# Random Projections, Topological Persistence, and the Johnson-Lindenstrauss Lemma

Michael Lamar and David Letscher
Department of Mathematics and Computer Science
Saint Louis University
{mlamar,letscher}@slu.edu

**Abstract**

We generalize the Johnson-Lindenstrauss lemma to the setting of topological persistence. The main result is that with positive probability a random linear projection to a sufficiently large Euclidean space will correctly calculate the $i$-dimensional persistent homology for a set of $n$ points up to a bounded multiplicative error.

The Johnson-Lindenstrauss lemma [4] has proven to be an extremely useful theoretical tool to address the "curse of dimensionality" when considering distances between points. It states that there is an embedding in logarithmically many dimensions that roughly preserves distances. In fact, the techniques used to prove it say that with reasonable probability, a particular random projection to this number of dimensions will have low distortion. Johnson-Lindenstrauss is then proved by projecting to a space where there is positive probability of distances being preserved within an acceptable error. Applications of this lemma and related techniques include nearest neighbor search, compressed sensing, and manifold learning.

**Johnson-Lindenstrauss Lemma.** *If $X \subset \mathbb{R}^m$ with $|X| = n$ then for every $\epsilon < \frac{1}{2}$ and $k \geq \frac{8}{\epsilon^2} \log n$ there exists a linear map $f : \mathbb{R}^m \to \mathbb{R}^k$ such that for every $x, y \in X$,*

$$(1-\epsilon)d^2(x,y) < d^2(f(x),f(y)) - d^2(x,y) < (1+\epsilon)d^2(x,y)$$

Persistent homology [2] has proven to be a useful tool in a variety of fields and a rich theory has been built. However, all applications have been in low-dimensional settings. A barrier to working with topological persistence for high dimensional data sets, is that both the size of the triangulations and bases for homology can be exponential in dimension. We prove an analog of Johnson-Lindenstrauss that shows that for sufficiently large data sets we can project to smaller dimensions and not significantly change the persistent homology groups of interest.

It is known that persistent homology is stable under the addition of noise (e.g. additive Gaussian noise) [1]. This stability can be expressed in terms of the bottleneck distance between persistence diagrams. Errors that come from linear projections, have multiplicative noise, similar to what is seen in Johnson-Lindenstrauss. We introduce a new measure of similarity of persistence diagrams that detects multiplicative errors and use it to prove a generalization of Johnson-Lindenstrass to persistent homology.

Before we state the main result, we provide some background on persistence diagrams and bottleneck distance. As a starting point, consider a finite point set $X \subset \mathbb{R}^m$ and let $X_\alpha = \cup_{x \in X} B(x, \alpha)$ be the union of balls centered at the points. The ordered set $\{X_\alpha\}_{\alpha \in \mathbb{R}}$ provides a filtration of $\mathbb{R}^m$ [3]. We will denote the persistence diagram for $H_i^p(X_\alpha; \mathbb{F})$ for a field $\mathbb{F}$ by $\mathcal{PH}_i(X_\alpha)$. A diagram is a multiset in $\{(x,y) \in \overline{\mathbb{R}}^2 \mid x \leq y\}$. Each point in this multiset corresponds to a

generator of the persistence module; the $x$-coordinate is the generator's birth time and the $y$-coordinate is it's death time.

To calculate the bottleneck distance, we add the line $y = x$ to the persistence diagram with infinite multiplicity. The bottleneck distance between two diagrams can be defined as

$$d_B(\mathcal{D}, \mathcal{D}') = \inf\{\epsilon \mid \exists \text{ bijection } f : \mathcal{D} \to \mathcal{D}' \text{ with } d(p, f(p)) \leq \epsilon \ \forall p \in \mathcal{D}\}$$

In essence log-bottleneck distance ensures that no birth or death time of a cycle is changed by a factor of more than $1 \pm \epsilon$, and is defined as $d_{LB}(\mathcal{D}, \mathcal{D}') = d_B(\log \mathcal{D}, \log \mathcal{D}')$ where $\log \mathcal{D} = \{(\log x, \log y) \mid (x, y) \in \mathcal{D}\}$. Note that this is the same as measuring the bottleneck distance on a log-log plot of the persistence diagram.

We have two versions of the main result; one with and one without sampling conditions. $X$ will be called $\eta$-general if for all points $S \subset X$ with $|S| \leq i$ and if the circumsphere of $S$ has center $c$ and radius $r$, then $d(c, x) \notin ((1-\eta)r, (1+\eta)r) \ \forall x \in X - S$.

**Theorem 1.** *Suppose that $X \subset \mathbb{R}^m$ with $|X| = n$. If $f(x) = \frac{1}{\sqrt{k}} Ax$ for $A$ an $m \times k$ matrix of independent standard normal random variables there exists $K_0$ such that for $k \geq K_0$ and $\epsilon < \frac{1}{10}$, then*

1. *$P\left(d_{LB}(\mathcal{PH}_i(f(X)_\alpha), \mathcal{PH}_i(X_\alpha)) < \epsilon\right) \geq 1 - \binom{n}{i+2}(9i+2)e^{-\frac{\epsilon_1^2 k}{8}}$ where $\epsilon_1 = \frac{\epsilon^{2^i}}{15^{2^{i+1}-2}}$*
2. *if $X$ is $\eta$-general then $P\left(d_{LB}(\mathcal{PH}_i(f(X)_\alpha), \mathcal{PH}_i(X_\alpha)) < \epsilon\right) \geq 1 - \binom{n}{i+2}(7i-3)e^{-\frac{(\epsilon\eta)^2 k}{8}}$*

**Corollary 2.**

1. *If $k > \frac{8}{\epsilon_1^2}(i+2)\log n$ then there exists a linear map $f : \mathbb{R}^m \to \mathbb{R}^k$ such that $d_{LB}(\mathcal{PH}_i(f(X)_\alpha)), \mathcal{PH}_i(X_\alpha)) < \epsilon$.*
2. *If $X$ is $\eta$-general and $k > \frac{8}{(\epsilon\eta)^2}(i+2)\log n$ then there exists a linear map $f : \mathbb{R}^m \to \mathbb{R}^k$ such that $d_{LB}(\mathcal{PH}_i(f(X)_\alpha)), \mathcal{PH}_i(X_\alpha)) < \epsilon$.*

**Future work** Unfortunately, the dimension requirements to apply this theorem are too large to lead to a practical algorithm for estimating persistent homology. We home to improve the dimensions bounds to $O(\frac{i \log n}{\epsilon^2})$. We are currently working on a theoretical algorithm that works in lower dimensions under certain sampling conditions. While this theoretical algorithm may prove to still be impractical, preliminary results indicate that this algorithm can be run in a handful of dimensions and provide meaningful results.

## References


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

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

[3] Herbert Edelsbrunner and Ernst P Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics (TOG)*, 13(1):43–72, 1994.

[4] William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1–147, 1984.