Nearly-Doubling Spaces of Persistence Diagrams

Donald R. Sheehy
Department of Computer Science, North Carolina State University, Raleigh, NC, USA

Siddharth S. Sheth
Department of Computer Science, North Carolina State University, Raleigh, NC, USA

Abstract

The space of persistence diagrams under bottleneck distance is known to have infinite doubling dimension. Because many metric search algorithms and data structures have bounds that depend on the dimension of the search space, the high-dimensionality makes it difficult to analyze and compare asymptotic running times of metric search algorithms on this space.

We introduce the notion of nearly-doubling metrics, those that are Gromov-Hausdorff close to metric spaces of bounded doubling dimension and prove that bounded k-point persistence diagrams are nearly-doubling. This allows us to prove that in some ways, persistence diagrams can be expected to behave like a doubling metric space. We prove our results in great generality, studying a large class of quotient metrics (of which the persistence plane is just one example). We also prove bounds on the dimension of the k-point bottleneck space over such metrics.

The notion of being nearly-doubling in this Gromov-Hausdorff sense is likely of more general interest. Some algorithms that have a dependence on the dimension can be analyzed in terms of the dimension of the nearby metric rather than that of the metric itself. We give a specific example of this phenomenon by analyzing an algorithm to compute metric nets, a useful operation on persistence diagrams.

2012 ACM Subject Classification Theory of computation → Computational geometry

Keywords and phrases Topological Data Analysis, Persistence Diagrams, Gromov-Hausdorff Distance

Digital Object Identifier 10.4230/LIPIcs.SoCG.2022.60

Funding This research was supported by the NSF under grant CCF-2017980.

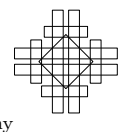
1 Introduction

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

A significant advantage of persistence diagrams 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 persistence diagrams, 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 [6, 10]. The metric space of persistence diagrams with the bottleneck distance is known to have infinite doubling dimension [8], making it unclear whether one ought to apply standard data structures such as cover trees [1] or net trees [10] to search in this space. Although the space of all persistence diagrams is infinite-dimensional, all hope is not lost. In this paper, we show that the bottleneck space of bounded persistence diagram (i.e., those whose points are in a bounded region) is close in a Gromov-Hausdorff sense to a finite-dimensional space. Our approach is to consider a very general class of quotient metrics generalizing the persistence plane and then bound the doubling dimension of bottleneck distances over such metrics. We also show that for some algorithms whose running time depends on the doubling dimension,


© Donald R. Sheehy and Siddharth S. Sheth;
licensed under Creative Commons License CC-BY 4.0



it can sometimes suffice to be close to a low-dimensional metric in order to achieve similar running times. Specifically, we show how to construct nets efficiently in these so-called nearly-doubling metrics.

As a first attempt at explaining why the bottleneck space of peristence diagrams appears to behave like a low-dimensional space in some experiments (see [16]), one might hope that "real-world" persistence diagrams naturally live in a low-dimensional subspace. Certainly, there are cases where data naturally live on a low-dimensional manifold and zooming in, one sees only the low-dimensional structure. However, this is not true of persistence diagrams. Zooming in can increase rather than decrease the apparent dimension. As a result, the key idea in this paper is not to look for a low-dimensional subspace, but rather a different low-dimensional space that is provably Gromov-Hausdorff close.

2 Related Work

There are numerous examples of metric search algorithms where search performance depends on the underlying space's doubling dimension. The performance guarantees of navigating nets in [14] depend on an exponential function of the doubling dimension. The same is true for Clarkson's *sb* data structure [6] and Har-Peled and Mendel's net-trees [10].

The bottleneck matching data structure of Efrat et al. [7] runs in time $O(n^{1.5}\log^d n)$ in $\mathbb{R}^d$ using ℓ_∞ distance. Kerber et al. [12] apply the geometric intuition of Efrat et al. [7] to the space of persistence diagrams and give the current state-of-the-art algorithm for computing the bottleneck distance between persistence diagrams. The running time is $O(n^{1.5}\log n)$. Kerber and Nigmetov also acknowledge the high dimensionality of some spaces as a problem when they build spanners that minimize distance computations for such spaces [13]. In their work, they explicitly mention persistence diagrams as a motivating example of an expensive to compute metric, but their theoretical results only apply to doubling metrics. Nigmetov [16] gave many experimental results showing that methods geared towards doubling spaces still work well on persistence diagrams. In this paper, we give some indication for why similar results could apply in the (non-doubling) setting of persistence diagrams.

Fasy et al. explore the infinite doubling dimension of persistence diagrams in [8] with a nearest neighbor data structure. They replace the bounded persistence plane with a grid to reduce the doubling dimension of the space of bounded persistence diagrams.

Our approach is based on the fact that the persistence plane is a quotient of the ℓ_∞ plane modulo the diagonal. This approach was first defined by Bubenik and Elchesen [2, 3]. They use this definition of the persistence plane in terms of quotient metrics to prove results on more general spaces of persistence diagrams.

Choudhary and Kerber [5] introduce the idea of a t-restricted doubling dimension where the dimension is computed only by focusing on balls of radius at most t. The notion of nearly-doubling metrics we introduce in this paper takes the opposite approach, capturing the doubling behavior at sufficiently large scales. This is more appropriate for persistence diagrams, because the high-dimensionality is present at arbitrarily small scales.

Huang et al. [11] present a similar result for clustering problems where they compute weighted approximations of subsets of doubling metrics in polynomial time.

3 Definitions

3.1 Metric Spaces

A *metric space*, (X, d) is a set X and a metric d. This is the default metric space used in this paper. 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 *$r$-ball centered at a*, denoted by $B(a, r)$, is the set of all points in X within distance at most r from a. The *spread* of a finite metric space is the ratio of its diameter to its smallest pairwise distance.

A collection of sets Y *covers* X if the union of the sets in Y contains X. An *r-cover* is a collection of sets of diameter at most $2r$ that covers X. A special case of an r-cover is a cover by metric balls of radius r. A *minimum r-cover* is an r-cover of X of minimum cardinality. The *covering number* of X is $\mathrm{N}_r(X) = |Y|$ where Y is a minimum r-cover of X. The *r-metric entropy* of X is defined as $\mathrm{H}_r(X) = \log_2 \mathrm{N}_r(X)$.

The *doubling dimension* of X, denoted $\dim(X)$, is the minimum number d such that every subset $S \subseteq X$ can be covered by 2^d sets of half the diameter of S. As observed in the original work on doubling dimension [14], a ball in a d-dimensional metric space can be covered with at most 2^{2d} balls of half the radius.[1] If $\dim(X)$ is finite, then X is a *doubling metric*. Throughout this paper, all mentions of dimension refer to the doubling dimension.

3.2 Packings and Coverings

A set $X_r \subset X$ is said to be *r-dense* or an *r-sample* of X if $X \subset \bigcup_{x \in X_r} B(x, r)$. A set $Z \subset X$ is said to be *r-separated* or an *r-packing* of X if $\mathrm{d}(z_i, z_j) > r$ for all distinct $z_i, z_j \in Z$. If $Z \subset X$ is both, r-dense and r-separated, then Z is an *r-net* of X.

The *packing number* of a set X, given by $\mathrm{M}_r(X)$, is the size of the maximum r-packing of X. The *sampling number* of X, given by $\mathrm{S}_r(X)$, is the size of the minimum r-sampling of X. There is a well-known relationship between the packing and covering numbers of a set X known from [15]. We present a proof for completeness.

▶ **Lemma 1** (Packing-Covering Duality). *If X is a metric set and r is some distance, then,*

$$\mathrm{M}_{2r}(X) \leq \mathrm{N}_r(X) \leq \mathrm{M}_r(X).$$

Proof. For the second inequality, let P be a maximum r-packing of X and $S = \bigcup_{p \in P} B(p, r)$ be such that S is not an r-cover of X. Thus, there exists $y \in X$ such that $\mathrm{d}(y, p) > r$ for all $p \in P$. Therefore, P is not a maximum r-packing of X and so $\mathrm{N}_r(X) \leq \mathrm{M}_r(X)$.

For the first inequality, let $Y = \{Y_1, \ldots Y_N\}$ be an r-cover of X of size $\mathrm{N}_r(X)$. Assume there exists P', a $2r$-packing of X, of size $\mathrm{N}_r(X) + 1$. By the pigeonhole principle there exists Y_i such that two elements of P', say p, p', are in Y_i because Y is an r-cover. Thus, $\mathrm{d}(p, p') \leq \mathrm{diam}(Y_i) \leq 2r$. Therefore, P' is not a 2ε-packing and so $\mathrm{M}_{2r}(X) \leq \mathrm{N}_r(X)$. ◀

A similar lemma holds for the covering number and sampling number.

▶ **Lemma 2.** *If X is a metric set and r is some distance, then*

$$\mathrm{S}_{2r}(X) \leq \mathrm{N}_r(X) \leq \mathrm{S}_r(X).$$

[1] In some prior work, the definition of doubling dimension is given in terms of coverage of metric balls rather than general covers. That definition suffers from several drawbacks; most notably, it is not monotone with respect to subsets.

Lemma 2 gives us a relationship between the doubling dimension computed using centered and uncentered balls of diameter $2r$.

Krauthgamer and Lee [14] say that the doubling dimension computed by covering a metric ball with balls of half the radius is a 2-approximation of the actual doubling dimension. The following lemma shows that the converse of that statement is also true.

▶ **Lemma 3.** *Let X be metric space. If, for any $r > 0$, there exists an $r/2$-sample of a ball $B(x, r)$ in X of cardinality 2^ρ, then* $\dim(X) \leq 2\rho$.

Proof. Let $Z \subset X$ be a set of diameter $2r$. Then $Z \subseteq B(z, 2r)$ for any $z \in Z$. So there exists Z', an r-sample of $B(z, 2r)$, of cardinality 2^ρ. Moreover, for every $z' \in Z'$ there exists an $r/2$-sample of cardinality 2^ρ of a ball $B(z', r)$. Therefore, there exists an $r/2$-sample of Z of cardinality at most $2^{2\rho}$. Thus, from Lemma 2 there exists an $r/2$-cover of Z of cardinality at most $2^{2\rho}$ and so $\dim(X) \leq 2\rho$. ◀

Krauthgamer and Lee [14] prove that an r-packing of an $O(r)$-ball has at most $2^{O(d)}$ points. A version of this lemma with more precise constants is the following.

▶ **Lemma 4** (Standard Packing Lemma)**.** *If X is a metric space of dimension d and $Z \subset B(x, r)$ for some $x \in X$ is an λ-packing then $|Z| \leq (2\Delta)^d$ where $\Delta \leq \frac{2r}{\lambda}$ is the spread of Z.*

Let X be a metric space and let Y be a subspace. The *quotient metric space* $(X/Y, \mathrm{d}_{X/Y})$ 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 $q(x) = [x]$.

3.3 Bottleneck Distance

Let X be a metric space and let A and B be two finite subsets of the same cardinality. A matching between A and B is bijection $m : A \to B$. The *bottleneck* of a matching m is

$$\max_{a \in A} \mathrm{d}(a, m(a)).$$

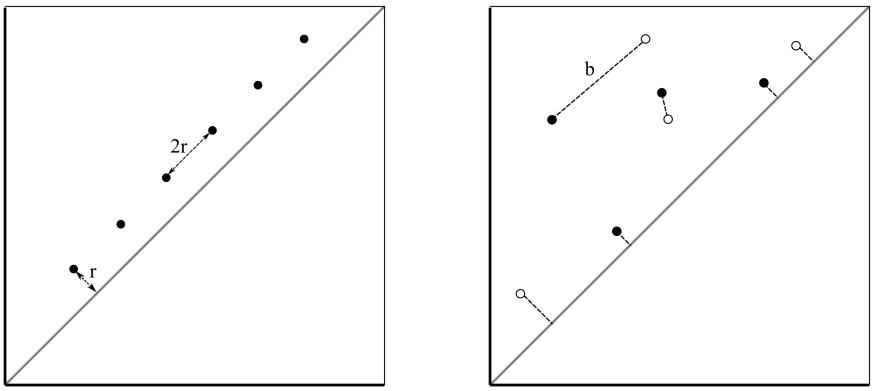
The *bottleneck distance* between A and B is the minimum of the bottleneck over all possible matchings between A and B.

3.4 The Persistence Plane

The *persistence plane* P is the quotient $(\mathbb{R}^2, \ell_\infty)$ modulo the diagonal $\{(x, x) \mid x \in \mathbb{R}\}$. The point associated with the equivalence class of the diagonal in the persistence plane is called the *diagonal point*. The dimension of P is infinite as shown in Figure 1a. This means that a quotient of two doubling metric spaces can be infinite-dimensional.

A *persistence diagram* is a multiset of points in the persistence plane. The natural metric on persistence diagrams is the bottleneck distance. To ensure diagrams A and B have the same cardinality, we augment A with $|B|$ copies of the diagonal point and we augment B with $|A|$ copies of the diagonal point. Then the *bottleneck distance for persistence diagrams* is the bottleneck distance between the augmented diagrams.

Treating the persistence plane as a quotient metric is due to Bubenik and Elchesen. [2]. Although this perspective is nonstandard, it provides several significant benefits. It simplifies algorithms for computing bottleneck distance, because having a single "point" representing the entire diagonal allows one to more easily perform augmentation compared to standard approaches [12]. It also simplifies sketching [17], in which one uses an approximate persistence diagram that has fewer distinct points with multiplicity.



(a) The Persistence Plane. **(b)** Bottleneck Matching.

Figure 1 (a) shows why the persistence plane has infinite doubling dimension. A ball of radius r centered at the diagonal would contain infinitely many points at distance r from the diagonal but a ball of radius $r/2$ centered off the diagonal can cover only one of them. (b) shows a bottleneck matching between two persistence diagrams.

3.5 Gromov-Hausdorff Distance

Given compact sets A and B in a metric space X, the *Hausdorff distance* between them is

$$\mathrm{d}_H(A,B) = \max\{\max_{a\in A}\min_{b\in B}\mathrm{d}(a,b), \max_{b\in B}\min_{a\in A}\mathrm{d}(a,b)\}.$$

For metric spaces (P,d_P) and (Q,d_Q), a *correspondence* between P and Q is a relation $\mathcal{R}\subseteq P\times Q$ such that for its canonical projections on P and Q, we have $\pi_P(\mathcal{R})=P$ and $\pi_Q(\mathcal{R})=Q$ respectively. The *distortion* of $\mathcal{R}$ is defined as

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

The *Gromov-Hausdorff distance*, d_{GH}, is a metric on compact metric spaces [9] defined as

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

In this paper we say two metric spaces are ε-close to mean that the Gromov-Hausdorff distance between them is at most ε. The Gromov-Hausdorff distance is a generalization of the Hausdorff distance in the sense that if P and Q are subsets of a common metric space, then their Gromov-Hausdorff distance is bounded by their Hausdorff distance. So if the Hausdorff distance between two subspaces of a metric space is bounded, the Gromov-Hausdorff distance between them is also bounded.

4 ε-Close Quotient Metric Spaces

A quotient metric space X/Y can have very high (or infinite) dimension even if X and Y are low-dimensional. A perfect example of this phenomenon is the persistence plane, which has infinite dimension despite being the quotient of a 2-dimensional space by a 1-dimensional

subspace. In this section, we 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 a finite subset.

▶ **Lemma 5.** *Let X be a d-dimensional metric space. If $Y \subset X$ is finite, then*

$$\dim(X/Y) \leq d + \log_2 |Y|.$$

Proof. Let $S \subseteq X/Y$ be such that $\operatorname{diam}(S) = 2\varepsilon$. Let $q : X \to X/Y$ be the quotient map. There exists a subset $S' \subseteq X$ such that $q(S') = S$. For $y \in Y$, define the Voronoi cell of y restricted to S' to be

$$\operatorname{Vor}_{|S'}(y) := \{x \in S' \mid \mathrm{d}(x, y) = \mathrm{d}(x, Y)\}.$$

Then, for each $y \in Y$, we have

$$\begin{aligned}
\operatorname{diam}(\operatorname{Vor}_{|S'}(y)) &:= \sup_{a,b \in \operatorname{Vor}_{|S'}(y)} \mathrm{d}(a, b) \\
&\leq \sup_{a,b \in \operatorname{Vor}_{|S'}(y)} \min\{\mathrm{d}(a, b), \mathrm{d}(a, y) + \mathrm{d}(b, y)\} \\
&= \sup_{a,b \in \operatorname{Vor}_{|S'}(y)} \min\{\mathrm{d}(a, b), \mathrm{d}(a, Y) + \mathrm{d}(b, Y)\} \\
&= \sup_{a,b \in \operatorname{Vor}_{|S'}(y)} \mathrm{d}_{X/Y}(q(a), q(b)) \\
&\leq \sup_{a,b \in S'} \mathrm{d}_{X/Y}(q(a), q(b)) \\
&= 2\varepsilon
\end{aligned}$$

So, $\operatorname{Vor}_{|S'}(y)$ is a set with diameter 2ε, and, by the definition of doubling dimension, has an $\varepsilon/2$-cover of size at most 2^d. Let C be the union of these covers for all $y \in Y$. Then C will $\varepsilon/2$-cover S' in X. Distances only decrease in the quotient, so the sets $\{q(U) \mid U \in C\}$ will $\varepsilon/2$-cover S in X/Y. So, we have an $\varepsilon/2$-cover of S of size at most $|Y|2^d$ and thus,

$$\dim(X/Y) \leq \log_2(|Y|2^d) = d + \log_2 |Y|. \qquad \blacktriangleleft$$

▶ **Theorem 6.** *Let X and Y be compact metric spaces such that $Y \subseteq X$ and $\dim(X) = d$. Then, X/Y is ε-close to a metric of dimension at most $d + \mathrm{H}_{\varepsilon/2}(Y)$.*

Proof. Let Y_ε be a minimum ε-sample of Y. Then by Lemma 2, the cardinality of the minimum $\varepsilon/2$-cover of Y is at least $|Y_\varepsilon|$. Therefore, $\mathrm{H}_{\varepsilon/2}(Y) \geq \log |Y_\varepsilon|$. So, Lemma 5 implies that X/Y_ε has dimension at most $d + \mathrm{H}_{\varepsilon/2}(Y)$. It will suffice to show that $\mathrm{d}_{GH}(X/Y, X/Y_\varepsilon) \leq \varepsilon$.

Let $q : X \to X/Y$ and $q_\varepsilon : X \to X/Y_\varepsilon$ denote the canonical quotient maps. Let $\mathcal{R} \subseteq X/Y \times X/Y_\varepsilon$ be the relation

$$\mathcal{R} = \{(q(x), q_\varepsilon(x)) \mid x \in X\}.$$

Quotient maps are surjective, so the canonical projections of $\mathcal{R}$ satisfy $\pi_{X/Y}(\mathcal{R}) = X/Y$ and $\pi_{X/Y_\varepsilon}(\mathcal{R}) = X/Y_\varepsilon$. Thus, $\mathcal{R}$ is a correspondence between X/Y and X/Y_ε.

Because Y_ε is an ε-sample of Y, for any $a \in X$, we have

$$\mathrm{d}(a, Y) \leq \mathrm{d}(a, Y_\varepsilon) \leq \mathrm{d}(a, Y) + \varepsilon.$$

It follows that

$$\begin{aligned}\mathrm{d}_{X/Y}(q(a),q(b)) &= \min\{\mathrm{d}(a,b),\mathrm{d}(a,Y)+\mathrm{d}(b,Y)\}\\ &\le \min\{\mathrm{d}(a,b),\mathrm{d}(a,Y_\varepsilon)+\mathrm{d}(b,Y_\varepsilon)\}\\ &= \mathrm{d}_{X/Y_\varepsilon}(q_\varepsilon(a),q_\varepsilon(b)),\end{aligned}$$

and also,

$$\begin{aligned}\mathrm{d}_{X/Y_\varepsilon}(q_\varepsilon(a),q_\varepsilon(b)) &= \min\{\mathrm{d}(a,b),\mathrm{d}(a,Y_\varepsilon)+\mathrm{d}(b,Y_\varepsilon)\}\\ &\le \min\{\mathrm{d}(a,b),\mathrm{d}(a,Y)+\mathrm{d}(b,Y)+2\varepsilon\}\\ &\le \mathrm{d}_{X/Y}(q(a),q(b))+2\varepsilon.\end{aligned}$$

We can then bound the distortion of $\mathcal{R}$ as follows.

$$\mathrm{distort}(\mathcal{R}) = \sup_{a,b\in X}\left|\mathrm{d}_{X/Y}(q(a),q(b)) - \mathrm{d}_{X/Y_\varepsilon}(q_\varepsilon(a),q_\varepsilon(b))\right| \le 2\varepsilon.$$

Because Y is compact, Y_ε is finite and X/Y_ε is the required ε-close space with doubling dimension at most $d+\mathrm{H}_{\varepsilon/2}(Y)$. ◀

Note that the preceding theorem does not directly apply to the persistence plane because it is not compact. We resolve this issue in Section 9 using bounded persistence diagrams.

5 Nearly-Doubling Metric Spaces

A metric space X is *ε-nearly-doubling* if there exists a doubling metric space Y such that $\mathrm{d}_{GH}(X,Y)\le\varepsilon$. In the previous section we showed that quotients of a doubling metric by a compact set are ε-nearly-doubling with a dimension that depends on ε. In later sections, we will show how bottleneck spaces are also nearly doubling with a focus on subsets of persistence diagrams. Before proceeding to those results, we explain the sense in which nearly doubling metrics share some of the properties of doubling metrics. In particular, they can behave like doubling metrics down to scale $O(\varepsilon)$. The most useful fact about doubling metrics is that they satisfy the packing property described in Lemma 4. The following lemma shows how to bound the size of packings of sufficiently large balls in nearly-doubling metrics.

▶ **Lemma 7** (Nearly-Doubling Packing Lemma). *Let $r,\lambda\in\mathbb{R}$ be such that $\lambda<r$. Let S be a λ-packing of a ball $B(c,r)$ in a metric space (X,d). Let (X',d') be a d-dimensional metric space such that $\mathrm{d}_{GH}(X,X')\le\varepsilon$. If $\lambda=\alpha\varepsilon$ for some $\alpha>2$, then $|X|\le\left(\frac{2\alpha+2}{\alpha-2}\Delta\right)^d$ where $\Delta\le\frac{2r}{\lambda}$ is the spread of S.*

Proof. Because $\mathrm{d}_{GH}(X,X')\le\varepsilon$ there exists a correspondence $\mathcal{R}$ between X and X' such that $|\mathrm{d}(a,b)-\mathrm{d}'(a',b')|\le 2\varepsilon$ for all $(a,a'),(b,b')\in\mathcal{R}$. For each $x\in S$, choose $f(x)\in X'$ to be a point such that $(x,f(x))\in\mathcal{R}$. Let $S'=\{f(x)\mid x\in S\}$. For any $a,b\in S$,

$$|\mathrm{d}(a,b)-\mathrm{d}'(f(a),f(b))|\le 2\varepsilon.$$

Because S is a λ-packing, we have that for all $a,b\in S$,

$$\begin{aligned}\mathrm{d}'(f(a),f(b)) &\ge \mathrm{d}(a,b)-2\varepsilon\\ &\ge \lambda-2\varepsilon.\end{aligned}$$

In other words, distinct points of S map to points of distance at least $\lambda - 2\varepsilon > 0$. It follows that f is a bijection and S' is a $(\lambda - 2\varepsilon)$-packing. The distortion bound on $\mathcal{R}$ implies that

$$\begin{aligned}\operatorname{diam}(S') &= \sup_{a,b \in S} \mathrm{d}'(f(a), f(b)) \\ &\leq \sup_{a,b \in S} \mathrm{d}(a,b) + 2\varepsilon \\ &= \operatorname{diam}(S) + 2\varepsilon \\ &\leq 2r + 2\varepsilon.\end{aligned}$$

So, the spread Δ' of S' is at most $\frac{2r+2\varepsilon}{\lambda - 2\varepsilon}$. Using the fact that $\alpha\varepsilon = \lambda < r$, we get the following bound on Δ' in terms of the spread Δ of S.

$$\Delta' \leq \frac{2r + 2\varepsilon}{\lambda - 2\varepsilon} = \frac{2r + \frac{2\lambda}{\alpha}}{\frac{\alpha-2}{\alpha}\lambda} < \frac{2r\alpha + 2r}{(\alpha-2)\lambda} \leq \frac{\alpha+1}{\alpha-2}\Delta.$$

We then use the fact that f is bijection and apply Lemma 4, to get

$$|S| = |S'| \leq \left(\left(\frac{2\alpha+2}{\alpha-2}\right)\Delta\right)^d. \qquad \blacktriangleleft$$

The nearly-doubling packing lemma explains why algorithms and data structures defined for doubling metrics work for nearly-doubling metrics down to some scale. We give a specific example and analysis in the following section.

6 Clarkson's Algorithm in Nearly-Doubling Spaces

The main theme of this paper is that although some metric spaces are high-dimensional, they are Gromov-Hausdorff close to low-dimensional metrics. We showed this is true for a wide class of compact quotient metrics in Section 4 and will extend these results to the bottleneck space of bounded persistence diagrams in Section 9. Before we tackle those problems, we will show that being close to a low-dimensional metric has some benefit. In particular, there are basic algorithms for doubling metrics that will also be efficient in nearly-doubling metrics.

In this section we analyze the performance of an algorithm for constructing a λ-net in a nearly-doubling metric space. The main result will be that as long as $\lambda \geq 3\varepsilon$, the running time can be bounded in terms of the dimension of an ε-close metric.

The algorithm we will consider for computing the net is sometimes called Clarkson's Algorithm. It is a variation of an algorithm originally due to Clarkson [6] with some simplifications due to Har-Peled and Mendel [10] and Sheehy [19]. The idea is to produce a net by greedy sampling (also known as farthest point sampling or Gonzalez ordering). Any point may be selected first and each subsequent point maximizes the distance to the points selected so far, stopping when the distance is less than the target scale λ. An open source Python implementation is available [18]. Given a finite subspace P of a doubling metric space X with cardinality n, the algorithm computes a net of P in time $O\left(n \log \frac{\operatorname{diam}(P)}{\lambda}\right)$. The big-O hides terms that are exponential in the dimension, but if the dimension is too high, the simpler upper bound of $O(n^2)$ applies. So, for inputs with polynomial spread in doubling metrics, the running time is $O(n \log n)$. Thus, our goal is to show that similar guarantees hold in nearly-doubling metrics.

The algorithm follows an incremental construction of the greedy sampling. The points in the net will be numbered $p_0, p_1, \ldots$. The first point p_0 is chosen arbitrarily. Let $P_i := \{p_0, \ldots, p_{i-1}\}$ be the ith *prefix*, and $\lambda_i := \mathrm{d}(p_i, P_i)$ be the *insertion radius* of p_i. For every

point $p \in P_i$ the algorithm maintains a list of $q \in P \backslash P_i$ that are the reverse nearest neighbors of p. Essentially, this is the Voronoi cell of p. A *neighbor graph* is defined on the Voronoi cells that is guaranteed to have an edge (p_i, p_j) if adding a point in the Voronoi cell of p_i can affect the Voronoi cell of p_j. At each step i the algorithm has the points of P_i in a max heap with the key of a point p_a given by the distance from p_a to the farthest point in its Voronoi cell. The algorithm simply pops a point p_a from the heap, and adds the farthest point p_i to the net. The Voronoi cells and the neighbor graph are updated. The neighbor graph stores exactly the cells that could change so one only needs to check the Voronoi cells of the neighbors of p_a. New edges in the neighbor graph incident to p_i can be found among the 2-hop neighbors (i.e., neighbors of neighbors) of p_a. A key insight to make the algorithm efficient is to keep some extra edges (p_a, p_b) in the graph as long as $\mathrm{d}(p_a, p_b) \leq 3\lambda_i$. Clarkson showed that the desired neighbors will all satisfy such a condition.

▶ **Theorem 8.** *Let ε and λ be such that $\lambda \geq 3\varepsilon$. If X is ε-close to a d-dimensional metric space, then Clarkson's algorithm computes a λ-net of X in $2^{O(d)} n \log_2(n\frac{\mathrm{diam}(X)}{\lambda})$ time.*

Proof. There are three aspects of the algorithm that must be analyzed: the update to the neighbor graph, the heap operations, and the update to the Voronoi cells. In the ith iteration, the points P_i form a λ_i-net. So, Lemma 7 and the condition that $\mathrm{d}(p_a, p_b) \leq 3\lambda_i$ for neighbors p_b of p_a imply that the degree of p_a is $2^{O(d)}$. This means that updating the neighbor graph takes constant time per point. It also means that the number of keys to update in the heap is constant per iteration. So, the heap operations take $2^{O(d)} n \log_2 n$ time in the worst case.

To analyze the number of distance computations performed when updating the Voronoi cells, we apply an analysis similar to that used by Har-Peled and Mendel [10]. For each point $x \in P$, we want to count how many times we compute the distance from q to the newly inserted point p_i (to see if it should change Voronoi cells). In such cases, we say p_i touches x.

Let $x \in \mathrm{Vor}(p_k)$ be touched by newly inserted point $p_i \in \mathrm{Vor}(p_j)$.

$$\begin{aligned}\mathrm{d}(x, p_i) &\leq \mathrm{d}(x, p_k) + \mathrm{d}(p_k, p_i) \\ &\leq \mathrm{d}(x, p_k) + \mathrm{d}(p_k, p_j) + \mathrm{d}(p_j, p_i) \\ &\leq \lambda_i + 3\lambda_i + \lambda_i = 5\lambda_i.\end{aligned}$$

For an integer m, define the annulus $A_m = \{p_i \mid 2^m \leq \lambda_i < 2^{m+1} \text{ and } p_i \text{ touches } x\}$. If $p_i \in A_m$ then $\mathrm{d}(x, p_i) \leq 5\lambda_i \leq 5 \cdot 2^{m+1}$. So $A_m \subset B(x, 5 \cdot 2^{m+1})$. Moreover A_m is 2^m-separated. Therefore, by Lemma 7, $|A_m| \leq 2^{O(d)}$. Thus, x is touched at most a constant number of times in each annulus. The algorithm stops as soon as λ_i is smaller than λ, so, the number of nonempty annuli that can contain x is at most $\log_2 \frac{\mathrm{diam}(P)}{\lambda}$. It follows that the total work of updating the Voronoi cells takes $2^{O(d)} n \log_2 \frac{\mathrm{diam}(P)}{\lambda}$ time.

Combining the running time of the graph update, the heap operations and the cell updates, we get a total running time of $2^{O(d)} n \log_2 \left(n\frac{\mathrm{diam}(P)}{\lambda}\right)$. ◀

7 Bottleneck Metrics

If the doubling dimension of X is d, then a *d-dimensional k-point diagram* is a set of k elements of X. Let $X^{(k)}$ be the space of k-point diagrams in X with the bottleneck metric.

▶ **Theorem 9.** *If X is a d-dimensional metric space, then for all integers $k \geq 1$, we have $\dim(X^{(k)}) \leq 4kd$.*

Proof. Let $D \in X^{(k)}$ and positive $r \in \mathbb{R}$ be chosen arbitrarily. It will suffice to construct an $r/2$-sample of $B(D, r)$ of size 2^{2kd}. For each point $p_i \in D$, there is an $r/2$-sample $\{x_{i,j}\}_{j\in[2^{2d}]}$ of $B(p_i, r)$ in X. For $j : [k] \to [2^{2d}]$, let

$$C_j := \{x_{i,j(i)} \mid i \in [k]\}.$$

Assume all diagrams $A = \{a_i\}_{i\in[k]}$ are indexed so the bottleneck matching with D has a_i matched to p_i. This means that each $a_i \in A$ is in $B(p_i, r)$. If $j(i)$ is the index of the nearest point in the sample of $B(p_i, r)$, then there is a matching $A \to C_j$ with bottleneck at most $r/2$. So, the set $C = \{C_j \mid j : [k] \to [2^{2d}]\}$ is an $r/2$-sample of $B(D, r)$. Clearly, $|C| = 2^{2kd}$, so the dimension of $X^{(k)}$ is at most $2\log_2(|C|) = 4kd$. ◀

If the bottleneck space is over a quotient metric, then Lemma 5 and Theorem 9 together yield the following corollary.

▶ **Corollary 10.** *Let X/Y be a quotient metric induced by a finite subspace Y over X. Then,* $\dim(X/Y^{(k)}) \leq 4k(d + \log_2 |Y|)$.

For many metric spaces such as ℓ_p-spaces, maximal sets with a fixed diameter are metric balls. In such metrics, or if the doubling dimension is defined in terms of metric balls (as opposed to general covers), there is no need for the factor of 4 in the dimension for the preceding two results. In particular this holds in the persistence plane.

For bottleneck spaces defined over nearly doubling metrics, it is useful to have the following theorem showing that the mapping from metric spaces to bottleneck spaces is Lipschitz.

▶ **Theorem 11.** *If X and Y are compact metric spaces, then for all integers $k \geq 1$,*

$$\mathrm{d}_{GH}(X^{(k)}, Y^{(k)}) \leq \mathrm{d}_{GH}(X, Y).$$

Proof. Let $\mathcal{R}$ be a minimum distortion correspondence between X and Y. Let 2ε be the distortion of $\mathcal{R}$. Let $[k] = \{0, \ldots k-1\}$. Let $\mathcal{R}^{(k)}$ denote the correspondence between $X^{(k)}$ and $Y^{(k)}$ defined as

$$\mathcal{R}^{(k)} = \{(\{a_i\}_{i\in[k]}, \{b_i\}_{i\in[k]}) \mid \exists \text{ bijection } m : [k] \to [k] \text{ s.t. } \forall i, (a_i, b_{m(i)}) \in \mathcal{R}\}.$$

To show that $\mathrm{d}_{GH}(X^{(k)}, Y^{(k)}) \leq \varepsilon$, it is sufficient to bound the distortion of $\mathcal{R}^{(k)}$.

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

$$\mathrm{d}_B(A, A') = \max_{i\in[k]} \mathrm{d}_X(a_i, a'_{\eta(i)}).$$

It follows that

$$\begin{aligned}
\mathrm{d}_B(B, B') &\leq \max_{j\in[k]} \mathrm{d}_Y(b_j, b'_{\eta(j)}) \\
&\leq \max_{j\in[k]} (\mathrm{d}_X(a_j, a'_{\eta(j)}) + 2\varepsilon) \\
&= \mathrm{d}_B(A, A') + 2\varepsilon.
\end{aligned}$$

Symmetrically, we have $\mathrm{d}_B(A, A') \leq \mathrm{d}_B(B, B') + 2\varepsilon$ and thus, $\mathrm{distort}(\mathcal{R}^{(k)}) \leq 2\varepsilon = \mathrm{distort}(\mathcal{R})$. To conclude, we observe that

$$\mathrm{d}_{GH}(X^{(k)}, Y^{(k)}) \leq \frac{1}{2}\mathrm{distort}(\mathcal{R}^{(k)}) \leq \frac{1}{2}\mathrm{distort}(\mathcal{R}) = \mathrm{d}_{GH}(X, Y).$$ ◀

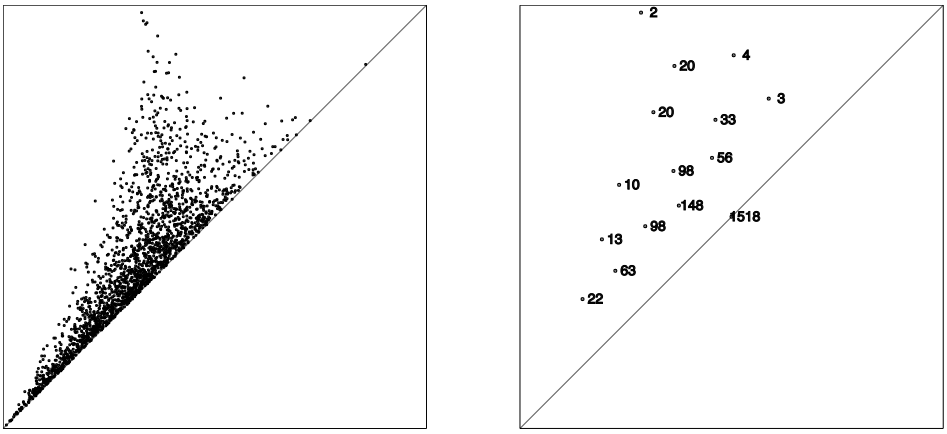


Figure 2 The image on the left shows a persistence diagram for points sampled on a sphere. The image on the right shows a sketch of that persistence diagram with first 14 points. The number to the right of each point shows its multiplicity in the sketch.

8 Bottleneck Spaces with Multiplicity

A k-point diagram D with multiplicity is a set $\underline{D} \subseteq X$ of cardinality k and a function $m_D : \underline{D} \to \mathbb{Z}_+$. The *total multiplicity* of D is $m_D = \sum_{p \in \underline{D}} m_D(p)$. In this section, we consider the space $X^{(k,N)}$ of k-point diagrams with total multiplicity N. This may be viewed as a subset of $X^{(N)}$, consisting of those diagrams with at most k distinct points. In Theorem 12, we show that $X^{(k,N)}$ has a dimension that depends only logarithmically on N.

The motivation for studying such diagrams with multiplicity again comes from persistence diagrams. It often happens that points in a persistence diagram have multiplicity. Recently, it was shown that actively seeking such multiplicity can lead to efficient sketches of persistence diagrams [17].

A simple sketching algorithm is to run Clarkson's Algorithm (see Section 6) on a persistence diagram starting with the diagonal point until k points have been added. The algorithm maintains the Voronoi cells of the points in the net and therefore one simply sets the multiplicity of each point to be the number of points in its Voronoi cell. The result is a k-point sketch, D_k, of a diagram D. It is then straightforward to show that $\mathrm{d}_B(D, D_k)$ is at most $\mathrm{d}_H(\underline{D}, \underline{D_k})$ [17]. The advantage of the sketch is that it is a guaranteed approximation and can be represented in much less size. In some cases (i.e., for $k = O(\log n)$) it is asymptotically faster to compute the bottleneck distance between sketches than the full diagrams. There is nothing special about persistence diagrams in this algorithm. An example of a sketch is shown in Figure 2.

If $D \in X^{(N)}$, then $D_k \in X^{(k,N)}$. Theorem 9 gives a bound of $4Nd$ on the dimension of $X^{(N)}$. However, as we show in the theorem below, the sketch will live in a lower dimensional space.

▶ **Theorem 12.** *Let X be a d-dimensional metric space. If k and N are positive integers such that $k \leq N$, then* $\dim(X^{(k,N)}) \leq \min\{4Nd, 2k(2d + \log_2(2Nk))\}$

Proof. Let $C \in X^{(k,N)}$ and $r \in \mathbb{R}$ be with $r > 0$ be chosen arbitrarily. We will construct an $r/2$-cover of $B(C, r)$ in $X^{(k,N)}$ by constructing an $r/2$-sample as follows. For each $p \in \underline{C}$ there exists an $r/2$-sample U_p of $B(p, r)$ of size at most 2^{2d}. This means that if $\mathrm{d}(x, p) \leq r$, then for some $u_i \in U_p$, we have $\mathrm{d}(x, u_i) \leq r/2$.

Let $\mathcal{U} = \cup_{p \in \underline{C}} U_p$. Because $|\underline{C}| = k$, we know that $|\mathcal{U}| \leq 2^{2d}k$. Let $S \subseteq X^{(k,N)}$ be defined as

$$S := \{D \mid \underline{D} \subset \mathcal{U}, |\underline{D}| \leq k, m_D = N\}.$$

For S to be an $r/2$-sample of $B(C,r)$ we will show that for all $E \in B(C,r)$ there exists $D \in S$ such that $\mathrm{d}_B(D,E) \leq r/2$. Let $E = (\underline{E}, m_E)$ be any diagram in $B(C,r)$. For every $q \in \underline{E}$, there exists $p \in \underline{C}$ such that $\mathrm{d}(p,q) \leq r$. So, there exists $q' \in U_p$ such that $\mathrm{d}(q,u_i) \leq r/2$ for some $u_i \in U_p$.

Consider a diagram $D = (\underline{D}, m_D)$ where $\underline{D} = \{q' \mid q \in \underline{E}\}$ and $m_D(q') = m_E(q)$ for all $q' \in \underline{D}$. By construction $D \in S$. The bottleneck distance is bounded as follows

$$\mathrm{d}_B(D,E) \leq \max_{q \in \underline{E}} \mathrm{d}(q,q') \leq r/2.$$

It follows that S is an $r/2$-sample.

We bound the size of S as follows. Because $|\mathcal{U}| \leq 2^{2d}k$, there are at most $\binom{2^{2d}k}{k} \leq k^k 2^{2kd}$ different choices of $\underline{D}$ for a diagram in S. The number of ways to distribute multiplicity N over the k points of $\underline{D}$ is $\binom{N+k-1}{k-1} \leq (2N)^k$, because $N \geq k$. It then follows that

$$|S| \leq k^k 2^{2kd} (2N)^k = (2Nk2^{2d})^k.$$

So, the doubling dimension is at most

$$2\log_2(|S|) \leq 2\log_2(2Nk2^{2d})^k = 2k(2d + \log_2(2Nk)).$$

On the other hand, treating the diagram as a collection of N points without multiplicity and applying the bounds for diagrams without multiplicity (Theorem 9) yields a dimension at most $4Nd$. Combining these two upper bounds on the dimension completes the proof. ◀

9 The 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.

The persistence plane is denoted by $\mathsf{P} = (\mathbb{R}^2, \ell_\infty)/\{(x,x) \mid x \in \mathbb{R}\}$. Let P_0 denote the bounded persistence plane obtained by restricting P to $[0,1] \times [0,1]$. Then, $\mathsf{P}_0^{(N)}$ is the bottleneck space of N-point *bounded* persistence diagrams.

The key to finding low-dimensional spaces near $\mathsf{P}_0^{(N)}$ is to first find a low-dimensional space near the persistence plane. Theorem 6 gives a recipe for doing so. There is an ε-sample of the diagonal of the bounded persistence plane of size $\lceil \frac{1}{2\varepsilon} \rceil$. So, one can consider the plane modulo the ε-sample rather than modulo the whole diagonal. The resulting metric space is denoted P_ε. It is a special case of the construction in Theorem 6, and thus the following lemma is immediate.

▶ **Lemma 13.** *For all $\varepsilon > 0$, $\dim(\mathsf{P}_\varepsilon) \leq 2 + \log_2 \lceil \frac{1}{2\varepsilon} \rceil$ and $\mathrm{d}_{GH}(\mathsf{P}_0, \mathsf{P}_\varepsilon) \leq \varepsilon$.*

▶ **Theorem 14.** *The bottleneck space of N-point bounded persistence diagrams, $\mathsf{P}_0^{(N)}$ is ε-close to a space of dimension at most $4N(2 + \log_2 \lceil \frac{1}{2\varepsilon} \rceil)$.*

Proof. By Theorem 12 and Lemma 13,

$$\dim(\mathsf{P}_\varepsilon^{(N)}) \leq 4N \dim(\mathsf{P}_\varepsilon) \leq 4N(2 + \log_2 \left\lceil \frac{1}{2\varepsilon} \right\rceil).$$

Moreover, Theorem 11 implies that

$$\mathrm{d}_{GH}(\mathsf{P}_0^{(N)}, \mathsf{P}_\varepsilon^{(N)}) \leq \mathrm{d}_{GH}(\mathsf{P}_0, \mathsf{P}_\varepsilon) \leq \varepsilon.$$ ◀

Thus, the space of bounded N-point persistence diagrams is nearly low-dimensional. We can further lower the dimension of the space using sketching. Having fewer points with multiplicity decreases the dimension.

▶ **Lemma 15.** *For all positive integers N, k such that $N \geq k$,*

$$\mathrm{d}_{GH}(\mathsf{P}_0^{(N)}, \mathsf{P}_0^{(k,N)}) \leq \sqrt{\frac{1}{2k}}.$$

Proof. Given an N-point diagram D, the greedy sketching algorithm produces a k-point diagram D_k with mass N. The bottleneck distance is well-defined for all persistence diagrams, so it will suffice to bound the Hausdorff distance. As $\mathsf{P}_0^{(k,N)}$ is a subspace of $\mathsf{P}_0^{(N)}$, the Hausdorff distance will be the maximum of $\mathrm{d}_B(D, D_k)$ over all bounded N-point persistence diagrams. The greedy sketch produces for each k, an ε_k-net of $\underline{D}$ with multiplicities so that $\mathrm{d}_B(D, D_k) = \varepsilon_k$. The maximum size of an ε_k-net in P_0 restricted to the region above the diagonal is $\frac{1}{2\varepsilon_k^2}$. It follows that $k \leq \frac{1}{2\varepsilon_k^2}$ and therefore, $\varepsilon_k \leq \sqrt{\frac{1}{2k}}$. So, for all bounded N-point persistence diagrams D, we have $\mathrm{d}_B(D, D_k) \leq \sqrt{\frac{1}{2k}}$ and so the Gromov-Hausdorff distance bound follows. ◀

We can now combine the previous results to prove the following theorem.

▶ **Theorem 16.** *The space $\mathsf{P}_0^{(N)}$ of bounded N-point persistence diagrams is $(\varepsilon + \sqrt{\frac{1}{2k}})$-close to a metric of dimension at most $2k(4 + 2\log_2 \lceil \frac{1}{2\varepsilon} \rceil + \log_2(2Nk))$.*

Proof. First, the triangle inequality, Lemma 15, and Theorem 6 that

$$\mathrm{d}_{GH}(\mathsf{P}_0^{(N)}, \mathsf{P}_\varepsilon^{(k,N)}) \leq \mathrm{d}_{GH}(\mathsf{P}_0^{(N)}, \mathsf{P}_0^{(k,N)}) + \mathrm{d}_{GH}(\mathsf{P}_0^{(k,N)}, \mathsf{P}_\varepsilon^{(k,N)}) \leq \sqrt{\frac{1}{2k}} + \varepsilon.$$

Then, Theorem 12 and Lemma 13 implies

$$\begin{aligned}\dim(\mathsf{P}_\varepsilon^{(k,N)}) &\leq 2k(2\dim(\mathsf{P}_\varepsilon) + \log_2(2Nk)) \\ &\leq 2k(2\dim(\mathsf{P}_\varepsilon) + \log_2(2Nk)) \\ &\leq 2k(4 + 2\log_2 \left\lceil \frac{1}{2\varepsilon} \right\rceil + \log_2(2Nk)).\end{aligned}$$ ◀

10 Conclusion

In this paper, we analyze several generalizations of metric spaces that arise naturally in topological data analysis, with the goal of bounding their dimension. Although the most significant of these, the bottleneck distance for persistence diagrams is infinite-dimensional, we show that in an important sense, it can behave like a low-dimensional space.

The idea of analyzing the running time of an algorithm in terms of the dimension of a nearby metric leads to many natural questions. For example, it should be possible to build linear-size spanners with ε (additive) slack if the input is ε-close to a doubling metric by a direct application of the ideas from Section 6. It is interesting to ask what other metric constructions that are known to be efficient in doubling metrics are also efficient in nearly-doubling metrics.

Although our general study of bottleneck spaces over quotient metrics was primarily motivated by the special case of persistence diagrams, this is not the only example. Other methods in topological data analysis produce different quotient metrics of the type studied in this paper, for example in the work of Carrière and Oudot on Mapper [4]. It remains to find more such examples. It also remains to consider more general quotient metrics, i.e., those defined by an arbitrary equivalence relation rather than just a subset.

Lastly, the results of this paper imply that in many cases, one could hope that metric analysis on collections of persistence diagrams is a reasonable thing to do. Not only will the entropy of the collection be bounded, many standard algorithms designed for doubling metrics should work well without change.

References

1 Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 97–104, New York, NY, USA, 2006. Association for Computing Machinery. doi:10.1145/1143844.1143857.

2 Peter Bubenik and Alex Elchesen. Universality of persistence diagrams and the bottleneck and wasserstein distances, 2021. arXiv:1912.02563.

3 Peter Bubenik and Alex Elchesen. Virtual persistence diagrams, signed measures, wasserstein distances, and banach spaces, 2021. arXiv:2012.10514.

4 Mathieu Carrière and Steve Oudot. Structure and stability of the 1-dimensional mapper. In *SoCG*, 2016.

5 Aruni Choudhary and Michael Kerber. Local doubling dimension of point sets. In *CCCG 2015 Proceedings*, 2015.

6 Kenneth L. Clarkson. Nearest neighbor searching in metric spaces: Experimental results for 'sb(s)'. Preliminary version presented at ALENEX99, 2003.

7 A. Efrat, A. Itai, and M. J. Katz. Geometry Helps in Bottleneck Matching and Related Problems. *Algorithmica*, 31(1):1–28, September 2001. doi:10.1007/s00453-001-0016-8.

8 Brittany Terese Fasy, Xiaozhou He, Zhihui Liu, Samuel Micka, David L. Millman, and Binhai Zhu. Approximate Nearest Neighbors in the Space of Persistence Diagrams. *arXiv:1812.11257 [cs]*, March 2021. arXiv:1812.11257.

9 Misha Gromov. *Metric Structure for Riemannian and Non-Riemannian Spaces*. Birkhauser, 1999.

10 Sariel Har-Peled and Manor Mendel. Fast Construction of Nets in Low-Dimensional Metrics and Their Applications. *SIAM Journal on Computing*, 35(5):1148–1184, January 2006. doi:10.1137/S0097539704446281.

11 Lingxiao Huang, Shaofeng H.-C. Jiang, Jian Li, and Xuan Wu. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 814–825, 2018. doi:10.1109/FOCS.2018.00082.

12 Michael Kerber, Dmitriy Morozov, and Arnur Nigmetov. Geometry Helps to Compare Persistence Diagrams. *ACM Journal of Experimental Algorithmics*, 22:1–20, December 2017. doi:10.1145/3064175.

13 Michael Kerber and Arnur Nigmetov. Metric spaces with expensive distances. *International Journal of Computational Geometry and Applications*, 30(02):141–165, June 2020. doi:10.1142/S0218195920500077.

14 Robert Krauthgamer and James R. Lee. Navigating nets: Simple algorithms for proximity search. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '04, pages 798–807, USA, 2004. Society for Industrial and Applied Mathematics.

15 G. G. Lorentz. Metric entropy and approximation. *Bulletin of the American Mathematical Society*, 72(6):903–937, 1966. doi:bams/1183528486.

16 Arnur Nigmetov. *Comparison of Topological Summaries*. PhD thesis, TU Graz, 2019.

17 Don Sheehy and Siddharth Sheth. Sketching persistence diagrams. In *SoCG*, 2021.

18 Donald R. Sheehy. greedypermutations, 2020. URL: https://github.com/donsheehy/greedypermutation.

19 Donald R. Sheehy. One hop greedy permutations. In *Proceedings of the 32nd Canadian Conference on Computational Geometry*, pages 221–225, 2020.