View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Warwick Research Archives Portal Repository

Dimensionality Reduction for k-Distance Applied to Persistent Homology

Shreya Arya
Duke University
[Durham, USA]
shreya.arya14@gmail.com

Jean-Daniel Boissonnat
Université Côte d'Azur, INRIA
[Sophia-Antipolis, France]
jean-daniel.boissonnat@inria.fr

Kunal Dutta
Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw
[Warsaw, Poland]
K.Dutta@mimuw.edu.pl

Martin Lotz
Mathematics Institute, University of Warwick
[Coventry, United Kingdom]
martin.lotz@warwick.ac.uk

Abstract

Given a set P of n points and a constant k, we are interested in computing the persistent homology of the Čech filtration of P for the k-distance, and investigate the effectiveness of dimensionality reduction for this problem, answering an open question of Sheehy [*Proc. SoCG, 2014*]. We show that *any* linear transformation that preserves pairwise distances up to a $(1 \pm \varepsilon)$ multiplicative factor, must preserve the persistent homology of the Čech filtration up to a factor of $(1-\varepsilon)^{-1}$. Our results also show that the Vietoris-Rips and Delaunay filtrations for the k-distance, as well as the Čech filtration for the approximate k-distance of Buchet et al. are preserved up to a $(1 \pm \varepsilon)$ factor.

We also prove extensions of our main theorem, for point sets (i) lying in a region of bounded Gaussian width or (ii) on a low-dimensional manifold, obtaining the target dimension bounds of Lotz [*Proc. Roy. Soc.* , 2019] and Clarkson [*Proc. SoCG, 2008*] respectively.

2012 ACM Subject Classification Theory of computation; Randomness, geometry and discrete structures; Computational geometry

Keywords and phrases Dimensionality reduction, Johnson-Lindenstrauss lemma, Topological Data Analysis, Persistent Homology, k-distance, distance to measure

Digital Object Identifier 10.4230/LIPIcs.SoCG.2020.00

Funding The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement No. 339025 GUDHI (Algorithmic Foundations of Geometry Understanding in Higher Dimensions).

1 Introduction

Persistent homology is one of the main tools used to extract information from 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 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

© Shreya Arya, Jean-Daniel Boissonnat, Kunal Dutta, Martin Lotz;
licensed under Creative Commons License CC-BY
36th International Symposium on Computational Geometry (SoCG 2020).
Editors: Sergio Cabello and Danny Z. Chen; Article No. 00; pp. 00:1–00:15
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

sets. However the usual distance function is highly sensitive to the presence of outliers and noise. One approach is to use distance functions that are more robust to outliers, such as the *distance-to-a-measure* and the related *k-distance* (for finite data sets), proposed recently by Chazal et al. [9] 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 [23, 7]. Other approaches involve using kernels [31], de-noising algorithms [8], [36], etc.

In all the above settings, the sub-routines required for computing persistent homology have exponential or worse dependence on the ambient dimension, and rapidly become unusable in real-time once the dimension grows beyond a few dozens - which is indeed the case in many applications, for example in image processing, neuro-biological networks, data mining (see e.g. [21]), a phenomenon often referred to as the *curse of dimensionality*.

The Johnson-Lindenstrauss Lemma. 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) [26]. The JL Lemma states that any set of n points in Euclidean space can be embedded in a space of dimension $O(\varepsilon^{-2}\log n)$ with $(1 \pm \varepsilon)$ distortion. Since the initial non-constructive proof of this fact by Johnson and Lindenstrauss [26], several authors have given successive improvements, e.g. Indyk and Motwani [24], Dasgupta and Gupta [15], Achlioptas [1], Ailon and Chazelle [2], Matoušek [29] and others, which address the issues of *efficient* constructivization and implementation, using random linear transformations. Dirksen [16] gave a unified theory for dimensionality reduction using subgaussian matrices.

In a different direction, variants of the Johnson-Lindenstrauss lemma with better target dimension have been given under several specific settings. For point sets lying in regions of bounded Gaussian width, a theorem of Gordon [22] implies that the target dimension can be reduced to a function of the Gaussian width, independent of the number of points. Sarlos [32] showed that points lying on a d-flat can be mapped on to $O(d/\varepsilon^2)$ dimensions independently of the number of points. Baraniuk and Wakin [5] proved an analogous result for points on a smooth manifold, which was subsequently sharpened by Clarkson [13]. Verma [34] gave a further improvement, directly preserving geodesic distances on the manifold. Other related results include those of Indyk and Naor [25] for sets of bounded doubling dimension, with additive errors, and Alon and Klartag [3] preserving general inner products, again with additive error only.

Dimension Reduction and Persistent Homology. The JL Lemma has also been used by Sheehy [33] and Lotz [27] to reduce the complexity of computing persistent homology. Both Sheehy and Lotz show that the persistent homology of a point cloud is approximately preserved under random projections [33, 27], up to a $(1 \pm \varepsilon)$ multiplicative factor, for any $\varepsilon \in [0, 1]$. Sheehy proves this for an n-point set, whereas Lotz's generalization applies to sets of bounded Gaussian width. However, their techniques involve only the usual distance to a point set and therefore remain 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 using the k-distance, is therefore a natural one, and has been raised by Sheehy [33], who observed that *"One notable distance function that is missing from this paper* [i.e. [33]] *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 the k-distance and show its applicability in computing persistent homology. It is not very hard to see that for a given point set P, the random Johnson-Lindenstrauss mapping preserves the pointwise k-distance to P (Theorem 13). However, this is not enough to preserve intersections of balls at varying scales of the radius parameter and thus does not suffice to preserve the persistent homology of Čech filtrations, as noted by Sheehy [33] and Lotz [27]. We show how the squared radius of a set of weighted points can be expressed as a convex combination of pairwise squared distances. From this, it follows that the Čech filtration under the k-distance, will be preserved by *any* linear mapping that preserves pairwise distances.

Extensions

Further, as our main result applies to any linear mapping that approximately preserves pairwise distances, the theorems of Lotz, Baraniuk and Wakin and others apply immediately. Thus, we give two extensions of our results. The first one, analogous to Lotz [27], shows that the persistent homology with respect to the k-distance, of point sets contained in regions having bounded Gaussian width, can be preserved via dimensionality reduction, with target dimension a function of the Gaussian width. Another result is that for points lying in a low-dimensional submanifold of a high-dimensional Euclidean space, the target dimension for preserving the persistent homology with k-distance depends linearly on the dimension of the manifold. Both these settings are commonly encountered in high-dimensional data analysis, machine learning, etc. (see e.g. the *manifold hypothesis* [19]).

▶ Remark 1. It should be noted that the approach of using dimensionality reduction for the k-distance, is complementary to denoising techniques such as [8] as we do not try to remove noise, only to be more robust to noise. Therefore, it can be used in conjunction with denoising techniques, as a pre-processing tool when the dimensionality is high.

Outline

The rest of this paper is as follows. In Section 2, we briefly summarize some basic definitions and background. Our theorems are stated in Section 3 and proved in Section 4. Some applications of our results are proved in Section 5. We end with some final remarks and open questions in Section 6.

2 Background

We begin with some preliminary background.

We shall need a well-known identity for the variance of bounded random variables, which will be crucial in the proof of our main theorem. Let $\lambda_1, \ldots, \lambda_k \geq 0$ be such that $\sum_{i=1}^k \lambda_i = 1$. Let $p_1, \ldots, p_k \in \mathbb{R}^D$ be given points. and let $b = \sum_{i=1}^k \lambda_i p_i$. Then for any point $x \in \mathbb{R}^D$, the following holds

$$\sum_{i=1}^{k} \lambda_i \|x - p_i\|^2 \quad = \quad \|x - b\|^2 + \sum_{i=1}^{k} \lambda_i \|b - p_i\|^2. \tag{1}$$

In particular, for $\lambda_i = 1/k$ for all i, we have

$$\frac{1}{k}\sum_{i=1}^{k}\|x-p_i\|^2 \quad = \quad \|x-b\|^2 + \sum_{i=1}^{k}\frac{1}{k}\|b-p_i\|^2. \tag{2}$$

2.1 Random Projections

The Johnson-Lindenstrauss lemma [26] states that any subset of n points of Euclidean space can be embedded in a space of dimension $O(\varepsilon^{-2}\log n)$ with $(1 \pm \varepsilon)$ distortion. In order to separate the technical aspects of our result from the issues of implementation, we use the notion of an *ε-distortion map with respect to P* (also commonly called a Johnson-Lindenstrauss map).

▶ **Definition 2.** *Given a point set $P \subset \mathbb{R}^D$, and $\varepsilon \in (0,1)$, a mapping $f : \mathbb{R}^D \to \mathbb{R}^d$ for some $d \leq D$ is an* ε-distortion map with respect to P, *if for all $x, y \in P$,*

$$(1-\varepsilon)\|x-y\| \leq \|f(x)-f(y)\| \leq (1+\varepsilon)\|x-y\|.$$

A random variable X with mean zero, is said to be *subgaussian* with *subgaussian norm* K if $\mathbb{E}\left[\exp X^2/K^2\right] \leq 2$. In this case, the tails of the random variable satisfy

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-t^2/2K^2\right).$$

We focus on the case where the Johnson-Lindenstrauss embedding is carried out via random subgaussian matrices, i.e. matrices where for some given $K > 0$, each entry is an independent subgaussian random variable with subgaussian norm K. This case is general enough to include the mappings of e.g. Achlioptas [1], Ailon-Chazelle [2], Dasgupta and Gupta [15], Indyk-Motwani [24], and Matoušek [29] (see e.g. Dirksen for a unified treatment [16]).

▶ **Lemma 3** (JL Lemma). *Given $0 < \varepsilon, \delta < 1$, and a finite point set $P \subset \mathbb{R}^D$ of size $|p| = n$. Then a random linear mapping $f\colon \mathbb{R}^D \to \mathbb{R}^d$ where $d = O(\varepsilon^{-2}\log n)$ given by $f(v) = \sqrt{\frac{D}{d}}Gv$ where G is a $d \times D$ subgaussian random matrix, is an ε-distortion map with respect to P, with probability at least $1-\delta$.*

2.2 k-Distance

The distance to a finite point set P is usually taken to be the minimum distance to a point in the set. For the computations involved in geometric and topological inference, however, this distance is extremely sensitive to outliers and noise. To handle this problem of sensitivity, Chazal et al. in [9] introduced the *distance to a probability measure* which, in the case of a uniform probability on P, is called the *k-distance*.

▶ **Definition 4** (k-distance). *For $k \in \{1, ..., n\}$ and $x \in \mathbb{R}^D$, the k-distance of x to P is*

$$d_{P,k}(x) = \min_{S_k \in \binom{P}{k}} \sqrt{\frac{1}{k}\sum_{p \in S_k}\|x-p\|^2} = \sqrt{\frac{1}{k}\sum_{p \in NN_P^k(x)}\|x-p\|^2} \tag{3}$$

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

It was shown in [4], that the k-distance can be expressed in terms of weighted points and power distance. A weighted point $\hat{p}$ is a point p of $\mathbb{R}^D$ together with a (not necessarily positive) real number called its weight and denoted by $w(p)$.The distance between two weighted points $\hat{p}_i = (p_i, w_i)$ and $\hat{p}_j = (p_j, w_j)$ is defined as $D(\hat{p}_i, \hat{p}_j) = \|p_i - p_j\|^2 - w_i - w_j$. This definition encompasses the case where the two weights are 0, in which case we have the squared euclidean distance and the case where one of the points has weight 0, in which case, we have the power distance of a point to a ball. We say that two weighted points are *orthogonal* when their distance is 0.
Let $B_{P,k}$ be the set of iso-barycentres of all subsets of k points in P. To each barycenter $b = (1/k)\sum_i p_i \in B_{P,k}$, we associate the weight $w(b) = -\frac{1}{k}\sum_i \|b - p_i\|^2$. Writing $\hat{B}_{P,k} = \{\hat{b} = (b, w(b)), b \in B_{P,k}\}$, we see from (2) that the k-distance is the square root of a power distance [4]

$$d_{P,k}(x) = \min_{\hat{b} \in \hat{B}_{P,k}} \sqrt{D(x, \hat{b})}. \tag{4}$$

Observe that in general the squared distance between a pair of weighted points can be negative, but the above assignment of weights ensures that the k-distance $d_{P,k}$ is a real function. Since $d_{P,k}$ is the square root of a non-negative power distance, the α-sublevel set of $d_{P,k}$, $d_{P,k}([-\infty, \alpha])$, $\alpha \in \mathbb{R}$, is the union of $\binom{n}{k}$ balls $B(b, \sqrt{\alpha^2 + w(b)})$, $b \in B_{P,k}$. However, some of the balls may be included in the union of others and be redundant. In fact, the number of barycenters (or equivalently of balls) required to define a level set of $d_{P,k}$ is equal to the number of the non-empty cells in the kth-order Voronoi diagram of P. Hence the number of non-empty cells is $\Omega\left(n^{\lfloor (D+1)/2 \rfloor}\right)$ [14] and computing them in high dimensions is intractable. It is then natural to look for approximations of the k-distance, e.g., the following definition has been proposed [7]:

▶ **Definition 5** (Approximation). *Let $P \subset \mathbb{R}^D$ and $x \in \mathbb{R}^D$. The approximate k-distance $\tilde{d}_{P,k}(x)$ is defined as*

$$\tilde{d}_{P,k}(x) \quad := \quad \min_{p \in P} \sqrt{D(x, \hat{p})} \tag{5}$$

where $\hat{p} = (p, w(p))$ with $w(p) = -d^2_{P,k}(p)$, the opposite of the squared k-distance of p.

As in the exact case, $\tilde{d}_{P,k}$ is the square root of a power distance and its α-sublevel set, $\alpha \in \mathbb{R}$, is a union of balls, specifically the balls $B(p, \sqrt{\alpha^2 - d^2_{P,k}(p)})$, $p \in P$. The major difference with the exact case is that, since we consider only balls around the points of P, their number is n instead of $\binom{n}{k}$ in the exact case (compare Eq. (5) and Eq. (4)). Still, $\tilde{d}_{P,k}(x)$ approximates the k-distance [7]:

$$\frac{1}{\sqrt{2}}\, d_{P,k} \leq \tilde{d}_{P,k} \leq \sqrt{3}\, d_{P,k}. \tag{6}$$

We now make an observation for the case when the weighted points are barycenters, which will be very useful in proving our main theorem.

▶ **Lemma 6.** *Given $b_1, b_2 \in B_{P,k}$, and $p_{i,1}, \ldots, p_{i,k} \in P$ for $i = 1, 2$, such that $b_i = \frac{1}{k}\sum_{l=1}^k p_{i,l}$, and $w(b_i) = \frac{1}{k}\sum_{l=1}^k \|b_i - p_{i,l}\|^2$ for $i = 1, 2$, then it holds that*

$$D(\hat{b}_1, \hat{b}_2) = \frac{1}{k^2} \sum_{l,s=1}^{k} \|p_{1,l} - p_{2,s}\|^2.$$

Proof of Lemma 6. We have

$$D(\hat{b}_1, \hat{b}_2) \;=\; \|b_1 - b_2\|^2 - w(b_1) - w(b_2) \;=\; \|b_1 - b_2\|^2 + \frac{1}{k}\sum_{l=1}^{k} \|b_1 - p_{1,l}\|^2 + \frac{1}{k}\sum_{l=1}^{k} \|b_2 - p_{2,l}\|^2.$$

Applying the identity (2), we get $\|b_1 - b_2\|^2 + \frac{1}{k}\sum_{l=1}^{k} \|b_2 - p_{2,l}\|^2 = \frac{1}{k}\sum_{l=1}^{k} \|b_1 - p_{2,l}\|^2$, so that

$$\begin{aligned} D(\hat{b}_1, \hat{b}_2) &= \frac{1}{k}\sum_{l=1}^{k} \|b_1 - p_{2,l}\|^2 + \frac{1}{k}\sum_{l=1}^{k} \|b_1 - p_{1,l}\|^2 \\ &= \frac{1}{k}\sum_{l=1}^{k} \|b_1 - p_{2,l}\|^2 + \frac{1}{k^2}\sum_{s=1}^{k}\sum_{l=1}^{k} \|b_1 - p_{1,l}\|^2 \\ &= \frac{1}{k}\sum_{l=1}^{k} \left(\|b_1 - p_{2,l}\|^2 + \frac{1}{k}\sum_{s=1}^{k} \|b_1 - p_{1,s}\|^2 \right) \\ &= \frac{1}{k}\sum_{l=1}^{k} \left(\frac{1}{k}\sum_{s=1}^{k} \|p_{1,s} - p_{2,l}\|^2 \right) \;=\; \frac{1}{k^2}\sum_{l,s=1}^{k} \|p_{1,s} - p_{2,l}\|^2, \end{aligned} \tag{7}$$

where in (7), we again applied (2) to each of the points $p_{2,s}$, with respect to the barycenter b_1. ◀

2.3 Persistent Homology

Simplicial Complexes and Filtrations Let V be a finite set. An (abstract) simplicial complex with vertex set V is a set K of finite subsets of V such that if $A \in K$ and $B \subseteq A$, then $B \in K$. The sets in K are called the simplices of K. A simplex $F \in K$ that is strictly contained in a simplex $A \in K$, is said to be a *face* of A.

A simplicial complex K with a function $f : K \to \mathbb{R}$ such that $f(\sigma) \leq f(\tau)$ whenever σ is a face of τ is a filtered simplicial complex. The sublevel set of f at $r \in \mathbb{R}$, $f^{-1}(-\infty, r]$, is a subcomplex of K. By considering different values of r, we get a nested sequence of subcomplexes (called a filtration) of K, $\emptyset = K^0 \subseteq K^1 \subseteq ... \subseteq K^m = K$, where K^i is the sublevel set at value r_i.

The Čech filtration associated to a finite set P of points in $\mathbb{R}^D$ plays an important role in Topological Data Analysis.

▶ **Definition 7** (Čech Complex)**.** *The Čech complex $\check{C}_\alpha(P)$ is the set of simplices $\sigma \subset P$ such that rad(σ) $\leq \alpha$, where* $\mathrm{rad}(\sigma)$ *is the radius of the smallest enclosing ball of σ, i.e.*

$$\mathrm{rad}(\sigma) \leq \alpha \Leftrightarrow \exists x \in \mathbb{R}^D,\ \forall p_i \in \sigma,\ \|x - p_i\| \leq \alpha.$$

When α goes from 0 to $+\infty$, we obtain the Čech filtration of P. $\check{C}_\alpha(P)$ can be equivalently defined as the nerve of the closed balls $\overline{B}(p, \alpha)$, centered at the points in P and of radius α:

$$\check{C}_\alpha(P) = \{\sigma \subset P | \cap_{p \in \sigma} \overline{B}(p, \alpha) \neq \emptyset\}.$$

By the nerve lemma, we know that the union of balls $U_\alpha = \cup_{p \in P} \overline{B}(p, \alpha)$, $p \in P$, and $\check{C}_\alpha(P)$ have the same homotopy type.

Persistence Diagrams. Persistent homology is a means to compute and record the changes in the topology of the filtered complexes as the parameter α increases from zero

to infinity. Edelsbrunner, Letscher and Zomorodian [17] gave an algorithm to compute the persistent homology, which takes a filtered simplicial complex as input, and outputs a sequence $(\alpha_{birth}, \alpha_{death})$ of pairs of real numbers. Each such pair corresponds to a topological feature, and records the values of α at which the feature appears and disappears, respectively, in the filtration. Thus the topological features of the filtration can be represented using this sequence of pairs, which can be represented either as points in the extended plane $\bar{\mathbb{R}}^2 = (\mathbb{R} \cup \{-\infty, \infty\})^2$, called the *persistence diagram* or as a sequence of barcodes (the *persistence barcode*) (see, e.g., [18]). A pair of persistence diagrams $\mathbb{G}$ and $\mathbb{H}$ corresponding to the filtrations (G_α) and (H_α) respectively, are *multiplicatively β-interleaved*, $(\beta \geq 1)$, if for all α, we have that $G_{\alpha/\beta} \subseteq H_\alpha \subseteq G_{\alpha\beta}$. We shall crucially rely on the fact that a given persistence diagram is closely approximated by another one if they are multiplicatively c-interleaved, with c close to 1 (see e.g. [10]).

The Persistent Nerve Lemma [12] shows that the persistent homology of the Čech complex is the same as the homology of the α-sublevel filtration of the distance function.

The Weighted Case. Our goal is to extend the above definitions and results to the case of the k-distance. As we observed earlier, the k-distance is a power distance in disguise. Accordingly, we need to extend the definition of the Čech complex to sets of weighted points.

▶ **Definition 8** (Weighted Čech Complex). *Let $\hat{P} = \{\hat{p}_1, ..., \hat{p}_n\}$ be a set of weighted points, where $\hat{p}_i = (p_i, w_i)$. The α-Čech complex of $\hat{P}$, $\check{C}_\alpha(\hat{P})$, is the set of all simplices σ satisfying*

$$\exists x,\ \forall p_i \in \sigma,\ \|x - p_i\|^2 \leq w_i + \alpha^2 \quad \Leftrightarrow \quad \exists x,\ \forall p_i \in \sigma,\ D(x, \hat{p}_i) \leq \alpha^2.$$

In other words, the α-Čech complex of $\hat{P}$ is the nerve of the closed balls $\overline{B}(p_i, r_i^2 = w_i + \alpha^2)$, centered at the p_i and of squared radius $w_i + \alpha^2$ (if negative, $\overline{B}(p_i, r_i^2)$ is imaginary).

The notions of weighted Čech filtrations and their persistent homology now follow naturally. Moreover, it follows from (4) that the Čech complex $\check{C}_\alpha(P)$ for the k-distance is identical to the weighted Čech complex $\check{C}_\alpha(\hat{B}_{P,k})$, where $\hat{B}_{P,k}$ is, as above, the set of iso-barycenters of all subsets of k points in P.

In the Euclidean case, we equivalently defined the α-Čech complex as the collection of simplices whose smallest enclosing balls have radius at most α. We can proceed similarly in the weighted case. Let $\hat{X} \subseteq \hat{P}$. We define the *radius of* $\hat{X}$ as $\mathrm{rad}^2(\hat{X}) = \min_{x \in \mathbb{R}^D} \max_{\hat{p}_i \in \hat{X}} D(x, \hat{p}_i)$, and the weighted center or simply the *center* of $\hat{X}$ as the point, noted $c(\hat{X})$, where the minimum is reached.

Our goal is to show that preserving smallest enclosing balls in the weighted scenario under a given mapping, also preserves the persistent homology. Sheehy [33] and Lotz [27], proved this for the unweighted case. Their proofs also work for the weighted case but only under the assumption that the weights stay unchanged under the mapping. In our case however, the weights need to be recomputed in $f(\hat{P})$. We therefore need a version of [27, Lemma 2.2] for the weighted case which does not assume that the weights stay the same under f. This is Lemma 12, which follows at the end of this section. The following lemmas will be instrumental in proving Lemma 12 and in proving our main result. Let $\hat{X} \subseteq \hat{P}$ and assume without loss of generality that $\hat{X} = \{\hat{p}_1, ..., \hat{p}_m\}$, where $\hat{p}_i = (p_i, w_i)$.

▶ **Lemma 9.** *$c(\hat{X})$ and $\mathrm{rad}(\hat{X})$ are uniquely defined.*

▶ **Lemma 10.** *Let I be the set of indices for which $D(c, \hat{p}_i) = \mathrm{rad}(\hat{X})$ and let $\hat{X}_I = \{\hat{p}_i, i \in I\}$. $c(\hat{X})$ is a convex combination of the points in X_I, i.e. $c(\hat{X}) = \sum_{i=1}^m \lambda_i p_i$ with $\sum_{i=1}^m \lambda_i = 1$, $\lambda_i \geq 0$ for all i, and $\lambda_i = 0$ for all $i \notin I$.*

Combining the above lemmas with [27, Lemma 4.2] gives the following lemma.

▶ **Lemma 11.** $\mathrm{rad}^2(\hat{X}) = \frac{1}{2}\sum_{i\in I}\sum_{j\in I}\lambda_i\lambda_j D(\hat{p}_i,\hat{p}_j)$.

Let $X \in \mathbb{R}^D$ be a finite set of points and $\hat{X}$ be the associated weighted points where the weights are computed according to a weighting function $w : X \to \mathbb{R}^-$. Given a mapping $f : \mathbb{R}^D \to \mathbb{R}^d$, we define $\widehat{f(X)}$ as the set of weighted points $\{(f(x), w(f(x))), x \in X\}$. Note that the weights are recomputed in the image space $\mathbb{R}^d$.

▶ **Lemma 12.** *In the above setting, if f is such that for some $\varepsilon \in (0,1)$ and for all subsets $\hat{S} \subseteq \hat{X}$ we have*

$$(1-\varepsilon)\mathrm{rad}^2(\hat{S}) \le \mathrm{rad}^2(\widehat{f(S)}) \le (1+\varepsilon)\mathrm{rad}^2(\hat{S}),$$

then the weighted Čech filtrations of $\hat{X}$ and $f(\hat{X})$ are multiplicatively $(1-\varepsilon)^{-1/2}$ interleaved.

3 Results

For the subsequent theorems, we denote by P a set of n points in $\mathbb{R}^D$.

Our first theorem shows that for the points in P, the pointwise k-distance $d_{P,k}$ is preserved by a random subgaussian matrix satisfying Lemma 3.

▶ **Theorem 13.** *Given $\varepsilon \in (0,1]$, an ε-distortion map with respect to P $f\colon \mathbb{R}^D \to \mathbb{R}^d$, where $d = O(\varepsilon^{-2}\log n)$, satisfies for all points $x \in P$:*

$$(1-\varepsilon)d^2_{P,k}(x) \le d^2_{f(P),k}(f(x)) \le (1+\varepsilon)d^2_{P,k}(x).$$

Moreover, given any $\delta \in (0,1)$, the above inequality holds with probability at least $1-\delta$ for a random function $f : \mathbb{R}^D \to \mathbb{R}^d$ given by $f(x) = \sqrt{D/d}Gx$, where G is a random subgaussian matrix, and $d = O\left(\frac{\log n}{\varepsilon^2}\right)$, where the constant in the O-notation depends on δ.

As mentioned previously, the preservation of the pointwise k-distance does not imply the preservation of the Čech complex formed using the points in P. Nevertheless, the following theorem shows that this can always be done in dimension $O(\log n/\varepsilon^2)$.

Let $\hat{B}_{P,k}$ be the set of iso-barycenters of every k-subset of P, weighted as in Section 2.2. Recall from Section 2.3 that the weighted Čech complex $\check{C}_\alpha(\hat{B}_{P,k})$ is identical to the Čech complex $\check{C}_\alpha(P)$ for the k-distance.

▶ **Theorem 14** (k-distance)**.** *Let $\hat{\sigma} \subseteq \hat{B}_{P,k}$ be a simplex in the weighted Čech complex $\check{C}_\alpha(\hat{B}_{P,k})$. Then, given $d \le D$ such that there exists a ε-distortion map with respect to P $f : \mathbb{R}^D \to \mathbb{R}^d$, the following holds:*

(i) $(1-\varepsilon)\mathrm{rad}^2(\hat{\sigma}) \le \mathrm{rad}^2(\widehat{f(\sigma)}) \le (1+\varepsilon)\mathrm{rad}^2(\hat{\sigma})$.

(ii) In particular, for a n-point set P, given $\delta \in (0,1)$, the function $f : \mathbb{R}^D \to \mathbb{R}^d$ given by $f(x) = \left(\sqrt{D/d}\right)Gx$, where G is a random $d\times D$ Gaussian matrix G where $d = O\left(\frac{\log n}{\varepsilon^2}\right)$, satisfies the above inequality with probability at least $1-\delta$.

For the approximation of the k-distance given by [7] also, we get an optimal target dimension, as the number of weighted points needed to compute the approximate k-distance, is just n.

▶ **Theorem 15** (Approximate k-distance). *Let $\hat{P}$ be the weighted points associated with P, introduced in Definition 5 (Equ. 5). Let, in addition, $\hat{\sigma} \subseteq \hat{P}$ be a simplex in the associated weighted Čech complex $\check{C}_\alpha(\hat{P})$. Then an ε-distortion mapping with respect to P, $f : \mathbb{R}^D \to \mathbb{R}^d$ satisfies: $(1-\varepsilon)\mathrm{rad}^2(\hat{\sigma}) \leq \mathrm{rad}^2(\widehat{f(\sigma)}) \leq (1+\varepsilon)\mathrm{rad}^2(\hat{\sigma})$. Moreover, the function $f : \mathbb{R}^D \to \mathbb{R}^d$ given by $f(x) = \left(\sqrt{D/d}\right) Gx$, where G is a random $d \times D$ Gaussian matrix G where $d = O(\log n/\varepsilon^2)$, satisfies the above inequality, with probability at least $1-\delta$.*

Applying Lemma 12 to the theorems 14 and 15, we get the following corollary.

▶ **Corollary 16.** *The persistent homology for the Čech filtrations of P and its image $f(P)$ under any ε-distortion mapping with respect to P, using the (i) exact k-distance, as well as the (ii) approximate k-distance, are multiplicatively $(1-\varepsilon)^{-1/2}$-interleaved with probability $1-\delta$.*

However, note that the approximation in Corollary 16 (ii) is with respect to the *approximate* k-distance, which is itself an $O(1)$ approximation of the k-distance (see (6)).

4 Proofs

We begin with the proofs of the auxiliary lemmas.

Proof of Lemma 9. The proof follows from the convexity of D (see Lemma 6). Assume, for a contradiction, that there exists two centers c_0 and $c_1 \neq c_0$ for $\hat{X}$. For convenience, write $r = \mathrm{rad}(\hat{X})$. By the definition of the center of $\hat{X}$, we have

$$\begin{aligned}
\exists \hat{p}_0, \forall \hat{p}_i : D(c_0, \hat{p}_i) &\leq D(c_0, \hat{p}_0) = \|c_0 - p_0\|^2 - w_0 = r^2 \\
\exists \hat{p}_1, \forall \hat{p}_i : D(c_1, \hat{p}_i) &\leq D(c_1, \hat{p}_1) = \|c_1 - p_1\|^2 - w_1 = r^2.
\end{aligned}$$

Consider $D_\lambda(\hat{p}_i) = (1-\lambda)D(c_0, \hat{p}_i) + \lambda D(c_1, \hat{p}_i)$ and write $c_\lambda = (1-\lambda)c_0 + \lambda c_1$. For any $\lambda \in (0,1)$, we have

$$\begin{aligned}
D_\lambda(\hat{p}_i) &= (1-\lambda)D(c_0, \hat{p}_i) + \lambda D(c_1, \hat{p}_i) \\
&= (1-\lambda)(c_0 - p_i)^2 + \lambda(c_1 - p_i)^2 - w_i \\
&= D(c_\lambda, \hat{p}_i) - c_\lambda^2 + (1-\lambda)c_0^2 + \lambda c_1^2 \\
&= D(c_\lambda, \hat{p}_i) + \lambda(1-\lambda)(c_0 - c_1)^2 \\
&> D(c_\lambda, \hat{p}_i).
\end{aligned}$$

Moreover, for any i,

$$D_\lambda(\hat{p}_i) = (1-\lambda)D(c_0, \hat{p}_i) + \lambda D(c_1, \hat{p}_i) \leq r^2.$$

Thus, for any i and any $\lambda \in (0,1)$, $D(c_\lambda, \hat{p}_i) < r^2$. Hence c_λ is a better center than c_0 and c_1, and r is not the minimal possible value for $\mathrm{rad}(\hat{X})$. We have obtained a contradiction. ◀

Proof of Lemma 10. We write for convenience $c = c(\hat{X})$ and $r = \mathrm{rad}(\hat{X})$ and prove that $c \in \mathrm{conv}(X_I)$ by contradiction. Let $c' \neq c$ be the point of $\mathrm{conv}(X_I)$ closest to c, and $\tilde{c} \neq c$ be a point on $[cc']$. Since $\|\tilde{c} - p_i\| < \|c - p_i\|$ for all $i \in I$, $D(\tilde{c}, \hat{p}_i) < D(c, \hat{p}_i)$ for all $i \in I$. For $\tilde{c}$ sufficiently close to c, $\tilde{c}$ remains closer to the weighted points $\hat{p}_j$, $j \notin I$, than to the $\hat{p}_i$, $i \in I$. We thus have

$$D(\tilde{c}, \hat{p}_j) < D(\tilde{c}, \hat{p}_i) < D(c, \hat{p}_i) = r^2.$$

It follows that c is not the center of $\hat{X}$, a contradiction.

◀

Proof of Lemma 11. From Lemma 10, and writing $c = c(\hat{X})$ for convenience, we have

$$\mathrm{rad}^2(\hat{X}) = \sum_{i\in I} \lambda_i \big(\|c - p_i\|^2 - w_i\big).$$

We use the following simple fact from [27, Lemma 4.5]

$$\sum_{i\in I} \lambda_i \|c - p_i\|^2 = \frac{1}{2} \sum_{i\in I} \sum_{j\in I} \lambda_i \lambda_j \|p_i - p_j\|^2.$$

Substituting in the expression for $\mathrm{rad}^2(\hat{X})$,

$$\begin{aligned}
\mathrm{rad}^2(\hat{X}) &= \frac{1}{2} \sum_{j\in I} \sum_{i\in I} \lambda_j \lambda_i \|p_i - p_j\|^2 - \frac{1}{2} \sum_{i\in I} 2\lambda_i w_i \\
&= \frac{1}{2} \sum_{i,j\in I} \lambda_j \lambda_i \|p_i - p_j\|^2 - \frac{1}{2} \sum_{i,j\in I} 2\lambda_i \lambda_j w_i \quad (\text{since} \quad \sum_{j\in I} \lambda_j = 1) \\
&= \frac{1}{2} \sum_{i,j\in I} \lambda_j \lambda_i \|p_i - p_j\|^2 - \frac{1}{2} \sum_{i,j\in I} \lambda_i \lambda_j (w_i + w_j) \\
&= \frac{1}{2} \sum_{i,j\in I} \lambda_i \lambda_j \left(\|p_i - p_j\|^2 - w_i - w_j\right) \\
&= \frac{1}{2} \sum_{i,j\in I} \lambda_i \lambda_j D(\hat{p}_i, \hat{p}_j).
\end{aligned}$$

◀

Proof of Theorem 13. The proof follows from the observation that the squared k-distance from any point $p \in P$ to the point set P, is a convex combination of the squares of the Euclidean distances to the k nearest neighbours of p. Since the mapping in the JL Lemma 3 is linear, and it $(1 \pm \varepsilon)$-preserves squared pairwise distances, their convex combinations also get $(1 \pm \varepsilon)$-preserved. ◀

Proof of Theorem 14. Let $\hat{\sigma} = \{\hat{b}_1, \hat{b}_2, ..., \hat{b}_m\}$, where $\hat{b}_i$ is the weighted point defined in Section 2.3, i.e. $\hat{b}_i = (b_i, w(b_i))$ with $b_i \in B_{P,k}$ and $w(b_i) = -\frac{1}{k} \sum_{l=1}^{k} \|b_i - p_{il}\|^2$, where $p_{i,1}, \ldots, p_{i,k} \in P$ are such that $b_i = \frac{1}{k} \sum_{j=1}^{k} p_{i,j}$. Applying Lemma 11 to $\hat{\sigma}$, we have that

$$\mathrm{rad}^2(\hat{\sigma}) = \frac{1}{2} \sum_{i,j\in I} \lambda_i \lambda_j D(\hat{b}_i, \hat{b}_j). \tag{8}$$

By Lemma 6, the distance between $\hat{p}_i$ and $\hat{p}_j$ is $D(\hat{b}_i, \hat{b}_j) = \frac{1}{k^2} \sum_{l,s=1}^{k} \|p_{i,l} - p_{j,s}\|^2$. As this last expression is a convex combination of squared pairwise distances of points in P, it is $(1 \pm \varepsilon)$-preserved by any ε-distortion map with respect to P, which implies that the convex combination $\mathrm{rad}^2(\hat{\sigma}) = \frac{1}{2} \sum_{i,j\in I} \lambda_i \lambda_j D(\hat{p}_i, \hat{p}_j)$ corresponding to the squared radius of σ in $\mathbb{R}^D$, will be $(1 \pm \varepsilon)$-preserved.

Let $f : \mathbb{R}^D \to \mathbb{R}^d$ be an ε-distortion map with respect to P, from $\mathbb{R}^D$ to $\mathbb{R}^d$, where d will be chosen later. By Lemma 11, the centre of $\widehat{f(\sigma)}$ is a convex combination of the points $(f(b_i))_{i=1}^m$. Let the centre $c(\widehat{f(\sigma)})$ be given by $c(\widehat{f(\sigma)}) = \sum_{i\in I} \nu_i D(\widehat{f(b_i)})$. where for $i \in I$, $\nu_i \geq 0$, $\sum_i \nu_i = 1$. Consider the convex combination of power distances $\sum_{i,j\in I} \nu_i \nu_j D(\hat{b}_i, \hat{b}_j)$. Since f is an ε-distortion map with respect to P, by Lemmas 6 and 3 we get

$$\frac{1}{2}(1-\varepsilon) \sum_{i,j\in I} \nu_i \nu_j D(\hat{b}_i, \hat{b}_j) \leq \frac{1}{2} \sum_{i,j\in I} \nu_i \nu_j D(\widehat{f(b_i)}, \widehat{f(b_j)}) = \mathrm{rad}^2(\widehat{f(\sigma)}). \tag{9}$$

On the other hand, since the squared radius is a minimizing function by definition, we get that

$$\begin{aligned}
\mathrm{rad}^2(\hat{\sigma}) &= \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j D(\hat{b}_i,\hat{b}_j) \;\leq\; \frac{1}{2}\sum_{i,j\in I}\nu_i\nu_j D(\hat{b}_i,\hat{b}_j) && (10)\\
&\leq \frac{1}{(1-\varepsilon)}\mathrm{rad}^2(f(\sigma)),\ \text{by (9)}\\
\mathrm{rad}^2(\widehat{f(\sigma)}) &= \frac{1}{2}\sum_{i,j\in I}\nu_i\nu_j D(\widehat{f(b_i)},\widehat{f(b_j)}) && (11)\\
&\leq \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j D(\widehat{f(b_i)},\widehat{f(b_j)}). && (12)
\end{aligned}$$

Combining the inequalities (9), (10), (12) gives

$$(1-\varepsilon)\mathrm{rad}^2(\hat{\sigma}) \;\leq\; \mathrm{rad}^2(\widehat{f(\sigma)}) \;\leq\; \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j D(\widehat{f(b_i)},\widehat{f(b_j)}) \;\leq\; (1+\varepsilon)\mathrm{rad}^2(\hat{\sigma}).$$

where the final inequality follows by Lemma 3, since f is an ε-distortion map with respect to P. Thus, we have that

$$(1-\varepsilon)\mathrm{rad}^2(\hat{\sigma}) \;\leq\; \mathrm{rad}^2(\widehat{f(\sigma)}) \;\leq\; (1+\varepsilon)\mathrm{rad}^2(\hat{\sigma}),$$

which completes the proof of the theorem. ◀

Proof of Theorem 15. Recall that, in Section 2.2, we defined the approximate k-distance to be $\tilde{d}_{P,k}(x) := \min_{p\in P}\sqrt{D(x,\hat{p})}$, where $\hat{p} = (p, w(p))$ is a weighted point, having weight $w(p) = -d^2_{P,k}(p)$. So, the Čech complex would be formed by the intersections of the balls around the weighted points in P. The proof follows on the lines of the proof of Theorem 14. Let $\hat{\sigma} = \{\hat{p}_1, \hat{p}_2, ..., \hat{p}_m\}$, where $\hat{p}_1, \ldots, \hat{p}_m$ are weighted points in $\hat{P}$, and let $c(\hat{\sigma})$ be the center of $\hat{\sigma}$. Applying again Lemma 11, we get

$$\mathrm{rad}^2(\hat{\sigma}) = \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j\|p_i-p_j\|^2 + \sum_{i\in I}\lambda_i w(p_i) = \sum_{i,j\in I; i<j}\lambda_i\lambda_j\|p_i-p_j\|^2 + \sum_{i\in I}\lambda_i w(p_i),$$

where $w(p) = d^2_{P,k}(p)$. In the second equality, we used the fact that the summand corresponding to a fixed pair of distinct indices $i < j$ is being counted twice and that the contribution of the terms corresponding to indices $i = j$ is zero. An ε-distortion map with respect to P preserves pairwise distances and the k-distance in dimension $O(\varepsilon^{-2}\log n)$. The result then follows as in the proof of Theorem 14. ◀

5 Extensions

In this section we state and prove some extensions of Theorem 14 for dimensionality reduction, obtaining better bounds for the target dimension than in Section 3, in certain settings like point sets contained in regions of bounded Gaussian width, or in low-dimensional submanifolds of Euclidean space.

5.1 Sets of Bounded Gaussian Width

The first result in this section, is analogous to a theorem [27] for point sets contained in a region of bounded Gaussian width.

▶ **Definition 17.** *Given a set $S \subset \mathbb{R}^D$, the* Gaussian width *of S is*

$$w(S) := \mathbb{E}\left[\sup_{x \in S} \langle x, g \rangle\right],$$

where $g \in \mathbb{R}^D$ is a random standard D-dimensional Gaussian vector.

In several areas like geometric functional analysis, compressed sensing, machine learning, etc. the Gaussian width is a very useful measure of the width of a set in Euclidean space (see e.g. [20] and the references therein). It is also closely related to the *statistical dimension* of a set (see e.g. [35, Chapter 7]).

▶ **Theorem 18.** *Let $P \subset \mathbb{R}^D$ be a finite set of points, and define $S := \{(x-y)/\|x-y\| : x, y \in P\}$. Let $w(S)$ denote the Gaussian width of S. Then, given any $\varepsilon, \delta \in (0,1)$, for any $d \geq \frac{\left(w(S)+\sqrt{2\log(2/\delta)}\right)^2}{\varepsilon^2} + 1$, the map from $\mathbb{R}^D \to \mathbb{R}^d$, given by $x \mapsto \sqrt{D/d}Gx$, where $d = O\left(\frac{\log n}{\varepsilon^2}\right)$ and G is a $d \times D$ random Gaussian matrix, preserves the persistent homology of the Čech filtration associated to P, up to a multiplicative factor of $(1-\varepsilon)^{-1/2}$, with probability at least $1-\delta$.*

Note that since the Gaussian width of an n-point set is at most $O(\log n)$ (using e.g. the Gaussian concentration inequality, see e.g. [6, Section 2.5]), Theorem 18 strictly generalizes Theorem 14 (*ii*).

Proof of Theorem 18. We state an analogue of the Johnson Lindenstrauss lemma for sets of bounded Gaussian width, given in [22, Theorem 3.1], which essentially follows from a result of Gordon [22].

▶ **Theorem 19** ([27], Theorem 3.1)**.** *Given $\varepsilon,\ \delta \in (0,1)$, $P \subset \mathbb{R}^D$, let $S := \{(x-y)/\|x-y\| : x, y \in S\}$. Then for any $d \geq \frac{\left(w(S)+\sqrt{2\log(2/\delta)}\right)^2}{\varepsilon^2} + 1$, the function $f : \mathbb{R}^D \to \mathbb{R}^d$ given by $f(x) = \left(\sqrt{D/d}\right)Gx$, where G is a random $d \times D$ Gaussian matrix G, is an ε-distortion map with respect to P, with probability at least $1-\delta$.*

By Theorem 19, the scaled random Gaussian matrix $f : x \mapsto \left(\sqrt{D/d}\right)Gx$ is an ε-distortion map with respect to P, with target dimension $d \geq \frac{\left(w(S)+\sqrt{2\log(2/\delta)}\right)^2}{\varepsilon^2} + 1$. Now applying the first statement in Theorem 14 to the point set P with the mapping f, immediately gives us that for any simplex $\hat{\sigma} \in \check{C}_\alpha(\hat{B}_{P,k})$, where $\check{C}_\alpha(\hat{B}_{P,k})$ is the weighted Čech complex with parameter α, the squared radius $\mathrm{rad}^2(\hat{\sigma})$ is preserved up to a multiplicative factor of $(1 \pm \varepsilon)$. By Lemma 12, this implies that the persistent homology for the Čech filtration is $(1-\varepsilon)^{-1/2}$-multiplicatively interleaved. ◀

5.2 Submanifolds of Euclidean Space

For point sets lying on a low-dimensional manifold in a high-dimensional Euclidean space, one can obtain a better target dimension using the bounds of Baraniuk and Wakin [5] or Clarkson [13], which will depend only on the parameters of the manifold.

▶ **Theorem 20.** *There exists an absolute constant $c > 0$ such that, given a finite point set P lying on a connected, compact, orientable, differentiable μ-dimensional manifold $M \subset \mathbb{R}^D$, and $\varepsilon, \delta \in (0,1)$, a random projection map $f : \mathbb{R}^D \to \mathbb{R}^d$ preserves the persistent homology of the Čech filtraton computed on P, using the k-distance, with probability at least $1 - \delta$, provided*

$$d \geq c\left(\frac{\mu \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon^2} + \frac{C(M)}{\varepsilon^2}\right),$$

where $C(M)$ depends only on M.

Proof of Theorem. The proof is a direct application of Clarkson's bound [13] to Theorem 14 (i). Clarkson's theorem is stated below.

▶ **Theorem 21** (Clarkson [13])**.** *There exists an absolute constant $c > 0$ such that, given a connected, compact, orientable, differentiable μ-dimensional manifold $M \subset \mathbb{R}^D$, and $\varepsilon, \delta \in (0,1)$, any random projection map $f : \mathbb{R}^D \to \mathbb{R}^d$, is an ε-distortion map with respect to P, with probability at least $1 - \delta$, for*

$$d \geq c\left(\frac{\mu \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon^2} + \frac{C(M)}{\varepsilon^2}\right),$$

where $C(M)$ depends only on M.

Now the statement of Theorem 20 follows directly by applying Clarkson's theorem to Theorem 14 (i).

◀

6 Conclusion and Future Work

Vietoris-Rips and Delaunay filtrations. Since the Vietoris-Rips filtration [30, Chapter 4] depends only on pairwise distances, it follows from Theorem 13 that this filtration is preserved upto a multiplicative factor of $(1-\varepsilon)^{-1/2}$, under a Johnson-Lindenstrauss mapping. Furthermore, the Delaunay and the Čech filtrations [30, Chapter 4] have the same persistent homology. Theorems 14 (i) therefore implies that the Delaunay filtration of a given finite point set P is also $(1-\varepsilon)^{-1/2}$-preserved under an ε-distortion map with respect to P. Thus, theorems 14 (ii), 15, 18 and 20 apply also to the Vietoris-Rips and Delaunay filtrations.

Kernels. Other distance functions defined using kernels have proved successful in overcoming issues due to outliers. Using a result analogous to Theorem 13, we can show that random projections preserve the persistent homology for kernels up to a $C(1-\varepsilon)^{-1/2}$ factor where C is a constant. We don't know if we can make $C = 1$ as for the k-distance.

Acknowledgement

We thank the reviewers for their helpful comments and suggestions.

References

1 Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671 – 687, 2003. Special Issue on PODS 2001.

2 Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, STOC '06, pages 557–563, New York, NY, USA, 2006. ACM.

3 Noga Alon and Bo'az Klartag. Optimal compression of approximate inner products and dimension reduction. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 639–650. IEEE Computer Society, 2017.

4 Franz Aurenhammer. A new duality result concerning Voronoi diagrams. *Discrete & Computational Geometry*, 5(3):243–254, Jun 1990.

5 Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

6 Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, UK, 2013.

7 Mickaël Buchet, Frédéric Chazal, Steve Y. Oudot, and Donald R. Sheehy. Efficient and robust persistent homology for measures. *Comput. Geom.*, 58:70–96, 2016.

8 Mickaël Buchet, Tamal K. Dey, Jiayuan Wang, and Yusu Wang. Declutter and resample: Towards parameter free denoising. *JoCG*, 9(2):21–46, 2018.

9 Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.

10 Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Y. Oudot. *The Structure and Stability of Persistence Modules*. Springer Briefs in Mathematics. Springer, 2016.

11 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.

12 Frédéric Chazal and Steve Yann Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the Twenty-fourth Annual Symposium on Computational Geometry*, SCG '08, pages 232–241, New York, NY, USA, 2008. ACM.

13 Kenneth L. Clarkson. Tighter bounds for random projections of manifolds. In *Proceedings of the Twenty-Fourth Annual Symposium on Computational Geometry*, SCG '08, page 39–48, New York, NY, USA, 2008. Association for Computing Machinery.

14 Kenneth L. Clarkson and Peter W. Shor. Applications of random sampling in computational geometry, ii. *Discrete & Computational Geometry*, 4(5):387–421, 1989.

15 Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

16 Sjoerd Dirksen. Dimensionality Reduction with Subgaussian Matrices: A Unified Theory. *Foundations of Computational Mathematics*, 16(5):1367–1396, Oct 2016.

17 Edelsbrunner, Letscher, and Zomorodian. Topological Persistence and Simplification. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002.

18 Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction*. American Mathematical Society, 2010.

19 Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

20 Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer Science & Business Media, 2013.

21 C. Giraud. *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2014.

22 Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in rn. In Joram Lindenstrauss and Vitali D. Milman, editors, *Geometric Aspects of Functional Analysis*, pages 84–106, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg.

23 Leonidas Guibas, Dmitriy Morozov, and Quentin Mérigot. Witnessed k-distance. *Discrete & Computational Geometry*, 49(1):22–45, Jan 2013.

24 Piotr Indyk, Rajeev Motwani, Prabhakar Raghavan, and Santosh Vempala. Locality-preserving hashing in multidimensional spaces. In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, STOC '97, pages 618–625, New York, NY, USA, 1997. ACM.

25 Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Trans. Algorithms*, 3(3):31–es, 2007.

26 William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(189-206):1, 1984.

27 Martin Lotz. Persistent homology for low-complexity models. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2230):20190081, 2019.

28 Arakaparampil M. Mathai and Serge B. Provost. *Quadratic forms in random variables: theory and applications*. Dekker, 1992.

29 Jiří Matoušek. On variants of the Johnson Lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.

30 Steve Y. Oudot. *Persistence Theory - From Quiver Representations to Data Analysis*, volume 209 of *Mathematical surveys and monographs*. American Mathematical Society, 2015.

31 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.

32 Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, page 143–152, USA, 2006. IEEE Computer Society.

33 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.

34 Nakul Verma. A note on random projections for preserving paths on a manifold. Technical report, UC San Diego, 2011.

35 Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

36 Ji Zhang. Advancements of outlier detection: A survey. *EAI Endorsed Trans. Scalable Information Systems*, 1(1):e2, 2013.