



Persistent Homology with Dimensionality Reduction: k-Distance vs Gaussian Kernels

Shreya Arya, Jean-Daniel Boissonnat, Kunal Dutta

► To cite this version:

Shreya Arya, Jean-Daniel Boissonnat, Kunal Dutta. Persistent Homology with Dimensionality Reduction: k-Distance vs Gaussian Kernels. 2018. hal-01950051v1

HAL Id: hal-01950051

<https://inria.hal.science/hal-01950051v1>

Preprint submitted on 10 Dec 2018 (v1), last revised 5 Dec 2019 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Persistent Homology with Dimensionality Reduction: k -Distance vs Gaussian Kernels

Shreya Arya

BITS Pilani, Goa Campus, India


shreya.arya14@gmail.com

 [orcid]

Jean-Daniel Boissonnat

Université Côte d’Azur, INRIA Sophia-Antipolis, France,


jean-daniel.boissonnat@inria.fr

 <https://orcid.org/0000-0002-1825-0097>

Kunal Dutta

Université Côte d’Azur, INRIA Sophia-Antipolis, France

kunal.dutta@inria.fr

 [orcid]

Abstract

We investigate the effectiveness of dimensionality reduction for computing the persistent homology for both k -distance and kernel distance. For k -distance, we show that the standard Johnson-Lindenstrauss reduction preserves the k -distance, which preserves the persistent homology upto a $(1 - \varepsilon)^{-1}$ factor with target dimension $O(k \log n / \varepsilon^2)$. We also prove a concentration inequality for sums of dependent chi-squared random variables, which, under some conditions, allows the persistent homology to be preserved in $O(\log n / \varepsilon^2)$ dimensions. This answers an open question of Sheehy. For Gaussian kernels, we show that the standard Johnson-Lindenstrauss reduction preserves the persistent homology up to an $4(1 - \varepsilon)^{-1}$ factor.

2012 ACM Subject Classification I.3.5 Computational Geometry and Object Modeling

Keywords and phrases Distance to measure, Gaussian Kernel, Persistent Homology, Dimension Reduction, Johnson-Lindenstrauss

Lines 485

1 Introduction

Persistent homology is one of the main tools used to extract information from the data in topological data analysis. Given a data set as a point cloud in some ambient space, the idea is to construct a filtration sequence of topological spaces from the point cloud, and extract the topological information from this sequence. The topological spaces are usually constructed by considering balls around the data points, in some given metric of interest, as the open sets. However the usual distance function is highly sensitive to the presence of outliers and noise. The notion of *distance-to-a-measure* and the related k -distance (for finite data sets), proposed recently by Chazal et al. [3] are more robust to noise and outliers. Although this is a promising direction, an exact implementation is extremely costly. To overcome this difficulty, approximations of the k -distance have been proposed recently that led to certified approximations of persistent homology [2]. In a different direction, Phillips, Wang and Zheng [16] have proposed using the *kernel density estimates* and the associated *kernel distance* to overcome the problem of outliers and noise. Further, they show that this approach has the advantage of allowing the construction of *coresets*, which retain the



© Shreya Arya, Jean-Daniel Boissonnat and Kunal Dutta;
licensed under Creative Commons License CC-BY

35th International Symposium on Computational Geometry (SOCG 2019).

Editors: John Q. Open and Joan R. Access; Article No. 23; pp. 23:1–23:17

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

geometric and topological information of the data up to any desired error factor, but have size depending only on the error factor and not on the initial data set.

However, in all the above settings, computing the persistent homology involves answering nearest-neighbour and range-membership queries for the data points. Although many standard algorithms are known for these problems, they have exponential or worse dependence on the ambient dimension, and rapidly become unusable once the dimension grows beyond a few tens - which is indeed the case in many applications, for example in image processing, neuro-biological networks, data mining (see e.g. [9]), a phenomenon often referred to as the *curse of dimensionality*. One of the simplest and most commonly used mechanisms to mitigate this curse, is that of *random projections*, as applied in the celebrated Johnson and Lindenstrauss lemma (JL lemma for short) [11]. The JL lemma has been used by Sheehy [18] and Lotz [13] to reduce the complexity of computing persistent homology, by showing that the persistent homology of a point cloud is approximately preserved under random projections [18, 13], up to a $[1 - \varepsilon, 1 + \varepsilon]$ multiplicative factor, for any $\varepsilon \in [0, 1]$. However, their method involves only the usual distance to a point set and therefore remains sensitive to outliers and noise as mentioned earlier. The question of adapting the method of random projections in order to reduce the complexity of computing persistent homology with k -distance, is therefore a natural one, and has been raised by Sheehy [18], who observes that “One notable distance function that is missing from this paper is the so-called distance to a measure or ... k -distance ... it remains open whether the k -distance itself is $(1 \pm \varepsilon)$ -preserved under random projection”.

Our contribution: In this paper, we combine the method of random projections with k -distance, and show its application in computing persistent homology. It can be shown (see Theorem 2) without too much trouble, that for a given point set P , the usual Johnson-Lindenstrauss mapping pointwise preserves the k -distance to P . However, this does not suffice to preserve the persistent homology, because, as Sheehy [18] and Lotz [13] have noted, one needs to preserve the filtration of the Čech complex, and this means preserving l -wise intersections of balls, at varying scales of the radius parameter. In Section 3, we show how Theorem 2 can be used to preserve the persistent homology with k -distance, but with target dimension $O(k \log n / \varepsilon^2)$. We prove, in Section 4, a result of possible independent interest - a general concentration bound for sums of certain dependent chi-squared random variables, which shows that under certain reasonable and naturally occurring conditions, they behave almost like sums of independent chi-squared variables. This enables us to get rid of the extra factor of k in the target dimension, yielding again a reduced dimensionality similar to the usual Johnson-Lindenstrauss bound, in Theorem 3. We also give a few examples where our stronger inequality holds, such as for random point sets and locally dense nets. Coming to the kernel distance, in Section 5, we show that under the usual Johnson-Lindenstrauss mapping, the persistent modules of the Čech complex are $2(1 - \varepsilon)^{-1}$ -interleaved, which implies the persistent homology is preserved up to the same factor.

The rest of this paper is as follows. In Section 2, we briefly summarize some basic definitions and background. We end with some final remarks and open questions in Section 6.

1.1 Results

We first prove a general concentration bound for the sum of k dependent chi-squared random variables having d degrees of freedom which, under some conditions, is almost as strong as a sum for the *independent* case.

► **Theorem 1.** *Given a matrix $V \in \mathbb{R}^{D \times k}$ having column vectors $v_1, v_2, \dots, v_k \in \mathbb{R}^D$, rank*

71 r , and singular values $\sigma_1 \geq \sigma_2 \dots \geq \sigma_r > \sigma_{r+1} = \dots \sigma_k = 0$, if there exists a constant c ,
 72 $0 < c \leq k/r$ such that $\sigma_1^2 \leq \frac{c}{k} \sum_{i=1}^k \sigma_i^2$, then for any $0 < \epsilon < 1$, there exists a mapping
 73 $f: \mathbb{R}^D \rightarrow \mathbb{R}^d$ such that,

$$74 \quad \mathbb{P} \left[\frac{1}{k} \sum_{i=1}^k \|f(v_i)\|^2 > (1 + \epsilon) \frac{1}{k} \sum_{i=1}^k \|v_i\|^2 \right] < e^{-rd\epsilon^2/4 + rd\epsilon^3/6}.$$

75 For the subsequent theorems, let P be a set of n points in \mathbb{R}^D . Our next theorem shows
 76 that for the points in P , the k -distance is pointwise preserved by the standard Johnson-
 77 Lindenstrauss random projection.

78 ► **Theorem 2.** *Given $\epsilon \in [0, 1]$, there exists a mapping $f: \mathbb{R}^D \rightarrow \mathbb{R}^d$, where $d = O(\epsilon^{-2} \log n)$
 79 such that for all points $x \in P$,*

$$80 \quad (1 - \epsilon)d_{P,k}^2(x) \leq d_{P,k}^2(f(x)) \leq (1 + \epsilon)d_{P,k}^2(x).$$

81 As mentioned previously, the pointwise preservation of the k -distance does not imply
 82 the preservation of the Čech complex formed using the points in P . The following theorem
 83 shows that this can always be done using $O(k \log n / \epsilon^2)$ dimensions. Further, under some
 84 reasonable extra conditions, $O(\log n / \epsilon^2)$ dimensions suffice.

85 Let $Bary_{P,k}$ be the set of barycenters of every k -subset of P and let $\text{rad}(S)$ be the radius
 86 of the minimum enclosing ball of set S . Define \mathcal{F} to be the family of sets of k vectors, formed
 87 by the difference of any barycenter in $Bary_{P,k}$ with any set of k points in P , that is

$$88 \quad \mathcal{F} := \{ \{p - b : p \in S \in \binom{P}{k}\}, b \in Bary_{P,k} \}.$$

89 Let r be the minimum dimension of the subspaces formed by the k -sets of vectors in \mathcal{F} .

90 ► **Theorem 3.** *Let $S \subseteq Bary_{P,k}$ define a simplex σ in the Čech complex $\check{C}_\alpha(X)$ defined by
 91 the k -distance, then there exists a map $f: \mathbb{R}^D \rightarrow \mathbb{R}^d$ such that*

$$92 \quad (1 - \epsilon)\text{rad}^2(\sigma) \leq \text{rad}^2(f(\sigma)) \leq (1 + \epsilon)\text{rad}^2(\sigma),$$

93 *where $d = O\left(\frac{k \log(n/k)}{\epsilon^2}\right)$. Further, if $r = O(k)$ then $d = O\left(\frac{\log(\frac{n}{k})}{\epsilon^2}\right)$.*

94 Using Lemma 14 and Theorem 3, we get the persistent homology modules of the Čech
 95 filtration are $(1 - \epsilon)^{-1}$ interleaved.

96 ► **Corollary 4.** *Under the assumption of Theorem 3 the persistent homology modules associ-
 97 ated to the Čech filtrations of P and $f(P)$ under the k -distance are multiplicatively $(1 - \epsilon)^{-1}$
 98 interleaved.*

99 For the approximation of the k -distance given by [2], we show the following.

100 ► **Theorem 5 (Approximate k -distance).** *Let P be a set of points in $X = \mathbb{R}^D$. Let $S \subseteq P$
 101 define a simplex σ in the Čech complex $\check{C}_\alpha(X)$ defined by the approximate k -distance, then
 102 \exists a map $f: \mathbb{R}^D \rightarrow \mathbb{R}^d$ be a map such that*

$$103 \quad (1 - \epsilon)\text{rad}^2(\sigma) \leq \text{rad}^2(f(\sigma)) \leq (1 + \epsilon)\text{rad}^2(\sigma) \quad \text{and} \quad d = O(\log n / \epsilon^2).$$

104 ► **Corollary 6.** *The persistent homology modules for the Čech filtrations of P and $f(P)$
 105 under the approximate k -distance are $(1 - \epsilon)^{-1}$ interleaved.*

For the kernel distance with Gaussian kernels, we show the following.

► **Theorem 7.** *Let $P \subset X = \mathbb{R}^D$. Let $S \subseteq P$ define a simplex σ in the Čech complex $\check{C}_{\alpha,K}(X)$ defined by the kernel power distance and let $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$ be a map such that for all $x, y \in P$ and $0 < \epsilon < 1/3$,*

$$(1 - \epsilon)\|x - y\|^2 \leq \|f(x) - f(y)\|^2 \leq (1 + \epsilon)\|x - y\|^2.$$

Then the following holds for the radius of the minimum enclosing ball of σ ,

$$\frac{1}{2}(1 - \epsilon)\text{rad}^2(\sigma) \leq \text{rad}^2(f(\sigma)) \leq 2(1 + \epsilon)\text{rad}^2(\sigma).$$

As a corollary, the persistent homology is $4(1 - \epsilon)^{-1}$ -preserved.

► **Corollary 8.** *Under the assumption of Theorem 7 the persistent modules of the Čech filtration are $\frac{4}{1-\epsilon}$ interleaved.*

2 Background

2.1 Distance to Measure

The distance to a point set P is usually taken to be the minimum distance to a point in the set, in other words $d_P(x)$ is the smallest r such that the closed ball $\bar{B}(x, r)$ intersects P . However, this distance is extremely sensitive to outliers. To handle the problem of outliers in geometric and topological inference, Chazal, Cohen-Steiner and Mérigot in [3] introduced the distance to a probability measure.

► **Definition 9** (Pseudo-distance to μ). Let μ be a probability measure on a metric space X and let $m \in [0, 1)$ be a mass parameter, the pseudo-distance to μ is defined as

$$\delta_{\mu,m} : x \in X \rightarrow \{\inf r \geq 0 \mid \mu(B(x, r)) > m\}.$$

The distance to the probability measure μ is then defined as follows

► **Definition 10** (Distance to a measure). Let μ be a probability measure on a metric space X and let $m \in [0, 1)$ be a mass parameter, the distance to measure μ is defined as

$$d_{\mu,m}(x) = \sqrt{\frac{1}{m} \int_0^m \delta_{\mu,l}^2(x) dl}. \quad (1)$$

Differently from the pseudo-distance above, this distance to a measure is stable with respect to perturbations of the measure μ under the Wasserstein distance. [3] If we take a set $P \subset \mathbb{R}^D$ of n points and the mass parameter to be $m = k/n$ where $k \in \{1, \dots, n\}$, then the distance to the uniform probability measure on P is called the k -distance, denoted $d_{P,k}$

► **Definition 11** (k -distance). For a set P of n points in \mathbb{R}^D and $k \in \{1, \dots, n\}$,

$$d_{P,k}(x) = \sqrt{\frac{1}{k} \sum_{p \in \text{NN}_P^k(x)} \|x - p\|^2}, \quad (2)$$

where $\text{NN}_P^k(x) \subset P$ denotes the k nearest neighbours in P to the point $x \in \mathbb{R}^D$

It was shown in [10], that the k -distance can be also written as a power distance. If $\text{Bary}_{P,k}$ is a set of iso-barycentres of any subset of k points in P then

$$d_{P,k}(x) = \min_{b \in \text{Bary}_{P,k}} \left(\|x - b\|^2 + w(b) \right)^{1/2}, \quad (3)$$

where the weight of a barycentre $b = \frac{1}{k} \sum_i p_i$ is given by $w(b) = \frac{1}{k} \sum_i \|b - p_i\|^2$. Since $d_{P,k}$ can be written as a power distance, the sub-level sets of $d_{P,k}$ is the union of at most $\binom{n}{k}$ balls. However, some of the balls can be included in others. In fact, the number of balls required corresponds to the non-empty cells of the k th order Voronoi diagram. The number of non empty cells is $O(n^{\lfloor \frac{D+1}{2} \rfloor k^{\lceil \frac{D+1}{2} \rceil}})$ [7]. Computing higher-order Voronoi diagrams in high dimensions is too costly to make computations tractable. It is then natural to look for approximations of the k -distance as proposed in [2]

► **Definition 12 (Approximation).** Let $P \subset \mathbb{R}^D$ and $x \in \mathbb{R}^D$. The approximate k -distance $\tilde{d}_{P,k}(x)$ is the power distance defined as

$$\tilde{d}_{P,k}(x) := \sqrt{\min_{p \in P} d_{P,k}^2(p) + \|p - x\|^2},$$

Here $d_{P,k}(x)$ is the k -distance of p and $\min_{p \in P} d_{P,k}^2(p)$ can be seen as the weight of point p .

So now the sublevel sets are union of balls around the points of P , which reduces the number of balls significantly, from $\binom{n}{k}$ to n . Still, $\tilde{d}_{P,k}(x)$ approximates the k -distance [2]

$$\frac{1}{\sqrt{2}} d_{P,k} \leq \tilde{d}_{P,k} \leq \sqrt{3} d_{P,k}$$

2.2 Random Projections

The Johnson Lindenstrauss lemma [11] states that any n point subset of the Euclidean space can be embedded in $O(\epsilon^{-2} \log n)$ dimension with $(1 \pm \epsilon)$ distortion by randomly projecting the points in a lower dimensional subspace.

► **Lemma 13 (JL Lemma).** Let $0 < \epsilon < 1$. Then, $\forall u, v \in P$, a finite n point subset in \mathbb{R}^D , \exists a mapping $f: \mathbb{R}^D \rightarrow \mathbb{R}^d$, where $d = O(\epsilon^{-2} \log n)$ such that,

$$(1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2.$$

2.3 Persistent Homology

Persistent Homology. Given a family of topological spaces $\{F_\alpha\}$, where α is a parameter in \mathbb{R} , persistent homology tracks the evolution of the homology of the spaces as the parameter α changes from $-\infty$ to $+\infty$. To each dimension h of the homology groups is associated a persistent diagram which encodes the homological information of dimension h in the form of a multiset of pairs $(\alpha_{\text{birth}}, \alpha_{\text{death}})$, where α_{birth} is the birth time of the topological h -dimensional feature and α_{death} is the death time of the same feature. This can be represented as a set of points $(\alpha_{\text{birth}}, \alpha_{\text{death}})$ in \mathbb{R}^2 or as sets of intervals called barcodes.

169 **Persistence Modules**

170 A filtration $\{F_\alpha\}_{\alpha \in \mathbb{R}}$ is a family of topological spaces F_α such that for any $\alpha \leq \beta$, $F_\alpha \subset$
 171 F_β . A persistence module is a family of vector spaces $\{U_\alpha\}$, $\alpha \in \mathbb{R}$ over a field \mathbb{F} and
 172 homomorphisms $u_\alpha^\beta : U_\alpha \rightarrow U_\beta$ such that for all $\alpha \leq \beta \leq \gamma$, $u_\alpha^\gamma = u_\beta^\gamma \cdot u_\alpha^\beta$ and $u_\alpha^\alpha =$
 173 Id . For the filtration $\{F_\alpha\}_{\alpha \in \mathbb{R}}$, $\alpha \leq \beta \implies F_\alpha \subset F_\beta$ induces a homomorphism at the
 174 homology level. These homology groups of F_α and the homomorphisms form a persistent
 175 module. Persistence diagrams can be compared by interleaving the persistence modules. If
 176 $\mathbb{U} = (U_\alpha, u_\alpha^\beta)$ and $\mathbb{V} = (V_\alpha, v_\alpha^\beta)$ are two persistence modules, then they are ϵ -interleaved if
 177 there exists a collection of homomorphism $\phi = \{\phi_\alpha : U_\alpha \rightarrow V_{\alpha+\epsilon}\}$ such that the following
 178 diagram commutes for all $\alpha \leq \beta$,

$$\begin{array}{ccc}
 U_\alpha & \xrightarrow{u_\alpha^\beta} & U_\beta \\
 \phi_\alpha \downarrow & & \downarrow \phi_\beta \\
 V_{\alpha+\epsilon} & \xrightarrow{v_{\alpha+\epsilon}^{\beta+\epsilon}} & V_{\beta+\epsilon}
 \end{array}$$

180 The interleaved persistence modules have persistence diagrams close to each other with
 181 respect to the bottleneck distance. The bottleneck distance is a metric defined on the
 182 persistent diagrams and the idea is that two persistence diagrams are close if their features
 183 with long lifespan have close birth and death times. In particular, the ϵ -interleaving of
 184 modules implies $d_B^{\ln}(\text{diagram}(U), \text{diagram}(V)) < \ln(\epsilon)$, where d_B^{\ln} is the bottleneck distance
 185 in the log scale [4]. We have:

186 ► **Lemma 14** (Lemma 2.3 [13]). *Let $P \subseteq \mathbb{R}^D$ be a finite set and $d_P : \mathbb{R}^D \rightarrow \mathbb{R}$ be a distance*
 187 *function. Assume we have a function $F : \mathbb{R}^D \rightarrow \mathbb{R}^d$ such that for all subsets $S \subseteq P$ we have*

$$(1 - \epsilon) \text{rad}(S) \leq \text{rad}(F(S)) \leq (1 + \epsilon) \text{rad}(S),$$

189 *where $\text{rad}(X)$ is the minimum enclosing ball of X . Then the persistent homology modules*
 190 *associated to the Čech filtrations of P and $F(P)$ are multiplicatively $(1 - \epsilon)^{-1}$ interleaved.*

 191 **Simplicial Complexes and Filtrations**

192 A k -simplex is the convex hull of $(k + 1)$ affinely independent points. Let σ be the k -simplex
 193 defined on $S = \{v_0, v_1, \dots, v_k\}$. A simplex τ defined on a subset of S is known as a face of σ .
 194 A simplicial complex K is a collection of simplices such that every face of a simplex in K
 195 is also a simplex in K and the non-empty intersection of two simplices in K must be a face
 196 of both simplices. An abstract simplicial complex is a collection K of finite non-empty sets
 197 such that if $A \in K$ and $B \subseteq A$, $B \neq \emptyset$, then $B \in K$. The sets in K are the simplices of K .

198 A simplicial complex K with a function $f : K \rightarrow \mathbb{R}$ such that $f(\sigma) \leq f(\tau)$ whenever
 199 σ is a face of τ is a filtered simplicial complex. Let the sublevel set at $r \in \mathbb{R}$ be defined
 200 as $f^{-1}(-\infty, r]$. This is a subcomplex of K . By considering different values of r , we get a
 201 nested sequence of subcomplexes of K , $\emptyset = K^0 \subseteq K^1 \subseteq \dots \subseteq K^m = K$, where K^i is the
 202 sublevel set at value r_i . Let P be a set of points in \mathbb{R}^D . The sublevel filtration of a distance
 203 function d_P is $d_P^{-1}(-\infty, \alpha]$. In this paper we consider d_P to be the k -distance.

 204 **Čech Filtration**

205 We know from the previous section that the k -distance can be written in the form of a power
 206 distance. Let (P, w) be a set of weighted points and d_P be the power distance. Recall that

the power distance to a point set P is

$$d_P(x) = \min_{p \in P} \sqrt{\|x - p\|^2 + w(p)}$$

where $x \in \mathbb{R}^D$ and $w(p)$ is the weight of the point p . Then the sublevel sets $d_P^{-1}(-\infty, \alpha]$ are the union of balls centered at points in P and of radius $r_p(\alpha) = \sqrt{\alpha^2 - w(p)}$ for $\alpha \geq 0$ and $r_p(\alpha) = 0$ when $\alpha < 0$.

► **Definition 15 (Weighted Čech Complex).** Let (P, w) be a weighted set of points in $X = \mathbb{R}^D$. The Čech complex at α is defined as

$$\check{C}_\alpha(X) = \{\sigma \in P \mid \cap_{p \in \sigma} \overline{B}(p, r_p(\alpha)) \neq \emptyset\}$$

where $\overline{B}(p, r_p(\alpha))$ is the closed ball of radius $r_p(\alpha)$ centered at p in X .

Equivalently, we can define the Čech complex $\check{C}_\alpha(X)$ as the set of simplices σ such that $\sigma \in \check{C}_\alpha(X)$ iff $\text{rad}(\sigma) \leq \alpha$, where $\text{rad}(\sigma)$ is the radius $\text{rad}(\sigma)$ of the smallest enclosing ball of σ which satisfies $\text{rad}^2(\sigma) = \max_{p \in \sigma} (\|c_P - p\|^2 + w(p))$, where the centre c_P is defined as $c_P = \text{argmin}_{x \in \mathbb{R}^D} \max_{p \in \sigma} (\|x - p\|^2 + w(p))$. The Čech complex \check{C}_α is the nerve of the balls $\overline{B}(p, r_p(\alpha)), p \in P$. By the nerve lemma, we know that the union of the balls and \check{C}_α have the same homotopy type. Moreover, the persistent nerve lemma shows that the persistent homology of the Čech complex is the same as the α -sublevel filtration of the distance function.

The Vietoris-Rips complex $\text{VR}_\alpha(X)$ can be viewed as a simplification of the Čech complex. The Weighted Rips complex $\text{VR}_\alpha(X)$ is the maximal complex whose 1-skeleton is the same as $\check{C}_\alpha(X)$. Naturally, $\check{C}_\alpha(X) \subseteq \text{VR}_\alpha(X)$, and by [10] we have $\text{VR}_\alpha(X) \subseteq \check{C}_{2\alpha}(X)$.

3 Application to Persistent Homology

In this section, we first prove Theorem 2 which shows that the standard Johnson Lindenstrauss Lemma preserves the k -distance in dimension $O(\log n)$.

Let $P = \{p_1, p_2, \dots, p_n\}$ be a set of n points in $X \in \mathbb{R}^D$.

Proof of Theorem 2. Let $x \in P$ and let p_1, p_2, \dots, p_k be its k nearest neighbours in P . From Lemma 13, we have $\forall p_i \in P$

$$(1 - \epsilon)\|x - p_i\|^2 \leq \|f(x) - f(p_i)\|^2 \leq (1 + \epsilon)\|x - p_i\|^2.$$

By summing up these inequalities we have,

$$(1 - \epsilon) \frac{1}{k} \sum_{i=1}^k \|x - p_i\|^2 \leq \frac{1}{k} \sum_{i=1}^k \|f(x) - f(p_i)\|^2 \leq (1 + \epsilon) \frac{1}{k} \sum_{i=1}^k \|x - p_i\|^2.$$

Since the k -distance of $f(x)$ is the root mean squared distance to its k -nearest neighbours, $d_{f(P),k}^2(f(x)) \leq \frac{1}{k} \sum_{i=1}^k \|f(x) - f(p_i)\|^2$. So, for the upper bound, we have,

$$d_{f(P),k}^2(f(x)) \leq \frac{1}{k} \sum_{i=1}^k \|f(x) - f(p_i)\|^2 \leq (1 + \epsilon) \frac{1}{k} \sum_{i=1}^k \|x - p_i\|^2 = (1 + \epsilon) d_{P,k}^2(x).$$

Let $q_1, q_2, \dots, q_k \in P$ be such that $f(q_1), f(q_2), \dots, f(q_k)$ are the k -nearest neighbours of $f(x)$. Then for the lower bound we have,

$$(1 - \epsilon) \frac{1}{k} \sum_{i=1}^k \|x - q_i\|^2 \leq \frac{1}{k} \sum_{i=1}^k \|f(x) - f(q_i)\|^2 = d_{f(P),k}^2(f(x)).$$

And since p_1, \dots, p_k are the nearest neighbours of x , $d_{P,k}^2(x) = \frac{1}{k} \sum_{i=1}^k \|x - p_i\|^2 \leq \frac{1}{k} \sum_{i=1}^k \|x - q_i\|^2$. This gives the required lower bound. \blacktriangleleft

So, we have that the standard Johnson Lindenstrauss mapping on Euclidean pairwise distances, preserves the k -distance. Let $Bary_{P,k}$ be the set of barycentres of subsets of k points of P . The k -distance can be written as a power distance i.e.

$$d_{P,k}(x) = \left(\min_{b \in Bary_{P,k}} \|x - b\|^2 + w(b) \right)^{1/2},$$

where $w(b) = 1/k \sum_i \|b - p_i\|^2$ for $b = 1/k \sum_i p_i$. So we have a weighted set of points, $Bary_{P,k}$ and balls around them of varying radii as sublevel sets.

Let $\text{rad}(S)$ be the radius of the minimum enclosing ball of weighted $S \subset X = \mathbb{R}^D$ and its center be c_S . Recall,

$$c_S = \text{argmin}_{x \in \mathbb{R}^D} \max_{s \in S} \|x - s\|^2 + w(s) \quad \text{and} \quad \text{rad}^2(S) = \max_{s \in S} \|c_S - s\|^2 + w(s).$$

To show that the persistent modules of the Čech complex are comparable, by lemma 14 we need to prove that the radius of the minimum enclosing ball of simplices in the Čech complex are comparable.

Since the simplices in the Čech complex are formed by the intersection of the balls centered at $Bary_{P,k}$, the standard Johnson Lindenstrauss lemma acting on the point set P will not be sufficient. It is showed in [13] and [18] that for the power distance for a point set P , the persistent modules of the Čech complexes are interleaved. The k -distance is a power distance for weighted barycenters. So, to interleave the modules we would have to augment the points in $Bary_{P,k}$ to the point set P increasing the number of points to $O\binom{n}{k}$ giving a target dimension $d = O(k \log n)$. Using the stronger concentration inequality we get a much better target dimension.

3.1 Proof of Theorem 3

In this section, we use the mappings in Theorems 2 and 1 to prove that persistent homology modules of the Čech complex are comparable after dimension reduction. This is formally stated in Theorem 3. To prove Theorem 3, we require a few preliminary results. Let $\sigma = \{b_1, b_2, \dots, b_r\}$, $b_i \in Bary_{P,k}$. Since we are dealing with a weighted point set $Bary_{P,k}$, we need to find the radius of the minimum enclosing ball of a set of balls centered at $b_i \in \sigma$. We infer the following lemmas from lemma 3.3 and lemma 3.19 in [8].

- **Lemma 16.** (i) *The center c_σ is a convex combination of points in σ , i.e. $c_\sigma = \sum_{i=1}^r \lambda_i b_i$ and $\sum_{i=1}^r \lambda_i = 1$ where λ_i 's are non-negative.*
 (ii) *In particular, either $\lambda_i = 0$ or, $\lambda_i > 0$ and the ball centered at b_i is internally tangent to the minimum enclosing ball.*

Proof of Theorem 3. We have $\sigma = \{b_1, b_2, \dots, b_r\}$ and $c_\sigma = \sum_{i=1}^r \lambda_i b_i$ where $\sum_{i=1}^r \lambda_i = 1$. From the above lemma we have, the balls around b_i are either internally tangent to the minimum enclosing ball or the corresponding $\lambda_i = 0$. This implies the radius of the minimum enclosing ball is the weighted distance between the center and b_i for which $\lambda_i \neq 0$. Since $\sum_{i=1}^r \lambda_i = 1$, we can write

$$\text{rad}^2(\sigma) = \sum_{i=1}^r \lambda_i (\|c_\sigma - b_i\|^2 + w(b_i)).$$

We use the following simple fact from [13]

$$\sum_{i=1}^r \lambda_i \|c_\sigma - b_i\|^2 = \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j \|b_i - b_j\|^2.$$

$$\text{rad}^2(\sigma) = \sum_{i=1}^r \lambda_i (\|c_\sigma - b_i\|^2 + w(b_i)) \quad (4)$$

$$= \frac{1}{2} \sum_{j=1}^r \sum_{i=1}^r \lambda_j \lambda_i \|b_i - b_j\|^2 + \frac{1}{2} \sum_{i=1}^r 2\lambda_i w(b_i) \quad (5)$$

$$= \frac{1}{2} \sum_{j=1}^r \lambda_j \sum_{i=1}^r \lambda_i (\|b_j - b_i\|^2 + w(b_i) + w(b_i)). \quad (6)$$

We can write $\|b_j - b_i\|^2 + w(b_i) = 1/k \sum_{l=1}^k \|b_j - p_{il}\|^2$ where $b_i = 1/k \sum_{l=1}^k p_{il}$ and $w(b_i) = 1/k \sum_{l=1}^k \|b_i - p_{il}\|^2$. These distances are in the form of the k -distance if we augment the barycenters to the point set P . Let $f(c_\sigma) = \sum_i \lambda_i f(b_i)$. So,

$$\text{rad}^2(\sigma) \leq \frac{1}{1-\epsilon} \left[\frac{1}{2} \sum_{j=1}^r \lambda_j \sum_{i=1}^r \lambda_i \left(\frac{1}{k} \sum_{l=1}^k \|f(b_j) - f(p_{il})\|^2 + \frac{1}{k} \sum_{l=1}^k \|f(b_i) - f(p_{il})\|^2 \right) \right] \quad (7)$$

$$\leq \frac{1}{1-\epsilon} \left[\frac{1}{2} \sum_{j=1}^r \lambda_j \sum_{i=1}^r \lambda_i (\|f(b_j) - f(b_i)\|^2 + w(f(b_i)) + w(f(b_i))) \right] \quad (8)$$

$$\leq \frac{1}{1-\epsilon} \left[\sum_{i=1}^r \lambda_i (\|f(c_\sigma) - f(b_i)\|^2 + w(f(b_i))) \right]. \quad (9)$$

The function $\sum_{i=1}^r \lambda_i (\|c - f(b_i)\|^2)$ is minimized at $c = f(c_\sigma)$. Let the center of $f(\sigma)$ be \hat{c} . So,

$$\text{rad}^2(\sigma) \leq \frac{1}{1-\epsilon} \left[\sum_{i=1}^r \lambda_i (\|\hat{c} - f(b_i)\|^2 + w(f(b_i))) \right] = \frac{1}{1-\epsilon} \text{rad}^2(f(\sigma)). \quad (10)$$

The other direction can be proved similarly.

Dimension analysis. We require that f must $(1 \pm \epsilon)$ preserve the distances of the form $\sum_{i=1}^k \|x - y_i\|^2$ where $x \in \text{Bary}_{P,k}$ and $y_i \in S \in \binom{P}{k}$. To prove the first statement in the theorem, we apply the Johnson-Lindenstrauss mapping in Theorem 2 for the set $\binom{P}{k} \cup \text{Bary}_{P,k}$. This gives a target dimension of $O\left(\frac{\log \binom{n}{k}^2}{\epsilon^2}\right) = O\left(\frac{k \log(n/k)}{\epsilon^2}\right)$, using $\binom{n}{k} \leq (en/k)^k$.

To prove the second statement of the theorem, we shall use the tail bounds in Theorem 1 which are $O(e^{-rd\epsilon^2/4})$. That is, the probability that the distortion $\sum_{i=1}^k \|f(x) - f(y_i)\|^2 / \sum_{i=1}^k \|x - y_i\|^2$ does not lie in the range $[1 - \epsilon, 1 + \epsilon]$ is at most $O(e^{-rd\epsilon^2/4})$. We want to find the target dimension required such that for all $x \in \text{Bary}_{P,k}$, $y_i \in S \in \binom{P}{k}$, the distortion lies in $[1 - \epsilon, 1 + \epsilon]$. A trivial union bound gives us $\binom{n}{k} \times \binom{n}{k} \times e^{-rd\epsilon^2/4} < 1$. Taking log on both sides and using $\binom{n}{k} < (ne/k)^k$ gives us $d = O\left(\frac{k}{r} \log\left(\frac{n}{k}\right)\right)$.

In section 2.1, we defined the approximate k -distance to be

$$\tilde{d}_{P,k}(x) := \sqrt{\min_{p \in P} d_{P,k}^2(p) + \|p - x\|^2},$$

where $w(p) = d_{P,k}^2(p)$. And so the Čech complex would be formed by the intersections of the balls around the weighted points in P .

Proof of Theorem 5. This proof follows Theorem 3. Let $\sigma = \{p_1, p_2, \dots, p_r\}$ and c_σ be the center of the minimum enclosing ball of the σ .

$$\text{rad}^2(\sigma) = \sum_{i=1}^r \lambda_i (\|c_\sigma - p_i\|^2 + w(p_i)) = \sum_{i < j} \lambda_i \lambda_j \|p_i - p_j\|^2 + \sum_{i=1}^k \lambda_i w(p_i),$$

where $w(p) = d_{P,k}^2(p)$. The standard Johnson Lindenstrauss lemma preserves pairwise distances and the k -distance in dimension $O(\log n)$. So following the proof of Theorem 3, we have the result. \blacktriangleleft

When is the assumption true? Suppose the points are distributed according to an i.i.d. Gaussian distribution. Then the distance vectors are also i.i.d. Gaussian. We required that the maximum singular value of the initial vector matrix is not too far from the average singular value. For random matrices of dimension $N \times n$ with independent columns we have that the singular values lie in $[\sqrt{N} - C\sqrt{n}, \sqrt{N} + C\sqrt{n}]$ for some constant C with probability $O(e^{-cN})$ (see e.g. [19, 17]). By our assumption, $N \gg n$, i.e. $D \gg k$. Therefore by a union bound over all sets of k points, it follows that the condition in the Theorem is satisfied, with $r = O(k)$.

4 Tail bounds for the sum of weighted $\chi_{(d)}^2$ random variables

Tail bounds for sums of independent unweighted chi-squared random variables, with arbitrary degrees of freedom, are standard in the literature [12]. In the more general case, when the random variables can be weighted and are not necessarily independent, they fall into the class of gamma distributions. However, the only tail bounds known for this case are general bounds for gamma distributions, which can be weaker and can be inconvenient to use, as they are usually stated in terms of the parameters of the gamma distribution. In this section, we prove tail bounds for sums of weighted, dependent chi-squared random variables, having d degrees of freedom. When the weights are close to uniform, and the covariance matrix has high rank, our bounds essentially behave like the unweighted, independent case. They degrade gradually with non-uniformity of the weight distribution, and the decreasing rank of the covariance matrix, eventually reducing to the case of a single $\chi_{(d)}^2$ random variable, when the rank is 1.

For the remainder of this section, we shall work with the following setup. Given a matrix $V \in \mathbb{R}^{D \times k}$ with column vectors $v_1, v_2, \dots, v_k \in \mathbb{R}^D$ and gaussian vectors $g_1, g_2, \dots, g_d \sim \mathcal{N}_D(0, I)$, define, $L_j = \sum_{i=1}^d \langle g_i, v_j \rangle^2$, for $1 \leq j \leq k$. By definition, the L_j s are $\chi_{(d)}^2$ random variables with weights $\|v_j\|^2$. Let L denote their sum, i.e. $L := \sum_{j=1}^k L_j$. In the following lemma we prove that L can be written as a sum of independent weighted chi-squared random variables.

► **Lemma 17.** *L is a sum of independent weighted chi-squared random variables of d degrees of freedom.*

Proof. First, observe that by changing the order of summation, we get $L = \sum_{i=1}^d \sum_{j=1}^k \langle g_i, v_j \rangle^2$. Define $Q_i := \sum_{j=1}^k \langle g_i, v_j \rangle^2$. Let X_i be the random vector defined as $X_i = [\langle g_i, v_1 \rangle, \dots, \langle g_i, v_k \rangle]^\top$. So, $Q_i = X_i^\top X_i$. Each $\langle g_i, v_j \rangle$ is normally distributed with mean 0 and variance $\|v_j\|^2$ i.e.

352 $\langle g_i, v_j \rangle \sim \mathcal{N}(0, \|v_j\|^2)$. So X_i follows a multivariate normal distribution, $X_i \sim \mathcal{N}_k(0, \Sigma)$
 353 where Σ is the $k \times k$ covariance matrix. To find Σ , we calculate

$$354 \quad \text{cov}[\langle g_i, v_j \rangle \langle g_i, v_l \rangle] = E[\langle g_i, v_j \rangle \langle g_i, v_l \rangle] - E[\langle g_i, v_j \rangle] E[\langle g_i, v_l \rangle] \quad (11)$$

$$355 \quad = E\left[\sum_{p=1}^D \sum_{m=1}^D g_{ip} v_{pj} g_{im} v_{ml}\right] = \sum_{p=1}^D \sum_{m=1}^D v_{pj} v_{ml} E[g_{ip} g_{im}] \quad (12)$$

$$356 \quad = \sum_{p=m=1}^D v_{pj} v_{pl} = v_j^\top v_l. \quad (13)$$

357 Thus, each entry of Σ is of the form $v_j^\top v_l$ with diagonal entries being $\|v_i\|^2$, in other words
 358 $\Sigma = V^\top V$. Σ is a $k \times k$ symmetric matrix, and by Cholesky decomposition $\Sigma = BB^\top$ where B
 359 is $k \times r$ matrix and $\text{rank}(\Sigma) = r \leq k$. We write $X_i = BY_i$ where Y_i is a standard multivariate
 360 normal distribution, i.e. $Y \sim \mathcal{N}_r(0, I)$. Now, we calculate $Q_i = X_i^\top X_i = Y_i^\top B^\top B Y_i$. Write
 361 $A = B^\top B$, so A is a $r \times r$ symmetric matrix. There exists an orthogonal $r \times r$ matrix P
 362 such that $P^\top A P = \text{diag}(\lambda_1, \dots, \lambda_r)$ where $\lambda_1, \dots, \lambda_r$ are the eigenvalues of matrix A . Let
 363 $Z_i = P^\top Y_i \implies Y_i = P Z_i$ as $PP^\top = I$. This means $Z_i \sim \mathcal{N}_r(0, I)$.

$$364 \quad Y_i^\top A Y_i = Z_i^\top P^\top A P Z_i = Z_i^\top \text{diag}(\lambda_1, \dots, \lambda_r) Z_i = \lambda_1 z_{i1}^2 + \lambda_2 z_{i2}^2 + \dots + \lambda_r z_{ir}^2,$$

365 where each z_{ij} is an independently distributed standard normal random variable and $z_{ij}^2 \sim$
 366 $\chi_{(1)}^2$, $j = 1, \dots, r$. Note that the eigenvalues of matrix $A = B^\top B$ is equal to the non-zero
 367 eigenvalues of Σ . And since $\Sigma = V^\top V$, the eigenvalues of Σ are equal to the singular values
 368 squared of matrix V , that is $\lambda_i = \sigma_i^2$ for all i . So, we have $Q_i = \sum_{j=1}^r \langle g_i, v_j \rangle^2 = \sum_{j=1}^r \sigma_j^2 z_{ij}^2$.

369 Since, $L = \sum_{i=1}^d Q_i$ and each Gaussian vector g_i , $i = 1, 2, \dots, d$, is independent we have

$$370 \quad L = \sum_{j=1}^r \sum_{i=1}^d \sigma_j^2 z_{ij}^2 = \sum_{j=1}^r \sigma_j^2 \tilde{z}_j^2,$$

371 where \tilde{z}_j^2 are independent chi-squared random variables of d degrees of freedom i.e. $\tilde{z}_j^2 \sim \chi_{(d)}^2$.
 372 \blacktriangleleft

373 **► Theorem 18.** Given Z_1, Z_2, \dots, Z_k independent chi-squared random variables of d degrees
 374 of freedom and weights $w_1, w_2, \dots, w_k \geq 0$. Let $Z = \sum_{i=1}^k w_i Z_i$. Then for $0 < \delta < 1$, the
 375 following holds,

$$376 \quad \mathbb{P}[|Z - \mathbb{E}[Z]| > \delta \mathbb{E}[Z]] \leq \exp\left(\frac{kd\delta}{2} - \frac{kd\delta^2}{4} + \frac{kd\delta^3}{6} - \frac{\delta d \sum_{i=1}^k w_i}{2w_m}\right).$$

377 **Proof.** We shall prove the bound for the upper tail; the proof of the lower tail follows
 378 similarly. We want to estimate the tail bounds for $Z = \sum_{i=1}^k w_i Z_i$. First, by linearity of
 379 expectation we have that $\mathbb{E}[Z] = \sum_{i=1}^k w_i \mathbb{E}[Z_i] = (\sum_{i=1}^k w_i)d$. Second, we find the moment
 380 generating function $\mathbb{E}[e^{tZ}]$. By independence of Z_i we get, $\mathbb{E}[e^{tZ}] = \mathbb{E}[\prod_{i=1}^k e^{tZ_i w_i}] =$
 381 $\prod_{i=1}^k \mathbb{E}[e^{tZ_i w_i}]$.

382 The m.g.f of $\chi_{(d)}^2$ random variable Y is $M(Y) = (1 - 2t)^{-d/2}$. So, $M(w_i Z_i) = (1 -$
 383 $2w_i t)^{-d/2}$. And,

$$384 \quad M(Z) = \prod_{i=1}^k (1 - 2w_i t)^{-d/2}.$$

385

$$386 \quad \mathbb{P}[Z > (1 + \delta) \mathbb{E}[Z]] = \mathbb{P}[e^{tZ} > e^{t(1+\delta) \mathbb{E}[Z]}]. \quad (14)$$

387 By the Markov inequality, (14) is at most

$$388 \quad \mathbb{P}[e^{tZ} > e^{t(1+\delta) \mathbb{E}[Z]}] \leq \frac{\mathbb{E}[e^{tZ}]}{e^{t(1+\delta) \mathbb{E}[Z]}} = \frac{\prod_{i=1}^k (1 - 2w_i t)^{-d/2}}{e^{t(1+\delta) \sum_{i=1}^k w_i}}. \quad (15)$$

389 We need to choose a t such that $(1 - 2w_i t) \geq 0$ for all w_i . So, substituting the value of
 390 $t = \frac{\delta}{2w_m(1+\delta)}$, where w_m is the maximum weight, and using $\log(1+x) < x - x^2/2 + x^3/3$
 391 we get (15) is at most

$$392 \quad (1 + \delta)^{kd/2} \exp\left(-\frac{\delta d \sum_{i=1}^k w_i}{2w_m}\right) \leq \exp\left(\frac{kd\delta}{2} - \frac{kd\delta^2}{4} + \frac{kd\delta^3}{6} - \frac{\delta d \sum_{i=1}^k w_i}{2w_m}\right). \quad (16)$$

393

394 Now, we prove Theorem 1, which shows that under certain conditions, the squared sum
 395 of k distances is more strongly concentrated around its mean than the Euclidean distance.

396 **Proof of Theorem 1.** Let $V = (v_1, \dots, v_k)$ be a set of k vectors in \mathbb{R}^D . Let G be a random
 397 Gaussian matrix of order $d \times D$ such that each entry $g_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ where $d < D$. Let
 398 $f : (\mathbb{R}^D)^k \rightarrow (\mathbb{R}^d)^k$ defined as follows

$$399 \quad f(V) = \frac{1}{\sqrt{d}} GV.$$

400 Let g_i denote the row of matrix G , so each entry of GV would be of the form $\langle g_i, v_j \rangle$. Let
 401 $f(v_i)$ denote the i th vector of $f(V)$. Recalling the proof setup at the beginning of this
 402 section, we get that

$$403 \quad \frac{1}{k} \sum_{i=1}^k \|f(v_i)\|^2 = \frac{1}{d} \cdot \frac{1}{k} \sum_{i=1}^d \sum_{j=1}^k \langle g_i, v_j \rangle^2 = \frac{1}{d} \cdot \frac{1}{k} L.$$

404 From Lemma 17 we have,

$$405 \quad L = dk \sum_{i=1}^k \|f(v_i)\|^2 = \sum_{j=1}^r \sigma_j^2 \tilde{z}_j^2,$$

406 where \tilde{z}_j^2 are independent chi-squared random variables of d degrees of freedom. Also, note
 407 that the sum of the singular value squared is the trace of the matrix $V^\top V$, so $\sum_{i=1}^r \sigma_i^2 =$
 408 $\sum_{i=1}^k \|v_i\|^2$.

409 Now, we estimate the probability of the event

$$410 \quad \left\{ \frac{1}{k} \sum_{i=1}^k \|f(v_i)\|^2 > (1 + \epsilon) \frac{1}{k} \sum_{i=1}^k \|v_i\|^2 \right\} = \left\{ \frac{L}{kd} > (1 + \epsilon) \frac{1}{k} \sum_{i=1}^k \sigma_i^2 \right\},$$

 411 where $\sigma_{r+1} = \dots = \sigma_k = 0$

412 Since L is the sum of r independent weighted chi-squared random variables of d degrees
 413 of freedom, using Theorem 18 we have

$$414 \quad \mathbb{P}\left[L > (1 + \epsilon) \left(\sum_{j=1}^k \sigma_j^2\right) d\right] \leq \exp\left(\frac{rd\epsilon}{2} - \frac{rd\epsilon^2}{4} + \frac{rd\epsilon^3}{6} - \frac{\epsilon k \sum_{i=1}^k \sigma_i^2}{2\sigma_1^2}\right)$$

415 By assumption, $\frac{\sum_{i=1}^k \sigma_i^2}{\sigma_1^2} \geq \frac{k}{c}$ and $c \leq k/r$ we get,

$$\begin{aligned}
 416 \quad \exp\left(\frac{rd\epsilon}{2} - \frac{rd\epsilon^2}{4} + \frac{rd\epsilon^3}{6} - \frac{\epsilon d \sum_{i=1}^k \sigma_i^2}{2\sigma_1^2}\right) &\leq \exp\left(\frac{rd\epsilon}{2} - \frac{rd\epsilon^2}{4} + \frac{rd\epsilon^3}{6} - \frac{rd\epsilon}{2}\right) \\
 417 \quad &\leq \exp\left(-\frac{rd\epsilon^2}{4} + \frac{rd\epsilon^3}{6}\right).
 \end{aligned}$$

418

419 5 Kernel Distance

420 We refer the reader to [15] for a background on kernels. A kernel $K : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ is a similarity function. In this section, we consider Gaussian Kernels of the form
 421 $K(x, y) = \sigma^2 \exp(-\|x - y\|^2 / 2\sigma^2)$. A similarity between measures μ and ν is $\kappa(\mu, \nu) =$
 422 $\int_{p \in \mathbb{R}^D} \int_{q \in \mathbb{R}^D} K(p, q) \mu(p) \nu(q) dp dq$.

423
 424 ► **Definition 19** (Kernel Distance). The kernel distance between two measures μ and ν is
 425 defined as

$$426 \quad D_K(\mu, \nu) = \sqrt{k(\mu, \mu) + k(\nu, \nu) - 2k(\mu, \nu)}.$$

427 The kernel distance between two measures is a metric. If we take the dirac measure with
 428 respect to x and y we can write $D_K(x, y)^2 = 2\sigma^2(1 - e^{-\|x - y\|^2 / 2\sigma^2})$. Now we define the kernel
 429 distance with respect to an underlying probability measure μ , $d_K^\mu : \mathbb{R}^D \rightarrow \mathbb{R}$.

$$430 \quad d_K^\mu(x) = D_K(\mu, x) = \sqrt{k(\mu, \mu) + k(x, x) - 2k(\mu, x)}.$$

431 The x in $D_K(\mu, x)$ represents the Dirac measure with respect to x . This distance function
 432 has been used in place of the distance to measure for geometric and topological inference.
 433 A power distance version of this distance function is given in [16].

$$434 \quad f_P^K(x) = \sqrt{\min_{p \in P} (D_K(p, x)^2 + d_K^\mu(p)^2)}, \quad (17)$$

435 where the $d_K^\mu(p)$ can be seen as the weight of the point p . To prove the stability properties
 436 of the persistence diagrams, the distance is approximated by $f_{P_+}^K(x)$. The construction of
 437 P_+ for the case of the empirical measure on μ is given in [16]. The following bounds hold
 438 for the power distance approximation of the kernel distance with respect to the measure

$$439 \quad \frac{1}{\sqrt{2}} d_K^\mu(x) \leq f_P^K(x) \leq \sqrt{14} d_K^\mu(x).$$

440 5.1 Dimension Reduction with Kernels

441 In this section we show that the standard Johnson Lindenstrauss transform preserves the
 442 Kernel Distance to a certain factor and we show that the corresponding Čech complexes are
 443 also comparable.

444 Using $1 - t \leq e^{-t} \leq 1 - t + (1/2)t^2$ when $t \geq 0$, for $\|x - y\| \leq \sqrt{3}\sigma$ we get

$$445 \quad \frac{1}{2} \|x - y\|^2 \leq D_K(x, y) \leq \|x - y\|^2. \quad (18)$$

► **Lemma 20.** Let $0 < \epsilon < 1/3$. Then, $\forall x, y \in P \subset \mathbb{R}^D$, $|P| = n$, \exists a mapping $f: \mathbb{R}^D \rightarrow \mathbb{R}^d$, where $d = O(\epsilon^{-2} \log n)$ such that,

$$\frac{(1-\epsilon)}{2} D_K(x, y)^2 \leq D_K(f(x), f(y))^2 \leq 2(1+\epsilon) D_K(x, y)^2.$$

Proof. Using (18) and Lemma 13 for $\|x - y\|^2 \leq 3\sigma^2$ we have,

$$(1-\epsilon) D_K(x, y)^2 \leq (1-\epsilon) \|x - y\|^2 \leq \|f(x) - f(y)\|^2 \leq 2 D_K(f(x), f(y))$$

We know that $D_K(x, y)^2 = 2\sigma^2 (1 - \exp(-\|x - y\|^2/2\sigma^2)) \leq 2\sigma^2$. Using $\|x - y\|^2 > 3\sigma^2$, $\epsilon \leq 1/3$ and Lemma 13,

$$D_K(f(x), f(y))^2 = 2\sigma^2 (1 - \exp(-\|f(x) - f(y)\|^2/2\sigma^2)) \quad (19)$$

$$\geq 2\sigma^2 (1 - \exp(-(1-\epsilon)\|x - y\|^2/2\sigma^2)) \quad (20)$$

$$\geq 2\sigma^2 (1 - \exp(-1)) \geq \sigma^2 \geq D_K(x, y)^2/2 \geq (1-\epsilon) D_K(x, y)^2/2 \quad (21)$$

The upper bound follows similarly from (18). ◀

► **Corollary 21.** Let μ be the empirical measure on P .

$$\frac{(1-\epsilon)}{2} d_K^\mu(p)^2 \leq d_K^\mu(f(p))^2 \leq 2(1+\epsilon) d_K^\mu(p)^2. \quad (22)$$

Proof. Since $d_K^\mu(p) = D_K(\mu, p) = \sqrt{k(\mu, \mu) + k(p, p) - 2k(\mu, p)}$. Using this we get,

$$d_K^\mu(p)^2 = \frac{1}{2n^2} \sum_{x \in P} \sum_{y \in P} -D_K(x, y)^2 + \frac{1}{n} \sum_{x \in P} D_K(x, p)^2.$$

By taking sums of the inequalities in Lemma 20, we have the result. ◀

Let $\text{rad}_K(S)$ be the radius of the minimum enclosing ball of a subset $S \in \mathbb{R}^D$ under the kernel power distance and let $\text{rad}(S)$ be the radius of the minimum enclosing ball of under the usual euclidean power distance. The proof of Theorem 7 is given in the Appendix.

6 Future Work

We conclude this paper with several open questions.

Relaxing the assumptions. We have shown that the k -distance is preserved under random projections and is more strongly concentrated under certain assumptions. We showed that for random point sets the result is generally true. It is left as an open question whether these assumptions can be relaxed in general.

General probability measures. We have considered the distance to measure for uniform measures. Do random projections preserve distances for general probability measures?

Manifolds. Baraniuk and Wakin [1] introduced random projections for the case of manifolds and showed a variant of the JL lemma where the target dimension depends on the intrinsic dimension of the underlying manifold. As shown in Theorem 2, we have that the k -distance is preserved under random projections for manifolds, but it remains to show if the target dimension can be reduced further using the stronger concentration inequality.

Kernel Distances. In Theorem 7 we get a constant factor of 3 and 8 depending on the value of σ , the noise factor. Instead of using the standard Johnson Lindenstrauss Lemma,

it is shown in [6], that an approximate embedding $\hat{\phi} : \mathbb{R}^D \rightarrow \mathbb{R}^d$ induces a kernel distance $D_{\hat{K}}(x, y) = \langle \hat{\phi}(x), \hat{\phi}(y) \rangle$ and for $0 < \epsilon < 1$, the kernel distance is interleaved as follows:

$$\frac{1}{1+\epsilon} D_K(x, y) \leq D_{\hat{K}}(x, y) \leq \frac{1}{1-\epsilon} D_K(x, y).$$

So using this kind of embedding we can expect something better, a reduction of a factor of 2. We intend to address this in the future.

References

- 1 Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.
- 2 Mickaël Buchet, Frédéric Chazal, Steve Y. Oudot, and Donald R. Sheehy. Efficient and robust persistent homology for measures. *Comput. Geom.*, 58:70–96, 2016. URL: <https://doi.org/10.1016/j.comgeo.2016.07.001>, doi:10.1016/j.comgeo.2016.07.001.
- 3 Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.
- 4 Frédéric Chazal, Vin De Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.
- 5 Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *The Journal of Machine Learning Research*, 18(1):5845–5884, 2017.
- 6 Di Chen and Jeff M. Phillips. Relative error embeddings of the gaussian kernel distance. In *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, pages 560–576, 2017. URL: <http://proceedings.mlr.press/v76/chen17a.html>.
- 7 Kenneth L Clarkson and Peter W Shor. Applications of random sampling in computational geometry, ii. *Discrete & Computational Geometry*, 4(5):387–421, 1989.
- 8 Kaspar Fischer. *Smallest enclosing balls of balls: Combinatorial structure & algorithms*. PhD thesis, ETH Zurich, 2005.
- 9 C. Giraud. *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2014. URL: <https://books.google.fr/books?id=qRuVoAEACAAJ>.
- 10 Leonidas Guibas, Dmitriy Morozov, and Quentin Mérigot. Witnessed k-distance. *Discrete & Computational Geometry*, 49(1):22–45, Jan 2013. URL: <https://doi.org/10.1007/s00454-012-9465-x>, doi:10.1007/s00454-012-9465-x.
- 11 William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.
- 12 M. G. Kendall, A. Stuart, and J. K. Ord, editors. *Kendall's Advanced Theory of Statistics*. Oxford University Press, Inc., New York, NY, USA, 1987.
- 13 Martin Lotz. Persistent homology for low-complexity models. *arXiv preprint arXiv:1709.01037*, 2017.
- 14 Arakaparampil M Mathai and Serge B Provost. *Quadratic forms in random variables: theory and applications*. Dekker, 1992.
- 15 Jeff M Phillips and Suresh Venkatasubramanian. A gentle introduction to the kernel distance. *arXiv preprint arXiv:1103.1625*, 2011.
- 16 Jeff M. Phillips, Bei Wang, and Yan Zheng. Geometric inference on kernel density estimates. In *31st International Symposium on Computational Geometry, SoCG 2015, June 22-25, 2015, Eindhoven, The Netherlands*, pages 857–871, 2015. URL: <https://doi.org/10.4230/LIPIcs.SOCG.2015.857>, doi:10.4230/LIPIcs.SOCG.2015.857.

- 17 Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739. **arXiv:** <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20294>.
- 18 Donald R Sheehy. The persistent homology of distance functions under random projection. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 328. ACM, 2014.
- 19 Jack W. Silverstein. The smallest eigenvalue of a large dimensional wishart matrix. *Ann. Probab.*, 13(4):1364–1368, 11 1985.
- 20 Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Appendix

Proof of Theorem 7. Let $B_K(x, r) = \{y \in X \mid D_K(x, y) \leq r\}$, $B(x, r) = \{y \in X \mid \|x - y\| \leq r\}$ and let $w(p) = d_K^\mu(p)^2 = d_K^P(p)^2$, since we consider the empirical measure on P . We know that

$$\sigma \in \check{C}_{\alpha, K}(X) \iff \cap_{p \in \sigma} B_K(p, \sqrt{\alpha^2 - w(p)}) \neq \emptyset$$

Let $x \in B_K(p, \sqrt{\alpha^2 - w(p)})$, then $D_K(p, x)^2 \leq \alpha^2 - w(p) \iff \|p - x\|^2 \leq -2\sigma^2 \ln\left(\frac{2\sigma^2 - (\alpha^2 - w(p))}{2\sigma^2}\right)$. So $x \in B(p, \sqrt{-2\sigma^2 \ln\left(\frac{2\sigma^2 - (\alpha^2 - w(p))}{2\sigma^2}\right)})$ where $\alpha^2 \geq w(p)$. So, in other words

$$\sigma \in \check{C}_{\alpha, K}(X) \iff \bigcap_{p \in \sigma} B\left(p, \sqrt{-2\sigma^2 \ln\left(\frac{2\sigma^2 - (\alpha^2 - w(p))}{2\sigma^2}\right)}\right) \neq \emptyset.$$

Note that since by definition, $D_K(x, p)^2 \leq 2\sigma^2$, therefore we have that the maximum α that ever occurs in the Čech filtration must satisfy $\alpha^2 = D_K^2(x, p) + w(p) \leq 2\sigma^2 + w(p)$, so that we always have $\frac{\alpha^2 - w(p)}{2\sigma^2} \leq 1$. Therefore, we can use the Taylor expansion for $\ln(1 - x)$ to get,

$$-2\sigma^2 \ln\left(\frac{2\sigma^2 - (\alpha^2 - w(p))}{2\sigma^2}\right) \leq -2\sigma^2 \ln\left(1 - \frac{\alpha^2 - w(p)}{2\sigma^2}\right) \leq 2\sigma^2 \left(\frac{\alpha^2 - w(p)}{2\sigma^2} + \frac{\alpha^2 - w(p)^2}{8\sigma^4} + \dots\right).$$

Fix $p \in P$, and let $x \in \mathbb{R}^D$ be an arbitrary point. We consider the following two cases.
Case I: $\|x - p\|^2 \leq (2 \ln 2)\sigma^2$. In this case, we have that by the monotonicity of $D_K^2(x, p)$ with $\|x - p\|^2$, $D_K^2(x, p) \leq \sigma^2$. Therefore, in the kernel distance, x lies in a ball of radius α centered at p , such that $\frac{\alpha^2 - w(p)}{2\sigma^2} = \frac{D_K^2(x, p)}{2\sigma^2} \leq 1/2$. For $|x| \leq 1/2$, the Taylor expansion can be approximated as $\ln(1 - x) \leq 3x/2$. Therefore, we get that

$$rad^2(\sigma) \leq D_K^2(x, p) + w(p) \leq D_K^2(x, p) + \frac{3w(p)}{2} \leq 3\alpha^2/2.$$

Case II: $\|x - p\|^2 > (2 \ln 2)\sigma^2$. In this case, we have $\sigma^2 < D_K^2(x, p) \leq 2\sigma^2$. Thus, $\sigma^2 > D_K^2(x, p)/2$. Therefore, we get

$$\frac{D_K^2(x, p)}{2} + w(p) \leq \sigma^2 + w(p) < \alpha^2, \text{ or,}$$

$$D_K^2(x, p) + w(p) \leq D_K^2(x, p) + 2w(p) < 2\alpha^2.$$

Thus, in both the above cases, we get that

$$rad(\sigma) \leq \sqrt{2}\alpha \iff \sigma \in \check{C}_{\sqrt{2}\alpha, K}(X).$$

563 Let $\sigma = \{p_1, p_2, \dots, p_r\}$ and $c_\sigma = \sum_i \lambda_i p_i$ be the center of the minimum enclosing ball of
 564 the σ . Following the proof of Theorem 3 we have that,

$$565 \quad \text{rad}^2(\sigma) = \sum_{i=1}^r \lambda_i (\|c_\sigma - p_i\|^2 + w(p_i)) = \sum_{i < j} \lambda_i \lambda_j \|p_i - p_j\|^2 + \sum_{i=1}^k \lambda_i w(p_i),$$

566 where $w(p) = d_K^P(p)^2$. Using Corollary 21 we have $\frac{(1-\epsilon)}{2}w(p) \leq w(f(p)) \leq 2(1+\epsilon)w(p)$. So,
 567 following the proof of Theorem 5 we get the desired result. \blacktriangleleft