Persistence weighted Gaussian kernel for topological data analysis

Genki Kusano[1] GENKSN@GMAIL.COM
Kenji Fukumizu[2] FUKUMIZU@ISM.AC.JP
Yasuaki Hiraoka[1] HIRAOKA@WPI-AIMR.TOHOKU.AC.JP
[1]Tohoku University, [2]The Institute of Statistical Mathematics

Abstract

Topological data analysis (TDA) is an emerging mathematical concept for characterizing shapes in complex data. In TDA, persistence diagrams are widely recognized as a useful descriptor of data, and can distinguish robust and noisy topological properties. This paper proposes a kernel method on persistence diagrams to develop a statistical framework in TDA. The proposed kernel satisfies the stability property and provides explicit control on the effect of persistence. Furthermore, the method allows a fast approximation technique. The method is applied into practical data on proteins and oxide glasses, and the results show the advantage of our method compared to other relevant methods on persistence diagrams.

1. Introduction

Recent years have witnessed an increasing interest in utilizing methods of algebraic topology for statistical data analysis. This line of research 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 & Ghrist, 2007), biology (Kasson et al., 2007; Xia & Wei, 2014), brain science (Lee et al., 2011; Petri et al., 2014; Singh et al., 2008), biochemistry (Gameiro et al., 2013), and material science (Nakamura et al., 2015a;b). In many of these applications, it is not straightforward to provide feature vectors or descriptors of data from their complicated geometric configurations. The aim of TDA is to detect informative topological properties (e.g., connected components, rings, and cavities) from such data, and use them as descriptors.

A key mathematical apparatus in TDA is *persistent homology*, which is an algebraic method for extracting robust topological information from data. To provide some intuition for the persistent homology, let us consider a typical way of constructing persistent homology from data points in a Euclidean space, assuming that the data lie on a submanifold. The aim is to make inference on the topology of the underlying manifold from finite data. 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, the union of r-balls consists simply of the disjoint r-balls. On the other hand, if r is too large, the union becomes a contractible space. *Persistent homology* (Edelsbrunner et al., 2002) can consider *all* r simultaneously, and provides an algebraic expression of topological properties together with their persistence over r. We give a brief explanation of persistent homology in Supplementary material A.3.

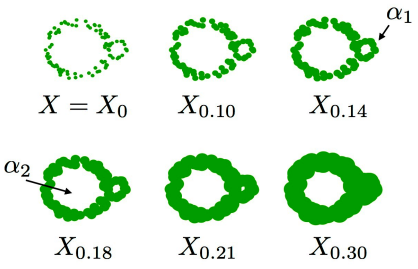


Figure 1. The union X_r of r-balls at points sampled from annuli with noise.

The persistent homology can be visualized in a compact form called a *persistence diagram* $D = \{(b_i, d_i) \in \mathbb{R}^2 \mid i \in I,\ b_i \leq d_i\}$, and this paper focuses on persistence diagrams, since the contributions of this paper can be fully explained in terms of persistence diagrams. Every point $(b_i, d_i) \in D$, called a *generator* of the persistent homology, represents a topological property (e.g., connected components, rings, and cavities) which appears at X_{b_i} and disappears at X_{d_i} in the r-ball model. Then, the *persistence*

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

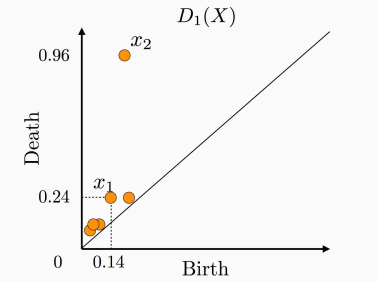


Figure 2. The persistence diagram of the r-ball model in Figure 1. The point x_1 represents the ring α_1, which is born at $r = 0.14$ and dies at $r = 0.24$. The noisy rings are plotted as the points close to the diagonal.

$d_i - b_i$ of the generator shows the robustness of the topological property under the radius parameter. As an example shown in Figure 1, the rings α_1, α_2 and other tiny ones are expressed as x_1, x_2, and the other points in the persistence diagram shown in Figure 2. A topological property with large persistence can be regarded as a reliable structure, while that with small persistence (points close to the diagonal) is likely to be noise. In this way, persistence diagrams encode topological and geometric information of data points.

While persistence diagrams nowadays start to be applied to various problems such as the ones listed in the beginning of this section, statistical or machine learning methods for analysis on persistence diagrams are still limited. In TDA, analysts often elaborate only one 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 statistical framework on persistence diagrams is a significant issue for further success of TDA.

To this aim, this paper discusses kernel methods for persistence diagrams (see Figure 3). 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. Here, 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, leading various new methods for nonparametric inference (Muandet et al., 2012; Smola et al., 2007; Song et al., 2013). It is known (Sriperumbudur et al., 2011) that, with an appropriate choice of kernels, a signed 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 non-negative measure, 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.

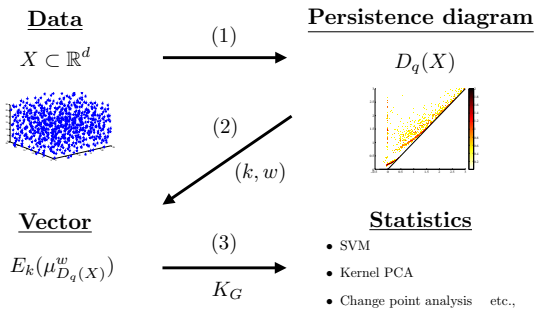


Figure 3. (1) A data X is transformed into a persistence diagram $D_q(X)$ (Section 2.1). (2) $D_q(X)$ is mapped to a vector $E_{k_G}(\mu^{w_{\mathrm{arc}}}_{D_q(X)})$, where k is a kernel and w is a weight controlling the effect of persistence (Section 3.1). (3) This vector provides statistical methods for persistence diagrams (Section 4).

For embedding persistence diagrams in an RKHS, we propose a useful class of positive definite kernels, called *persistence weighted Gaussian kernel* (PWGK). It is important that the PWGK can discount the contributions of generators close to the diagonal (small persistence), since in many applications those generators are likely to be noise. The advantages of this kernel are as follows. (i) We can explicitly control the effect of persistence, 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 & Recht, 2007), and thus it is applicable to persistence diagrams with a large number of generators, which are seen in practical examples (Section 4).

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

There are already some relevant works on statistical approaches to persistence diagrams. Some studies discuss how to transform a persistence diagram to a vector (Bubenik, 2015; Cang et al., 2015; Carriere et al., 2015; Robins & Turner, 2015). 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 and support vector machines. The most relevant to our method is Reininghaus et al. (2015) (see also Kwitt et al. (2015)), where they vectorize a persistence diagram by using the difference of two Gaussian kernels evaluated at symmet-

ric points with respect to the diagonal so that it vanishes on the diagonal. We will show detailed comparisons between this method and ours. Additionally, there are some works discussing statistical properties of persistence diagrams for random data points: Chazal et al. (2014b) 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.

2. Background

We review the concepts of persistence diagrams and kernel methods. For readers who are not familiar with algebraic topology and homology, we give a brief summary in Supplementary material. See also Hatcher (2001) as an accessible introduction to algebraic topology.

2.1. Persistence diagram

Let $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ be a finite subset in a metric space (M, d_M). To analyze topological properties of X, let us consider a fattened ball model $X_r = \bigcup_{i=1}^n B(\boldsymbol{x}_i; r)$ consisting of balls $B(\boldsymbol{x}_i; r) = \{\boldsymbol{x} \in M \mid d_M(\boldsymbol{x}_i, \boldsymbol{x}) \leq r\}$ with radius r, and use the *homology* $H_q(X_r)$ to describe the topology of X_r. Here, for a topological space S, its *q-th homology* $H_q(S)$ $(q = 0, 1, \dots)$ is defined as a vector space, and its dimension $\dim H_q(S)$ counts the number of connected components $(q = 0)$, rings $(q = 1)$, cavities $(q = 2)$, and so on[1]. For the precise definition of homology, see Supplementary material. For example, $X_{0.21}$ in Figure 1 consists of one connected component and two rings, and hence $\dim H_0(H_{0.21}) = 1$ and $\dim H_1(X_{0.21}) = 2$.

Because of $X_r \subset X_s$ for $r \leq s$, the set $\mathbb{X} = \{X_r \mid r \geq 0\}$ becomes a filtration[2]. When the radius changes as in Figure 1, a new generator $\alpha_i \in H_q(X_r)$ appears at some radius $r = b_i$ and disappears at a radius $r = d_i$ larger than b_i (called *birth* and *death*, respectively). By gathering all generators α_i $(i \in I)$ in the filtration $\mathbb{X}$, we obtain the collection of these birth-death pairs $\underline{D}_q(X) = \{(b_i, d_i) \in \mathbb{R}^2 \mid i \in I\}$ as a multi-set[3]. The *persistence diagram* $D_q(X)$ is defined by the disjoint union of $\underline{D}_q(X)$ and the diagonal set $\Delta = \{(a, a) \mid a \in \mathbb{R}\}$ counted with infinite multiplicity. A point $x = (b, d) \in D_q(X)$ is also called a *generator* of the persistence diagram. The *persistence* $\mathrm{pers}(x) := d - b$ of x is its lifetime and measures the robustness of x in the filtration. We will see shortly that the diagonal set is included in a persistence diagram to simplify the definition of a distance on persistence diagrams.

Figure 2 shows the persistence diagram $D_1(X)$ of X given in Figure 1. The generators x_1 and x_2 correspond to the rings α_1 and α_2 in Figure 1, respectively. The persistence of x_2 is the longest, while the other generators including x_1 have small persistences, implying that they can be seen as noisy rings. Although there are no topological rings in X itself, the persistence diagram $D_1(X)$ shows that there is a robust ring α_2 and several noisy rings in $\mathbb{X}$. In this way, the persistence diagram provides an informative topological summary of X_r over all r.

We remark that, in the finite fattened ball model, there is only one generator in $D_0(X)$ which does not disappear in the filtration; its lifetime is ∞. Thus, from now on, we deal with $D_0(X)$ by removing this infinite lifetime generator in order to simplify the notation[4]. We also note that the cardinality of $\underline{D}_q(X)$ obtained from the finite fattened ball model is finite.

2.2. Stability with respect to d_B

Any statistical data involve noise or stochasticity, and thus it is desired that the persistence diagrams are stable under perturbation of data. A popular measure to study the similarity between two persistence diagrams D and E is the *bottleneck distance*

$$d_B(D, E) := \inf_{\gamma} \sup_{x \in D} \|x - \gamma(x)\|_\infty ,$$

where γ ranges over all multi-bijections[5] from D to E[6]. Note that the cardinalities of D and E are equal by considering the diagonal set Δ with infinite multiplicity. As a distance between finite sets X, Y in a metric space M, let us recall the *Hausdorff distance* given by

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

Then, we have the following stability property (for more general settings, see Chazal et al. (2014a)).

[1] Throughout this paper we use a field coefficient for homology.

[2] A *filtration* is a family of subsets $\{X_a \mid a \in A\}$ indexed by a totally ordered set A such that $X_a \subset X_b$ for $a \leq b$.

[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 is called the *reduced persistence diagram*.

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

[6] For $z = (z_1, z_2) \in \mathbb{R}^2$, $\|z\|_\infty$ denotes $\max(|z_1|, |z_2|)$.

Proposition 2.1. *(Chazal et al., 2014a; Cohen-Steiner et al., 2007) Let X and Y be finite subsets in a metric space (M, d_M). Then the persistence diagrams satisfy*

$$d_B(D_q(X), D_q(Y)) \leq d_H(X, Y).$$

Proposition 2.1 provides a geometric intuition of the stability of persistence diagrams. Assume that X is the true location of points and Y is a data obtained from skewed measurement with $\varepsilon = d_H(X, Y)$ (Figure 4). If there is a point $(b, d) \in D_q(Y)$, then we can find at least one generator in X which is born in $(b-\varepsilon, b+\varepsilon)$ and dies in $(d-\varepsilon, d+\varepsilon)$. Thus, the stability guarantees the similarity of two persistence diagrams, and hence we can infer the true topological features from one persistence diagram.

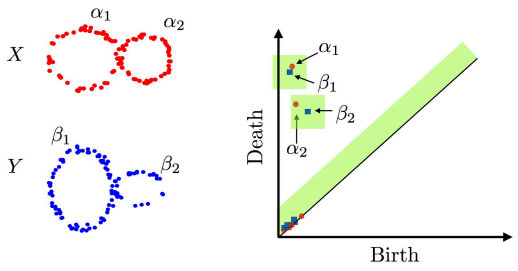


Figure 4. Two data X and Y (left) and their persistence diagrams (right). The green region is an ε-neighborhood of $D_q(Y)$.

2.3. Kernel methods for representing signed measures

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 $k_G(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ $(\sigma > 0)$, where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. 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 (Smola et al., 2007; Sriperumbudur et al., 2011). Let Ω be a locally compact Hausdorff space, $M_b(\Omega)$ be the space of all finite signed Radon measures on Ω, and k be a bounded measurable kernel on Ω. Then we define a mapping from $M_b(\Omega)$ to $\mathcal{H}_k$ by

$$E_k : M_b(\Omega) \to \mathcal{H}_k, \quad \mu \mapsto \int k(\cdot, x) d\mu(x). \tag{1}$$

The integral should be understood as the Bochner integral (Diestel & Uhl, 1977), which exists here, since $\int \|k(\cdot, x)\|_{\mathcal{H}_k} d\mu(x)$ is finite.

For a locally compact Hausdorff space Ω, let $C_0(\Omega)$ denote the space of continuous functions vanishing at infinity[7]. A kernel k on Ω is said to be C_0-kernel if $k(x, x)$ is of $C_0(\Omega)$ as a function of x. 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, by the mapping (1), the vector $E_k(\mu)$ in the RKHS uniquely determines the finite signed measure μ, and thus serves as a representation of μ.

Proposition 2.2 ((Sriperumbudur et al., 2011))**.** *If k is C_0-universal, the mapping E_k is injective. Thus,*

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

defines a distance on $M_b(\Omega)$.

3. Kernel methods for persistence diagrams

We propose a kernel for persistence diagrams, called the *Persistence Weighted Gaussian Kernel* (PWGK), to embed the diagrams into an RKHS. This vectorization of persistence diagrams enables us to apply any kernel methods to persistence diagrams. We show the stability theorem with respect to the distance defined by the embedding, and discuss efficient computation of the PWGK.

3.1. Persistence weighted Gaussian kernel

We propose a method for vectorizing persistence diagrams using the kernel embedding (1) by regarding a persistence diagram as a discrete measure. In vectorizing persistence diagrams, it is important to discount the effect of generators located near the diagonal, since they tend to be caused by noise. To this end, we explain slightly different two ways of embeddings, which turn out to introduce the same inner products for two persistence diagrams.

First, for a persistence diagram D, we introduce a weighted 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. The weight function $w(x)$ discounts the effect of generators close to the diagonal, and a concrete choice will be discussed later. As discussed in Section 2.3, given a C_0-universal kernel k on $\mathbb{R}^2_{ul} := \{(b, d) \in \mathbb{R}^2 \mid b < d\}$, the measure μ_D^w can be embedded as an element of the RKHS $\mathcal{H}_k$ via

$$\mu_D^w \mapsto E_k(\mu_D^w) := \sum_{x \in D} w(x) k(\cdot, x). \tag{2}$$

From Proposition 2.2, this mapping does not lose any information about persistence diagrams, and $E_k(\mu_D^w) \in \mathcal{H}_k$ serves as a representation of the persistence diagram.

[7] A function f is said to vanish at infinity if for any $\varepsilon > 0$ there is a compact set $K \subset \Omega$ such that $\sup_{x \in K^c} |f(x)| \leq \varepsilon$.

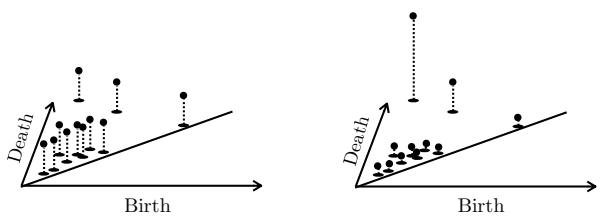


Figure 5. Unweighted (left) and weighted (right) measures.

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, and consider the mapping

$$E_{k^w} : \mu_D \mapsto \sum_{x\in D} w(x)w(\cdot)k(\cdot,x) \in \mathcal{H}_{k^w}. \tag{3}$$

This also defines vectorization of persistence diagrams, and it is essentially equivalent to the first one, as seen from the next proposition (See Supplementary material for the proof.).

Proposition 3.1. *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.

Note that under the identification of Proposition 3.1, we have

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

and thus the two constructions introduce the same similarity (and hence distance) among persistence diagrams. We apply methods of data analysis to vector representations $E_k(\mu_D^w)$ or $E_{k^w}(\mu_D)$. The first construction may be more intuitive by the direct weighting of a measure, while the second one is also practically useful since all the parameter tuning is reduced to kernel choice.

For a practical purpose, we propose to use the Gaussian kernel $k_G(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ $(\sigma > 0)$ for k, and $w_{\mathrm{arc}}(x) = \arctan(C\mathrm{pers}(x)^p)$ $(C, p > 0)$ for a weight function. The corresponding positive definite kernel is

$$k_{PWG}(x,y) = w_{\mathrm{arc}}(x)w_{\mathrm{arc}}(y)e^{-\frac{\|x-y\|^2}{2\sigma^2}}. \tag{4}$$

We call it *Persistence Weighted Gaussian Kernel* (PWGK). Since the Gaussian kernel is C_0-universal and $w > 0$ on $\mathbb{R}^2_{ul}$, $d_{k_G}^{w_{\mathrm{arc}}}(D,E) := \|E_{k_G}(\mu_D^{w_{\mathrm{arc}}}) - E_{k_G}(\mu_E^{w_{\mathrm{arc}}})\|_{\mathcal{H}_{k_G}}$ defines a distance on the persistence diagrams. We also note that w_{arc} is an increasing function with respect to persistence. Hence, a noisy (resp. essential) generator x gives a small (resp. large) value $w_{\mathrm{arc}}(x)$. By adjusting the parameters C and p, we can control the effect of the persistence.

3.2. Stability with respect to $d_{k_G}^{w_{\mathrm{arc}}}$

Given a data X, we vectorize the persistence diagram $D_q(X)$ as an element $E_{k_G}(\mu_{D_q(X)}^{w_{\mathrm{arc}}})$ of the RKHS. Then, for practical applications, this map $X \mapsto E_{k_G}(\mu_{D_q(X)}^{w_{\mathrm{arc}}})$ should be stable with respect to perturbations to the data as discussed in Section 2.2. The following theorem shows that the map has the desired property (See Supplementary material for the proof.).

Theorem 3.2. *Let M be a compact subset in $\mathbb{R}^d$, $X, Y \subset M$ be finite subsets and $p > d+1$. Then*

$$d_{k_G}^{w_{\mathrm{arc}}}(D_q(X), D_q(Y)) \le L(M,d;C,p,\sigma)d_H(X,Y),$$

where $L(M,d;C,p,\sigma)$ is a constant depending on M, d, C, p, σ.

Let $\mathcal{P}_{\mathrm{finite}}(M)$ be the set of finite subsets in a compact subset $M \subset \mathbb{R}^d$. Since the constant $L(M,d;C,p,\sigma)$ is independent of X and Y, Theorem 3.2 concludes that the map

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

is Lipschitz continuous. To the best of our knowledge, a similar stability result has not been obtained for the other Gaussian type kernels (e.g., Reininghaus et al. (2015) does not deal with the Hausdorff distance.). Our stability result is achieved by incorporating the weight function w_{arc} with appropriate choice of p.

3.3. Kernel methods on RKHS

Once persistence diagrams are represented by the vectors in an RKHS, we can apply any kernel methods to those vectors. The simplest choice is to consider the linear kernel

$$\begin{aligned} K_L(D,E) &= \langle E_{k_G}(\mu_D^{w_{\mathrm{arc}}}), E_{k_G}(\mu_E^{w_{\mathrm{arc}}})\rangle_{\mathcal{H}_{k_G}} \\ &= \sum_{x\in\underline{D}}\sum_{y\in\underline{E}} w_{\mathrm{arc}}(x)w_{\mathrm{arc}}(y)k_G(x,y) \end{aligned}$$

on the RKHS. We can also consider a nonlinear kernel on the RKHS, such as the Gaussian kernel:

$$K_G(D,E) = \exp\left(-\frac{d_{k_G}^{w_{\mathrm{arc}}}(D,E)^2}{2\tau^2}\right), \tag{5}$$

where τ is a positive parameter and

$$\begin{aligned} d_{k_G}^{w_{\mathrm{arc}}}(D,E)^2 &:= \|E_{k_G}(\mu_D^{w_{\mathrm{arc}}}) - E_{k_G}(\mu_E^{w_{\mathrm{arc}}})\|^2_{\mathcal{H}_{k_G}} \\ &= \sum_{x\in\underline{D}}\sum_{x'\in\underline{D}} w_{\mathrm{arc}}(x)w_{\mathrm{arc}}(x')k_G(x,x') \\ &\quad + \sum_{y\in\underline{E}}\sum_{y'\in\underline{E}} w_{\mathrm{arc}}(y)w_{\mathrm{arc}}(y')k_G(y,y') \\ &\quad - 2\sum_{x\in\underline{D}}\sum_{y\in\underline{E}} w_{\mathrm{arc}}(x)w_{\mathrm{arc}}(y)k_G(x,y). \end{aligned}$$

Note that we can observe better performance with nonlinear kernels for some complex tasks (Muandet et al., 2012) and the RKHS Gaussian kernel K_G is universal (Christmann & Steinwart, 2010). In Section 4, we apply K_G for SVM, kernel PCA, and kernel change point detection.

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 4.3, for example, the number of generators is 30000, while $n = 80$.

If the persistence diagrams contain at most m points, each element of the Gram matrix $(K_G(D_i, D_j))_{i,j=1,\dots,n}$ involves $O(m^2)$ evaluation 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, since in many applications n is relatively small

We solve this computational issue by using the random Fourier features (Rahimi & Recht, 2007). To be more precise, let $z_1, \dots, z_M$ 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 $\frac{1}{M}\sum_{a=1}^{M} e^{\sqrt{-1}z_a x}(e^{\sqrt{-1}z_a y})^*$, where $*$ denotes the complex conjugation. Then, $\sum_{x\in\underline{D}_i}\sum_{y\in\underline{D}_j} w(x)w(y)k_G(x,y)$ is approximated by $\frac{1}{M}\sum_{a=1}^{M} B_i^a (B_j^a)^*$, where $B_\ell^a = \sum_{x\in\underline{D}_\ell} w(x)e^{\sqrt{-1}z_a x}$. As a result, the computational complexity of the approximated Gram matrix is $O(mnM + n^2M)$.

We note that 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^{-3} with $M = 10^3$. As a whole, these m^2 errors may cause a critical error to the approximation. Moreover, if σ is largely deviated from the ensemble $\|x-y\|$ for $x \in \underline{D}_i, y \in \underline{D}_j$, 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, choice of parameter is important. For supervised learning such as SVM, we use the cross-validation (CV) approach. For unsupervised case, we follow the heuristics proposed in Gretton et al. (2007). In Section 4.3, we set $\sigma = \mathrm{median}\{\sigma(D_\ell) \mid \ell = 1, \dots, n\}$, where $\sigma(D) = \mathrm{median}\{\|x_i - x_j\| \mid x_i, x_j \in \underline{D},\ i < j\}$, so that σ takes close values to many $\|x-y\|$. For the parameter C, we also set $C = (\mathrm{median}\{\mathrm{pers}(D_\ell) \mid \ell = 1, \dots, n\})^{-p}$, where $\mathrm{pers}(D) = \mathrm{median}\{\mathrm{pers}(x_i) \mid x_i \in \underline{D}\}$. Similarly, τ is defined by $\mathrm{median}\{d_{k_G}^{w_{\mathrm{arc}}}(D_i, D_j) \mid 1 \le i < j \le n\}$. In this paper, since all points of data are in $\mathbb{R}^3$, we set $p = 5$ from the assumption $p > d + 1$ in Theorem 3.2.

4. Experiments

We demonstrate the performance of the PWGK using synthesized and real data. In this section, all persistence diagrams are 1-dimensional (i.e., rings) and computed by CGAL (Da et al., 2015) and PHAT (Bauer et al., 2014).

4.1. Comparison to the persistence scale space kernel

The most relevant work to our method is Reininghaus et al. (2015). They propose a positive definite kernel called *persistence scale space kernel* (PSSK for short) K_{PSS} on the persistence diagrams:

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

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

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 must control the global bandwidth and the discounting effect. (ii) The approximation by the random Fourier features is not applicable to the PSSK, since it is not shift-invariant in total. We also note that, in Reininghaus et al. (2015), only the linear kernel is considered on the RKHS, while our approach involves a nonlinear kernel on the RKHS.

Regarding the approximation of the PSSK, Nyström method (Williams & Seeger, 2001) or incomplete Cholesky factorization (Fine & Scheinberg, 2001) can be applied. In evaluating the kernels, we need to calculate $(k(x,y))_{x\in\underline{D}_i, y\in\underline{D}_j}$, which is not symmetric. We then need to apply Nyström or incomplete Cholesky to the symmetric but big positive definite matrix $(k(x,y))_{\substack{x\in\underline{D}_i, y\in\underline{D}_j \\ i,j=1,\dots,n}}$ of size $O(nm)$. Either, we need to apply incomplete Cholesky to the non-symmtric matrix $(k(x,y))_{x\in\underline{D}_i, y\in\underline{D}_j}$ for all the combination of (i,j), which requires considerable computational cost for large n. In contrast, the random Fourier features can be applied to the kernel *function* irrespective to evaluation points, and the same Fourier expansion can be applied to any (i,j). This guarantees the

Table 1. Results of SVM with PWGK, PSSK, and Gaussian. Average classification rates (%) for 99 test data sets are shown.

	RKHS-Linear	RKHS-Gauss
PWGK	60.0	83.0
PSSK	49.5	54.5
Gauss	57.6	69.7

efficient computational cost.

The detailed comparisons will be experimentally verified in Sections 4.2 and 4.3. With respect to the parameter t in the PSSK, since $e^{-\frac{\|x-\bar{y}\|^2}{8t}}$ is only used to vanish the value of the feature map $\Phi_t(D)$ on the diagonal, we set $t = \frac{\sigma^2}{4}$ by using the same σ defined in Section 3.4.

4.2. Classification with synthesized data

We first use the proposed method for a classification task with SVM, and compare the performance with the PSSK. The synthesized data are generated as follows. Each data set assumes one or two circles, and data points are located at even spaces along the circle(s). It always contains one larger circle S_1 of radius r_1^o raging from 1 to 10, and it may have a smaller circle S_2 of radius 0.2 (10 points) with probability 1/2. Roughly speaking, the class label Y is made by $\mathbf{XOR}(z_0, z_1)$, where z_i $(i = 0, 1)$ is a binary variable: $z_0 = 1$ if the smaller S_2 exists, and $z_1 = 1$ if the birth and death (b^o, d^o) of the generator corresponding to S_1 satisfies $b^o < A_B$ and $d^o > A_D$ for fixed thresholds A_B, A_D. We can control b^o and d^o by choosing the number of points N_1^o along S_1 (for birth) and the radius r_1^o (for death). To generate the data points, we add noise effects to make the classification harder: the radius r_1 and sample size N_1 are in fact given by adding noise to r_1^0 and N_1^o, and S_1 are made according to the shifted r_1 and N_1, while the class label is given by the non-shifted r_1^o and N_1^o. For the precise description of the data generation procedure, see Supplementary material. By this construction, the classifier needs to look at both of the location of the generator and the existence of the generator for the smaller one around the diagonal.

SVMs are trained with persistence diagrams given by 100 data sets, and evaluated with 99 independent test data sets. For the kernel on RKHS, we used both of the linear and Gaussian kernels. 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 the weight of the PWGK is set to be 5. The variance parameter in the RKHS-Gaussian kernel is set by the median heuristics. We also apply the Gaussian kernel (without any weights) for embedding persistence diagrams to RKHS. In Table 1, we can see that the PSSK does not work well for this problem, even worse than the Gaussian kernel, and the classification rate by the linear RKHS kernel used originally in Reininghaus et al. (2015) is almost the chance level. This must be caused by the difficulty in handling the global location of generators and close look around the diagonal simultaneously. This classification task involves strong nonlinearity on the RKHS, as seen in the large improvement by PWGK+Gauss kernel.

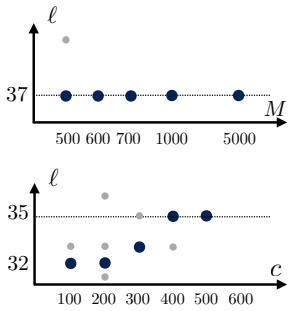
4.3. Analysis of SiO_2

In this experiment, we compare the PWGK and the PSSK to the non-trivial problem of glass transition on SiO_2, focusing also on their computational efficiency.

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 & Sen, 2007). For estimating the glass transition temperature by simulations, we first prepare atomic configurations of SiO_2 for a certain range of temperatures, 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 the two lines are close to each other, determining the interval is a subtle problem, and usually the rough estimate of the interval is only available. Hence, it is desired to develop a mathematical framework to detect the glass transition temperature.

Our strategy is to regard the glass transition temperature as the change point and detect it from a collection $\mathcal{D} = \{D_\ell \mid \ell = 1, \dots, n\}$ of persistence diagrams made by atomic configurations of SiO_2, where ℓ is the index of the temperatures listed in the decreasing order. We use the kernel Fisher discriminant ratio $\mathrm{KFDR}_{n,\ell,\gamma}(\mathcal{D})$ (Harchaoui et al., 2009) as a statistical quantity for the change point detection. Here, we set $\gamma = 10^{-3}$ in this paper, and the index ℓ achieving the maximum of $\mathrm{KFDR}_{n,\ell,\gamma}(\mathcal{D})$ corresponds to the estimated change point. The KFDR is calculated by the Gram matrix $(K(D_i, D_j))_{i,j=1,\dots,n}$ with respect to the kernel K.

We compute $\mathcal{D}$ with $n = 80$ from the data used in (Nakamura et al., 2015a;b). Since the persistence diagrams of SiO_2 contain huge amount of points, we apply the random Fourier features and the Nyström methods (Drineas & Mahoney, 2005) for the approximations of the PWGK with the Gaussian RKHS and the PSSK, respectively. The sample sizes used in both approximations are denoted by M and c, where c is the number of chosen columns. Figrue 6 summarizes the plots of the change points for several sample sizes and the computational time.



Computational time

PWGK		PSSK	
M	sec	c	sec
500	40	100	130
700	60	200	350
1000	90	300	700
3000	220	400	1200
5000	410	500	1530

Figure 6. (Left) Estimated change points by the approximated PWGK (top) and the PSSK (bottom). The parameters M and c are the sample numbers used in both approximations. At each parameter, both methods are tested multiple times and the black dots and the gray dots mean the majority and the minority of estimated change points, respectively. (Right) Computational time.

The interval of the glass transition temperature T estimated by the conventional method explained above is $2000K \leq T \leq 3500K$, which corresponds to $35 \leq \ell \leq 50$.

The computational complexity of the random Fourier features with respect to the sample size is $O(M)$, while that of the Nyström method involves matrix inversion of $O(c^3)$. For this reason, the PSSK with $c > 500$ cannot be performed in reasonable time, and hence we cannot check the convergence of the change points with respect to the sample size as shown in Figure 6. On the other hand, the PWGK plot shows the convergence to $\ell = 37$, implying that $\ell = 37$ is the true change point. We here emphasize that the computation to obtain $\ell = 37$ by the PWGK is much faster than the PSSK.

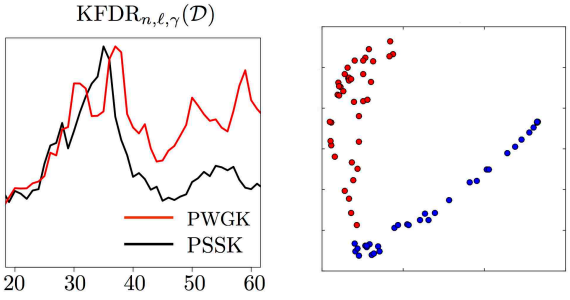


Figure 7. (Left) The KFDR plots of the PWGK ($M = 1000$) and the PSSK ($c = 500$). (Right) The 2-dimensional KPCA plot of the PWGK.

Figure 7 shows the normalized plots of KFDR$_{n,\ell,\gamma}(\mathcal{D})$ and the 2-dimensional plot given by KPCA (the color is given by the result of the change point detection by the PWGK). As we see from the figure, the KPCA plot shows the clear phase change between before (red) and after (blue) the change point. This strongly suggests that the glass transition occurs at the detected change point.

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

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

4.4. Protein classification

We apply the PWGK to two classification tasks studied in Cang et al. (2015). They use the molecular topological fingerprint (MTF) as a feature vector 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 (e.g., the longest, second longest, etc.) in persistence diagrams. We compare the performance of the PWGK with the Gaussian RKHS kernel and the MTF method under the same setting of the SVM reported in Cang et al. (2015).

The first task is a protein-drug binding problem, and we classify the binding and non-binding of drug to the M2 channel protein of the influenza A virus. For each form, 15 data were obtained by NMR experiments, in which 10 data are used for training and the remaining for testing. We randomly generated 100 ways of partitions, and calculated the 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

In this paper, we have proposed a kernel framework for analysis with persistence diagrams, and the persistence weighted Gaussian kernel as a useful kernel for the framework. As a significant advantage, our kernel enables one to control the effect of persistence in data analysis. We have also proven the stability result with respect to the kernel distance. 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 using the PWGK derived meaningful results in physics and biochemistry. From the viewpoint of computations, our kernel provides an accurate and efficient approximation to compute the Gram matrix, suitable for practical applications of TDA.

Acknowledgement

We thank Takenobu Nakamura for providing experimental and simulation data used in Section 4.3. This work is partially supported by JST Mathematics CREST (15656429) and JSPS KAKENHI Grant Number 26540016.

References

Bauer, U., Kerber, M., Reininghaus, J., and Wagner, H. *Mathematical Software – ICMS 2014: 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings*, chapter PHAT – Persistent Homology Algorithms Toolbox, pp. 137–143. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

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

Cang, Z., Mu, L., Wu, K., Opron, K., Xia, K., and Wei, G. W. A topological approach for protein classification. *Molecular Based Mathematical Biology*, 3(1), 2015.

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

Carlsson, G., Ishkhanov, T., de Silva, V., and Zomorodian, A. On the local behavior of spaces of natural images. *International journal of computer vision*, 76(1): 1–12, 2008.

Carriere, M., Oudot, S., and Ovsjanikov, M. Local signatures using persistence diagrams. preprint, 2015.

Chazal, F., de Silva, V., and Oudot, S. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1): 193–214, 2014a.

Chazal, F., Glisse, M., Labruère, C., and Michel, B. Convergence rates for persistence diagram estimation in topological data analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 163–171, 2014b.

Christmann, A. and Steinwart, I. Universal kernels on non-standard input spaces. In *Advances in Neural Information Processing Systems 23*, pp. 406–414. Curran Associates, Inc., 2010.

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

Da, T.K.F., Loriot, S., and Yvinec, M. 3D alpha shapes. In *CGAL User and Reference Manual*. CGAL Editorial Board, 4.7 edition, 2015.

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

Diestel, J. and Uhl, J. J. *Vector measures*. American Mathematical Soc., 1977.

Drineas, P. and Mahoney, M. W. On the nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002.

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

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

Fine, S. and Scheinberg, K. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

Gameiro, M., Hiraoka, Y., Izumi, S., Kramar, M., Mischaikow, K., and Nanda, V. A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics*, 32(1):1–17, 2013.

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

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pp. 585–592, 2007.

Harchaoui, Z., Moulines, E., and Bach, F. R. Kernel change-point analysis. In *Advances in Neural Information Processing Systems*, pp. 609–616, 2009.

Hatcher, A. *Algebraic Topology*. Cambridge University Press, 2001.

Kasson, P. M., Zomorodian, A., Park, S., Singhal, N., Guibas, L. J., and Pande, V. S. Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23(14):1753–1759, 2007.

Kwitt, R., Huber, S., Niethammer, M., Lin, W., and Bauer, U. Statistical topological data analysis - a kernel perspective. In *Advances in Neural Information Processing Systems 28*, pp. 3052–3060. Curran Associates, Inc., 2015.

Lee, H., Chung, M. K., Kang, H., Kim, B.-N., and Lee, D. S. Discriminative persistent homology of brain networks. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pp. 841–844. IEEE, 2011.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pp. 10–18, 2012.

Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., Matsue, K., and Nishiura, Y. Description of medium-range order in amorphous structures by persistent homology. *arXiv:1501.03611*, 2015a.

Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., and Nishiura, Y. Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26(304001), 2015b.

Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P. J., and Vaccarino, F. Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101):20140873, 2014.

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

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. A stable multi-scale kernel for topological machine learning. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 4741–4748, 2015.

Robins, V. and Turner, K. Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *arXiv:1507.01454*, 2015.

Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., and Ringach, D. L. Topological analysis of population activity in visual cortex. *Journal of vision*, 8(8):11, 2008.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *In Algorithmic Learning Theory: 18th International Conference*, pp. 13–31. Springer, 2007.

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

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

Williams, C. K. I. and Seeger, M. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, pp. 682–688. MIT Press, 2001.

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