_{x,y\in\mathcal{M}} \left| \mathrm{Re}\left(\frac{1}{M_{\mathrm{rff}}} \sum_{a=1}^{M_{\mathrm{rff}}} \xi_{z_a}(x)\xi_{z_a}(y)^* \right) - e^{-\frac{\|x-y\|^2}{2\sigma^2}} \right| \leq \delta$$

where $\xi_{z_a}(x) = e^{\sqrt{-1}z_a^T x}$ and $M_{\mathrm{rff}} = \Omega(\frac{2}{\delta^2}\log\frac{\sqrt{2}\mathrm{diam}(\mathcal{M})}{\sigma\delta})$. For a collection $\mathcal{D} = \{D_\ell \mid \ell = 1, \ldots, n\}$ of persistence diagrams, the absolute error between the (k_{G}, w)-linear kernel $K_{\mathrm{L}}(D_i, D_j; k_{\mathrm{G}}, w)$ and its random Fourier feature approximation is given by

$$\begin{aligned} &\left| \sum_{x\in D_i}\sum_{y\in D_j} w(x)w(y)\mathrm{Re}\left(\frac{1}{M_{\mathrm{rff}}} \sum_{a=1}^{M_{\mathrm{rff}}} \xi_{z_a}(x)\xi_{z_a}(y)^* \right) - \sum_{x\in D_i}\sum_{y\in D_j} w(x)w(y)e^{-\frac{\|x-y\|^2}{2\sigma^2}} \right| \\ &\leq \sum_{x\in D_i} w(x) \sum_{y\in D_j} w(y) \left| \mathrm{Re}\left(\frac{1}{M_{\mathrm{rff}}} \sum_{a=1}^{M_{\mathrm{rff}}} \xi_{z_a}(x)\xi_{z_a}(y)^* \right) - e^{-\frac{\|x-y\|^2}{2\sigma^2}} \right|. \end{aligned} \tag{30}$$

In order to make the Equation (30) bounded by an arbitrary $\varepsilon > 0$, M_{rff} is given by $\Omega(\frac{2W_{i,j}^2}{\varepsilon^2}\log\frac{\sqrt{2}W_{i,j}\mathrm{diam}(\mathcal{M})}{\sigma\varepsilon})$ where $W_{i,j} := \sum_{x\in D_i} w(x) \sum_{y\in D_j} w(y)$. In this case, we can define the subset $\mathcal{M}$ by $\bigcup_{\ell=1}^n D_\ell$. When we calculate several $K_{i,j} := K_{\mathrm{L}}(D_i, D_j; k_{\mathrm{G}}, w)$ of Section 4.3 without approximation[19], we observed $K_{i,j} \approx 10^8$. Since the true values are huge, we consider 5% relative error for Equation (30) and set $\varepsilon := 0.05K_{i,j}$. Then,

$$\frac{2W_{i,j}^2}{\varepsilon^2}\log\frac{\sqrt{2}W_{i,j}\mathrm{diam}(\mathcal{M})}{\sigma\varepsilon} = \frac{800W_{i,j}^2}{K_{i,j}^2}\log\frac{20\sqrt{2}W_{i,j}\mathrm{diam}(\mathcal{M})}{\sigma K_{i,j}} =: M_{i,j}.$$

In Section 4.3 and Section 4.4, we observed $\frac{W_{i,j}}{K_{i,j}} \approx 2.5$ and $\frac{\mathrm{diam}(\mathcal{M})}{\sigma} \approx 10$ from several computation without approximation, and $M_{i,j} \approx 800 \cdot (2.5)^2 \log(20\sqrt{2} \cdot 3 \cdot 10) \approx 3 \cdot 10^4$. Thus, the approximation (30) with $M_{\mathrm{rff}} = 10^5$, which is used in our experiments, gives 95% accuracy in the sense of the relative error.

19. Even though a single $K_{i,j}$ can be computed in several hours on MacBook Pro, 2.6 GHz Intel Core i5, 8 GB 1600 MHz DDR3, the total $\frac{1}{2}n(n-1)$ computations of $K_{i,j}$ for the Gram matrix cause huge computational time.

References

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

Anonymous. What is random packing? *Nature*, 239:488–489, 1972.

Ulrich Bauer, Michael Kerber, Jan Reininghaus, and Hubert Wagner. *Mathematical Software – ICMS 2014: 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings*, chapter PHAT – Persistent Homology Algorithms Toolbox, pages 137–143. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015.

Peter Bubenik and Paweł Dłotko. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78:91–114, 2017.

Peter Bubenik and Jonathan A Scott. Categorification of persistent homology. *Discrete & Computational Geometry*, 51(3):600–627, 2014.

Zixuan Cang, Lin Mu, Kedi Wu, Kristopher Opron, Kelin Xia, and Guo-Wei Wei. A topological approach for protein classification. *Molecular Based Mathematical Biology*, 3 (1), 2015.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46 (2):255–308, 2009.

Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International journal of computer vision*, 76(1): 1–12, 2008.

Mathieu Carrière, Steve Oudot, and Maks Ovsjanikov. Local signatures using persistence diagrams. preprint, 2015.

Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014.

Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16:3603–3635, 2015.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have l_p-stable persistence. *Foundations of computational mathematics*, 10(2): 127–139, 2010.

Tran Kai Frank Da, Sébastien Loriot, and Mariette Yvinec. 3D alpha shapes. In *CGAL User and Reference Manual*. CGAL Editorial Board, 4.7 edition, 2015. URL `http://doc.cgal.org/4.7/Manual/packages.html#PkgAlphaShapes3Summary`.

Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358, 2007.

Joseph Diestel and J Jerry Uhl Jr. Vector measures. with a foreword by bj pettis. mathematical surveys, no. 15. *American Mathematical Society, Providence, RI*, 56:12216, 1977.

Pietro Donatini, Patrizio Frosini, and Alberto Lovato. Size functions for signature recognition. In *Vision Geometry VII*, volume 3454, pages 178–184. International Society for Optics and Photonics, 1998.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002.

Stephen R. Elliott. *Physics of amorphous materials (2nd)*. Longman London; New York, 1990.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

Nicolas Francois, Mohammad Saadatfar, R Cruikshank, and A Sheppard. Geometrical frustration in amorphous and partially crystallized packings of spheres. *Physical review letters*, 111(14):148001, 2013.

Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramar, Konstantin Mischaikow, and Vidit Nanda. A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics*, 32(1):1–17, 2015.

Neville G Greaves and Sabyasachi Sen. Inorganic glasses, glass-forming liquids and amorphizing solids. *Advances in Physics*, 56(1):1–166, 2007.

Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592, 2007.

Zaid Harchaoui, Eric Moulines, and Francis R. Bach. Kernel change-point analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2009.

Allen Hatcher. *Algebraic topology*. Cambridge University Press, 2002.

Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 113(26):7035–7040, 2016.

Peter M Kasson, Afra Zomorodian, Sanghyun Park, Nina Singhal, Leonidas J Guibas, and Vijay S Pande. Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23(14):1753–1759, 2007.

Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence weighted gaussian kernel for topological data analysis. In *International Conference on Machine Learning*, pages 2004–2013, 2016.

Hyekyoung Lee, Moo K Chung, Hyejin Kang, Bung-Nyun Kim, and Dong Soo Lee. Discriminative persistent homology of brain networks. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 841–844. IEEE, 2011.

David Lopez-Paz, Krikamol Muandet, and Benjamin Recht. The randomized causation coefficient. *Journal of Machine Learning Research*, 16(16):2901–2907, 2015.

Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pages 10–18, 2012.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

Takenobu Nakamura, Yasuaki Hiraoka, Akihiko Hirata, Emerson G Escolar, and Yasumasa Nishiura. Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26(304001), 2015.

Giovanni Petri, Paul Expert, Federico Turkheimer, Robin Carhart-Harris, David Nutt, Peter J Hellyer, and Francesco Vaccarino. Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101):20140873, 2014.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2007.

Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 4741–4748, 2015.

Vanessa Robins and Katharine Turner. Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Physica D: Nonlinear Phenomena*, 334:99–117, 2016.

Mohammad Saadatfar, Hiroshi Takeuchi, Vanessa Robins, Nicolas Francois, and Yasuaki Hiraoka. Pore configuration landscape of granular crystallization. *Nature Communications*, 8:15082 EP, 2017.

Gurjeet Singh, Facundo Memoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L Ringach. Topological analysis of population activity in visual cortex. *Journal of vision*, 8(8):11, 2008.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *In Algorithmic Learning Theory: 18th International Conference*, pages 13–31. Springer, 2007.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *The Journal of Machine Learning Research*, 12:2389–2410, 2011.

Zoltán Szabó, Bharath Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

Salvatore Torquato, Thomas M Truskett, and Pablo G Debenedetti. Is random close packing of spheres well defined? *Physical review letters*, 84(10):2064, 2000.

Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International journal for numerical methods in biomedical engineering*, 30 (8):814–844, 2014.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.