LOW DIMENSIONAL SPACES OF PERSISTENCE DIAGRAMS

DONALD R. SHEEHY AND SIDDHARTH S. SHETH
NORTH CAROLINA STATE UNIVERSITY

Abstract. The space of persistence diagrams under bottleneck distance is known to be high-dimensional. Because many metric search algorithms and data structures have bounds that depend on the dimension of the search space, the high-dimensionality of this space makes it difficult to analyze and compare asymptotic running times of metric search algorithms on this space. In this paper, we explore the dimension of a generalization of the space of PDs. We show that for a class of quotient metrics that includes the bounded persistence plane (i.e., a bounded region of the plane modulo the diagonal) the metric is close in a Gromov-Hausdorff sense to a metric of bounded dimension. We also show how to bound the dimension of bottleneck metrics over k-point subsets of doubling metrics. Finally, we put these together to show that the space of bounded, k-point persistence diagrams is close to a low-dimensional space.

1. Introduction

A persistence diagram (PD) is a topological invariant commonly used in topological data analysis (TDA). Ever since their introduction, PDs have been a popular tool to compare the shapes of point clouds, metric spaces, and real-valued functions.

A significant advantage of PDs over many other topological invariants is that they come equipped with a natural metric, the bottleneck distance, and thus topological features are rendered not only qualitative, but also quantitative. This opens the possibility of doing metric analysis on PDs, such as (approximate) nearest neighbor search or range search.

Many metric proximity search algorithms and data structures have asymptotic running time bounds in terms of the doubling dimension of the search space [2, 3]. The metric space of PDs with the bottleneck distance is known to have infinite doubling dimension[4], making it unclear whether one ought to apply standard data structures such as cover trees[1] or net trees[5] to search in this space. Although the space of all persistence diagrams is infinite-dimensional, there may be reasonable subspaces of persistence diagrams that either are or at least behave like they are low-dimensional. This paper describes some general classes of persistence diagrams that are nearly low-dimensional in the sense that they are close in Gromov-Hausdorff distance to a low-dimensional space (see Theorem 5.1). Along the way, we give general techniques for proving quotient metrics are nearly low-dimensional (Section 3) and a general bound on the doubling dimension of bottleneck spaces over doubling metrics (Section 4)

2. Definitions

2.1. **Metric, Cover, and Dimension.** A *metric space*, $\mathsf{X} = (X, \mathrm{d})$ is a set X and a metric d. The distance between $a \in X$ and a set Y is given by $\mathrm{d}(x, Y) := \inf_{b \in Y} \mathrm{d}(a, b)$. The *diameter* of a set X is $\mathrm{diam}(X) = \sup_{a,b \in X} \mathrm{d}(a, b)$. An *$\varepsilon$-ball centered at a*, denoted by $B(a)$, is the set of all points within ε distance of a.

A collection of sets Y is said to cover X if the union of the sets in Y contains X. A cover Y is an *ε-cover* of X if every set in Y has diameter at most 2ε. An ε-cover Y of X of minimum cardinality is a *minimum ε-cover* and $\mathrm{N}_\varepsilon(X) = |Y|$ is the *covering number* of X. The *ε-metric entropy* of X is defined as $\mathrm{H}_\varepsilon(X) = \log_2 \mathrm{N}_\varepsilon(X)$.

The *doubling constant*, λ, of X is defined as,

$$\lambda = \max_{Z \subseteq X} \mathrm{N}_{\mathrm{diam}(Z)/2}(Z).$$

The *doubling dimension* of X is $\dim(\mathsf{X}) := \log_2(\lambda)$. If $\dim(\mathsf{X})$ is finite, then X is a *doubling metric.* Throughout this paper, all mentions of dimension are referring to the doubling dimension.

2.2. **Quotient Metric Spaces.** Let X be a metric space and let Y be a subspace. The *quotient space* $\mathsf{X}/\mathsf{Y} = (X/Y, \mathrm{d}_{X/Y})$ is defined so that $\mathrm{d}_{X/Y}([a],[b]) := \min\{\mathrm{d}(a,b), \mathrm{d}(a,Y) + \mathrm{d}(b,Y)\}$. There also exists a surjective quotient map, $q : X \to X/Y$ such that $q(x) = [x]$.

The *persistence plane* is the quotient of $(\mathbb{R}^2, \ell_\infty)$ modulo the diagonal. A persistence diagram is a multiset of points in the persistence plane. The dimension is infinite. There is an interesting observation here: a quotient of two low-dimensional metric spaces can be infinite dimensional.

2.3. **Gromov-Hausdorff Distance and Nearly Low-Dimensional Spaces.** For metric spaces $\mathsf{P} = (P, \mathrm{d}_P)$ and $\mathsf{Q} = (Q, \mathrm{d}_Q)$, a *correspondence* between P and Q is a relation $\mathrm{R} \subseteq P \times Q$ such that for its canonical projections on P and Q, we have $\pi_P(\mathrm{R}) = P$ and $\pi_Q(\mathrm{R}) = Q$ respectively. The *distortion* of R is defined as

$$\mathrm{distort}(\mathrm{R}) := \sup_{(p_1,q_1),(p_2,q_2) \in \mathrm{R}} |\mathrm{d}_P(p_1,p_2) - \mathrm{d}_Q(q_1,q_2)|.$$

The *Gromov-Hausdorff distance d_{GH}* is a metric on metric spaces defined as

$$\mathrm{d}_{GH}(\mathsf{P},\mathsf{Q}) := \frac{1}{2} \inf\{\mathrm{distort}(\mathrm{R}) \mid \mathrm{R} \subseteq P \times Q \text{ is a correspondence}\}.$$

A metric space P is *ε-nearly low-dimensional* if there exists a doubling metric space Q such that $\mathrm{d}_{GH}(\mathsf{P},\mathsf{Q}) \le \varepsilon$.

3. ε-Close Quotient Metric Spaces

In this section show how to approximate a quotient space with a lower dimensional quotient space. We first present a lemma on the dimension of a quotient of a doubling metric modulo finite subset.

Lemma 3.1. *Let* X *be a metric space with* $\dim(\mathsf{X}) = d$. *If* $Y \subset X$ *is finite, then* $\dim(\mathsf{X}/\mathsf{Y}) \le d + \log|Y|$

Proof. Let $S \subseteq X$ be such that $\mathrm{diam}_{X/Y}(S) = 2\varepsilon$. Let $I : Y \to 2^S$ be a cover of S indexed by the points of Y.

For all $y \in Y$ and $a, b \in I(y)$,

$$\begin{aligned} 2\varepsilon &\ge \mathrm{d}_{X/Y}([a],[b]) \\ &:= \min\{\mathrm{d}(a,b), \mathrm{d}(a,Y) + \mathrm{d}(b,Y)\} \\ &= \min\{\mathrm{d}(a,b), \mathrm{d}(a,y) + \mathrm{d}(b,y)\} \\ &= \mathrm{d}(a,b). \end{aligned}$$

Thus,

$$\mathrm{diam}(I(y)) = \sup_{a,b \in V(y)} \mathrm{d}(a,b) \le 2\varepsilon.$$

By the definition of doubling dimension, for each $I(y)$ there exists an ε-cover of size at most 2^d sets. Thus, an ε-cover for S can be constructed using the $|Y|$ set $\{I(y) \mid y \in Y\}$ separately. So, we have an ε-cover of S of size at most $|Y|2^d$ and thus,

$$\dim(\mathsf{X}/\mathsf{Y}) \le d + \log|Y|.$$

□

Lemma 3.2. *For any quotients* X/Y *and* $\mathsf{X}/\mathsf{Y}_\varepsilon$ *over* X*, if* Y_ε *is an* ε*-cover of* Y*, then there exists a correspondence between* X/Y *and* $\mathsf{X}/\mathsf{Y}_\varepsilon$*.*

Proof. Let q and q_ε denote the canonical quotient maps for X/Y and X/Y_ε respectively. Let $\mathrm{R} \subseteq X/Y \times X/Y_\varepsilon$ be a relation such that $\mathrm{R} = \{([a],[b]) \mid \exists x \in X : q(x) = [a], q_\varepsilon(x) = [b]\}$. Because the quotient maps are surjective, it is easy to verify for the canonical projections of R that $\pi_{X/Y}(\mathrm{R}) = X/Y$ and $\pi_{X/Y_\varepsilon}(\mathrm{R}) = X/Y_\varepsilon$. Thus, R is a correspondence. □

Theorem 3.3. *For every quotient* X/Y *of a compact metric space* X *with doubling dimension* d*, there exists a space of doubling dimension* $d + \mathrm{H}_\varepsilon(Y)$ *that is* ε*-close to* X/Y *in terms of the Gromov-Hausdorff distance.*

Proof. Let $Y_\varepsilon \subseteq Y$ be such that $\bigcup_{y_i \in Y_\varepsilon} B(y_i)$ is a minimal ε-cover of Y. Then, $\mathrm{H}_\varepsilon(Y) = \log |Y_\varepsilon|$. So, by Lemma 3.1, we know that $\mathsf{X}/\mathsf{Y}_\varepsilon$ has doubling dimension at most $d + \mathrm{H}_\varepsilon(Y)$. It suffices to show that $\mathrm{d}_{GH}(\mathsf{X}/\mathsf{Y}, \mathsf{X}/\mathsf{Y}_\varepsilon) \leq \varepsilon$.

By lemma 3.2, we have a correspondence, R, between X/Y and $\mathsf{X}/\mathsf{Y}_\varepsilon$. For an arbitrary pair $([a],[a]'),([b],[b]') \in \mathrm{R}$, let

$$\begin{aligned}\delta &= |\mathrm{d}_{X/Y}([a],[b]) - \mathrm{d}_{X/Y_\varepsilon}([a]',[b]')| \\ &= |\min\{\mathrm{d}(a,b), \mathrm{d}(a,Y) + \mathrm{d}(b,Y)\} - \min\{\mathrm{d}(a,b), \mathrm{d}(a,Y_\varepsilon) + \mathrm{d}(b,Y_\varepsilon)\}|.\end{aligned}$$

Because $Y_\varepsilon \subset Y$, we have $\mathrm{d}(a,Y) \leq \mathrm{d}(a,Y_\varepsilon) \leq \mathrm{d}(a,Y) + \varepsilon$. Thus, computing distort(R) reduces to the following two cases:

Case. $\mathrm{d}(a,b) \leq \mathrm{d}(a,Y) + \mathrm{d}(b,Y)$

In this case, we have

$$\mathrm{d}_{X/Y}([a],[b]) = \mathrm{d}(a,b), \text{ and } \mathrm{d}_{X/Y_\varepsilon}([a]',[b]') = \mathrm{d}(a,b).$$

Thus, $\delta = 0$.

Case. $\mathrm{d}(a,Y) + \mathrm{d}(b,Y) < \mathrm{d}(a,b)$

On the other hand, if $\mathrm{d}(a,Y) + \mathrm{d}(b,Y) < \mathrm{d}(a,b)$, then,

$$\mathrm{d}_{X/Y}([a],[b]) = \mathrm{d}(a,Y) + \mathrm{d}(b,Y) \text{ and}$$

$$\mathrm{d}_{X/Y_\varepsilon}([a]',[b]') \leq \mathrm{d}(a,Y_\varepsilon) + \mathrm{d}(b,Y_\varepsilon) \leq \mathrm{d}(a,Y) + \mathrm{d}(b,Y) + 2\varepsilon.$$

Thus, $\delta \leq 2\varepsilon$.

Thus, the distortion of R is at most 2ε and hence,

$$\mathrm{d}_{GH}(\mathsf{X}/\mathsf{Y}, \mathsf{X}/\mathsf{Y}_\varepsilon) \leq \frac{1}{2}\mathrm{distort}(\mathrm{R}) \leq \varepsilon.$$

Because Y is compact, Y_ε is finite and $\mathsf{X}/\mathsf{Y}_\varepsilon$ is the required ε-close space with doubling dimension $d + \mathrm{H}_\varepsilon(Y)$.

□

4. d-Dimensional k-Point Diagrams

If the doubling dimension of X is d, then a *d-dimensional k-point diagram* is a set of k distinct elements of X. Let $A = \{a_i\}_{i \in I_k}$ and $B = \{b_i\}_{i \in I_k}$ be d-dimensional k-point diagrams where $I_k = \{1, \ldots, k\}$. The bottleneck of a bijection $\eta : A \to B$ is $\max_{a_i \in A} \mathrm{d}(a_i, \eta(a_i))$. An optimal matching minimizes the bottleneck and the *bottleneck distance* between A and B is $\mathrm{d}_B(a,b) = \min_\eta \max_{a_i \in a} \mathrm{d}(a_i, \eta(a_i))$. Thus, $\mathrm{C}_{\mathsf{k}}^{\mathsf{X}} = (\mathrm{C}_k^X, \mathrm{d}_B)$ is a metric space defined over the set of all d-dimensional k point diagrams and the bottleneck distance of their matchings, d_B.

Theorem 4.1. *Let $\mathsf{X} = (X, \mathrm{d})$ be a d-dimensional doubling metric. Let $\mathrm{C}_{\mathsf{k}}^{\mathsf{X}}$ denote the metric space of k-point diagrams in X with the bottleneck metric. Then,* $\dim(\mathrm{C}_{\mathsf{k}}^{\mathsf{X}}) \leq kd$*.*

Proof. Let T be an ε-ball in C_k^X. So by definition, for matchings η,

$$\sup_{A,B \in T} \inf_{\eta : A \to B} \max_i \mathrm{d}(a_i, \eta(a_i)) \leq \varepsilon$$

Therefore, $T = \{\{a_1, \ldots, a_k\} \mid a_i \in B_i\}$ where each B_i is an ε-ball in X. There exists an ε-cover, U_i, for every B_i such that $|U_i| \leq 2^d$.

Moreover, for every $\{a_1, \ldots, a_k\} \in T$ there exists $\{V_1, \ldots, V_k\}$ where each set $V_i \in U_i$ is such that $a_i \in V_i$. Thus $C = \{\{V_1 \ldots V_k\} \mid V_i \in U_i\}$ is an ε-cover of T of size at most 2^{kd}. So $\dim(\mathrm{C}_{\mathsf{k}}^{\mathsf{X}}) \leq kd$. □

5. Space of Bounded Persistence Diagrams

From the preceding two sections we get an approximation of single-class quotient spaces and a bound on the doubling dimension of finite point bottleneck spaces respectively. These results come together in the space of bounded persistence diagrams to form a nearly low dimensional subspace of persistence diagrams.

Let $\mathsf{D} = (\mathbb{R}^2, \ell_\infty)$. Denoting G as the diagonal from $(0, 0)$ to (N, N), let $\mathsf{D_N/G} = (D_N/G, \mathrm{d}_{D_N/G})$ be the N-bounded persistence plane. By definition, $\mathrm{d}_{D_N/G}([a], [b]) = \min\{\ell_\infty(a, b), \ell_\infty(a, G) + \ell_\infty(b, G)\}$. Let $\mathrm{C}_{\mathsf{k}}^{\mathsf{D_N/G}} = (\mathrm{C}_k^{D_N/G}, \mathrm{d}_B)$ be the bottleneck space of all k-point persistence diagrams bounded by N and let $\mathrm{C}_{\mathsf{k}}^{\mathsf{D_N/G_\varepsilon}} = (\mathrm{C}_k^{D_N/G_\varepsilon}, \mathrm{d}_B)$ denote the k-point bottleneck space in the approximate N-bounded k-point persistence plane where G_ε is the minimum ε-cover of G.

Theorem 5.1. *The space of k-point N-bounded persistence diagram is ε-close to a space of doubling dimension at most $(2 + \log\lceil N/2\varepsilon\rceil)k$ in terms of the Gromov-Hausdorff distance.*

Proof. Because $\dim(\mathsf{D}) = 2$ and $|G_\varepsilon| = \lceil N/2\varepsilon\rceil$, from theorem 4.1, we know that $\dim(\mathrm{C}_{\mathsf{k}}^{\mathsf{D_N/G_\varepsilon}}) \leq (2 + \log\lceil N/2\varepsilon\rceil)k$. To show that this approximate persistence plane is ε-close the bounded persistence plane we use a technique similar to theorem 3.3.

By lemma 3.2, we know that the correspondence R between $\mathsf{D_N/G}$ and $\mathsf{D_N/G_\varepsilon}$ has distortion at most 2ε. Let R^k denote the correspondence between $\mathrm{C}_{\mathsf{k}}^{\mathsf{D_N/G}}$ and $\mathrm{C}_{\mathsf{k}}^{\mathsf{D_N/G_\varepsilon}}$ defined as follows:

$$\mathrm{R}^k = \{(\{a_1, \ldots, a_k\}, \{b_1, \ldots, b_k\}) \mid \exists m : I_k \to I_k, \forall i : (a_i, b_{m(i)}) \in \mathrm{R}\}.$$

To show that $\mathrm{d}_{GH}(\mathrm{C}_{\mathsf{k}}^{\mathsf{D_N/G_\varepsilon}}, \mathrm{C}_{\mathsf{k}}^{\mathsf{D_N/G}}) \leq \varepsilon$, it is sufficient to bound the distortion of R^k.

Let (A, B) and (A', B') be arbitrary pairs in the R^k, where $A = \{a_i\}_{i \in I_k}$, $A' = \{a'_i\}_{i \in I_k}$, $B = \{b_i\}_{i \in I_k}$, and $B' = \{b'_i\}_{i \in I_k}$. Without loss of generality, we may assume they are indexed so that for all j, we have $(a_j, b_j) \in \mathrm{R}$ and $(a'_j, b'_j) \in \mathrm{R}$. Let $\eta : I_k \to I_k$ be the permutation of indices that gives the bottleneck matching between A and A', i.e.,

$$\mathrm{d}_B(A, A') = \max_{i \in I_k} \mathrm{d}_{D^N/G}(a_i, a'_{\eta(i)}).$$

It follows from the distortion bound on R that

$$\begin{aligned} \mathrm{d}_B(B, B') &\leq \max_{j \in I_k} \mathrm{d}_{D^N/G_\varepsilon}(b_j, b'_{\eta(j)}) \\ &\leq \max_{j \in I_k} (\mathrm{d}_{D^N/G}(a_j, a'_{\eta(j)}) + 2\varepsilon) \\ &= \mathrm{d}_B(A, A') + 2\varepsilon. \end{aligned}$$

Symmetrically, we have $\mathrm{d}_B(A, A') \leq \mathrm{d}_B(B, B') + 2\varepsilon$ and thus, $\mathrm{distort}(R^k) \leq 2\varepsilon$ as desired. □

Thus, the space of bounded k-point persistence diagrams is nearly low-dimensional.

References

[1] A. Beygelzimer and J. Langford. Cover trees for nearest neighbor. pages 97–104, 2006.

[2] K. Clarkson. Nearest-Neighbor Searching and Metric Space Dimensions. In G. Shakhnarovich, T. Darrell, and P. Indyk, editors, *Nearest-Neighbor Methods in Learning and Vision*. The MIT Press, 2006.

[3] A. Efrat, A. Itai, and M. J. Katz. Geometry Helps in Bottleneck Matching and Related Problems. *Algorithmica*, 31(1):1–28, Sept. 2001.

[4] B. T. Fasy, X. He, Z. Liu, S. Micka, D. L. Millman, and B. Zhu. Approximate Nearest Neighbors in the Space of Persistence Diagrams. *arXiv:1812.11257 [cs]*, Mar. 2021. arXiv: 1812.11257.

[5] S. Har-Peled and M. Mendel. Fast Construction of Nets in Low-Dimensional Metrics and Their Applications. *SIAM Journal on Computing*, 35(5):1148–1184, Jan. 2006.