The Persistent Homology of Distance Functions under Random Projection

Donald R. Sheehy
University of Connecticut
don.r.sheehy@gmail.com

Abstract

Given n points P in a Euclidean space, the Johnson-Linden-strauss lemma guarantees that the distances between pairs of points is preserved up to a small constant factor with high probability by random projection into $O(\log n)$ dimensions. In this paper, we show that the persistent homology of the distance function to P is also preserved up to a comparable constant factor. One could never hope to preserve the distance function to P pointwise, but we show that it is preserved sufficiently at the critical points of the distance function to guarantee similar persistent homology. We prove these results in the more general setting of weighted kth nearest neighbor distances, for which $k = 1$ and all weights equal to zero gives the usual distance to P.

1 Introduction

What can we say about the topology of a space from just a finite sample, and when can we have any confidence in our answers? These are the questions driving the growing field of topological inference and topological data analysis [11, 10]. Given a set of n points P assumed to be sampled on or near some underlying set $S \subset \mathbb{R}^d$, we want to provide some information about the topology of S.

The most natural means of endowing P with some interesting topology is to look at the sublevel sets of the distance function d_P, which measures the minimum distance from any point x to a point p in P. The α-sublevel set of d_P is the union of balls of radius α centered at the points of P. Niyogi, Smale, and Weinberger [34] showed that when S is a smooth manifold embedded in $\mathbb{R}^D$ and P is a sufficiently dense sample, the homology of S can be inferred from the homology of a sublevel set of d_P. Their method assumes some knowledge of the "right" sampling density required and this is necessary as it is possible for a set to exhibit different topology at different sampling resolutions. Persistent homology gives a truly multiscale view of the topology of S; it describes the changes in topology at different scales. The persistence algorithm of Edelsbrunner, Letscher, and Zomorodian [21] was explicitly developed to track the changes in homology of the sublevel sets of d_P. Since then, many other uses of persistent homology have emphasized the distance to a point set as the primary object of study.

However, computing the persistent homology of the distance to a point set becomes harder as the dimension increases because the complexity of the discrete representation of the sublevel sets can be exponential in the ambient dimension. Thus, it would be nice to reduce the dimension of the input set while approximately preserving the persistent homology. The Johnson-Lindenstrauss Lemma implies that a random, scaled projection of P into $O(\log n/\varepsilon^2)$ dimensions preserves the distance between pairs of points up to a distortion of $1 \pm \varepsilon$ with high probability, but no projection can preserve the distance function over the whole space. Despite this, we show that the distance function is approximately preserved at the critical points of the distance function and thus the persistent homology is also approximately preserved by the random projection. This was known only up to a constant factor of about $\sqrt{2} + \varepsilon$ (see section 2 for a definition of approximate persistent homology and a summary of this result). In this paper, we show that any linear transformation that preserves distances and inner products like those in the Johnson-Lindenstrauss Lemma, also preserves the persistent homology of d_P up to a factor of $1 + O(\varepsilon)$. We prove this result more generally to show that it also applies to weighted distance functions and weighted kth nearest neighbor distances, which have been used in geometric inference to add robustness to outliers [38, 9].

This paper may be viewed as contributing to two previously disjoint lines of research. The first concerns the search for more general geometric properties that are preserved under random projection. For example the Sarlos's work on projection of affine spaces [37]; Agarwal, Har-Peled, and Yu's work on projection of curves, surfaces, and moving points [2]; Baraniuk and Wakin [5] and later Clarkson's [17] work on projections of manifolds; and Magen's work on projection of volumes [30, 31]. The second concerns the search for more efficient computation of the persistent homology of the distance to a point cloud. This line of research has also advanced on many different fronts. Sheehy [39] and Dey, Fan, and Wang [20] looked at sparse constructions of Vietoris-Rips filtrations which give constant factor approximations to the persistent homology of a distance function; Oudot and Sheehy [36] used the theory of zigzag persistence to achieve a similar sparsification with strong guarantees on noise removal. Recently, Kerber and Sharathkumar [28] employed coresets for minimum enclosing balls to reduce the number of input points required. Lamar and Letscher use random projection in a different way to reconstruct low-dimensional skeletons of high-dimensional Delaunay triangulations for use in computing $\mathrm{Pers}(\mathrm{d}_P)$ exactly up to fixed dimensional homology groups [29]. These methods focus on shrinking the input to the persistence algorithm. Another line of work attempts to speed up the computation of the algorithm directly using discrete Morse theory or related reductions as in the work of Mischaikow and Nanda [33] and Bauer, Kerber and Reininghaus [6].

Dealing with noise and outliers is another major challenge in topological data analysis. This challenge has spawned several different research directions including statistical approaches ([35, 7, 4]) and the use of alternative distance functions that are less sensitive to outliers (see for example the work of Chazal, Cohen-Steiner, and Mérigot [13], Guibas, Mérigot, and Morozov [23], and Buchet et al. [9]). To accommodate approximations to more general distance functions, we prove our results about random projections for weighted kth nearest neighbor distances. That is, the distance at any point is the kth smallest weighted distance to a point of P, where each point has a nonnegative weight that increases its distance to the other points. Recently, Buchet et al. [9] showed that weighted distances with $k = 1$ suffice to give a good approximation to the so-called distance to the empirical measure of P. Choosing k larger than 1 allows the distance function to ignore up to $k-1$ outliers locally [38]. Setting all weights to zero and $k = 1$ gives the usual distance to P.

2 Background

Distance Functions We will deal with several different distance functions induced by the input point set P. We assume a nonnegative weight $w(p)$ for each point $p \in P$. The power distance from a point x to a weighted point p is defined as

$$\pi_p(x) := \sqrt{\|x-p\|^2 + w(p)^2},$$

where $\|\cdot\|$ denotes the Euclidean norm. Note that this differs by a sign change from the notion of power distance used in weighted Delaunay triangulations, but this version makes more sense for geometric inference (see [9]). Intuitively a point moves away from the rest of the space as its weight increases.

The *distance to the set P of weighted points* is defined as

$$\mathrm{d}_P(x) := \min_{p \in P} \pi_p(x).$$

Similarly, the *weighted kth nearest neighbor distance* is defined as

$$\mathrm{d}_P^k(x) := \min_{S \in \binom{P}{k}} \max_{p \in S} \pi_p(x),$$

where $\binom{P}{k}$ denotes the set of subsets of P of size k.

Minimum enclosing ball The *minimum enclosing ball* of a point set P is the closed ball B of minimum radius that contains all of P. We adapt this definition to be meaningful also for a weighted point set P. The center of the minimum enclosing ball of $P \subset \mathbb{R}^D$ is

$$\mathrm{center}(P) := \operatorname*{argmin}_{x \in \mathbb{R}^D} \max_{p \in P} \pi_p(x).$$

The *radius of the minimum enclosing ball* of P then is

$$\mathrm{rad}(P) = \max_{p \in P} \pi_p(\mathrm{center}(P)).$$

When the weights are all zero, this definition matches that of the standard case of unweighted points. We will make use of the fact that $\mathrm{center}(S) \in \mathrm{conv}(S)$, where $\mathrm{conv}(S)$ denotes the convex closure of S (see Fischer and Gärtner [22] for a proof of this fact in a more general form). Note that $\mathrm{d}_P^1 = \mathrm{d}_P$, so we will work only with this more general class of functions d_P^k.

Minimum enclosing balls are especially relevant to our problem because the critical points of the distance function are the centers of minimum enclosing balls of subsets of the input points [15, 12, 11].

Filtrations and Persistent Homology A *filtration* $\{F_\alpha\}_{\alpha \geq 0}$ is a sequence of topological spaces such that $F_\alpha \subseteq F_\beta$ whenever $\alpha \leq \beta$. In this paper, all filtrations are parameterized by a real number. The two types of filtrations we consider are *sublevel filtrations* of distance functions and *filtered simplicial complexes*, yielding, respectively, continuous and discrete representations. The sublevel filtration of a distance function d_P^k is $\{F_\alpha\}_{\alpha \geq 0}$, where

$$F_\alpha = (\mathrm{d}_P^k)^{-1}[0, \alpha].$$

From here on, we omit the index set in our notation for filtrations, writing $\{F_\alpha\}$ instead of $\{F_\alpha\}_{\alpha\geq 0}$.

For any set S, an *abstract simplicial complex* X is a family of subsets of S that is closed under taking subsets, i.e. if $\sigma \in X$ and $\tau \subset \sigma$ then $\tau \in X$. The sets in X are called *simplices* and the elements of S are called *vertices*. Given a subset $V \subseteq S$, the *induced subcomplex* on V is the set of simplices $\sigma \in X$ such that $\sigma \subseteq V$. A *filtered simplicial complex* is a filtration $\{X_\alpha\}$ where each space X_α is a simplicial complex. We say that the *birth time* of a simplex σ in a filtration is the minimum α such that $\sigma \in X_\alpha$.

The *barycentric decomposition* of a simplicial complex X, denoted $\mathrm{bary}(X)$, has one vertex for each simplex in X and a simplex for every subset of simplices in X that are totally ordered by inclusion, i.e. $\{\sigma_1, \ldots, \sigma_r\} \in \mathrm{bary}(X)$ if and only if $\sigma_1 \subset \cdots \subset \sigma_r \in X$. For a filtered complex $\{X_\alpha\}$, $\{\mathrm{bary}(X_\alpha)\}$ is also a filtration, the birth time of a vertex σ in $\mathrm{bary}(X_\alpha)$ is the birth time of σ in $\{X_\alpha\}$ and the birth time of a simplex in $\mathrm{bary}(X_\alpha)$ is the maximum of the birth times of its vertices. Given a positive integer k, the *k-barycentric decomposition* of X, denoted k-$\mathrm{bary}(X)$ is the induced subcomplex of $\mathrm{bary}(X)$ on the simplices $\sigma \in X$ with $|\sigma| \geq k$. That is, to form k-$\mathrm{bary}(X)$, we remove vertices from $\mathrm{bary}(X)$ corresponding to simplices in X with fewer than k. In particular, k-$\mathrm{bary}(X) = \mathrm{bary}(X)$ when $k = 1$.

Persistent Homology gives a description of the changes in the topology of the spaces in filtration as α grows. The output is represented as a multiset of pairs $(\alpha_{\mathrm{birth}}, \alpha_{\mathrm{death}})$, where each pair describes the lifespan of a topological feature in a filtration $\mathcal{F} = \{F_\alpha\}$. This set of pairs is denoted $\mathrm{Pers}(\mathcal{F})$. We overload this notation and use $\mathrm{Pers}(\mathrm{d}_P^k)$ to denote the persistence of the sublevel filtration of the function d_P^k. The persistence algorithm of Edelsbrunner et al. [21] takes a filtered simplicial complex $\mathcal{F}$ as input and outputs the multiset of pairs $\mathrm{Pers}(\mathcal{F})$. Since the elements of $\mathrm{Pers}(\mathcal{F})$ are pairs of real numbers, they can be drawn in the plane as a set of points called a *persistence diagram* or as a set of intervals called a *persistence barcode*.

Approximate Persistent Homology For filtrations $\mathcal{F}$ and $\mathcal{G}$ and any constant $c \in \mathbb{R}$, a *c-matching* between $\mathrm{Pers}(\mathcal{F})$ and $\mathrm{Pers}(\mathcal{G})$ is an undirected, partial matching between the sets of pairs such that

1. every pair $(\alpha_{\mathrm{birth}}, \alpha_{\mathrm{death}})$ with $\alpha_{\mathrm{death}}/\alpha_{\mathrm{birth}} > c$ is matched, and
2. if $(\alpha_{\mathrm{birth}}, \alpha_{\mathrm{death}})$ is matched to $(\beta_{\mathrm{birth}}, \beta_{\mathrm{death}})$ then
   $\alpha_{\mathrm{birth}}/\beta_{\mathrm{birth}} \leq c$ and $\alpha_{\mathrm{death}}/\beta_{\mathrm{death}} \leq c$ (and vice versa).

We say that $\mathrm{Pers}(\mathcal{F})$ is a *c-approximation* to $\mathrm{Pers}(\mathcal{G})$ if there exists a c-matching between them.[1]

Using results on the stability of persistence [14], the easiest way to prove that $\mathrm{Pers}(\mathcal{F})$ is a c-approximation to $\mathrm{Pers}(\mathcal{G})$ is to show a *c-interleaving*, i.e. for all $\alpha \geq 0$,

$$F_{\alpha/c} \subseteq G_\alpha \subseteq F_{c\alpha}.$$

This is the main tool that we use throughout this paper.

[1] The notion of c-approximation presented here is related to the so-called L_∞-bottleneck distance [18], denoted d_B, when writing the persistence pairs on the log-scale. If $\mathrm{Pers}(\mathcal{F})$ is a c-approximation to $\mathrm{Pers}(\mathcal{G})$ then $\mathrm{d}_B(\log(\mathrm{Pers}(\mathcal{F})), \log(\mathrm{Pers}(\mathcal{F}))) \leq \log c$, where $\log(\mathrm{Pers}(\mathcal{F}))$ denotes the set $\{(\log b, \log d) \mid (b,d) \in \mathrm{Pers}(\mathcal{F})\}$. We prefer the above definition as it obviates the need to augment persistence diagrams with infinite multiplicity on the diagonals as is required in the definition of the bottleneck distance.

Čech Complexes and Filtrations For a fixed parameter α, the α-Čech complex $\mathcal{C}_\alpha$ is the set of simplices σ such that $\mathrm{rad}(\sigma) \leq \alpha$. Equivalently,

$$\sigma \in \mathcal{C}_\alpha \quad \text{if and only if} \quad \bigcap_{p \in \sigma} \mathrm{ball}(p, \sqrt{\alpha^2 - w(p)^2}) \neq \emptyset.$$

A complex formed in this way is called the *nerve* of the cover of the union of balls. The Nerve Theorem [25, Cor. 4G.3] implies that $\mathcal{C}_\alpha$ is homotopy equivalent to the α-sublevel sets of d_P.

The sequence of Čech complexes $\{\mathcal{C}_\alpha\}$ for all $\alpha \geq 0$ is called the *Čech filtration*. The Persistent Nerve Lemma of Chazal and Oudot [16] implies that the Čech filtration has the same persistent homology as the sublevel filtration of the distance function to P, i.e. $\mathrm{Pers}(\{\mathcal{C}_\alpha\}) = \mathrm{Pers}(\mathrm{d}_P)$. In recent work, we showed that this can be extended to kth nearest neighbor distances by replacing the Čech filtration with the k-barycentric decomposition of the Čech filtration [38], i.e. $\mathrm{Pers}(\{k\text{-bary}(\mathcal{C}_\alpha)\}) = \mathrm{Pers}(\mathrm{d}_P^k)$. Note that both of these results are statements about nerves of good covers and therefore extend to weighted distances.

Random Projection For high dimensional input sets $P \subset \mathbb{R}^D$, projecting P into a random linear subspace of lower dimension (followed by a rescaling) is a popular method for reducing the dimension while preserving some underlying structure. Following the definitions from Clarkson [17], we say that a *d-map* from $\mathbb{R}^D$ to $\mathbb{R}^d$ is an orthogonal projection onto a random linear subspace of dimension d scaled by a factor of $\sqrt{D/d}$. Given a set of vectors V and a function f, we say f *ε-preserves squared lengths* of V if for every $v \in V$,

$$(1-\varepsilon)\|v\|^2 \leq \|f(v)\|^2 \leq (1+\varepsilon)\|v\|^2$$

The Johnson-Lindenstrauss Lemma [26] says that with high probability a d-map preserves squared lengths of n vectors V for $d = \Theta(\log n/\varepsilon^2)$ dimensions.

Lemma 1 (JL Lemma [26])**.** *There is a constant c such that for a given set $U \subset \mathbb{R}^D$ of n vectors and $\varepsilon, \delta > 0$, with probability at least $1-\delta$, a $(c\log(n/\delta)/\varepsilon^2)$-map ε-preserves squared lengths of U.*

Random projection can also be shown to preserve inner products (angles) up to an additive error. For example, let U be a set of unit vectors and let f be a linear map that ε-preserves squared lengths of $\bigcup_{u,v \in U}\{u+v, u-v\}$, then for all $u, v \in U$,

$$\left| f(u)^\top f(v) - u^\top v \right| \leq \varepsilon. \tag{1}$$

One direction of this bound follows from the observation that

$$\begin{aligned}
f(u)^\top f(v) &= \frac{1}{4}\left(\|f(u)+f(v)\|^2 - \|f(u)-f(v)\|^2\right) \\
&\leq \frac{1}{4}\left((1+\varepsilon)\|u+v\|^2 - (1-\varepsilon)\|u-v\|^2\right) \\
&= \frac{1}{4}\left(4u^\top v + \varepsilon(\|u\|^2 + \|v\|^2)\right) \\
&= u^\top v + \varepsilon.
\end{aligned}$$

A similar argument shows that $f(u)^\top f(v) \geq u^\top v - \varepsilon$.

When dealing with points, we will use the following definition.

Definition 2. *An* ε-JL projection *of a point set* $P \subseteq \mathbb{R}^D$ *is a linear map* $f : \mathbb{R}^D \to \mathbb{R}^d$ *such that for all* u, v, *and* w *in* P,

$$|(v-u)^\top (w-u) - (f(v)-f(u))^\top (f(w)-f(u))| \le \varepsilon \|v-u\| \|w-u\|.$$

In particular, if $w = v$ *then*

$$(1-\varepsilon)\|u-v\|^2 \le \|f(u)-f(v)\|^2 \le (1+\varepsilon)\|u-v\|^2.$$

Corollary 3. *There is a constant* c *such that for a given set* $P \subset \mathbb{R}^D$ *of* n *points and* $\varepsilon, \delta > 0$, *with probability at least* $1-\delta$, *a* $(c\log(n^3/\delta)/\varepsilon^2)$*-map is an* ε-JL *projection for* P.

Proof. Consider the set of vectors V defined as

$$V = \bigcup_{\substack{u,v,w \in P \\ u \neq w, v \neq w}} \left\{ \frac{v-u}{\|v-u\|} + \frac{w-u}{\|w-u\|}, \frac{v-u}{\|v-u\|} - \frac{w-u}{\|w-u\|} \right\}.$$

Lemma 1 says that with probability at least $1-\delta$, a d-map ε-preserves squared lengths of V, where $d = (c\log(n^3/\delta)/\varepsilon^2)$. Let f be such a map. Using the observation in (1) about inner products, it follows that

$$\left| \left(\frac{v-u}{\|v-u\|}\right)^\top \left(\frac{w-u}{\|w-u\|}\right) - f\left(\frac{v-u}{\|v-u\|}\right)^\top f\left(\frac{w-u}{\|w-u\|}\right) \right| \le \varepsilon \tag{2}$$

It follows from the linearity of f and (2) that f is an ε-JL projection. □

Rather than stating all our results as probabilistic statements with high probability, we will assume that an ε-JL *projection* is given. This way, we do not constrain the method of producing the projection, whether it be sparse [27], fast [3], database-friendly [1], or otherwise. Corollary 3 implies that $d = O(\log n/\varepsilon^2)$ suffices for any P, though it may be that projection into even lower dimensions still gives an ε-JL projection as is the case with subsets of affine subspaces [37] or smooth submanifolds [17].

Throughout, when we speak of projecting a set of weighted points, we assume that the weights of the points are unchanged by the projection. That is, $w(f(p)) = w(p)$ for any projection f. There are useful methods that modify weights of points when projecting to a low dimensional subspace to help correct for the inevitable error (see for example the work of Boissonat and Ghosh [8]), however, we do not consider such methods in this work.

Three Results Using existing theory, there are some easy to prove facts about persistent homology under random projection. They are listed here as a partial map of the limits of what can be shown by direct application of known results.

- First, if one is willing to settle for the persistence diagram of the so-called (Vietoris-)Rips filtration rather than the distance function, random projection clearly gives a good approximation, because the Rips filtration only depends on pairwise distances. Thus, one can easily get a $1/(1-\varepsilon)$-interleaving of the Rips filtrations of P and $f(P)$ after identifying their vertex sets.

- Second, the Rips filtration is known to give a $\sqrt{2}$-approximation for the persistence diagram of the distance function, and this can be combined with the previous statement about Rips filtrations to guarantee that the persistent homology of the distance function is $(\sqrt{2}/(1-\varepsilon))$-preserved under random projection. If the dimension is small, a result of de Silva and Ghrist [19] on the interleaving of Čech and Rips filtrations gives a slightly stronger bound of $\sqrt{2d/(d+1)}/(1-\varepsilon)$.

- Third, if one is willing to use more dimensions, the r-skeleton of the Čech filtration is preserved under random projection into $O(r \log n/\varepsilon^2)$ dimensions. A lemma of Agarwal, Har-Peled, and Yu shows that the radius of the minimum enclosing ball of a point set is preserved up to $1 \pm \varepsilon$ if you are allowed to add the center to the set [2]. There are $n^{O(r)}$ simplices in the r-skeleton of the Čech filtration and each has a birth time equal to the radius of the minimum enclosing ball of its vertices. By including all $n^{O(r)}$ centers of these minimum enclosing balls in the point set, an ε-JL projection f into $O(r \log n/\varepsilon^2)$ dimensions will preserve the radii of all subsets of size at most $r+1$. Thus, there will be a $1/(1-\varepsilon)$-interleaving of the r-skeletons of the Čech filtrations of P and $f(P)$ after identifying their vertex sets.

The main objective of this paper is to improve on the latter two results by showing that the persistent homology in every dimension is $(1+\varepsilon)$-preserved under random projection into $O(\log n)$ dimensions.

3 The Main Results

Our main results about the persistent homology of distance functions will proceed by showing an interleaving between the Čech filtrations of the points before and after the projection. This will also extend to an interleaving between the barycentric decompositions of these two Čech filtrations. Since the birth time of a simplex $\sigma \subset P$ in a Čech filtration is equal to $\mathrm{rad}(\sigma)$, we first show in Section 3.1 that the radius of every subset of P is approximately preserved by random projection. Then we prove the main theorems about persistent homology in Section 3.2.

3.1 Minimum Enclosing Balls under Random Projection

Lemma 4. *Let S be a set of weighted points and let $f : \mathbb{R}^D \to \mathbb{R}^d$ be an ε-JL projection of S. If $x = \mathrm{conv}(S)$ and p is any point in S then*

$$\left| \|x-p\|^2 - \|f(x)-f(p)\|^2 \right| \leq 4\varepsilon \, \mathrm{rad}(S)^2$$

Proof. Fix any $p \in S$. Label the points of S as $p_1, \ldots, p_r$ such that $p = p_1$. Since $x \in \mathrm{conv}(S)$, we can write x as an affine combination of the points of S as follows.

$$x = \sum_{i=1}^{r} \lambda_i p_i, \text{ where } \sum_{i=1}^{r} \lambda_i = 1.$$

It follows that

$$\|x-p\|^2 = \left\| \sum_{i=1}^{r} \lambda_i (p_i - p) \right\|^2 = \sum_{i=1}^{r} \sum_{j=1}^{r} \lambda_i \lambda_j (p_i - p)^\top (p_j - p). \tag{3}$$

By the linearity of f, the projection of x is

$$f(x) = \sum_{i=1}^{r} \lambda_i f(p_i).$$

So, by the same derivation as in (3), we get that

$$\|f(x) - f(p)\|^2 = \sum_{i=1}^{r} \sum_{j=1}^{r} \lambda_i \lambda_j (f(p_i) - f(p))^\top (f(p_j) - f(p)).$$

Since f is an ε-JL projection, for all i and j,

$$|(p_i - p)^\top (p_j - p) - (f(p_i) - f(p))^\top (f(p_j) - f(p))| < \varepsilon \|p_i - p\| \|p_j - p\|.$$

Let y denote center(S). By the triangle inequality, for all $p_i \in S$,

$$\|p_i - p\| \leq \|p_i - y\| + \|p - y\| \leq \pi_{p_i}(y) + \pi_p(y) \leq 2 \operatorname{rad}(S).$$

It now follows that

$$\begin{aligned}
&\big|\|p - x\|^2 - \|f(p) - f(x)\|^2\big| \\
&= \sum_{i=1}^{r} \sum_{j=1}^{r} \lambda_i \lambda_j \Big|(p_i - p)^\top (p_j - p) - (f(p_i) - f(p))^\top (f(p_j) - f(p))\Big| \\
&\leq \sum_{i=1}^{r} \sum_{j=1}^{r} \lambda_i \lambda_j \varepsilon \|p_i - p\| \|p_j - p\| \\
&\leq \sum_{i=1}^{r} \sum_{j=1}^{r} \lambda_i \lambda_j 4\varepsilon \operatorname{rad}(S)^2 \\
&= 4\varepsilon \operatorname{rad}(S)^2.
\end{aligned}$$

□

Theorem 5. *Let P be a set of weighted points in $\mathbb{R}^D$ and let $f : \mathbb{R}^D \to \mathbb{R}^d$ be an ε-JL projection for P. For every subset S of P,*

$$(1 - 4\varepsilon)\operatorname{rad}(S)^2 \leq \operatorname{rad}(f(S))^2 \leq (1 + 4\varepsilon)\operatorname{rad}(S)^2.$$

Proof. Fix any subset S of P. Let $x = $ center(S). For any $p \in P$, the definition of the radius of the minimum enclosing ball implies that

$$\operatorname{rad}(f(S))^2 \leq \max_{p \in P} \pi_{f(p)}(f(x))^2 = \max_{p \in P}(\|f(x) - f(p)\|^2 + w(p)^2).$$

Applying Lemma 4 and the observation that $\pi_p(x) \leq \operatorname{rad}(S)$ for all $p \in S$, we get the following.

$$\begin{aligned}
\operatorname{rad}(f(S))^2 &\leq \max_{p \in P}(\|x - p\|^2 + 4\varepsilon \operatorname{rad}(S)^2 + w(p)^2) \\
&= \max_{p \in P}(\pi_p(x)^2 + 4\varepsilon \operatorname{rad}(S)^2) \\
&\leq \max_{p \in P}((1 + 4\varepsilon)\operatorname{rad}(S)^2) \\
&= (1 + 4\varepsilon)\operatorname{rad}(S)^2.
\end{aligned}$$

Now, we need to prove the lower bound on $\mathrm{rad}(f(S))$. If T is the subset of points p in S such that $\pi_p(x) = \mathrm{rad}(S)$, then $x \in \mathrm{conv}(T)$ and $f(x) \in \mathrm{conv}(f(T))$. Similarly, $\mathrm{center}(f(S)) \in \mathrm{conv}(f(T))$. The perpendicular bisecting hyperplane between $f(x)$ and $\mathrm{center}(f(S))$ must have points of $f(T)$ on both sides. Thus, for some point q in T, $f(q)$ is closer to $f(x)$ than $\mathrm{center}(f(S))$. So, by applying this fact and Lemma 4, we derive the following bound.

$$\begin{aligned}
\mathrm{rad}(f(S))^2 &\geq \pi_{f(q)}(\mathrm{center}(f(S)))^2 \\
&= \|f(q) - \mathrm{center}(f(S))\|^2 + w(q)^2 \\
&\geq \|f(q) - f(x)\|^2 + w(q)^2 \\
&\geq \|q - x\|^2 + w(q)^2 - 4\varepsilon\,\mathrm{rad}(S)^2 \\
&= \pi_q(x)^2 - 4\varepsilon\,\mathrm{rad}(S)^2 \\
&= (1 - 4\varepsilon)\mathrm{rad}(S)^2.
\end{aligned}$$

□

3.2 Persistent Homology under Random Projection

We are now ready to prove the main theorem of this paper, which guarantees that the persistent homology of weighted kth nearest neighbor distance functions are all preserved up to a $1+\varepsilon$ factor by an ε-JL projection of P.

Theorem 6. *Let $P \in \mathbb{R}^D$ be a set of weighted points, let k be a positive integer, and let $f : \mathbb{R}^D \to \mathbb{R}^d$ be an ε-JL projection for P. Then $\mathrm{Pers}(\mathrm{d}^k_{f(P)})$ is a $1/\sqrt{1-4\varepsilon}$-approximation to $\mathrm{Pers}(\mathrm{d}^k_P)$.*

Proof. Let $\{\mathcal{C}_\alpha\}$ and $\{\mathcal{C}'_\alpha\}$ denote the Čech filtrations of P and $f(P)$ respectively, both realized with P as a vertex set. That is, a subset $\sigma \subseteq P$ is in $\mathcal{C}_\alpha$ if $\mathrm{rad}(\sigma) \leq \alpha$ and $\sigma \in \mathcal{C}'_\alpha$ if $\mathrm{rad}(f(\sigma)) \leq \alpha$. By Theorem 10 of [38],

$$\mathrm{Pers}(\{k\text{-bary}(\mathcal{C}_\alpha)\}) = \mathrm{Pers}(\mathrm{d}^k_P),$$

and similarly,

$$\mathrm{Pers}(\{k\text{-bary}(\mathcal{C}'_\alpha)\}) = \mathrm{Pers}(\mathrm{d}^k_{f(P)}).$$

So, it will suffice to prove that $\{k\text{-bary}(\mathcal{C}_\alpha)\}$ and $\{k\text{-bary}(\mathcal{C}'_\alpha)\}$ are $(1+\varepsilon)$-interleaved. These filtrations have a simplex for every nested sequence $\sigma_1 \subset \cdots \subset \sigma_r$ of subsets of points of P where $|\sigma_1| \geq k$. The birth time of such a simplex is $\mathrm{rad}(\sigma_r)$. By Theorem 5, for every $\sigma \in \mathcal{C}_\alpha$,

$$\sqrt{1-4\varepsilon}\,\mathrm{rad}(\sigma) \leq \mathrm{rad}(f(\sigma)) \leq \sqrt{1+4\varepsilon}\,\mathrm{rad}(\sigma).$$

It follows that for all $\alpha \geq 0$,

$$k\text{-bary}(\mathcal{C}_{\sqrt{1-4\varepsilon}\,\alpha}) \subseteq k\text{-bary}(\mathcal{C}'_\alpha) \subseteq k\text{-bary}(\mathcal{C}_{\sqrt{1+4\varepsilon}\,\alpha}).$$

This interleaving implies desired approximation of guarantee. □

Remark 7. *If one is only interested in the case of $k = 1$, it is not necessary to pass through the barycentric decomposition. In that case, the proof is identical while reasoning directly about the Čech filtration rather than its barycentric decomposition.*

4 Discussion

We have shown that the persistent homology of the distance, the weighted distance, and the weighted kth nearest neighbor distance to a point set P are $(1 \pm \varepsilon)$-preserved under random projection into $O(\log n/\varepsilon^2)$ dimensions. The key idea is to show that the squared radius of the minimum enclosing ball of every subset of P is preserved up to a factor of $1 \pm 4\varepsilon$ by the projection. Similar to classic theorems on random projection, these results hold regardless of the input dimension. This may seem surprising given that the homology of the distance function is trivial in dimensions above the ambient dimension. This implies that for $r = \omega(\log n/\varepsilon^2)$, any pairs (α, β) in the r-dimensional persistent homology for the distance to n points must have $\beta/\alpha < 1 + \varepsilon$. That is, the persistence is negligible in higher dimensions.

Projecting more aggressively This work has assumed the usual context of random projection in that the target dimension is $O(\log n/\varepsilon^2)$ so that pairwise distances between points are preserved. There are several instances for which it is known that projection into even fewer dimensions can similarly preserve pairwise distances. For example, if P is sampled from a r-dimensional affine subspace, then Sarlos showed that projection into $O(r/\varepsilon^2)$ dimensions suffice. Similar results hold for projections of manifolds by Baraniuk and Wakin [5] and later by Clarkson [17]. The Clarkson result gives a target dimension that only depends on the underlying manifold. As shown by Verma [40], if one is only interested in preserving geodesics on the manifold, there is a simple description of this target dimension in terms of a covering number of the manifold.

All of these examples highlight a slight friction in the perspectives of geometric inference and metric embeddings, which might be called the difference between the finite sample view and the finite metric view. In geometric inference and its various forms in unsupervised learning (of which manifold learning is but one), one assumes there is an underling object of fixed complexity and having more samples, makes the inference problem correspondingly easier. For many metric problems where random projection has been found useful, such as in approximate nearest neighbor searching [24], the hardness of instances is generally assumed to grow as the number of points increases.

Open Problems One notable distance function that is absent from this paper is the so-called distance to a measure [13] or, in the finite sample case, the k-distance [23, 32]. Although recent results on approximating the persistence diagram of the k-distance using weighted distance functions makes our new random projection results applicable to k-distances [9], it remains open whether the k-distance itself is $(1 \pm \varepsilon)$-preserved under random projection. A possible first step towards proving this would be to observe that the so-called k-means energy E of every k-tuple S of points is preserved, where this energy is defined as the average squared distance from the points to their centroid. This is a corollary of the following identity, where x is the centroid of S.

$$E = \frac{1}{n} \sum_{p \in S} \|p - x\|^2 = \frac{1}{2n^2} \sum_{p,q \in S} \|p - q\|^2.$$

The k-distance is a weighted distance on the set of centroids of k-tuples of input points with weights defined as the defined as the k-means energy.

It also remains open whether the constant factors in Theorem 6 are tight. Some preliminary work indicates that the factor of 4 coming from Lemma 4 may not be necessary. Requiring distortion

$1 \pm \varepsilon/4$ rather than just $1 \pm \varepsilon$ demands a target dimension that is 16 times larger. This is a constant, but it is a significant one. Stated in terms of minimum enclosing balls, this asks if an ε-JL projection preserves $\mathrm{rad}(S)^2$ up to $1 \pm \varepsilon$ for every subset S of P. This would be a strengthening of Theorem 5 that would propagate improved constants through all of the later results.

Lastly, the linearity of the projection is sufficient, but is not necessary. We conjecture that any map that preserves pairwise distances up to $1 \pm \varepsilon$ should also preserve the radii of of every subset up to $1 \pm \varepsilon$ as well. We intend to address this more general question in future work.

Acknowledgements

I would like to thank David Letscher and Alex Russell for highly motivating conversations.

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003. 2

[2] Pankaj K. Agarwal, Sariel Har-Peled, and Hai Yu. Embeddings of surfaces, curves, and moving points in euclidean space. In *Proceedings of the 24th ACM Symposium on Computational Geometry*, pages 381–389, 2008. 1, 2

[3] Nir Ailon and Bernard Chazelle. Faster dimension reduction. *Commun. ACM*, 53(2):97–104, 2010. 2

[4] Sivaraman Balakrishnan, Alessandro Rinaldo, Don Sheehy, Aarti Singh, and Larry A. Wasserman. Minimax rates for homology inference. *Journal of Machine Learning Research - Proceedings Track*, 22:64–72, 2012. 1

[5] Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009. 1, 4

[6] Ulrich Bauer, Michael Kerber, and Jan Reininghaus. Clear and compress: Computing persistent homology in chunks. In *TopoInVis*, 2013. 1

[7] Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, and Carlos Rodríguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electron. J. Statist.*, 5:204–237, 2011. 1

[8] Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using Tangential Delaunay Complexes. In *In Proc. 26th Annual Symposium on Computational Geometry*, 2010. 2

[9] Mickaël Buchet, Frédéric Chazal, Steve Y. Oudot, and Donald R. Sheehy. Efficient and robust topological data analysis on metric spaces. *arXiv preprint arXiv:1306.0039*, 2013. 1, 2, 4

[10] Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc.*, 46:255–308, 2009. 1

[11] Frédéric Chazal and Date Cohen-Steiner. Geometric inference. In *Tesselations in the Sciences*. Springer-Verlag, 2013. To appear. 1, 2

[12] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in euclidean space. *Discrete & Computational Geometry*, 41:461–479, 2009. 2

[13] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11:733–751, 2011. 1, 4

[14] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *preprint arXiv:1207.3674*, 2012. 2

[15] Frédéric Chazal and André Lieutier. Weak feature size and persistent homology: Computing homology of solids in R^n from noisy data samples. In *Proc. of the 21st ACM Symposium on Computational Geometry*, pages 255–262, 2005. 2

[16] Frédéric Chazal and Steve Y. Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the 24th ACM Symposium on Computational Geometry*, pages 232–241, 2008. 2

[17] Kenneth L. Clarkson. Tighter bounds on random projection of manifolds. In *Proceedings of the 24th ACM Symposium on Computational Geometry*, 2008. 1, 2, 2, 4

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

[19] Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algorithmic & Geometric Topology*, 7:339–358, 2007. 2

[20] Tamal K. Dey, Fengtao Fan, and Yusu Wang. Computing topological persistence for simplicial maps. *arXiv preprint arXiv:1208.5018*, 2013. 1

[21] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 4(28):511–533, 2002. 1, 2

[22] Kaspar Fischer and Bernd Gärtner. The smallest enclosing ball of balls: Combinatorial structure and algorithms. *Int. J. Comput. Geometry Appl.*, 14(4–5):341–378, October 2004. 2

[23] Leonidas Guibas, Dmitriy Morozov, and Quentin Mérigot. Witnessed k-distance. *Discrete & Computational Geometry*, 49(1):22–45, 2013. 1, 4

[24] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012. 4

[25] Allen Hatcher. *Algebraic Topology.* Cambridge University Press, 2001. 2

[26] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.*, 26:189–206, 1984. 2, 1

[27] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. In *SODA*, pages 1195–1206, 2012. 2

[28] Michael Kerber and R. Sharathkumar. Approximate cech complexes in low and high dimensions. In *ISAAC*, 2013. 1

[29] Michael Lamar and David Letscher. Random projections for calculating topological persistence. *in submission*, 2013. 1

[30] Avner Magen. Dimensionality reductions in $\ell 2$ that preserve volumes and distance to affine spaces. *Discrete & Computational Geometry*, 38(1):139–153, 2007. 1

[31] Avner Magen and Anastasios Zouzias. Near optimal dimensionality reductions that preserve volumes. In *RANDOM*, pages 523–534, 2008. 1

[32] Quentin Mérigot. Lower bounds for k-distance approximation. In *Proceedings of the Twenty-ninth Annual Symposium on Computational Geometry*, SoCG '13, pages 435–440, New York, NY, USA, 2013. ACM. 4

[33] Konstantin Mischaikow and Vidit Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete Computational Geometry*, 50(2):330–353, 2013. 1

[34] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008. 1

[35] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. A topological view of unsupervised learning from noisy data. *SIAM J. Comput.*, 40(3):646–663, 2011. 1

[36] Steve Y. Oudot and Donald R. Sheehy. Zigzag zoology: Rips zigzags for homology inference. In *Proceedings of the 29th annual Symposium on Computational Geometry*, pages 387–396, 2013. 1

[37] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006. 1, 2

[38] Donald R. Sheehy. A multicover nerve for geometric inference. In *CCCG: Canadian Conference in Computational Geometry*, 2012. 1, 2, 3.2

[39] Donald R. Sheehy. Linear-size approximations to the Vietoris-Rips filtration. *Discrete & Computational Geometry*, 49(4):778–796, 2013. 1

[40] Nakul Verma. A note on random projections for preserving paths on a manifold. Technical Report CS2011-0971, UC San Diego, 2011. 4