DIFFERENTIATING GEOMETRIC STRUCTURE OF POINT CLOUD DISTRIBUTIONS USING PERSISTENT HOMOLOGY

JASON MAO

Abstract. The field of topological data analysis aims to characterize datasets by their topological structures. In particular, representative tools such as persistent homology and persistence landscapes are used to condense the information provided by a shape or a point cloud into a more compact form that highlights structural properties of the data. These tools have previously been used to analyze global properties of shapes, such as their connectivity or their genus. Our work shows the potential of these tools to capture the local, geometric properties of shapes, such as the sharpness of its angles. Furthermore, we prove a theoretical result on how numerical metrics on persistence landscapes can capture geometric distinctions between point cloud distributions with arbitrarily high probability.

1. INTRODUCTION

Topological data analysis (TDA) determines and compares structural properties of large amounts of data that would typically be difficult to analyze through classical algorithmic means. One such tool used in TDA is the concept of persistent homology, in which point clouds are analyzed by considering how their homology groups change with respect to a parameter, such as time or a radius [4]. TDA is very successful in determining the shape of data, and information extracted from TDA may be translated into objects such as persistence diagrams [6, 7] and decorated merge trees [2], upon which we can more precisely determine, say, a numerical value representing how different two clouds of data might be.

Most prior work in TDA has focused on how machinery in TDA can reveal global, topological distinctions between point clouds, such as differences in the number of topological "holes" formed by the general shapes of point clouds [2]. In contrast, this paper seeks to identify how TDA can reveal local, geometric distinctions between point clouds, such as differences in the sharpness of angles formed by the general shapes of point clouds. We consider how tools such as persistence diagrams and persistence landscapes can provide useful metrics for numerically qualifying these geometric differences.

We begin with experimental results. We evaluate the efficacy of persistence landscapes of point cloud samples in distinguishing point cloud distributions in two regimes: in one, we consider point cloud distributions taken noisily from the perimeters of isosceles triangles with varying base angles, and in the second, we consider point cloud distributions taken noisily from the perimeters of regular polygons with varying numbers of sides. This allows us to deduce that metrics on persistence landscapes are indeed capable of distinguishing geometric figures based solely on the angle measurements present within their point cloud distributions. We also evaluate how certain statistics of birth and death times (including averages, standard deviations, and sums) derived from persistence diagrams offer useful distinguishing information about the distributions from which point clouds were sampled.

We then prove a theoretical result concerning the relation between topological metrics on persistence landscapes and geometric metrics on point cloud distributions (specifically, the Wasserstein distance W_∞). In particular, we show that if X and Y are sufficiently large point clouds sampled from distributions μ and ν, and λ_X and λ_Y are the persistence landscapes associated with X and

Y, then $\|\lambda_X - \lambda_Y\|_\infty \leq O(W_\infty(\mu, \nu))$ holds with high probability. This theoretical result provides concrete justification for the high-probability effectiveness of persistence landscapes as a tool for computationally capturing geometric differences in point cloud distributions.

This paper is organized as follows. In Section 2, we provide an expository discussion of topological definitions and background that form the foundation behind TDA. In Section 3, we discuss the tools of persistence diagrams and persistence landscapes. Experimental results are then provided in Section 4, which justify the practical robustness of persistence diagrams and persistence landscapes in distinguishing point clouds. In Section 5, we state and prove a theorem on how metrics between persistence landscapes relate to the Wasserstein distance between point cloud distributions. Finally, in Section 6, we discuss some potential directions for further investigation of this topic.

2. BACKGROUND: SIMPLICIAL COMPLEXES AND HOMOLOGY

The most useful tool in identifying topological features of spaces will be the computation of homology groups. Throughout this paper, homology groups will be considered within the context of simplicial complexes.

Definition 2.1. Consider a set of affinely independent points $v_0, v_1, \ldots, v_k \in \mathbb{R}^d$, i.e. $\{v_i - v_0 : i \in \{1, \ldots, k\}\}$ is a set of linearly independent vectors in $\mathbb{R}^d$. We define a *k-simplex* $\sigma = \langle v_0, v_1, \ldots, v_k \rangle$ to be the convex hull of its *vertices* $\{v_0, v_1, \ldots, v_k\}$. k is the *dimension* of σ, and a *face* of a simplex σ is any simplex generated by some non-empty subset of its vertices.

In particular, a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron. Simplices can be used to model more complex topological objects by effectively gluing them together, thus forming a *simplicial complex*. For our purposes, we prefer to view simplicial complexes from a more abstract perspective, as described in Definition 2.2.

Definition 2.2. An *abstract simplicial complex* K is a finite collection of sets such that if $\sigma \in K$, and $\tau \subseteq \sigma$ is non-empty, then $\tau \in K$.

Any abstract simplicial complex may be realized geometrically by taking each set $\sigma \in K$ of size k to be the convex hull of k affinely independent points in some high-dimensional Euclidean space. Thus, we will provide geometric realizations of abstract simplicial complexes to provide intuition, whereas we will lean more toward abstract realizations when computationally working with simplicial complexes.

Definition 2.3. Let K be a simplicial complex, and let p be a positive integer. Then a *p-chain* is a formal sum $\sum c_i \sigma_i$, where $c_i \in \mathbb{F}_2$ and σ_i is a p-simplex in K. Together with the addition operation, these p-chains form the *chain group* $C_p(K)$ (or simply C_p when the context is understood).

Remark 2.4. More generally, p-chains may be defined as formal sums of p-simplices over any field. However, our paper operates entirely under the field $\mathbb{F}_2$, so certain results throughout are specific to this field.

Example 2.5. If K is a simplicial complex consisting of a 3-simplex and all of its faces, then $C_3 \cong \mathbb{F}_2$. Also, $C_2 \cong \mathbb{F}_2^4$, since a 3-simplex has four 2-simplices as faces.

Definition 2.6. The *boundary operator* $\partial_p : C_p \to C_{p-1}$ is the unique group homomorphism for which any p-simplex $\sigma = \langle v_0, \ldots, v_p \rangle \in C_p$ satisfies $\partial_p(\sigma) = \sum_{i=0}^{p} \langle v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_p \rangle$. The image of any p-chain under ∂_p is then deduced from the images of the p-simplices that make up the p-chain. Let $B_p \subseteq C_p$ and $Z_p \subseteq C_p$ denote the image of ∂_{p+1} and the kernel of ∂_p, respectively.

Example 2.7. Consider a simplicial complex K consisting of

- the 0-simplices $\langle v_0 \rangle$, $\langle v_1 \rangle$, and $\langle v_2 \rangle$,

- the 1-simplices $\langle v_0, v_1\rangle$, $\langle v_1, v_2\rangle$, $\langle v_0, v_2\rangle$, and $\langle v_3\rangle$, and
- the 2-simplex $\langle v_0, v_1, v_2\rangle$.

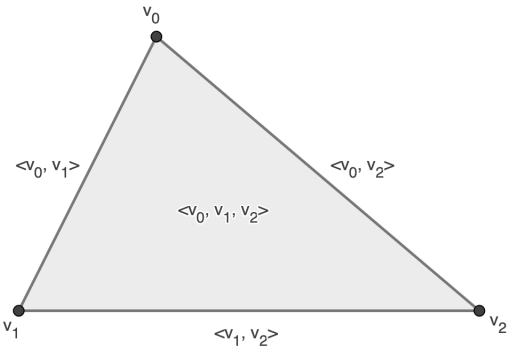

FIGURE 1. The simplicial complex K consisting of one labeled 2-simplex, three labeled 1-simplices, and three labeled 0-simplices forms the interior, edges, and vertices of a triangle.

Then $\partial_2(\langle v_0, v_1, v_2\rangle) = \langle v_0, v_1\rangle + \langle v_1, v_2\rangle + \langle v_0, v_2\rangle$. Also,

$$\partial_1(\langle v_0, v_1\rangle + \langle v_1, v_2\rangle + \langle v_0, v_2\rangle) = (\langle v_0\rangle + \langle v_1\rangle) + (\langle v_1\rangle + \langle v_2\rangle) + (\langle v_2\rangle + \langle v_0\rangle) = 0.$$

It then follows that $\partial_1(\partial_2(\langle v_0, v_1, v_2\rangle)) = 0$.

As alluded to by the previous example, the proposition below describes one crucial property of the boundary operator ∂_p.

Proposition 2.8. [4, Chapter 4] *Let K be a simplicial complex, and let p be a positive integer. Then for all $(p+1)$-chains $c \in C_{p+1}$, we have $\partial_p(\partial_{p+1}(c)) = 0$.*

In particular, B_p is always a subgroup of Z_p, and thus B_p must be a normal subgroup of Z_p as Z_p is abelian. This allows us to produce the following definition:

Definition 2.9. Let K be a simplicial complex, and let p be a positive integer. Then the *p^{th} homology group of K* is the group $H_p = Z_p/B_p$.

Proposition 2.10. [4, Chapter 4] *For any simplicial complex K and positive integer p, the p^{th} homology group of K is isomorphic to $\mathbb{F}_2^{\beta_p}$ for some nonnegative integer β_p, called the p^{th}* Betti number *of K.*

Example 2.11. For any simplicial complex K, the rank β_0 of the 0^{th} homology group H_0 is equal to the number of connected components of K. This follows from the observation that any two vertices correspond to the same equivalence class in $H_0 = Z_0/B_0$ if and only if they are connected by a path of 1-simplices.

Recall that one goal of topological data anslysis (TDA) is to analyze point clouds of data (i.e. finite collections of points in $\mathbb{R}^d$) by computing topological properties of its implied shape. In order to apply topological methods to collections of points, however, we must first establish some method of transforming point clouds into simplicial complexes. The most common simplicial complexes associated with point clouds are the *Čech Complex* and the *Vietoris-Rips Complex*.

Definition 2.12. Consider a finite collection S of points in $\mathbb{R}^d$. Then for any real number $\epsilon > 0$, the *Čech Complex* C_ϵ is the abstract simplicial complex with vertices $\{v_0, v_1, \ldots, v_{|S|-1}\}$, where $\sigma = \langle v_{i_0}, v_{i_1}, \ldots, v_{i_k}\rangle$ is a simplex in C_ϵ if and only if there exists some point $p \in \mathbb{R}^d$ such that $|p - v_{i_k}| < \frac{\epsilon}{2}$ for all $i \in \{0, \ldots, k\}$.

Intuitively, one can think of the Čech Complex for some point cloud S and some $\epsilon > 0$ as the result of constructing one k-simplex for every k-wise intersection of balls of radius $\frac{\epsilon}{2}$ centered at each point in S.

Example 2.13. Given the point cloud $S = \{(-1, 0), (1, 0), (0, \sqrt{3})\} \subseteq \mathbb{R}^2$, its Čech complex $C_{21/10}$ consists of the three vertices $v_0 = (-1, 0)$, $v_1 = (1, 0)$, and $v_2 = (0, \sqrt{3})$, along with the one-simplices $\langle v_0, v_1 \rangle$, $\langle v_1, v_2 \rangle$, and $\langle v_2, v_0 \rangle$. Notably, $C_{21/10}$ does not contain the 2-simplex $\langle v_0, v_1, v_2 \rangle$.

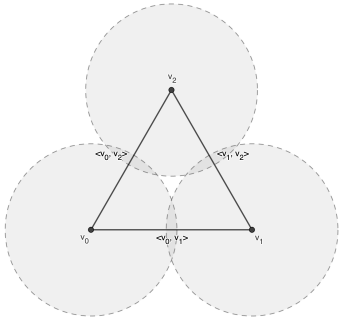

FIGURE 2. Čech complex of $C_{21/10}$ for $S = \{(-1, 0), (1, 0), (0, \sqrt{3})\} \subseteq \mathbb{R}^2$

The Nerve Theorem [4] states that the simplicial complex C_ϵ is homotopy equivalent to the topological space formed by the union of all balls of radius $\frac{\epsilon}{2}$ centered at each $v_i \in S$. Thus, we can be confident that the Čech complex is a fair representation of S as a topological space.

Definition 2.14. Consider a finite collection S of points in $\mathbb{R}^d$. Then for any real number $\epsilon > 0$, the *Vietoris-Rips Complex* VR_ϵ is the abstract simplicial complex whose vertices $\{v_0, v_1, \ldots, v_{|S|-1}\}$ consist of the point cloud S, where $\sigma = \langle v_{i_0}, v_{i_1}, \ldots, v_{i_k} \rangle$ is a simplex in VR_ϵ if and only if $|v_{i_a} - v_{i_b}| < \epsilon$ for all $a, b \in \{0, \ldots, k\}$.

Example 2.15. Given the point cloud $S = \{(-1, 0), (1, 0), (0, \sqrt{3})\} \subseteq \mathbb{R}^2$, its Vietoris-Rips complex $VR_{21/10}$ consists of the three vertices $v_0 = (-1, 0)$, $v_1 = (1, 0)$, and $v_2 = (0, \sqrt{3})$, along with the one-simplices $\langle v_0, v_1 \rangle$, $\langle v_1, v_2 \rangle$, and $\langle v_2, v_0 \rangle$, and the 2-simplex $\langle v_0, v_1, v_2 \rangle$.

Though the Vietoris-Rips complex and Čech complex are different, the following proposition suggests that the two complexes capture roughly the same amount of information.

Proposition 2.16. [3, Theorem 2.5] *For any point cloud S and any $\epsilon > 0$, we have*

$$C_\epsilon \subseteq VR_\epsilon \subseteq C_{\sqrt{2}\epsilon}.$$

Also of concern is the means through which the Čech complex and Vietoris-Rips complex may be computed. Although such computations generally have exponential time complexity in the size of the point cloud, the Vietoris-Rips complex VR_ϵ is notably easier to compute, as it may be summarized by the graph whose vertices are the point cloud, with an edge between two vertices whose distance is less than ϵ. Thus, in practice the Vietoris-Rips complex is easier to use than the Čech complex.

3. PERSISTENT HOMOLOGY

A key feature of our method of transforming point clouds into simplicial complexes is that the precise simplicial complex that we arrive at is heavily dependent on our choice of ϵ. This motivates the concept of persistent homology, in which we analyze how the homology groups of the simplicial complexes corresponding to a given point cloud change as ϵ is changed.

Definition 3.1. Let p be a positive integer. Given an interval $I \subseteq \mathbb{R}$, a *persistent complex* is a collection of chain groups $\mathcal{C} = \{C_p^i : i \in I\}$, together with maps $\phi_{i\to j} : C_p^i \to C_p^j$ for all $i, j \in I$ with $i \leq j$. Throughout this paper, we work exclusively with persistent complexes with chain groups C_p^ϵ derived precisely from the Vietoris-Rips complexes VR_ϵ, and with maps $\phi_{i\to j}$ defined as inclusion maps.

Definition 3.2. Let p be a positive integer, and consider a persistent complex $\mathcal{C} = (C_p^i)_i$. Recall that for any $C_p^x, C_p^y \in (C_p^i)_i$ with $x \leq y$, there exists a map $\phi_{x,y} : C_p^x \to C_p^y$. This map thereby induces a homomorphism $\varphi_{x,y} : H_p(C_p^x) \to H_p(C_p^y)$. Then the (x, y)*-persistent homology group of* $\mathcal{C}$, denoted by $H_p^{x\to y}(\mathcal{C})$ is defined to be the image of $\varphi_{x,y}$.

Example 3.3. See Figure 3. Consider the point cloud $S = \{(-4, 0), (0, 4), (4, 0), (0, -4), (9, 0)\} \subseteq \mathbb{R}^2$, and take the persistent complex with chain groups C_p^1, C_p^2, and C_p^3 resulting from the Vietoris-Rips complexes VR_6, $VR_{7.5}$, and VR_9, respectively. Then $H_1^{1\to 2}$ has rank 1, since the hole in VR_6 persists into $VR_{7.5}$. However, $H_1^{2\to 3}$ has rank 0, since the hole in $VR_{7.5}$ does not persist into VR_9.

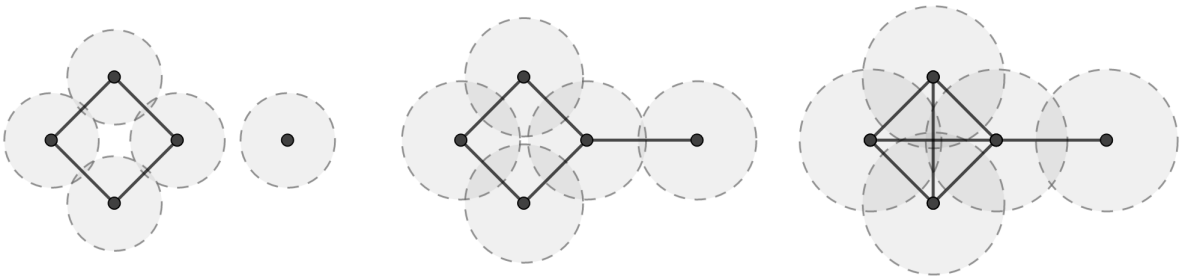

FIGURE 3. Vietoris-Rips complexes for $S = \{(-4, 0), (0, 4), (4, 0), (0, -4), (9, 0)\} \subseteq \mathbb{R}^2$

At this point, we remark that persistence complexes can be derived more generally from any *monotone function* f on a simplicial complex X.

Definition 3.4. Given a simplicial complex X, a *monotone function* on X is a map $f : X \to \mathbb{R}$ such that if σ is a simplicial complex in X, and τ is a face of σ, then $f(\tau) \leq f(\sigma)$. Given a monotone function $f : X \to \mathbb{R}$, we may construct a persistent complex whose chain groups C_p^ϵ are precisely the simplicial complexes $f^{-1}((-\infty, \epsilon])$, together with maps $\phi_{i\to j} : C_p^i \to C_p^j$ defined as inclusion maps.

3.1. **PERSISTENCE DIAGRAMS.** Since the collection of all persistent homology groups of a persistent complex substantiates an immensely large amount of information, we use *persistence diagrams* to condense this information into a visually cleaner format.

Definition 3.5. Let S be a point cloud, and let n be a positive integer. The n^{th} *degree persistence diagram of* S (for all relevant n) is a collection of ordered pairs $(b_i, d_i) \in \mathbb{R}^2$ corresponding to basis elements c_i of the homology groups $H_n(V_\epsilon)$, for all real $\epsilon > 0$. In particular, for all $\epsilon > 0$, the homology group $H_n(V_\epsilon)$ has a basis made up of all c_i for which $b_i < \epsilon < d_i$. These ordered pairs are typically displayed graphically as a collection of points on the coordinate plane, with the x and y axes denoting *birth time* b_i and *death time* d_i, respectively.

Remark 3.6. In the n^{th} degree persistence diagram of a point cloud S (for relevant positive integers n), we use the term H_n *feature* to refer to an ordered pair (b_i, d_i) in the persistence diagram of S.

Example 3.7. Consider the point cloud and persistent complex described in Example 4.3. Shown below are its 0^{th} and 1^{st} degree persistence diagrams. Note that the dotted horizontal line at the top of the graph is used for ordered pairs (b_i, d_i) representing features that never die i.e. persist in $H_n(V_\epsilon)$ for arbitrarily large $\epsilon \in \mathbb{R}$. In this case, the simplicial complex V_ϵ will always consist of a single connected component that never dies, so the 0^{th} persistence diagram contains one point on this dotted line.

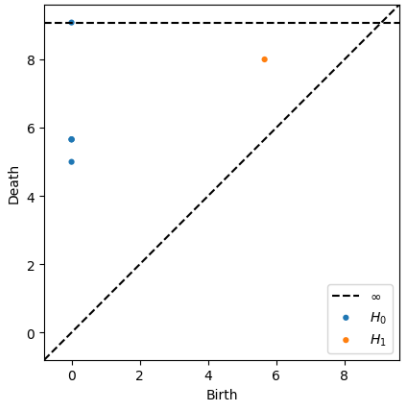

FIGURE 4. Persistence diagrams of $S = \{(-4, 0), (0, 4), (4, 0), (0, -4), (9, 0)\} \in \mathbb{R}^2$.

Persistence diagrams summarize the birth and death times of topological features in the most simple form possible: a collection of (possibly repeated) ordered pairs $\{(b_i, d_i)\}$ in $\mathbb{R}^2$.

3.2. **PERSISTENCE LANDSCAPES.** In addition to persistence diagrams, we now define the persistence landscape, which in contrast summarizes birth and death time pairs through a function $\lambda : \mathbb{N} \times \mathbb{R} \to [0, \infty]$.

Definition 3.8. Consider a point cloud S whose n^{th} degree persistence diagram consists of ordered pairs $(b_i, d_i) \in \mathbb{R}^2$. For each ordered pair (b_i, d_i), define $f_{(b_i,d_i)} : \mathbb{R} \to [0, \infty]$ by

$$f_{(b_i,d_i)}(x) := \max\{0, \min\{x - b_i, d_i - x\}\}.$$

Then the *n^{th} degree persistence landscape of S* is the function $\lambda : \mathbb{N} \times \mathbb{R} \to [0, \infty]$, defined so that $\lambda(k, x)$ is the k^{th} largest value of $f_{(b_i,d_i)}(x)$ across all i. If $f_{(b_i,d_i)}(x)$ attains less than k distinct values, we take $\lambda(k, x) = 0$ by convention.

Visually, the graphs of persistence landscapes in the coordinate plane consist of line segments that either form 45° angles with the x-axis or lie on the x-axis. Taller peaks in these graphs correlate with longer-lasting, isolated homological features.

Example 3.9. The plot in Figure 5 below shows the graphs of the 1^{st} degree persistence landscape $\lambda(k, x)$ of a point cloud S noisily sampled from the perimeter of an isosceles triangle triangle with vertices $\{(-10, 0), (0, 4), (10, 0)\}$. In particular, the plot shows the two functions $\lambda_k : \mathbb{R} \to [0, \infty]$ for $k = 0, 1$ defined by $\lambda_k(x) = \lambda(k, x)$. Both λ_0 and λ_1 feature a single prominent hump near larger values of x, along with several smaller humps surrounding smaller values of x. The singular prominent hump indicates that the shape of S contains only one notable degree-one homological feature, whereas the smaller humps reflect the noise in the arrangement of points in S.

The general shape of persistence landscapes can sometimes indicate strong topological differences between point clouds as well.

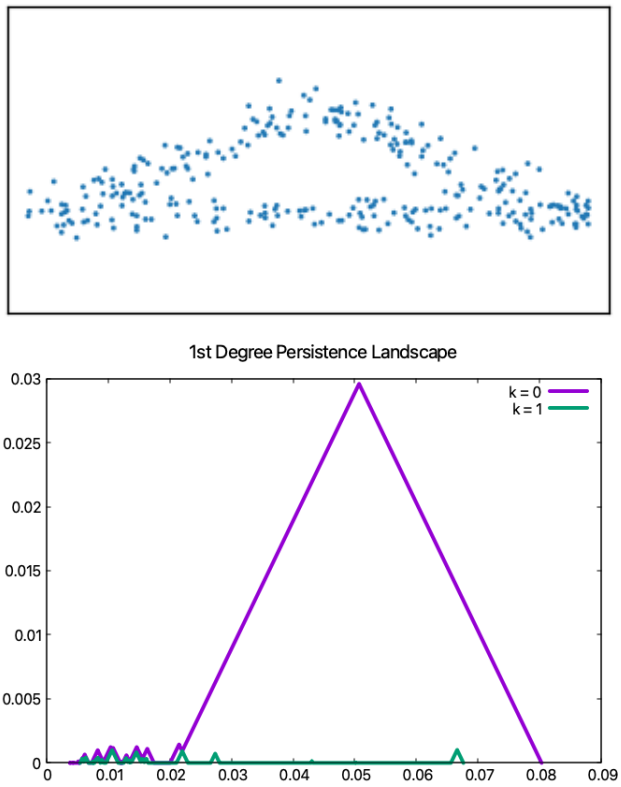

FIGURE 5. The top plot shows a point cloud S noisily sampled from the perimeter of a triangle with vertices $\{(-10,0),(0,4),(10,0)\}$. The bottom plot shows the 1^{st} degree persistence landscape of S.

Example 3.10. The plots in Figure 6 below show the graphs of the 1^{st} degree persistence landscapes of two different point clouds. One point cloud is noisily sampled from the circumferences of two tangent circles of radii $\frac{2}{3}$ and $\frac{1}{3}$, whereas the other point cloud is noisily sampled from the circumference of a single circle of radius 1. The graphs of $\lambda(1,x)$ for the two point clouds differ noticeably; the graph of the former point cloud features a single prominent hump, whereas the graph of the latter point cloud does not feature any prominent humps. This difference in general shape of persistence landscapes reflects how the topology of the former point cloud has two holes, whereas the topology of the latter point cloud has only a single hole.

We may define a metric on the space of persistence landscapes as follows.

Definition 3.11. Consider two persistence landscapes $\lambda, \lambda' : \mathbb{N} \times \mathbb{R} \to [0,\infty]$. Then for any real number $p > 0$, the *p-landscape distance* between λ and λ' is defined as follows:

$$\|\lambda - \lambda'\|_p := \left(\sum_{k=0}^{\infty} \int_{\mathbb{R}} (\lambda(k,x) - \lambda'(k,x))^p \, \mathrm{d}x \right)^{1/p}$$

For $p = \infty$, we define $\| \bullet \|_\infty$ as follows:

$$\|\lambda - \lambda'\|_\infty := \sup_{(k,x) \in \mathbb{N} \times \mathbb{R}} |\lambda(k,x) - \lambda'(k,x)|$$

In previous examples, we have loosely qualified the differences between persistence landscapes by identifying their geometric qualities as graphs in $\mathbb{R}^2$. This approach, however, only works most

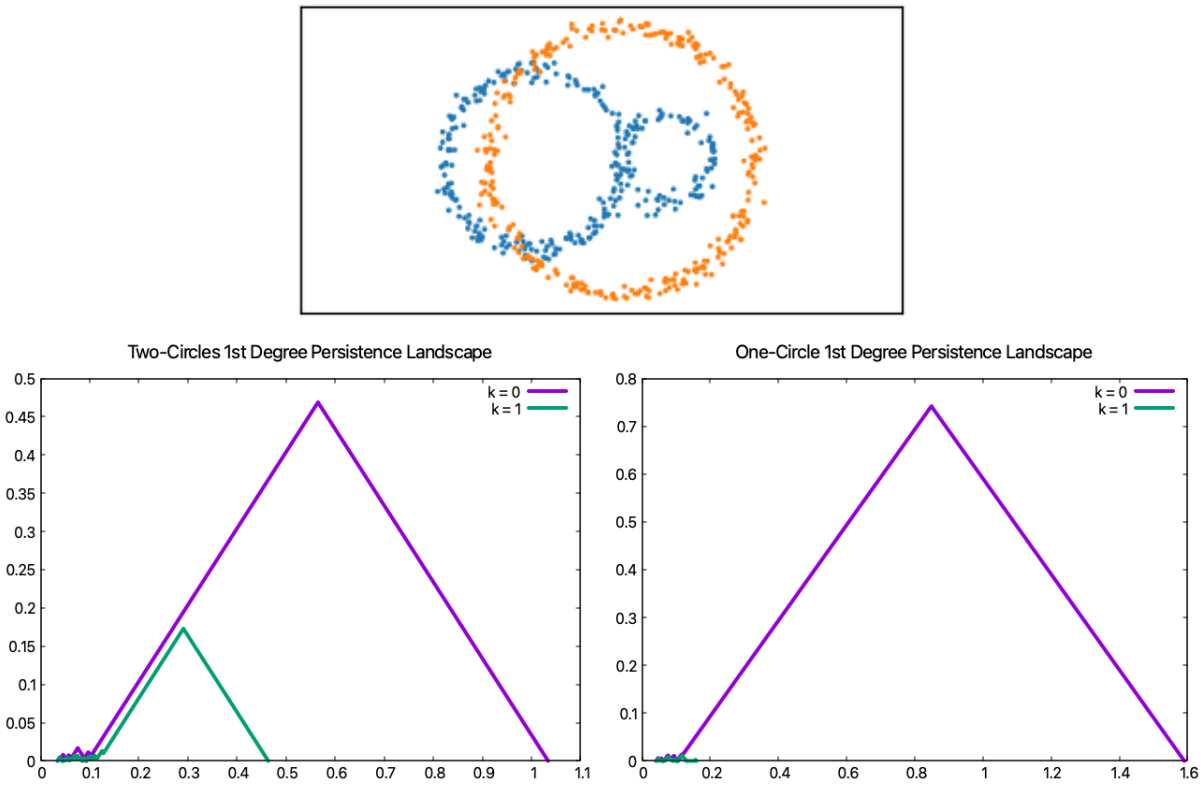

FIGURE 6. The topmost plot shows the two sampled point clouds overlaid on top of each other. The leftmost and rightmost plots show the 1^{st} degree persistence landscapes of the blue and orange point clouds, respectively.

effectively in distinguishing persistence landscape of point clouds with very blatant topological differences; the persistence landscapes of point clouds with more local, geometric differences are harder to observe by eye. The p-landscape distances defined above allow us to more rigorously assign quantitative measurements to the differences between persistence landscapes, allowing for a more robust and statistical analysis of point clouds.

4. STATISTICAL EXPERIMENTS

In this section, we provide experimental evidence to demonstrate the utility of persistence landscapes and persistence diagrams in distinguishing point cloud distributions.

4.1. PERSISTENCE LANDSCAPES.

First, we compare the persistence landscapes of isosceles triangles with varying base angles $\theta \in \{5°, 10°, \ldots, 85°\}$. In particular, we perform the following procedure:

- For each of the 17 values of $\theta \in \{5°, 10°, \ldots, 85°\}$, consider the isosceles triangle Δ_θ with two base angles of measure θ and perimeter 1.
- Produce 25 different point cloud samples consisting of 300 points taken uniformly at random from the perimeter Δ_θ. Then adjust each sample by adding two-dimensional Gaussian noise with standard deviation 0.01.
- For each of these 25 different point cloud samples, construct its 1^{st} degree persistence landscape $\{\lambda_i\}_{i=1}^{25}$. Then take the empirical average of these 25 persistence landscapes.
- For each of the 17 average persistence landscapes λ_θ over $\theta \in \{5°, 10°, \ldots, 85°\}$, compute the pairwise ∞-landscape distances between these persistence landscapes.

Figure 7 displays visualizations of some of the point cloud samples produced through this procedure, along with the set of all computed distances derived from this procedure.

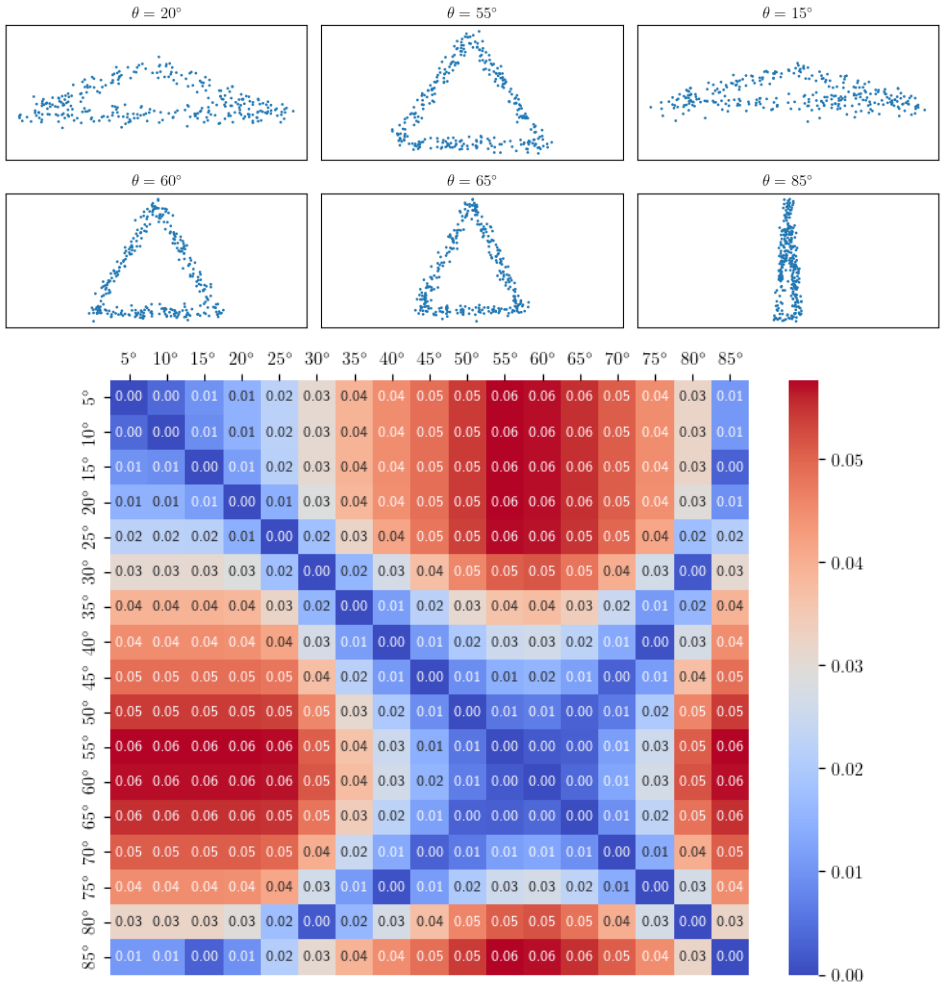

FIGURE 7. The top diagram shows six point cloud samples from acute triangles with base angles $\theta \in \{20°, 55°, 15°, 60°, 65°, 85°\}$. In the bottom diagram, the entry labeled (a, b) indicates the ∞-landscape distance between the persistence landscapes λ_a and λ_b.

We make the following observations:

- The ∞-landscape distance between the persistence landscapes for $\theta = 20°$ and $\theta = 65°$ is relatively large. This reflects the fact that their respective triangles have very strong geometric dissimilarities; the former is an obtuse isosceles triangle very sharp acute angles, whereas the latter is nearly equilateral and lacks obtuse angles or sharp acute angles.
- The ∞-landscape distance between the persistence landscapes for $\theta = 60°$ and $\theta = 65°$ is relatively small. This reflects the fact that their respective triangles are very similar, as each of their corresponding angles differ by $10°$ at most.
- The ∞-landscape distance between the persistence landscapes for $\theta = 15°$ and $\theta = 85°$ is relatively small, even though $15°$ and $85°$ differ greatly in numerical value. This relatively

small distance reflects the geometric similarities in their respective triangles yet again; indeed, both triangles feature very sharp acute angles.

Next, we compare the persistence landscapes of regular polygons with varying numbers of sides. In particular, we perform the following procedure:

- For each integer $n \in \{4, 5, 6, 7, 8, 9, 10, \infty\}$, consider the regular polygon Δ_n with n sides and perimeter 1. We take Δ_∞ to be the circle of perimeter 1.
- Produce 25 different point cloud samples consisting of 300 points taken uniformly at random from the perimeter of Δ_n. Then adjust each sample by adding two-dimensional Gaussian noise with standard deviation 0.01.
- For each of these 25 different point cloud samples, construct each of their 1^{st} degree persistence landscapes $\{\lambda_i\}_{i=1}^{25}$. Then take the empirical average of these 25 persistence landscapes λ_n.
- For each of the 8 average persistence landscapes λ_n over $n \in \{4, 5, 6, 7, 8, 9, 10, \infty\}$, compute the pairwise ∞-landscape distances between these persistence landscapes.

The set of all computed distances derived from this procedure is summarized in Figure 8.

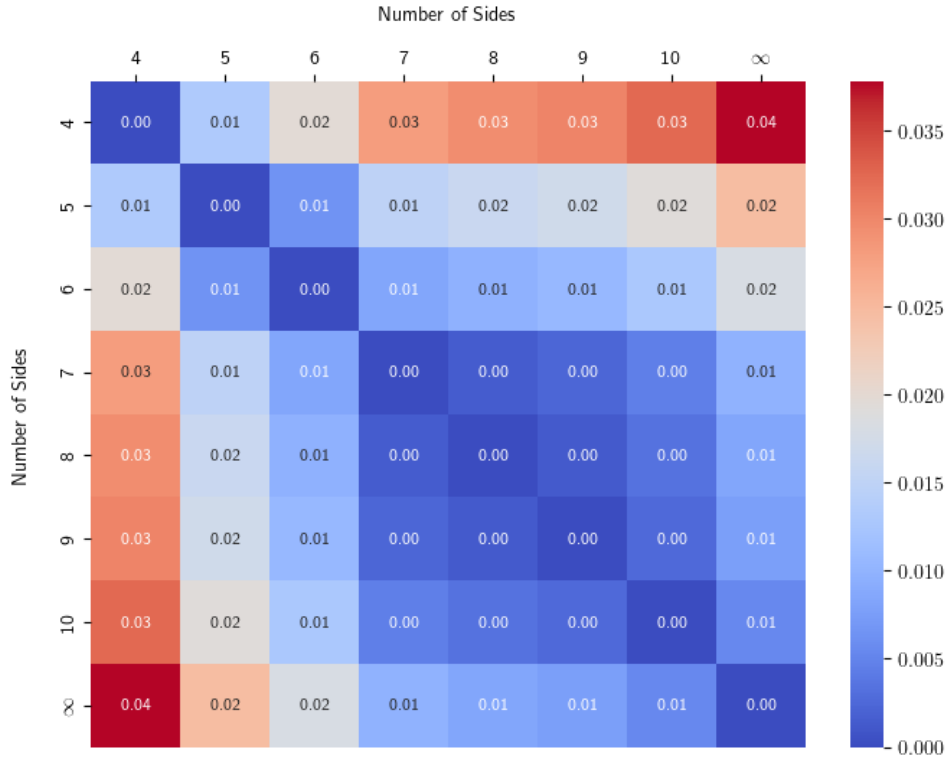

FIGURE 8. The entry labeled (a, b) indicates the ∞-landscape distance between the persistence landscapes λ_a and λ_b.

Contrary to the previous example, the heat map shown here indicates that the point clouds are comparatively less distinguishable. This is to be expected, as the geometric differences between regular polygons become less apparent when considering regular polygons with much greater quantities of sides, as shown in Figure 9.

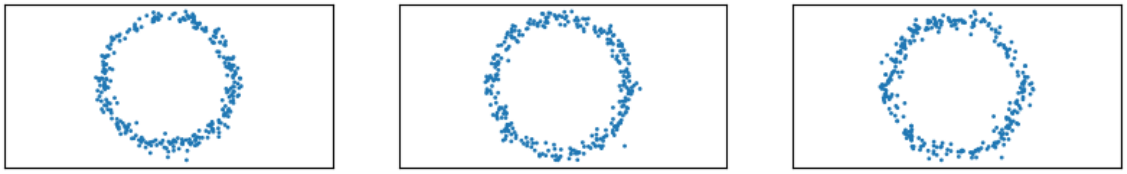

FIGURE 9. The point clouds (from left to right) are sampled from the perimeters of a regular decagon, a regular octagon, and a regular hexagon. Their geometric features are only barely distinguishable.

4.2. **PERSISTENCE DIAGRAMS.** To further demonstrate the utility of persistent homology in distinguishing point cloud distributions, we discuss trends in the following empirical metrics on persistence diagrams taken from point clouds noisily sampled from the perimeters of isosceles triangles.

- `avgDH0`: average time of death of H_0 features,
- `avgBH1`: average time of birth of H_1 features,
- `avgDH1`: average time of death of H_1 features,
- `devDH0`: standard deviation of time of death of H_0 features,
- `devBH1`: standard deviation of time of birth of H_1 features,
- `devDH1`: standard deviation of time of death of H_1 features,

For any given point cloud distribution, we repeatedly compute each statistic for several point clouds sampled from that distribution, keeping sample size constant. For example, we might take $k = 200$ random point cloud samples of a given figure, each of which contains $n = 500$ points. Then, our computed statistic for that point cloud distribution is taken to be the average of all k statistics derived from each of the k point clouds.

See Figure 10. Here, we consider point cloud distributions determined by noisily sampling points from the perimeters of isosceles triangles with bases 20 and heights h, for real numbers h varying from 0 to 20. These isosceles triangles are then scaled to have a normalized perimeter of 1. Using $k = 200$ random point cloud samples each containing $n = 500$ points, we derive the following plots of these statistics against h.

Remark 4.1. To be precise about the sampling procedure, these point clouds are formed by selecting 500 points uniformly at random along the perimeter of the triangle, then adjusting each of them in a direction chosen at random, by a distance chosen uniformly at random between 0 and 1.5% of the perimeter of the triangle. In particular, these experiments do not implement Gaussian noise. This method of constructing point cloud samples has appeared similarly in prior research [2].

The `devDH1` statistic is notably very effective at distinguishing triangles for $h > 5$ as it stabilizes only for values of h much closer to 20. In contrast, the `avgDH0` and `devDH0` statistics do not distinguish triangles for $h > 5$ very well; this is expected, as these statistics effectively reflect the density of a point cloud, which will always remain constant between different triangles. Other notable features of these graphs include the local maxima in the graphs of `avgBH1`, `avgDH1`, and `devBH1` at roughly $h = 3$.

See Figure 11. Recall from Remark 4.1 that the process of creating these point clouds involved the introduction of random noise with a distance bounded by 1.5% of the perimeter of the triangle. If we instead generate point clouds to have no noise at all, the resulting point clouds will feature points that are largely collinear with each other, resulting in TDA statistics with very different behaviors.

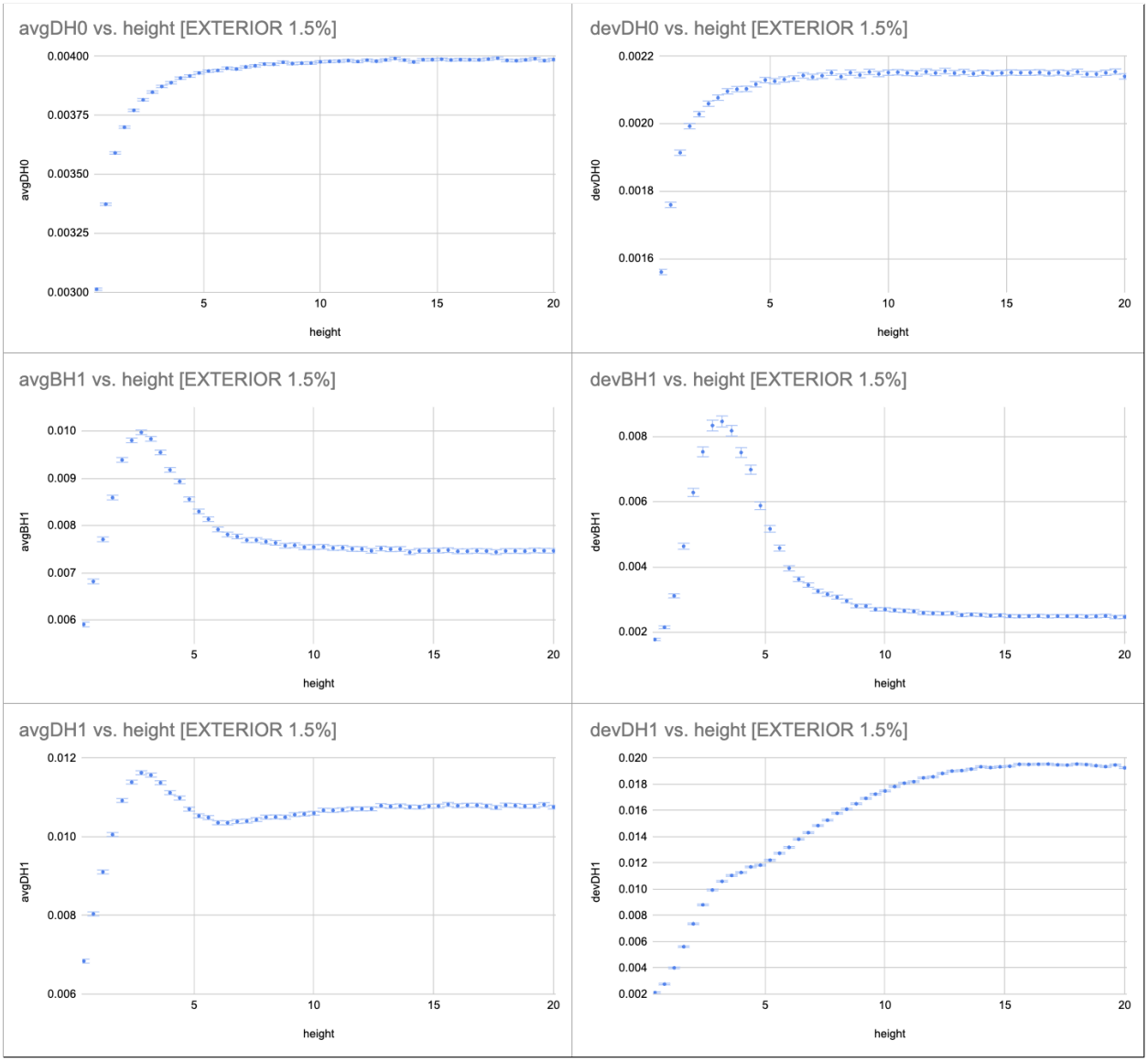

FIGURE 10. Plots of `avgDH0`, `avgBH1`, `avgDH1`, `devDH0`, `devBH1`, and `devDH1` with respect to height. Each statistic is an average of $k = 200$ measurements of point cloud samples of $n = 500$ points from the perimeters of isosceles triangles of varying shapes, with added noise of at most 1.5% the triangle's perimeter. Triangles are normalized to have perimeter 1.

The `avgDH0` and `devDH0` statistics behave similar to as before. However, the `avgBH1` statistic instead appears to have a local minimum at roughly $h = 3.5$ before stabilizing for values of $h > 10$. In the `devBH1` statistic, we observe that the standard deviation of birth times tends to zero for values of $h > 10$. This is because for values of $h > 10$, there is only one H_1 feature, that being the hole formed by the perimeter of the triangle; only rarely do other H_1 features appear since many points in the point cloud are collinear. This also explains the trend in the `devDH1` statistic.

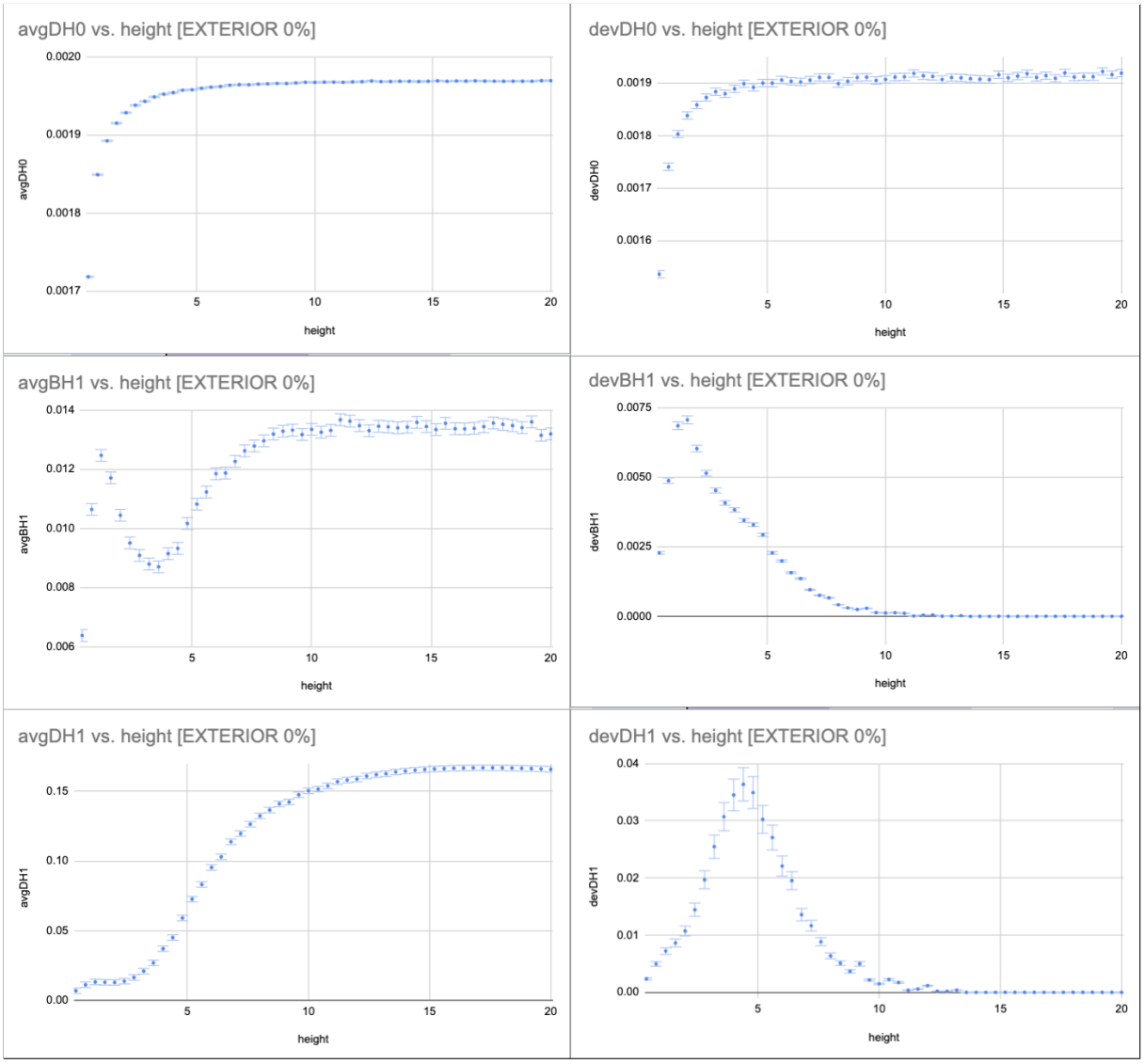

FIGURE 11. Plots of `avdH0`, `avgBH1`, `avgDH1`, `devDH0`, `devBH1`, and `DH1` with respect to height. Each statistic is an average of $k = 200$ measurements of point cloud samples of $n = 500$ points from the exteriors of isosceles triangles of varying shapes, with no added noise. Triangles normalized to have perimeter 1.

5. A THEORETICAL RESULT

We may formally describe point cloud distributions over $\mathbb{R}^2$ with probability measures μ on $\mathbb{R}^2$, in which a point cloud of size N may be viewed as a collection of N points sampled from $\mathbb{R}^2$ in accordance with μ. Given two probability measures μ and ν and any real $p > 1$, we take the Wasserstein distance W_p to be as defined in [5]; that is, we say

$$W_p(\mu,\nu) := \left(\inf\left\{\left(\int_{\mathbb{R}^2\times\mathbb{R}^2} |x-y|^p \,\xi(\mathrm{d}x,\mathrm{d}y)\right) : \xi \in \mathcal{H}(\mu,\nu)\right\}\right)^{\frac{1}{p}},$$

where $\mathcal{H}(\mu,\nu)$ denotes the set of probability measures on $\mathbb{R}^2\times\mathbb{R}^2$ with marginals μ and ν. We also make the following well-known remark about the Wasserstein distance.

Proposition 5.1. *For any two reals $p_1 < p_2$ and any probability measure μ on $\mathbb{R}^2$, we have $W_{p_1}(\mu) \geq W_{p_2}(\mu)$.*

In this section, we seek to prove a relationship between the ∞-landscape distance between persistence landscapes and the ∞-Wasserstein distance between the point cloud distributions from which those landscapes were derived.

Theorem 5.2. *Let μ and ν be probability measures on $\mathbb{R}^2$, and let $\epsilon > 0$ be a real number. For positive integers N, take X_N and Y_N to be sets of N points sampled independently from the probability distributions μ and ν, respectively. Furthermore, let λ_μ and λ_ν denote the respective landscapes of the Vietoris-Rips complexes of X_N and Y_N, respectively. Then the inequality $\|\lambda_\mu - \lambda_\nu\|_\infty \leq 2W_\infty(\mu, \nu) + \epsilon$ holds with arbitrarily high probability for sufficiently large N.*

In particular, the above theorem states that if persistence landscapes sampled from two different distributions have a large ∞-landscape distance, then their corresponding point cloud distributions must have a large W_∞ distance. In this sense, persistence landscape metrics can detect geometric differences in point cloud distributions.

Roughly speaking, we will approach our proof to Theorem 5.2 by showing that if $W_\infty(\mu, \nu)$ is small, then $\|\lambda_\mu - \lambda_\nu\|_\infty$ must be small. We begin by proving Lemma 5.3, which states that if μ_N and ν_N are the empirical measures on $\mathbb{R}^2$ derived from X_N and Y_N, then $W_\infty(\mu, \nu)$ being small must imply $W_\infty(\mu_N, \nu_N)$ is small. Separately, we describe how point clouds X_N and Y_N may produce monotone functions f_X and f_Y. In Lemma 5.5, we use the result from Lemma 5.3 to argue that these monotone functions f_X and f_Y take on similar values, which allows us to conclude that λ_X and λ_Y also take on similar values by invoking a result from [1].

In the statement of Lemma 5.3 below, recall that μ_N and ν_N denote the empirical measures on $\mathbb{R}^2$ derived from X_N and Y_N.

Lemma 5.3. *For any real $\epsilon > 0$ and for sufficiently large N, the inequality $W_\infty(\mu_N, \nu_N) \leq W_\infty(\mu, \nu) + \epsilon$ must hold with arbitrarily high probability.*

Our proof requires the following result on the empirical convergence of the Wasserstein distance, taken as an immediate corollary from results in [5].

Theorem 5.4. [5] *Given any probability distribution μ on $\mathbb{R}^2$ and any real $\epsilon > 0$, the inequality $\mathbb{P}(W_1(\mu_N, \mu)) \leq \epsilon$ must hold for sufficiently large N with arbitrarily high probability.*

Proof of Lemma 5.3. Using Theorem 5.4, we may write the following chain of inequalities.

$$\begin{aligned} W_\infty(\mu_N, \nu_N) &\leq W_\infty(\mu, \nu) + W_\infty(\mu, \mu_N) + W_\infty(\nu, \nu_N) && \text{by the Triangle Inequality} \\ &\leq W_\infty(\mu, \nu) + W_1(\mu, \mu_N) + W_1(\nu, \nu_N) && \text{by Proposition 5.1} \\ &\leq W_\infty(\mu, \nu) + \epsilon && \text{with high probability, by Theorem 5.4} \end{aligned}$$

This concludes the proof. □

In essence, Lemma 5.3 states that if $W_\infty(\mu, \nu)$ is small, then $W_\infty(\mu_N, \nu_N)$ must also be comparably small with high probability. Furthermore, Lemma 5.3 is equivalent to the existence (with high probability) of a bijection $\pi : X_N \to Y_N$ such that $d(x, \pi(x)) \leq W_\infty(\mu, \nu) + \epsilon$ for all $x \in X_N$.

In order to consider the persistence landscapes derived from point clouds X_N and Y_N, we refocus our attention to the monotone functions derived from point clouds X_N and Y_N. In particular, we take our simplicial complex K to be the simplicial complex on N points such that nonempty subset of points in K form a simplex. Take π_X to be an arbitrary bijection between from X_N to the points in K, and take π_Y to be the bijection from Y_N to the points in K so that $\pi_Y^{-1} \circ \pi_X = \pi$. We may then define a monotone function $f_X : K \to \mathbb{R}$ for the point cloud X_N by defining $f(k)$ for $k \in K$ to be the diameter of $\pi_X^{-1}(k)$ in X_N; we may define f_Y similarly.

Lemma 5.5. *For all $k \in K$, we have $|f_X(k) - f_Y(k)| \leq 2W_\infty(\mu, \nu) + 2\epsilon$.*

Proof. Take $x := \pi_X^{-1}(k) \subseteq X_N$ and $y := \pi_Y^{-1}(k) \subseteq Y_N$. By construction, we have $y = \pi(x)$, so by Lemma 5.3, each point $x_i \in x$ is paired with a point $y_i \in y$ for which $d(x_i, y_i) \leq W_\infty(x_i, y_i) + \epsilon$. Take the diameter of x to be $d(x_i, x_j)$ for the appropriate $x_i, x_j \in x$, and assume without loss of generality that the diameter of y is less than or equal to the diameter of y. It follows that

$$\begin{aligned}
(\text{diameter of } Y) &\leq d(y_i, y_j) \\
&\leq d(x_i, x_j) + d(x_i, y_i) + d(x_j, y_j) \qquad \text{by the Triangle Inequality} \\
&\leq d(x_i, x_j) + 2W_\infty(\mu, \nu) + 2\epsilon \\
&\leq (\text{diameter of } X) + 2W_\infty(\mu, \nu) + 2\epsilon
\end{aligned}$$

It follows that absolute difference in the diameters of x and y is at most $2W_\infty(\mu, \nu) + 2\epsilon$. Yet $|f_X(k) - f_Y(k)|$ is precisely this absolute difference, so the lemma is proven. □

We now conclude with a proof of Theorem 5.2.

Proof of Theorem 5.2. Recall that we seek to analyze the persistence landscapes of point clouds X_N and Y_N of size N sampled from probability measures μ and ν on $\mathbb{R}^2$, respectively. To do so, we define filtrations f_X and f_Y on a simplicial complex K of size n in a manner dependent on a bijection $\pi : X_N \to Y_N$, which exists with the properties we desire with arbitrarily high probability according to Lemma 5.3. This desired property of π allows us to conclude $|f_X(k) - f_Y(k)| \leq 2W_\infty(\mu, \nu) + \epsilon$ for all $k \in K$ by Lemma 5.5. We now conclude by invoking a theorem from [1].

Theorem 5.6. [1] *Let f and g be filtrations on a simplicial complex X. (Recall Definition 3.4.) Furthermore, let λ_f and λ_g denote the persistence landscapes derived from the persistence complexes corresponding to f and g. Then $\|\lambda_f - \lambda_g\|_\infty \leq \sup_{x \in X} |f(x) - g(x)|$.*

It follows that if λ_μ and λ_ν are the persistence landscapes derived from X_N and Y_N, we have

$$\|\lambda_f - \lambda_g\|_\infty \leq \sup_{k \in K} |f_X(k) - f_Y(k)| \leq 2W_\infty(\mu, \nu) + \epsilon.$$

This concludes the proof of Theorem 5.2. □

Theorem 5.2 ultimately demonstrates that large ∞-landscape distances between persistence landscapes is indicative of large geometric differences in the probability distributions from which those persistence landscapes were constructed.

6. FURTHER DISCUSSIONS

Future directions could attempt to refine the inequality provided in Theorem 5.2 to achieve a tighter bound more representative of the true capabilities of persistence landscapes. Furthermore, Theorem 5.2 only provides asymptotic guarantees on the relationship between the sample size N and the probability of success, so future work could attempt to give estimates of the quantitative values of N needed for Theorem 5.2 to apply in specific practical, experimental settings.

Alternative directions may also consider a geometric metric other than the Wasserstein distance W_∞; in particular, the Wasserstein distance is unfortunately not invariant under rigid transformations of probability measures, so two probability measures that model the same point cloud distributions up to congruence may still have very large Wasserstein distances in spite of their geometric similarities.

7. ACKNOWLEDGEMENTS

I deeply thank my mentor Jonathan Rodríguez Figueroa for introducing me to this area of research and providing invaluable guidance, feedback, and resources. I also thank MIT PRIMES for providing the support that has made this research possible.

References

[1] Peter Bubenik. Statistical topological data analysis using persistence landscapes, 2015.
[2] Justin Curry, Haibin Hang, Washington Mio, Tom Needham, and Osman Berat Okutan. Decorated merge trees for persistent topology. *Journal of Applied and Computational Topology*, 6(3):371–428, 2022.
[3] Vin De Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358, 2007.
[4] Herbert Edelsbrunner and John L Harer. *Computational topology: an introduction*. American Mathematical Society, 2022.
[5] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure, 2013.
[6] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1), August 2017.
[7] Renata Turkeš, Jannes Nys, Tim Verdonck, and Steven Latré. Noise robustness of persistent homology on greyscale images, across filtrations and signatures. *PLOS ONE*, 16(9):e0257215, September 2021.