Dimensionality Reduction for k-Distance Applied to Persistent Homology

Shreya Arya, Jean-Daniel Boissonnat, Kunal Dutta

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

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 first show using the Johnson-Lindenstrauss lemma, that the persistent homology can be preserved up to a $(1 \pm \varepsilon)$ factor while reducing dimensionality to $O(k \log n/\varepsilon^2)$. Our main result shows that the target dimension can be improved to $O(\log n/\varepsilon^2)$ under a reasonable and naturally occuring condition. The proof involves a multi-dimensional variant of the Hanson-Wright inequality for subgaussian quadratic forms and works with subgaussian random matrices in the Johnson-Lindenstrauss mapping, which includes the Gaussian matrices of Indyk-Motwani, the sparse random matrices of Achlioptas and the Ailon-Chazelle fast Johnson-Lindenstrauss transform. To provide evidence that our condition encompasses quite general situations, we show that it is satisfied when the points are independently distributed (i) in $\mathbb{R}^D$ under a subgaussian distribution, or (ii) on a spherical shell in $\mathbb{R}^D$ with a minimum angular separation, using Gershgorin's theorem. Our results also show that the JL-mapping preserves up to a $(1 \pm \varepsilon)$ factor, the Rips and Delaunay filtrations for the k-distance, as well as the Čech filtration for the approximate k-distance of Buchet et al.

2012 ACM Subject Classification Theory of computation → Computational geometry

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

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).

Lines 500

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 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. [6] 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 [18, 4]. Other approaches involve using kernels [26], de-noising algorithms [5], [30], 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. [17]), 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) [20]. 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 [20], several authors have given successive improvements, e.g. Indyk and Motwani [19], Dasgupta and Gupta [11], Achlioptas [1], Ailon and Chazelle [2], Matoušek [24] and others, which address the issues of *efficient* constructivization and implementation. The constructive versions use random matrices for the embedding guaranteed by the JL Lemma, such as random Gaussian or ± 1 matrices, sparse random matrices, variants of the discrete Fourier transform, etc. A unified treatment of Johnson-Lindenstrauss transforms using subgaussian random matrices, (which covers most of the above-mentioned types of random matrices) can be found in Dirksen [13].

Dimension Reduction and Persistent Homology. The JL Lemma has also been used by Sheehy [28] and Lotz [22] 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 [28, 22], 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 [28], who observed that *"One notable distance function that is missing from this paper* [i.e. [28]] *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 16). 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 [28] and Lotz [22]. We first show how the JL transform can be used to $(1 \pm \varepsilon)$-preserve the persistent homology for the k-distance with target dimension $O(k \log n/\varepsilon^2)$ (Theorem 17), and $O(\log n/\varepsilon^2)$ for the approximate k-distance (Theorem 18). However, since the approximate k-distance is itself an $O(1)$ approximation of the k-distance, the persistent homology is actually preserved up to an $O(1)$ factor. A stronger result can be obtained under a reasonable and naturally occuring condition, enabling us to remove the extra k factor in the target dimension and to obtain the optimal dimensionality of the usual JL Lemma (Theorem 20).

This last result is proved using a variant of the Hanson-Wright inequality, which gives a concentration bound for sums of certain weighted correlated subgaussian random variables. Specifically, we show that, if the matrix given by the position coordinates of a set of k points has high *stable rank*, then under random projections the sum of the k squared Euclidean norms of these points, is *more strongly* concentrated than their squared Euclidean norms individually (Lemma 24).

The stable rank of a matrix (see Section 2 for a formal definition and geometric interpretations), first introduced in [27], has since then become popular as a robust version of the classical rank of a matrix, with several applications in numerical analysis, compressed sensing, and related areas [29]. The stable rank is more resistant to small perturbations than the classical rank, which can increase significantly under slight perturbations.

The extra condition in Theorem 20 is satisfied in quite general settings and we show that it holds for points distributed independently (i) in $\mathbb{R}^D$ according to a subgaussian distribution (Theorem 22), and (ii) arbitrary points on a spherical shell which is not necessarily thin, under a separation condition (Theorem 23). Our proofs use bounds on the operator and Frobenius norm of certain “random-like" matrices, and Gershgorin's theorem.

It should be noted that the approach of using dimensionality reduction for the k-distance, is complementary to denoising techniques such as [5] 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 extremely high.

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. We end with some final remarks and open questions in Section 5. All missing proofs are provided in the Appendix.

2 Background

2.1 Subgaussian Distributions; Concentration Inequalities

▶ **Definition 1.** Given a probability space Ω, a zero-mean random variable $X : \Omega \to \mathbb{R}$ is said to be *subgaussian* with *variance proxy* v^2 if its tails satisfy $\mathbb{P}\left[|X| > t\right] \leq 2\exp\left(-\frac{t^2}{2v^2}\right)$ for all $t \geq 0$. The *subgaussian norm* of a random variable X is defined to be the least K, such that $\mathbb{E}\left[\exp\left(X^2/K^2\right)\right] \leq 2$. A $D \times d$ matrix is a *subgaussian random matrix* if its entries are independent subgaussian random variables, having subgaussian norm K, for some $K > 0$.

In other words, a random variable is subgaussian if its tails decay as fast as those of a normal random variable with variance v^2. The variance proxy is closely related to the subgaussian norm and it can be shown that for subgaussian random variables, they differ at most by constant factors (e.g. [29](Proposition 2.5.2)).

Many practically occuring, as well as theoretically studied distributions are subgaussian, e.g. the Gaussian and Bernoulli distributions, as well as distributions having bounded support. Further, mixtures of subgaussian distributions are also subgaussian. Thus, the subgaussian is a very general class of distributions. For more information on subgaussian distributions, equivalent characterizations etc., we refer the reader to [29, Chapter 2.5].

The following multi-dimensional variant of the *Hanson-Wright inequality* [29][Chapter 6.3, see also exercise 6.2.7], gives concentration bounds for certain quadratic forms of subgaussian random variables. More general quadratic forms are discussed in e.g. Latała [21].

▶ **Theorem 2.** *Let X be a $D \times n$ random matrix with entries being independent subgaussian random variables G_{ij} with subgaussian norm K. Let A be an $n \times n$ matrix with entries $a_{ij} \in \mathbb{R}$. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\mathrm{Tr}(XAX^\top) - \mathbb{E}\left[\mathrm{Tr}(XAX^\top)\right]\right| \geq t\right] \leq 2 \cdot \exp\left(-c \cdot \min\left(\frac{t^2}{K^4 D \|A\|_F^2}, \frac{t}{K^2\|A\|}\right)\right).$$

2.2 Random Projections

The Johnson-Lindenstrauss lemma [20] 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 shall stick to the case where the Johnson-Lindenstrauss embedding is carried out via random subgaussian matrices. This case is general enough to include the mappings of e.g. Achlioptas [1], Ailon-Chazelle [2] and Matoušek [24], thus allowing sparse random matrices.

▶ **Lemma 3** (JL Lemma). *Let $0 < \varepsilon < 1$. Given a finite point set $P \subset \mathbb{R}^D$ of size $|p| = n$, 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, satisfies with probability at least $1 - o\left(\frac{1}{n}\right)$:*

$$\forall u, v \in P, \;\; (1-\varepsilon)\|u-v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2,$$

(where u, v are column vectors).

The constructive proofs of the JL Lemma rely essentially on the following tail bound for sums of squares of subgaussian random variables.

▶ **Lemma 4** (JL Tail Bounds). *Given independent subgaussian random variables $X_1, \ldots, X_d$ having subgaussian norm 1, and $\varepsilon \in (0,1)$, there exists a constant $c > 0$ such that the sum $X := \sum_{i=1}^d X_i^2$ satisfies : $\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \varepsilon\,\mathbb{E}[X]\right] \leq 2 \cdot \exp\left(-cd\varepsilon^2\right)$.*

Notation. Throughout this paper, we shall use the term *subgaussian JL mapping* to refer to a linear mapping from $\mathbb{R}^D$ to $\mathbb{R}^d$, given by a subgaussian random matrix scaled by $\sqrt{\frac{D}{d}}$.

2.3 Stable Rank and Operator Norm of a Matrix

▶ **Definition 5.** The *stable rank* of a given $m \times n$ real $m \times n$ matrix A is the ratio of the squares of its Frobenius norm with its operator ℓ_2 norm, i.e. $r_{st}(A) := \frac{\|A\|_F^2}{\|A\|_2^2} = \frac{\sum_i s_i^2}{s_1^2}$, where $s_1 \geq s_2 \geq \ldots \geq 0$ are the singular values of A.

Geometrically, it may be considered as the ratio of the squared mean width of the ellipsoid generated by the columns of M (i.e. the set $E_M := \{x \,:\, \|Mx\|^2 \leq 1\}$), to the square of its largest axis. The stable rank is also related to the notions of Gaussian width and statistical dimension [29, Chapter 6] .

▶ **Definition 6.** For an $m \times n$ real matrix M, its *operator norm* $\|M\|$ is given by $\max_{x:\mathbb{R}^n} \frac{\|Mx\|}{\|x\|}$.

2.4 k-Distance

The distance to a 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 [6] introduced the *distance to a probability measure* which, in the case of a uniform probability on P, is called the *k-distance.*

▶ **Definition 7** (k-distance)**.** For $k \in \{1, ..., n\}$ and $x \in \mathbb{R}^D$, the k-distance of x to P is

$$d_{P,k}(x) = \min_{S_k \in \binom{P}{k}} \sqrt{\frac{1}{k} \sum_{p \in S_k} \|x - p\|^2} = \sqrt{\frac{1}{k} \sum_{p \in \mathrm{NN}_P^k(x)} \|x - p\|^2} \qquad (1)$$

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

It was shown in [3], 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 real number called its weight and denoted by $w(b)$.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 $w(b) = -(1/k) \sum_i \|b - p_i\|^2$. Writing $\hat{B}_{P,k} = \{\hat{b} = (b, w(b)), b \in B_{P,k}\}$, we see that the k-distance is in fact the square root of a power distance [3]

$$d_{P,k}(x) = \min_{\hat{b} \in \hat{B}_{P,k}} \sqrt{D(x, \hat{b})}. \qquad (2)$$

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 of an unweighted point in $\mathbb{R}^D$ and a weighted k-barycenter in P, is always real. Since $d_{P,k}$ is the square root of a 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 $B(b, \sqrt{\alpha^2 + w(p)})$, $b \in B_{P,k}$. 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 $\Omega\left(n^{\lfloor (D+1)/2 \rfloor}\right)$ [10] 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 [4]:

▶ **Definition 8** (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

$$\tilde{d}_{P,k}(x) \quad := \quad \min_{p \in P} \sqrt{D(x, \hat{p})} \qquad (3)$$

where $\hat{p} = (p, w(p))$ with $w(p) = -d_{P,k}^2(p)$, 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 $B(p, \sqrt{\alpha^2 - d_{P,k}^2(p)})$, $p \in P$. 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. (3) and Eq. (2)). Still, $\tilde{d}_{P,k}(x)$ approximates the k-distance [4]:

$$\frac{1}{\sqrt{2}}\, d_{P,k} \leq \tilde{d}_{P,k} \leq \sqrt{3}\, d_{P,k}. \tag{4}$$

2.5 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 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 9** (Čech Complex). The Čech complex $\check{C}_\alpha(P)$ is the set of simplices σ such that $\text{rad}(\sigma) \leq \alpha$, where $\text{rad}(\sigma)$ is the radius of the smallest enclosing ball of σ, i.e.

$$\text{rad}(\sigma) = \min_{x \in \mathbb{R}^D} \max_{p \in \sigma} \|x - p\|.$$

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.

Persistent Homology. 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 [14] 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. [15]). 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 if they are multiplicatively c-interleaved, with c close to 1 (see e.g. [7]).

The Persistent Nerve Lemma [9] 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 10** (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 Equ. (2) 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 $\text{rad}^2(\hat{X}) = \min_{x \in \mathbb{R}^D} \max_{\hat{p}_i \in \hat{X}} D(x, \hat{p}_i)$, and the weighted center or simply the center of $\hat{X}$ as the point, noted $c(\hat{X})$, where the minimum is reached. The following lemma follows from the definition.

▶ **Lemma 11.** *A simplex $\hat{\sigma} \subset \hat{P}$ belongs to $\check{C}_\alpha(\hat{P})$ if and only if* $\text{rad}(\hat{\sigma}) \leq \alpha$.

Our goal is to show that preserving smallest enclosing balls in the weighted scenario under a given mapping, also preserves the persistent homology. Sheehy [28] and Lotz [22], 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 Lotz's lemma [22, Lemma 2.2] for the weighted case which does not assume that the weights stay the same under F. This is Lemma 15, which follows at the end of this section. The following lemmas will be instrumental in proving Lemma 15 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 12.** $c(\hat{X})$ *and* $\text{rad}(\hat{X})$ *are uniquely defined.*

▶ **Lemma 13.** *Let I be the set of indices for which $D(c, \hat{p}_i) = \text{rad}(\hat{X})$ 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 an observation of Lotz [22, Lemma 4.2] gives the following lemma whose proof follows Lotz [22, Lemma 2.2].

▶ **Lemma 14.** $\text{rad}^2(\hat{X}) = \dfrac{1}{2} \sum_{j=1}^m \lambda_j \left(\sum_{i=1}^m \lambda_i \left(\|p_i - p_j\|^2 - 2w_i \right) \right)$.

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 15.** *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* ($\text{rad}(\hat{S})$ *is, as usual, the radius of the minimum enclosing ball of* $\hat{S}$)

$$(1 - \varepsilon)\text{rad}(\hat{S}) \leq \text{rad}(\widehat{F(S)}) \leq (1 + \varepsilon)\text{rad}(\hat{S}),$$

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

3 Results

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

3.1 Dimensionality Reduction with k-distance : a first result

Our first theorem shows that for the points in P, the pointwise k-distance $d_{P,k}$ is preserved by a subgaussian Johnson-Lindenstrauss mapping.

▶ **Theorem 16.** *Given $\varepsilon \in (0,1]$, a subgaussian JL mapping $f\colon \mathbb{R}^D \to \mathbb{R}^d$, where $d = O(\varepsilon^{-2}\log n)$, satisfies with probability at least $1 - o\left(\frac{1}{n}\right)$ and for all points $x \in P$:*

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

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. The following theorem shows that this can always be done using $O(k\log n/\varepsilon^2)$ dimensions.

Let $\hat{B}_{P,k}$ be the set of iso-barycenters of every k-subset of P, weighted as in Section 2.4. Recall from Section 2.5 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 17** (k-distance)**.** *Let P be a set of points in $\mathbb{R}^D$. Let $\hat{\sigma} \subseteq \hat{B}_{P,k}$ be a simplex in the weighted Čech complex $\check{C}_\alpha(\hat{B}_{P,k})$ which is identical to the Čech complex $\check{C}_\alpha(P)$ for the k-distance. Then a subgaussian JL mapping $f : \mathbb{R}^D \to \mathbb{R}^d$, where $d = O\left(\frac{k\log n}{\varepsilon^2}\right)$, satisfies with probability at least $1 - o\left(\frac{1}{n}\right)$, $(1-\varepsilon)\mathrm{rad}^2(\hat{\sigma}) \le \mathrm{rad}^2(\widehat{f(\sigma)}) \le (1+\varepsilon)\mathrm{rad}^2(\hat{\sigma})$.*

For the approximation of the k-distance given by [4], we get an optimal target dimension, as the number of weighted points needed to compute the approximate k-distance, is just n.

▶ **Theorem 18** (Approximate k-distance)**.** *Let P be a set of points in $\mathbb{R}^D$ and let $\hat{P}$ be the associated weighted points introduced in Definition 8 (Equ. 3). Let, in addition, $\hat{\sigma} \subseteq \hat{P}$ be a simplex in the weighted Čech complex $\check{C}_\alpha(\hat{P})$ which is identical to the Čech complex $\check{C}_\alpha(P)$ defined by the approximate k-distance. Then a subgaussian JL mapping $f : \mathbb{R}^D \to \mathbb{R}^d$ satisfies: $(1-\varepsilon)\mathrm{rad}^2(\hat{\sigma}) \le \mathrm{rad}^2(\widehat{f(\sigma)}) \le (1+\varepsilon)\mathrm{rad}^2(\hat{\sigma})$ and $d = O(\log n/\varepsilon^2)$.*

The next corollary follows from Lemma 15 and Theorem 18.

▶ **Corollary 19.** *The persistent homology for the Čech filtrations of P and its image $f(P)$ under a subgaussian JL mapping, using the approximate k-distance, are multiplicatively $(1-\varepsilon)^{-1/2}$-interleaved with probability $1 - o(1)$.*

However, note that the above approximation is with respect to the *approximate* k-distance, which is itself an $O(1)$ approximation of the k-distance (see (4)). To get a $(1-\varepsilon)^{-1}$-interleaving, we therefore need to apply the JL lemma to the exact k-distance, which motivates the following theorem.

3.2 An improved result

For the exact k-distance, the target dimension in Theorem 17 can be reduced to $O(\log n/\varepsilon^2)$ under an extra condition, expressed in terms of the stable rank of certain matrices. Later in Lemma 24 we shall prove a concentration result on sums of certain weighted dependent subgaussian random variables, which will be crucial in proving our main result. We then use the mapping in Theorem 16 and tail bounds proved in Lemma 24 to obtain the following theorem that improves on Theorem 17.

For a set S of k points in P and $b \in B_{P,k}$, define $\mathcal{F}(S,b)$ as the $k \times D$ matrix with rows consisting of the vectors $p - b \in \mathbb{R}^D$, $p \in S$. Let $\mathcal{F}$ denote the family of all possible $\mathcal{F}(S,b)$:

$$\mathcal{F} := \{\mathcal{F}(S,b), S \in \binom{P}{k}, \; b \in B_{P,k}\}. \quad (5)$$

We write $r_{\mathcal{F}}$ for the smallest stable rank of the matrices $\mathcal{F}(S,b)$ over all $S \in \binom{R}{k}$, $b \in B_{P,k}$.

▶ **Theorem 20** (k-distance under large stable rank assumption)**.** *Under the same conditions as in Theorem 17 and the extra condition that $r_{\mathcal{F}} = \Omega(k)$, then, with probability at least $1 - o(1)$, a subgaussian JL mapping $f : \mathbb{R}^D \to \mathbb{R}^d$, where $d = O\left(\frac{\log n}{\varepsilon^2}\right)$, satisfies*

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

For each set of k vectors in $\mathcal{F}$, Lemma 24 guarantees good concentration of the image under a random Gaussian projection, provided the stable rank is high (i.e. $\Omega(k)$). Thus, if all k-sets of vectors in $\mathcal{F}$ have high stable rank (i.e. $r_{\mathcal{F}}$ is high), then the squared k-distance will be highly concentrated for most points, and so the required dimensionality of the image space would be less, and vice versa.

The next corollary follows from Lemma 15 and Theorem 20.

▶ **Corollary 21.** *Under the assumption of Theorem 20, the Čech filtrations of P and its image $f(P)$ under a subgaussian JL mapping f, using the k-distance are multiplicatively $(1-\varepsilon)^{-1/2}$-interleaved with probability $1 - o(1)$.*

3.3 Subgaussian Point Distributions and Point Sets with Angular Separation

We now give examples of cases where the stable rank condition in Theorem 20 is satisfied.

▶ **Theorem 22.** *Let $P = \{p_1, \dots, p_n\} \subset \mathbb{R}^D$ be a set of points distributed randomly, with $p_i \equiv (X_{i1}, \dots, X_{iD}) \in \mathbb{R}^D$, where X_{ij}, $i = 1, \dots, n$, $j = 1, \dots, D$, are independent subgaussian random variables with subgaussian norm K. Assume further that $D > \frac{3k \ln n}{C_2}$. Then, there exist constants $C_1, C_2 > 0$ such that, with high probability and for all $4 \leq k \leq (C_2 D / 3 \ln n)$, the minimum stable rank $r_{\mathcal{F}} \geq \frac{k-4}{16 C_1^2}$, and therefore the conclusion of Theorem 20 holds.*

Further, we show that for uniform point distributions on several commonly occuring geometric surfaces of bounded fatness, such as spheres, and ellipsoids and polytopes of bounded elongation, the conclusion of Theorem 20 follows as a consequence of Theorem 22. The above observations are stated precisely in Corollary 29.

The next theorem shows that $r_{\mathcal{F}}$ is lower bounded for any point set in a spherical shell, which can even have constant thickness, requiring only an angular separation condition.

▶ **Theorem 23.** *Suppose for all $p \in P$, $\frac{1}{2} \leq \|p\|^2 \leq \frac{3}{2}$, and for all $p, q \in P$, $|\langle p, q \rangle| \leq \eta \|p\| \cdot \|q\|$, with $\eta := \frac{1}{k-1}$. Then the minimum stable rank condition satisfies $r_{\mathcal{F}} \geq \frac{k-6}{24}$, and so the conclusion of Theorem 20 holds.*

For sufficiently large D, the above theorem is stronger than having points with a spherical subgaussian distribution, because for $D \gg k^2/\log n$, the concentration inequality for the squared norms of subgaussians [29, Corollary 2.8.3] implies that with high probability, (i)

all points lie in the shell with inner and outer radii $\sqrt{\frac{1}{2}}$ and $\sqrt{\frac{3}{2}}$ respectively, and (ii) the maximum cosine between any pair of n points from this spherical shell is $O\left(\sqrt{\frac{\log n}{D}}\right)$, which is much smaller than $1/k$. The proof is via an application of Gershgorin's theorem [16].

3.4 Tail bounds for the sum of dependent weighted squares of subgaussian random variables

The proof of Theorem 20 is based on a concentration inequality for sums of *dependent* subgaussian random variables. The inequality is an application of a modified version of the Hanson-Wright inequality (recall Section 2.1).

Let G be a random $d \times D$ matrix with row vectors $G_i \in \mathbb{R}^D$ having coordinates independent, subgaussian random variables with mean zero, unit variance, and subgaussian norm at most K. Let V denote a $D \times k$ matrix with column vectors $v_i \in \mathbb{R}^D$, $1 \leq i \leq k$.

▶ **Lemma 24.** *Given G and V as above and $L := \|GV\|_F^2$, then for $d \leq K^2 D$, we have*

$$\mathbb{P}\left[|L - \mathbb{E}\left[L\right]| \geq \varepsilon \, \mathbb{E}\left[L\right]\right] \quad \leq \quad 2 \cdot \exp\left(-c\left(\frac{\varepsilon^2 d \|V\|_F^2}{K^2 \|V\|^2}\right)\right).$$

▶ Remark. The inequality in the above bound should be compared with the bound in Lemma 4. Lemma 24 has an extra factor of $\frac{\|V\|_F^2}{\|V\|^2} =: r_{st}(V)$ in the exponent, (ignoring the K^2 term, which can be assumed to be 1, by replacing X by X/K). Recall that $r_{st}(V)$ is always at least 1 and can be as large as k. When the column vectors of V are orthogonal, V has full stable rank, which gives a much stronger concentration of $\exp\left(-c_1 \varepsilon^2 dk\right)$ for the projected k-distance than the usual Euclidean distance which only gives tail bounds of $\exp\left(-c_2 \varepsilon^2 d\right)$ (where c_1, c_2 are independent of V, D and d). The stable rank degrades smoothly as the column vectors of V move away from being orthogonal, approaching 1 in the case when the column vectors are parallel to each other. In this case, the concentration bounds obtained from Lemma 24 are the same as the bounds obtained from Lemma 4.

4 Proofs

4.1 Proof of Lemma 24

Proof. Oberve that $L = \|GV\|_F^2 = \mathrm{Tr}(GVV^\top G^\top)$. Now we apply Theorem 2, with G as the $d \times D$ random matrix, and $A = VV^\top$ as a $D \times D$ matrix. The following well-known observations can be easily derived (see e.g. Lemma 27 in the Appendix): $\|A\| = \|V\|^2$, and $\|A\|_F^2 \leq \|V\|^2 \cdot \|V\|_F^2$. This gives

$$\begin{aligned}
\mathbb{P}\left[|L - \mathbb{E}\left[L\right]| \geq t\right] \quad &\leq \quad 2 \cdot \exp\left(-c \cdot \min\left(\frac{t^2}{K^4 D \|A\|_F^2}, \frac{t}{K^2 \|A\|}\right)\right) \\
&\leq \quad 2 \cdot \exp\left(-c \cdot \min\left(\frac{t^2}{K^4 D \|A\|_F^2}, \frac{t}{K^2 \|V\|^2}\right)\right).
\end{aligned}$$

Observe that

$$\begin{aligned}
\mathbb{E}\left[L\right] \quad &= \quad \mathbb{E}\left[\mathrm{Tr}(GVV^\top G^\top)\right] = \mathbb{E}\left[\mathrm{Tr}(V^\top G^\top GV)\right] = d \cdot \mathrm{Tr}(V^\top V) \qquad (6) \\
&= \quad d \sum_{r=1}^{k} v_r^\top v_r = \sum_{r=1}^{D} d\|v_r\|^2 = d\|V\|_F^2. \qquad (7)
\end{aligned}$$

where the equality (6) followed from the linearity of expectation. Now taking $t = \varepsilon\,\mathbb{E}\,[L] = \varepsilon d\|V\|_F^2$, we get

$$\begin{aligned}\mathbb{P}\left[\left|L - d\|V\|_F^2\right| \geq \varepsilon d\|V\|_F^2\right] &\leq 2\cdot\exp\left(-c\cdot\min\left(\frac{\varepsilon^2 d^2\|V\|_F^4}{K^4 D\|V\|^2\|V\|_F^2}, \frac{\varepsilon d\|V\|_F^2}{K^2\|V\|^2}\right)\right)\\ &= 2\cdot\exp\left(-c\cdot\min\left(\frac{\varepsilon^2 d^2\|V\|_F^2}{K^4 D\|V\|^2}, \frac{\varepsilon d\|V\|_F^2}{K^2\|V\|^2}\right)\right)\\ &\leq 2\cdot\exp\left(-c\left(\frac{\varepsilon^2 d\|V\|_F^2}{K^2\|V\|^2}\right)\right),\end{aligned}$$

where the last line follows by observing that $d/(K^2D) \leq 1$ for $d \leq K^2 D$. ◀

4.2 Proof of Theorem 16

The proof follows from the observation that the squared k-distance from any point $p \in P$ to the point set P, is a linear 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 linear combinations also get $(1 \pm \varepsilon)$-preserved.

4.3 Proof of Theorems 17–18

Proof of Theorem 17. Let $\hat{\sigma} = \{\hat{b}_1, \hat{b}_2, ..., \hat{b}_m\}$, where $\hat{b}_i$ is the weighted point defined in Section 2.5, 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 $p_{i1}, \ldots, p_{ik} \in P$ are such that $b_i = \frac{1}{k}\sum_{j=1}^k p_{ij}$. Let c_σ be the center of $\hat{\sigma}$. Applying Lemma 14 to $\hat{\sigma}$, we have that

$$\mathrm{rad}^2(\hat{\sigma}) = \frac{1}{2}\sum_{j=1}^m \lambda_j \left(\sum_{i=1}^m \lambda_i \left(\|b_j - b_i\|^2 - 2w(b_i)\right)\right). \tag{8}$$

Observe that $\frac{1}{k}\sum_{l=1}^k (b_j - p_{il})^2 - \frac{1}{k}\sum_{l=1}^k \|b_i - p_{il}\|^2 = \|b_i - b_j\|^2$. Using $w_i = -\frac{1}{k}\sum_{l=1}^k \|b_i - p_{il}\|^2$, the term inside the summation in Eq. (8) becomes

$$\left(\|b_j - b_i\|^2 - w(b_i)\right) - w(b_i) = \frac{1}{k}\sum_{l=1}^k \|b_j - p_{il}\|^2 + \frac{1}{k}\sum_{l=1}^k \|b_i - p_{il}\|^2.$$

Let $f : \mathbb{R}^D \to \mathbb{R}^d$ be a subgaussian JL mapping from $\mathbb{R}^D$ to $\mathbb{R}^d$, where d will be chosen later. The mapping f will preserve the radius $\mathrm{rad}(\hat{\sigma})$ up to an approximation factor of $(1 \pm \varepsilon)$ if f preserves all expressions of the form $\frac{1}{k}\sum_{i=1}^k \|b - p\|^2$, where $p \in P$ and $b \in B_{P,k}$ with the same approximation factor. Further, since f is linear, it preserves barycenters and, if $b_i = \frac{1}{k}\sum_{l=1}^k p_{il}$ as above, we have

$$f(b_i) = f\left(\frac{1}{k}\sum_{l=1}^k p_{il}\right) = \frac{1}{k}\sum_{l=1}^k f(p_{il}) \quad \text{and} \quad w(f(b_i)) = -\frac{1}{k}\sum_{l=1}^k \|f(b_i) - f(p_{il})\|^2.$$

Using the above expression of $\mathrm{rad}(\hat{\sigma})$, and by our choice of f, we get

$$
\begin{aligned}
\mathrm{rad}^2(\hat{\sigma}) &= \frac{1}{2}\sum_{j=1}^{m}\lambda_j\sum_{i=1}^{m}\lambda_i\left(\frac{1}{k}\sum_{l=1}^{k}\|b_j-p_{il}\|^2+\frac{1}{k}\sum_{l=1}^{k}\|b_i-p_{il}\|^2\right)\\
&\le \frac{1}{2(1-\varepsilon)}\sum_{j=1}^{m}\lambda_j\sum_{i=1}^{m}\lambda_i\left(\frac{1}{k}\sum_{l=1}^{k}\|f(b_j)-f(p_{il})\|^2+\frac{1}{k}\sum_{l=1}^{k}\|f(b_i)-f(p_{il})\|^2\right).\\
&\le \frac{1}{2(1-\varepsilon)}\left[\sum_{i=1}^{m}\lambda_i\big(\|c(\widehat{f(\sigma)})-f(b_i)\|^2-w(f(b_i))\big)\right] \qquad \text{(by definition of } c(\sigma)\text{, the center of }\sigma)\\
&\le \frac{1}{2(1-\varepsilon)}\mathrm{rad}^2(\widehat{f(\sigma)}), \qquad (\text{with } \widehat{f(\sigma)}=\{(f(b_i),w(f(b_i))\}_{i=1}^{m}))
\end{aligned}
$$

The last inequality comes from the following fact. Using the fact that $c(\sigma)$ is a convex combination of the b_i (Lemma 13) and the linearity of f, we have $c(\widehat{f(\sigma)}) = \sum_i \lambda_i f(b_i)$. The function $A : x \mapsto A(x) = \sum_{i=1}^{m}\lambda_i\big(\|x-f(b_i)\|^2-w(f(b_i))\big)$ is minimized at $x = c(\widehat{f(\sigma)})$. Hence $\min_x A(x) = A(c(\widehat{f(\sigma)})) = \mathrm{rad}^2(\widehat{f(\sigma)})$. The other inequality can be proved similarly.

Dimension analysis. Consider the augmented set of points $\tilde{P} := P \cup B_{P,k}$. By the JL lemma, any distance of the form $\|b-p\|$ is $(1\pm\varepsilon)$-preserved with probability $1-\exp(-d\varepsilon^2)$. Applying a union bound to all pairs of points in $\tilde{P}$, all distances of the form $\|p-q\|, p,q \in \tilde{P}$ are $(1\pm\varepsilon)$-preserved with probability $1-O\left(\binom{n}{k}^2\right)\times\exp(-d\varepsilon^2) = 1-O(\exp(2k\log n - d\varepsilon^2))$. We thus need to take as the target dimension $d = \Omega\left(\frac{k\log n}{\varepsilon^2}\right)$, which proves Theorem 17. ◀

Proof of Theorem 20. To get a better target dimension and prove Theorem 20, instead of the usual tail bound 4 used in JL-type inequalities, we need the stronger tail bounds of Lemma 24. Let us fix $S \in \binom{P}{k}$ and $b \in B_{P,k}$, so that $\mathcal{F}(S,b)$ is a fixed set of $k+1$ points. Taking the target dimension as d, we apply Lemma 24 with V as $\mathcal{F}(S,b)$ and G as a random subgaussian $d\times D$ matrix. We get a tail bound of $2\cdot\exp\left(-\frac{c\varepsilon^2 d\|V\|_F^2}{K^2\|V\|^2}\right)$. By our assumption, the stable rank of $\mathcal{F}(S,b)$ is $r_{st} = \frac{\|V\|_F^2}{\|V\|^2} = \Omega(k)$, say at least k/C for some constant C independent of k, D, n. Taking f to be the linear transform given by $f(p) = \sqrt{\frac{D}{d}}Gp$, we get that the probability that the sum of the squared distances $\left(\sum_{p\in S}\|p-b\|^2\right)$ is not $(1\pm\varepsilon)$-preserved, is at most $2\exp\left(-\frac{c\varepsilon^2 k}{K^2 C}d\right)$. Thus, choosing $d = \frac{3CK^2\log n}{c\varepsilon^2}$, where K is the subgaussian norm of the random matrix G, and c is as in Lemma 24, the probability that $\sum_{p\in S}\|p-b\|^2$ is not $(1\pm\varepsilon)$-preserved, is

$$
\begin{aligned}
\mathbf{P} &:= \mathbb{P}\left[\left|\sum_{p\in S}\|f(p)-f(b)\|^2-\sum_{p\in S}\|p-b\|^2\right| \ge \varepsilon\sum_{p\in S}\|p-b\|^2\right]\\
&\le 2\exp\left(\left(\frac{-c\varepsilon^2 k}{K^2 C}\right)\left(\frac{3CK^2\log n}{c\varepsilon^2}\right)\right) \le 2\exp(-(3k)\log n) = 2n^{-3k}.
\end{aligned}
$$

Taking a union bound over all $\mathcal{F}(S,b)\in\mathcal{F}$, i.e. over all $S\in\binom{P}{k}$ and all $b\in B_{P,k}$, we get

$$
\begin{aligned}
\mathbf{P} &= \mathbb{P}\left[\exists\mathcal{F}(S,b) : \left|\sum_{p\in S}\|f(p)-f(b)\|^2-\sum_{p\in S}\|p-b\|^2\right| > \varepsilon\sum_{p\in S}\|p-b\|^2\right]\\
&\le \left|\binom{P}{k}\right| \times |B_{P,k}|\cdot 2n^{-3k} \le \binom{n}{k}^2\cdot 2n^{-3k} \le 2n^{2k}\cdot n^{-3k} \le 2n^{-k}.
\end{aligned}
$$

Hence, with probability at least $(1 - O(n^{-k})) = 1 - o(1/n)$, taking $d = 3\frac{CK^2 \log n}{c\varepsilon^2}$ suffices to preserve the Čech complex $\check{C}_\alpha(X)$ for the k-distance. ◀

Proof of Theorem 18. Recall that, in Section 2.4, 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 $w(p) = -d^2_{P,k}(p)$. And 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 20. 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 14, we get

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

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. A subgaussian JL mapping preserves pairwise distances and the k-distance in dimension $O(\varepsilon^{-2} \log n)$. The result then follows as in the proof of Theorem 20. ◀

4.4 Proof of Theorem 22

Proof of Theorem 22. Recall that the elements of the set $\mathcal{F}$ are $k \times D$ matrices $\mathcal{F}(S, b)$, with $S \in \binom{P}{k}$ and $b \in B_{P,k}$. Let us again fix $\mathcal{F}(S, b)$, that is, we choose an arbitrary $b \in B_{P,k}$, and $S = \{p_1, \ldots, p_k\} \in \binom{P}{k}$. We shall show that with high probability, the stable rank $r_{st} = r_{st}(\mathcal{F}(S, b))$ is $\Omega(k)$. Finally we'll apply the union bound over all elements of $\mathcal{F}$ to prove the theorem.

Let $X, \bar{Y}$ be $k \times D$ matrices, such that the row vectors of X are given by the position vectors of the points in S, and the row vectors of $\bar{Y}$ are simply the position vector of b repeated k times. Then $\mathcal{F}(S, b) = X - \bar{Y}$. Therefore we have

$$r_{st}(\mathcal{F}(S, b)) \quad = \quad r_{st}(X - \bar{Y}) = \frac{\|X - \bar{Y}\|_F^2}{\|X - \bar{Y}\|^2} \quad \ge \quad \frac{\|X - \bar{Y}\|_F^2}{\left(\|X\| + \|\bar{Y}\|\right)^2}, \tag{9}$$

where the last inequality follows by applying the triangle inequality on the operator norm of a matrix. Note that the matrix $\mathcal{F}(S, b)$ is not a typical random matrix, as the entries are not independent. Instead, we use the following lemma to lower bound the Frobenius norm.

▶ **Lemma 25.** *There exists a constant $C > 0$, such that for $k \ge 4$, the squared Frobenius norm of $\mathcal{F}(S, b)$ satisfies*

$$\mathbb{P}\left[\|\mathcal{F}(S, b)\|_F^2 \le \left(\frac{k}{2} - 2\right) DK^2\right] \le \exp\left(-CkD\right). \tag{10}$$

Applying the union bound over all $\mathcal{F}(S, b) \in \mathcal{F}$, we get that

$$\begin{aligned}\mathbf{P} \quad &= \quad \mathbb{P}\left[\exists \mathcal{F}(S, b) \in \mathcal{F} \ : \ \|\mathcal{F}(S, b)\|_F^2 \le \left(\frac{k}{2} - 2\right) K^2\right] \le |\mathcal{F}| \cdot \exp\left(-CkD\right) \\ &= \quad \left|\binom{P}{k}\right| \times |B_{P,k}| \cdot \exp\left(-CkD\right) \quad = \quad \binom{n}{k}^2 \cdot \exp\left(-CkD\right).\end{aligned}$$

To bound the operator norm of X and $\bar{Y}$ in (9), we shall require the following well-known concentration inequality [29].

▶ **Lemma 26** (Tail bounds for operator norm of random $k \times D$ matrix [29](Theorem 4.4.5))**.** *Given any $k \times D$ real matrix T consisting of independently random zero-mean subgaussian entries with subgaussian norm K, for any $s > 0$, the operator norm of T satisfies* $\mathbb{P}\left[\|T\| \geq C_1(\sqrt{k}+\sqrt{D}+s)K\right] \leq \exp\left(-C_2 s^2\right)$*, where $C_1, C_2 > 0$ are independent of k, D.*

Let C_1, C_2 denote the constants in Lemma 26, and $Q := \mathbb{P}\left[\|X\| + \|\bar{Y}\| \geq 2C_1 K\left(s+\sqrt{D}+\sqrt{k}\right)\right]$.

$$Q \leq \mathbb{P}\left[\|X\| \geq C_1 K\left(s+\sqrt{D}+\sqrt{k}\right)\right] + \mathbb{P}\left[\|\bar{Y}\| \geq C_1 K\left(s+\sqrt{D}+\sqrt{k}\right)\right] =: t_1 + t_2.$$

For t_1, applying Lemma 26 with $s = \sqrt{D}$, we get $t_1 \leq \exp\left(-C_2 D\right)$. For t_2, observe that $\|\bar{Y}\|^2 = k\|b\|^2$, since for any $x \in \mathbb{R}^D$, $\|x\| = 1$, we have $\|Yx\|^2 = k\langle b, x\rangle^2 \leq k\|b\|^2$, as $\bar{Y}$ has k repeated rows, and $|\langle b, x\rangle|$, for any unit vector x, is at most $\|b\|$. Furthermore, b being the mean of k points, is itself a subgaussian random vector in $\mathbb{R}^D$ with mean zero and subgaussian norm $\leq K/\sqrt{k}$. Now applying Lemma 26 to the row matrix b with $s = \sqrt{D}$ and $k = 1$, we get $t_2 = \mathbb{P}\left[\sqrt{k}\|b\| \geq 2\sqrt{k}\frac{K}{\sqrt{k}}C_1\left(1+\sqrt{D}+s\right)^{1/2}\right] \leq \exp\left(-C_2 D\right)$. Applying again the union bound over $\mathcal{F}$, we get that for each element of $\mathcal{F}$, the squared operator norm is at most $\left(2C_1(s+\sqrt{D}+\sqrt{k})K\right)^2 \leq 8C_1^2 K^2(s^2+D)$, for $k \ll D$, with probability at least $1 - \binom{n}{k}^2 \cdot 2 \cdot \exp\left(-C_2 D\right)$. This probability is $1 - o(1)$ if D is at least $\frac{3k \ln n}{C_2}$. Thus the stable rank of each member of $\mathcal{F}$ is at least $\frac{(k-4)D}{8C_1^2(s^2+D)} = \frac{k-4}{16C_1^2}$, with high probability. ◀

5 Conclusion and Future Work

Vietoris-Rips and Delaunay filtrations. By the standard interleaving and equivalence results of the Čech filtration with the Vietoris-Rips filtration [12, Theorem 2.5] and the Delaunay filtration [25, Chapter 4], Theorem 16 and Theorem 20 respectively immediately imply that both these filtrations are preserved under the Johnson-Lindenstrauss mapping, up to a $(1-\varepsilon)^{-1/2}$ multiplicative factor.

Kernels. Other distance functions defined using kernels have proved successful in overcoming issues due to outliers. Using a result analogous to Theorem 16, we can show that the JL mapping preserves 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.

Relaxing the assumption. We have shown that the k-distance is preserved under random projections and is more strongly concentrated under a certain assumption, which holds for random or angularly separated point sets. 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 (non-uniform and / or non-discrete) probability measures, perhaps under some extra conditions?

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. URL: http://www.sciencedirect.com/science/article/pii/S0022000003000254, doi:https://doi.org/10.1016/S0022-0000(03)00025-4.

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. URL: http://doi.acm.org/10.1145/1132516.1132597, doi:10.1145/1132516.1132597.

3 Franz Aurenhammer. A new duality result concerning Voronoi diagrams. *Discrete & Computational Geometry*, 5(3):243–254, Jun 1990. URL: https://doi.org/10.1007/BF02187788, doi:10.1007/BF02187788.

4 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.

5 Mickaël Buchet, Tamal K. Dey, Jiayuan Wang, and Yusu Wang. Declutter and resample: Towards parameter free denoising. *JoCG*, 9(2):21–46, 2018. URL: https://doi.org/10.20382/jocg.v9i2a3, doi:10.20382/jocg.v9i2a3.

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

7 Frédéric Chazal, Vin De Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

8 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.

9 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. URL: http://doi.acm.org/10.1145/1377676.1377719, doi:10.1145/1377676.1377719.

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

11 Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.10073, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.10073, doi:10.1002/rsa.10073.

12 Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebr. Geom. Topol.*, 7(1):339–358, 2007. URL: https://doi.org/10.2140/agt.2007.7.339, doi:10.2140/agt.2007.7.339.

13 Sjoerd Dirksen. Dimensionality reduction with subgaussian matrices: A unified theory. *Foundations of Computational Mathematics*, 16(5):1367–1396, Oct 2016. URL: https://doi.org/10.1007/s10208-015-9280-x, doi:10.1007/s10208-015-9280-x.

14 Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002. URL: https://doi.org/10.1007/s00454-002-2885-2, doi:10.1007/s00454-002-2885-2.

15 Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction.* American Mathematical Society, 2010. URL: http://www.ams.org/bookstore-getitem/item=MBK-69.

16 S. Geršgorin (S. Gerschgorin). Über die Abgrenzung der Eigenwerte einer Matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, pages 749–754, 1931.

17 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.

18 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`.

19 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. URL: `http://doi.acm.org/10.1145/258533.258656`, `doi:10.1145/258533.258656`.

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

21 Latała, Rafal. Estimates of moments and tails of gaussian chaoses. *Ann. Probab.*, 34(6):2315–2331, 11 2006. URL: `https://doi.org/10.1214/009117906000000421`, `doi:10.1214/009117906000000421`.

22 Martin Lotz. Persistent homology for low-complexity models. *arXiv preprint arXiv:1709.01037*, 2017.

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

24 Jiří Matoušek. On variants of the Johnson Lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20218`, `arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.20218`, `doi:10.1002/rsa.20218`.

25 Steve Y. Oudot. *Persistence Theory - From Quiver Representations to Data Analysis*, volume 209 of *Mathematical surveys and monographs*. American Mathematical Society, 2015. URL: `http://bookstore.ams.org/surv-209/`.

26 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`.

27 Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4), July 2007. URL: `http://doi.acm.org/10.1145/1255443.1255449`, `doi:10.1145/1255443.1255449`.

28 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.

29 Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. `doi:10.1017/9781108231596`.

30 Ji Zhang. Advancements of outlier detection: A survey. *EAI Endorsed Trans. Scalable Information Systems*, 1(1):e2, 2013. URL: `http://dblp.uni-trier.de/db/journals/sis/sis1.html#Zhang13`.

A Appendix

A.1 Proofs from Section 2

Proof Sketch of Theorem 2 . Firstly, we normalize X to X/K, so it is safe to assume $K = 1$ in the remainder of the proof. Let $X_1, \ldots, X_D \in \mathbb{R}^n$ denote the rows of X. Observe that $\mathrm{Tr}(XAX^\top) = \sum_{i=1}^{D} X_i A X_i^\top$. Let $\mathrm{diag}(A)$ denote the $n \times n$ matrix with its diagonal entries same as those of A, and zeroes in the non-diagonal entries, and let $B := A - \mathrm{diag}(A)$.

Separating the summands corresponding to the diagonal terms of A from those corresponding to the non-diagonal terms, we get for a fixed $i \in [D]$,

$$X_i A X_i^\top = X_i \text{diag}(A) X_i^\top + X_i B X_i^\top.$$

Thus we get

$$XAX^\top = \sum_{i=1}^{D} \left(X_i \text{diag}(A) X_i^\top + X_i B X_i^\top \right) := S_1 + S_2.$$

Therefore the required probability can be upper bounded as

$$\mathbb{P}\left[\left| \text{Tr}(XAX^\top) - \mathbb{E}\left[\text{Tr}(XAX^\top) \right] \right| \geq t \right] \leq \mathbb{P}\left[\left| S_1 - \mathbb{E}\left[S_1 \right] \right| \geq t/2 \right] + \mathbb{P}\left[\left| S_2 - \mathbb{E}\left[S_2 \right] \right| \geq t/2 \right].$$

For the diagonal terms we have $S_1 = \sum_{i=1}^{D} X_i \text{diag}(A) X_i^\top = \sum_{i=1}^{D} \sum_{j=1}^{n} a_{jj}^2 X_{ij}^2$. Note that $(X_{ij}^2 - \mathbb{E}\left[X_{ij}^2 \right])_{i,j}$ is a collection of Dn independent mean-zero sub-exponential random variables. Thus, we have a weighted sum of independent mean-zero sub-exponential variables, and we can use Bernstein's inequality (see e.g. [29][Theorem 2.8.2]), to get

$$\begin{aligned} \mathbb{P}\left[\left| S_1 - \mathbb{E}\left[\text{Tr}(XAX^\top) \right] \right| \geq t \right] &\leq 2 \exp\left(-c \min\left(\frac{t^2}{\sum_{j=1}^{n} a_{jj}^2}, \frac{t}{\max_j |a_{jj}|} \right) \right) \\ &\leq 2 \exp\left(-c \min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|} \right) \right). \qquad (11) \end{aligned}$$

For the non-diagonal terms, first observe that $\mathbb{E}\left[S_2 \right] = 0$, by expanding the terms and using the linearity of expectation. Let $Z = \exp\left(sS_2 \right)$, where $s > 0$ is a variable parameter. By Markov's inequality, for the upper tail we have

$$\mathbb{P}\left[S_2 - \mathbb{E}\left[S_2 \right] \geq t/2 \right] = \mathbb{P}\left[S_2 \geq t/2 \right] = \mathbb{P}\left[Z \geq e^{st/2} \leq \mathbb{E}\left[Z \right] e^{-st/2} \right].$$

We now compute a bound on the moment-generating function $\mathbb{E}\left[\exp\left(sS_2 \right) \right]$. Let $Y_i := X_i B X_i^\top$, $i = 1, \ldots, D$. Observe that the Y_is are independent random variables, since they depend only on the corresponding X_is. Therefore we have

$$\mathbb{E}\left[Z \right] = \mathbb{E}\left[\exp\left(\sum_i sY_i \right) \right] = \mathbb{E}\left[\prod_i \exp\left(sY_i \right) \right] = \prod_i \mathbb{E}\left[\exp\left(sY_i \right) \right].$$

Now the individual terms $\mathbb{E}\left[\exp\left(sY_i \right) \right]$ can be bounded as in the proof of the Hanson-Wright inequality, (see [29][Chapter 6.2]) as

$$\begin{aligned} \mathbb{E}\left[\exp\left(sY_i \right) \right] &\leq \mathbb{E}\left[\exp\left(4sX_i B X_i'^\top \right) \right], \text{ using [29, Theorem 6.1.1]}, \\ &\leq \mathbb{E}\left[Csg_i B g_i'^\top \right], \text{ using [29, Comparison Lemma 6.2.3]}, \\ &\leq \exp\left(C' s^2 \|B\|_F^2 \right), \text{ using [29, Lemma 6.2.2, Gaussian chaos]}, \end{aligned}$$

assuming $|s| \leq c'/\|A\|$. Taking the product over $i = 1, \ldots, D$, we get

$$\mathbb{P}\left[Z \geq e^{st/2} \right] \leq \exp\left(-st/2 + C' s^2 D \|B\|_F^2 \right).$$

Optimizing over the value of s, we get

$$\begin{aligned} \mathbb{P}\left[S_2 \geq t/2 \right] &\leq \exp\left(-c \min\left(\frac{t^2}{D\|B\|_F^2}, \frac{t}{D\|B\|} \right) \right) \\ &\leq \exp\left(-c \min\left(\frac{t^2}{D\|A\|_F^2}, \frac{t}{D\|A\|} \right) \right). \qquad (12) \end{aligned}$$

The lower tail can be bounded similarly, by using $Z = \exp(-sS_2)$. The complete inequality now follows by adding the bounds in the inequalities (11) and (12) and de-normalizing with respect to K. ◀

Proof of Lemma 12. Assume, for a contradiction, that there exists two centers c_0 and $c_1 \neq c_0$ for $\hat{X}$. For convenience, write $r = \text{rad}(\hat{X})$. 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,

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

Hence, 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 $\text{rad}(\hat{X})$. We have obtained a contradiction. ◀

Proof of Lemma 13. We write for convenience $c = c(\hat{X})$ and $r = \text{rad}(\hat{X})$ and prove that $c \in \text{conv}(X_I)$ by contradiction. Let $c' \neq c$ be the point of $\text{conv}(X_I)$ 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 14. From Lemma 13, and writing $c = c(\hat{X})$ for convenience, we have

$$\text{rad}^2(\hat{X}) = \sum_{i=1}^{m} \lambda_i \big(\|c - p_i\|^2 - w_i\big).$$

We use the following simple fact from [22](Lemma 4.5)

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

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

$$\begin{aligned} \text{rad}^2(\hat{X}) &= \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{m} \lambda_j \lambda_i \|p_i - p_j\|^2 - \frac{1}{2} \sum_{i=1}^{m} 2\lambda_i w_i \\ &= \frac{1}{2} \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{m} \lambda_i \left(\|p_i - p_j\|^2 - 2w_i \right) \right) \qquad \left(\text{since } \sum_{i=1}^{m} \lambda_j = 1\right). \end{aligned}$$

◀

A.2 Proof from Section 4.2

Recall that $P = \{p_1, p_2, ...p_n\}$ is a set of n points in $X \in \mathbb{R}^D$. and let $x \in P$. From Lemma 3, we have for all $p \in P$

$$(1-\varepsilon)\|x-p\|^2 \leq \|f(x)-f(p)\|^2 \leq (1+\varepsilon)\|x-p\|^2. \tag{13}$$

In particular, the above inequality holds for $p \in \{p_1, \ldots, p_k\}$, the k nearest neighbours of x in P. By summing up the k inequalities for $p = p_1, \ldots, p_k$ and dividing by k, we have

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

Since the k-distance of $f(x)$ is the square root of the mean squared distance of $f(x)$ to its k-nearest neighbours,

$$d^2_{f(P),k}(f(x)) = \frac{1}{k}\sum_{q \in NN^k_{f(P)}(f(x))}\|f(x)-f(q)\|^2 \leq \frac{1}{k}\sum_{i=1}^{k}\|f(x)-f(p_i)\|^2.$$

So, for the upper bound, we have,

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

Let $q_1, q_2, ..., q_k \in P$ be such that $f(q_1), f(q_2), ..., f(q_k) \in NN^k_{f(P)}(f(x))$. Then for the lower bound we have, using Equ. (13),

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

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

A.3 Proofs from Section 4.4

▶ **Lemma 27.** *Let M be a $m \times n$ real matrix, and let A denote $MM^\top$. Then (i) $\|A\| = \|M\|^2$, and (ii) $\|A\|_F \leq \|M\| \cdot \|M\|_F$.*

Proof of Lemma 27. For the statement (i), consider the singular value decomposition of M, say $M = U\Sigma W^\top$, where U and W are respectively $m \times m$ and $n \times n$ orthogonal matrices, and Σ is an $m \times n$ diagonal matrix, whose entries are the singular values of M. Then we have that $A = MM^\top = W\Sigma U^\top U\Sigma W^\top = W\Sigma^2 W^\top$. Rotating the coordinate axes using the matrix W, we see that for any unit vector $x \in \mathbb{R}^n$, $\|Ax\|$ is bounded by the maximum entry of Σ^2, that is the square of the operator norm of M.

For (ii), let $M_1, \ldots M_n \in \mathbb{R}^m$ be the row vectors of M. By the definition of the operator norm, for each $i \in 1, \ldots, n$, we have $\|MM_i\|^2 \leq \|M\|^2\|M_i\|^2$. Therefore

$$\|A\|_F^2 = \|MM^\top\|_F^2 = \sum_{i=1}^{n}\|MM_i\|^2 \leq \sum_{i=1}^{n}\|M\|^2\|M_i\|^2 = \|M\|^2 \cdot \|M\|_F^2.$$

◀

We shall need the following well-known concentration inequality for the proof.

▶ **Lemma 28** (see e.g. Corollary 2.8.3 in [29]). *There exists a constant $c > 0$ such that, given independent zero-mean subgaussian random variables X_i, $i = 1, \ldots, m$ with subgaussian norm K, and any $s > 0$, the sum Y of the squares of the X_i, $Y := \sum_{i=1}^m X_i^2$ satisfies $\mathbb{E}[Y] \leq mK^2$ and*

$$\mathbb{P}\left[|Y - \mathbb{E}[Y]| \geq smK^2\right] \leq 2\exp\left(-c\min\left(s^2, s\right)m\right).$$

Proof of Lemma 25. Let $S_1 := \{p_1, \ldots, p_k\} \subset P$, Let $S_2 := \{q_1, \ldots, q_k\} \in \binom{P}{k}$ be a set of k points in P whose mean is b. Let p_{ij} denote the j-th coordinate of p_i, $j = 1, \ldots, D$, and write $\bar{p}_j = \sum_{i=1}^k p_{ij}$. Similarly, let q_{ij} denote the j-th coordinate of q_i, $j = 1, \ldots, D$ and write $\bar{q}_j = \sum_{i=1}^k q_{ij}$. Then M is the $k \times D$ matrix whose elements are $p_{ij} - \frac{1}{k}\sum_{i=1}^k q_{ij} = p_{ij} - \frac{1}{k}\bar{q}_j$. Let us bound the squared Frobenius norm Z of M. Writing $Z = \sum_{j=1}^D Z_j$ where the Z_j are the squared norms of the column vectors of M, we have:

$$\begin{aligned}
Z_j &= \sum_{i=1}^k \left(p_{ij} - \frac{1}{k}\bar{q}_j\right)^2 \\
&= \sum_{i=1}^k p_{ij}^2 - \frac{2}{k}\bar{p}_j\bar{q}_j + \frac{1}{k}\,\bar{q}_j^2 \\
&= \sum_{i=1}^k p_{ij}^2 + \frac{1}{k}\,\left((\bar{p}_j - \bar{q}_j)^2 - \bar{p}_j^2\right) \\
&\geq \sum_{i=1}^k p_{ij}^2 - \frac{1}{k}\,\bar{p}_j^2.
\end{aligned}$$

Since the random variables p_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, D$ are subgaussian, with subgaussian norm at most K, and the variables $\bar{p}_j$, $j = 1, \ldots, D$, are the sums of k independent subgaussian variables, they are themselves subgaussian (see e.g. [29][Chapter 2.6, Sums of independent subgaussians]), with mean zero and subgaussian norm at most $\sqrt{k}K$. Now since, $Z = \sum_{j=1}^D Z_j$, we can apply Lemma 28 with $m = kD$, $s = 1/2$, to lower bound Z and get

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^k\sum_{j=1}^D p_{ij}^2 \leq kDK^2/2\right] &\leq \mathbb{P}\left[\left|\sum_{i=1}^k\sum_{j=1}^D p_{ij}^2 - \mathbb{E}\left[\sum_{i=1}^k\sum_{j=1}^D p_{ij}^2\right]\right| \geq kDK^2/2\right] \\
&\leq 2\exp\left(-ckD/4\right),
\end{aligned}$$

Similarly, we upper bound $\sum_{j=1}^k \bar{p}_j$ using Lemma 28 with $m = D$, $s = 1$, to get

$$\begin{aligned}
\mathbb{P}\left[\sum_{j=1}^D \bar{p}_j^2 \geq 2kDK^2\right] &\leq \mathbb{P}\left[\left|\sum_{j=1}^D \bar{p}_j^2 - \mathbb{E}\left[\sum_{j=1}^D \bar{p}_j^2\right]\right| \geq kDK^2\right] \\
&\leq 2\exp\left(-ckD\right),
\end{aligned}$$

where c is the constant in Lemma 28. Thus, we get that $Z \geq \left(\frac{k}{2} - 2\right)DK^2$ with probability at least $1 - 2\cdot\exp\left(-ckD/4\right)$. ◀

▶ **Corollary 29.** *For the following subsets of $\mathbb{R}^D$, Theorem 20 holds:*

(i) Uniformly distributed points on the unit sphere in $\mathbb{R}^D$,

(ii) Uniformly distributed points on ellipsoids in $\mathbb{R}^D$, having bounded ratio of the sum of the squared axes, to the square of the largest axis,

(iii) Uniformly distributed points on symmetric polytopes in $\mathbb{R}^D$, with bounded ratio of the square of the average thickness of the intersecting slabs, to their maximum thickness.

Proof Sketch (Corollary 29). *(i).* The statement follows by using a Gaussian point distribution and normalizing by the ℓ_2 norm. If $g \sim \mathcal{N}(0, I_D)$ is a Gaussian random vector in $\mathbb{R}^D$, then the normalized random vector $\tilde{g} := \frac{g}{\|g\|}$ has the uniform distribution on the unit sphere S^{D-1} centered at the origin. By the Gaussian concentration inequality (see e.g. [29, Theorem 5.2.2]), for each point $p \in P$, its norm is at most $\sqrt{3D/2}$, with probability at least $1 - \exp(-cD)$, for some $c > 0$.

Let $\tilde{\mathcal{F}}(S, b)$ denote the row-normalized form of the matrix $\mathcal{F}(S, b)$. Thus for a given $S \in \binom{P}{k}$, $b \in B_{P,k}$, normalizing each row of the matrix $\mathcal{F}(S, b)$, reduces the square of the Frobenius norm by a factor of at most $3D/2$, with probability at least $1-\exp(-cD)$. Taking the union bound over all choices of $S \in \binom{P}{k}$ and $b \in B_{P,k}$, the squared Frobenius norm of any $\tilde{\mathcal{F}}(S, b)$ is at least $\frac{\|\mathcal{F}(S,b)\|_F^2}{3D/2}$, with probability at least $1 - \binom{n}{k}^2 \exp(-cD)$.

Similarly, applying the Gaussian concentration inequality, the ℓ_2 norm of each row of $\mathcal{F}(S, b)$ is at least $\sqrt{D/2}$ with probability at least $1 - \exp(-c'D)$, and thus the squared operator norm of $\tilde{\mathcal{F}}(S, b)$ is at most $\frac{\|\mathcal{F}(S,b)\|^2}{D/2}$ Thus, for $D = \Omega(k \ln n)$, with high probability, for any $S \in \binom{P}{k}$ and $b \in B_{P,k}$, the stable rank of $\tilde{\mathcal{F}}(S, b)$ is at least $1/3$ of the stable rank of $\mathcal{F}(S, b)$, and so remains $\Omega(k)$ (Theorem 22). Applying Theorem 20 now completes the proof.

(ii). This statement follows by applying the linear transform corresponding to the given ellipsoid, on the uniform distribution on the sphere.

(iii). This statement follows by first applying Theorem 22 to the uniform distribution on the surface of the unit cube $\mathcal{C}_m := [-1, 1]^m$ in $\mathbb{R}^m$, where m is the number of slabs representing the given polytope P, and then applying the linear transform corresponding to P, to the uniform point distribution on $\mathcal{C}_m$. ◀

Proof of Theorem 23. Let $S_1 = \{p_1, \dots, p_k\} \subset P$ and $S_2 := \{q_1, \dots, q_k\} \subset P$ be two sets of k points, not necessarily disjoint. Let X be the $D \times k$ matrix with columns as the position vectors of $p_1, \dots, p_k$. Let $\bar{Y}$ be the $D \times k$ matrix with columns as the vector $\frac{1}{k}\sum_{i=1}^{k} q_i$ repeated k times. We shall prove a lower bound on the stable rank of the matrix $X - \bar{Y}$, for all possible choices of S_1 and S_2. For a fixed S_1 and S_2, the stable rank is given by

$$r_{st}(X - \bar{Y}) = \frac{\|X - \bar{Y}\|_F^2}{\|X - \bar{Y}\|^2} \geq \frac{\|X - \bar{Y}\|_F^2}{\left(\|X\| + \|\bar{Y}\|\right)^2}. \tag{14}$$

First, we prove the following lower bound on the Frobenius norm.

▶ **Lemma 30.** $\|X - \bar{Y}\|_F^2 \geq \|X\|_F^2 - \frac{1}{k}\|\sum_{i=1}^{k} p_i\|^2$.

Next, we observe that

▶ **Lemma 31.** $\|\sum_{i=1}^{k} p_i\|^2 \leq 3k$.

Therefore, by Lemmas 30 and 31 we get that the Frobenius norm

$$\|X - \bar{Y}\|_F^2 \geq \|X\|_F^2 - \frac{1}{k}\|\sum_{i=1}^{k} p_i\|^2 \geq \frac{k}{2} - 3. \tag{15}$$

Secondly, we upper bound the operator norm of $X - \bar{Y}$. To bound $\|\bar{Y}\|$, observe that since the columns of $\bar{Y}$ are simply the vector $\frac{1}{k}\sum_{i=1}^{k} q_i$ repeated k times, we have

$$\|\bar{Y}\|^2 = k\frac{\|\sum_{i=1}^{k} q_i\|^2}{k^2} = \frac{\|\sum_{i=1}^{k} q_i\|^2}{k}.$$

Now applying Lemma 31, we get that

$$\|\bar{Y}\|^2 \quad \leq \quad 3. \tag{16}$$

Finally, to upper bound $\|X\|$, we shall use the following Lemma, which follows from a direct application of Gershgorin's theorem [16] to the matrix $X^\top X$:

▶ **Lemma 32.** *Let M be a $D \times k$ matrix whose column vectors $M_i \in \mathbb{R}^D$, $i \in [k]$, have norm at most $\sqrt{\frac{3}{2}}$ and satisfy $\max_{1\leq i\neq j\leq k} |\langle M_i, M_j\rangle| \leq \eta\|M_i\|\cdot\|M_j\|$. Then $\|M\|^2 \leq \frac{3}{2}+(k-1)\frac{3\eta}{2}$.*

Applying Lemma 32 to the matrix X with $\eta = \frac{1}{k-1}$, we immediately get that

$$\|X\| \quad \leq \quad \sqrt{\frac{3}{2} + (k-1)\frac{3\eta}{2}} = \sqrt{3}. \tag{17}$$

Putting together the bounds from Eqns. (18), (17), and (16), and substituting in Eqn (14) we get

$$r_{st} \geq \frac{k/2 - 3}{(2\sqrt{3})^2} \geq \frac{k-6}{24}.$$

Since the above lower bound holds regardless of the choice of the sets S_1 and S_2, this proves the theorem.

◀

Proof of Lemma 30. Let p_{ji} denote the j-th coordinate of p_i, $j = 1, \ldots, D$, and write $\bar{p}_j = \sum_{i=1}^{k} p_{ji}$. Similarly, let q_{ji} denote the j-th coordinate of q_i, $j = 1, \ldots, D$ and write $\bar{q}_j = \sum_{i=1}^{k} q_{ji}$. Then $X - \bar{Y} =: M$ is the $D \times k$ matrix whose elements are $p_{ji} - \frac{1}{k}\sum_{i=1}^{k} q_{ji} = p_{ji} - \frac{1}{k}\bar{q}_j$. Let us bound the squared Frobenius norm Z of M. Writing $Z = \sum_{j=1}^{D} Z_j$ where the Z_j are the squared norms of the row vectors of M, we have:

$$Z_j \quad \geq \quad \sum_{i=1}^{k} p_{ji}^2 - \frac{1}{k}\bar{p}_j^2,$$

as in the proof of Lemma 25.

Therefore summing over j, we have

$$\|M\|_F^2 = \sum_{j=1}^{D} Z_j \geq \sum_{i=j=1}^{i=k,j=D} p_{ji}^2 - \frac{1}{k}\sum_{j=1}^{D}\bar{p}_j^2 = \|X\|_F^2 - \left\|\sum_{i=1}^{k} p_i\right\|^2.$$

◀

Proof of Lemma 31.

$$\left\|\sum_{i=1}^{k} p_i\right\|^2 \quad = \quad \sum_{i=1}^{k}\|p_i\|^2 + \sum_{1\leq i\neq j\leq k}\langle p_i, p_j\rangle.$$

Since $\|p_i\|^2 = 1$ for all $p_i \in P$, and $\langle p_i, p_j \rangle \leq \frac{\|p_i\| \cdot \|p_j\|}{k-1}$ by the assumptions in the statement of the Theorem, we have,

$$\begin{aligned} \|\sum_{i=1}^{k} p_i\|^2 &\leq \frac{3k}{2} + k(k-1)\frac{1}{k-1} \cdot \frac{3}{2} \\ &= 3k. \end{aligned} \tag{18}$$

◀

Proof of Lemma 32. The proof follows from Gershgorin's theorem, stated below.

▶ **Theorem 33** (Gershgorin [16]). *Given a complex $k \times k$ matrix B having entries B_{ij}, $i, j \in [k]$, and columns B_i, $i \in [k]$, every eigenvalue of B is contained in the union of the closed discs in the complex plane, centered at the entries B_{ii}, having radius at most $R_i := \sum_{j \neq i} |\langle B_i, B_j \rangle|$.*

From the definition of the operator norm, we have that $\|M\|^2 = \sigma_1^2$, where σ_1 is the largest eigenvalue of $M^\top M$. Since $M^\top M$ is a positive semidefinite matrix, all its eigenvalues are real. Now applying Gershgorin's theorem with $B = M^\top M$, we have that all eigenvalues of B are contained in the union of the intervals $\{x \in \mathbb{R} : \|B_{ii} - x\| \leq R_i\}$, where $R_i = \sum_{j:\, j \neq i} |\langle M_i, M_j \rangle| \leq (k-1)\|M_i\| \cdot \|M_j\| \cdot \eta$. Therefore, the maximum eigenvalue of $M^\top M$ has absolute value at most $\frac{3}{2} + (k-1)\eta \cdot \frac{3}{2}$. ◀