Subsampling Methods for Persistent Homology

Frédéric Chazal, frederic.chazal@inria.fr
Brittany Terese Fasy, brittany.fasy@alumni.duke.edu
Fabrizio Lecci, lecci@cmu.edu
Bertrand Michel, bertrand.michel@upmc.fr
Alessandro Rinaldo, arinaldo@stat.cmu.edu
Larry Wasserman, larry@stat.cmu.edu

Abstract

Persistent homology is a multiscale method for analyzing the shape of sets and functions from point cloud data arising from an unknown distribution supported on those sets. When the size of the sample is large, direct computation of the persistent homology is prohibitive due to the combinatorial nature of the existing algorithms. We propose to compute the persistent homology of several subsamples of the data and then combine the resulting estimates. We study the risk of two estimators and we prove that the subsampling approach carries stable topological information while achieving a great reduction in computational complexity.

1 Introduction

Topological Data Analysis (TDA) refers to a collection of methods for finding topological structure in data (Carlsson, 2009). The input is a dataset drawn from a probability measure supported on an unknown low-dimensional set $\mathbb{X}$. The output is a collection of data summaries that are used to estimate the topological features of $\mathbb{X}$.

One approach to TDA is persistent homology (Edelsbrunner and Harer, 2010), which is a method for studying the topology at multiple scales simultaneously. For example, let A be a set and let $f(x) = \inf_{y \in A} ||x - y||$ be the *distance function.* The lower-level sets $\{x : f(x) \leq t\}$ change as t increases from $-\infty$ to ∞. Persistent homology summarizes the evolution of $\{x : f(x) \leq t\}$ as a function of t. In particular, the *persistence diagram* represents the birth and death time of each topological feature as a point in the plane. Thanks to stability properties (Cohen-Steiner et al., 2007; Chazal et al., 2009, 2012a,b), persistence diagrams provide relevant multi-scale topological information about the data; see Section 2. The persistence diagram can be converted into a summary function called a landscape (Bubenik, 2012).

Contribution and Related Work. The time and space complexity of persistent homology algorithms is one of the main obstacles in applying TDA techniques to high-dimensional problems. To overcome the problem of computational costs, we propose the following strategy: given a large point cloud, take several subsamples, compute the landscape for each subsample and then combine the information. More precisely, let λ be a random persistence landscape from Ψ_μ^m, a measure on the space of landscape functions induced by a sample of size m from a metric measure space $(\mathbb{X}, \rho, \mu)$. We show that the average landscape is stable with respect to perturbations of the underlying measure μ in the Wasserstein metric; see Theorem 5. The empirical counterpart of the average landscape is $\overline{\lambda_n^m} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i$, where $\lambda_1, \ldots, \lambda_n \sim \Psi_\mu^m$. The empirical average landscape can be used as an unbiased estimator of $\mathbb{E}_{\Psi_\mu^m}[\lambda]$ and as a biased estimator of $\lambda_{\mathbb{X}_\mu}$, the computationally expensive persistence landscape associated to the support of the measure μ. Unlike $\lambda_{\mathbb{X}_\mu}$, the estimator $\overline{\lambda_n^m}$ is robust to the presence of outliers. In the same spirit, we propose a different estimator constructed by choosing a sample of m points of $\mathbb{X}$ as close as possible to $\mathbb{X}_\mu$, and then computing its persistent homology to approximate $\lambda_{\mathbb{X}_\mu}$. See Section 3 for more details.

Closely related to our approach, the distribution of persistence diagrams associated to subsamples of fixed size has also been proposed in Blumberg et al. (2014). There, the authors show that the distribution of persistence diagrams associated to subsamples of fixed size is stable with respect to perturbations of the underlying measure in the Gromov-Prohorov metric. Though similar in spirit, our approach relies on different techniques and, in particular, leads to easily computable summaries of the persistent homology of a given space. These summaries are particularly useful when the exact computation of the persistent homology is unfeasbile, as in the case of large point clouds.

Software. We plan to release the R package **persistence**, which provides efficient algorithms for the computation of persistent homology from **Dionysus** and **GUDHI**, and makes them available with the user-friendly R interface. **Dionysus**[1] is a C++ library written and maintained by Dmitriy Morozov; **GUDHI**[2] is new born project hosted by INRIA and whose goal is the development and implementation of new algorithms for geometric understanding in high dimensions. Preliminary results show that **GUDHI** outperforms its major competitors; see Boissonnat et al. (2013). Our package includes a series of tools for the statistical analysis of persistent homology, including the methods described in Fasy et al. (2013), Chazal et al. (2014a), and this paper.

Outline. Background on persistent homology is presented in Section 2. Our approach is introduced in Section 3, with a formal definition of the estimators briefly described in this introduction. Section 4 contains the stability result of the average landscape. Section 5 is devoted to the risk analysis of the proposed estimators. In Section 6, we apply our methods to two examples. We conclude with some remarks in Section 7 and defer proofs and technical details to the appendices.

2 Background

Computing persistent homology requires building a nested sequence of geometric complexes indexed by a real parameter. In this section, we briefly introduce these families and topological summaries of them, but refer the reader to Section 4.2. of Chazal et al. (2012b) for a complete definition of these geometric filtered complexes and their use in TDA, to Edelsbrunner and Harer (2010) for the definition of persistence diagrams, and to Bubenik (2012) for the definition and properties of persistence landscapes.

2.1 Geometric Complexes

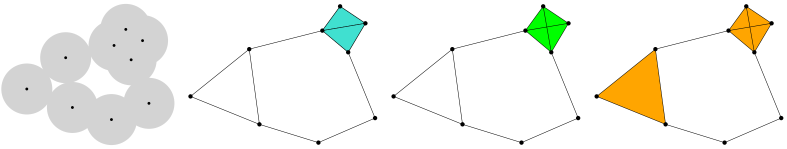

Figure 1: From left to right: the α sublevelset of the distance function to a point set $\mathbb{X}$ in $\mathbb{R}^2$, the α-complex, $\mathrm{Cech}_\alpha(\mathbb{X})$, and $\mathrm{Rips}_{2\alpha}(\mathbb{X})$. The last two complexes include a tetrahedron.

To compute the persistent homology from a set of data, we need to construct a set of structures called simplicial complexes. A simplicial complex $\mathcal{C}$ is a set of simplices (points, segments, triangles, etc) such that any face from a simplex in $\mathcal{C}$ is also in $\mathcal{C}$ and the intersection of any two simplices of $\mathcal{C}$ is a (possibly empty) face of these simplices.

Given a metric space $\mathbb{X}$, we define three simplicial complexes whose vertex set is $\mathbb{X}$; see Figure 1 for illustrations. The *Vietoris-Rips complex* $\mathrm{Rips}_\alpha(\mathbb{X})$ is the set of simplices $[x_0, \ldots, x_k]$ such that $d_{\mathbb{X}}(x_i, x_j) \leq \alpha$ for all (i, j). The *Čech complex* $\mathrm{Cech}_\alpha(\mathbb{X})$ is similarly defined as the set of simplices $[x_0, \ldots, x_k]$ such that there exists a point $x \in \mathbb{X}$ for which $d_{\mathbb{X}}(x, x_i) \leq \alpha$ for all i. Note that these two complexes are related by $\mathrm{Rips}_\alpha(\mathbb{X}) \subseteq \mathrm{Cech}_\alpha(\mathbb{X}) \subseteq \mathrm{Rips}_{2\alpha}(\mathbb{X})$ and that their definition does not require $\mathbb{X}$ to be finite. When $\mathbb{X} \subset \mathbb{R}^d$, we also define the *$\alpha$-complex* as the set of simplices $[x_0, \ldots, x_k]$ such that there exists a ball of radius at most α containing $x_0, \ldots, x_k$ on its boundary and whose interior does not intersect $\mathbb{X}$.

[1] http://www.mrzv.org/software/dionysus/

[2] https://project.inria.fr/gudhi/

Each family described above is non-decreasing with α: for any $\alpha \leq \beta$, there is an inclusion of $\mathrm{Rips}_\alpha(\mathbb{X})$ in $\mathrm{Rips}_\beta(\mathbb{X})$, and similarly for the Čech and Alpha complexes. These sequences of inclusions are called *filtrations*. In the following, we let $\mathrm{Filt}(\mathbb{X}) := (\mathrm{Filt}_\alpha(\mathbb{X}))_{\alpha \in \mathcal{A}}$ denote a filtration corresponding to one of the parameterized complexes defined above.

2.2 Persistence Diagrams

The topology of $\mathrm{Filt}_\alpha(\mathbb{X})$ changes as α increases: new connected components can appear, existing connected components can merge, cycles and cavities can appear or be filled, etc. Persistent homology tracks these changes, identifies *features* and associates an *interval* or *lifetime* (from b to d) to them. For instance, a connected component is a feature that is born at the smallest α such that the component is present in $\mathrm{Filt}_\alpha(\mathbb{X})$, and dies when it merges with an older connected component. Intuitively, the longer a feature persists, the more relevant it is. The lifetime of a feature can be represented as a point in the plane with coordinates (b, d). The obtained set of points (with multiplicity) is called the *persistence diagram* $D(\mathrm{Filt}(\mathbb{X}))$ (and we will abuse terminology slightly by denoting it $D_{\mathbb{X}}$). Note that the diagram is entirely contained in the half-plane above the diagonal Δ defined by $y = x$, since death always occurs after birth. Chazal et al. (2012a) shows that this diagram is still well-defined under very weak hypotheses, and in particular $D(\mathrm{Filt}(\mathbb{X}))$ is well-defined for any compact metric space $\mathbb{X}$ Chazal et al. (2012b). The most persistent features (supposedly the most important) are those represented by the points furthest from the diagonal in the diagram, whereas points close to the diagonal can be interpreted as (topological) noise.

To avoid (minor) technical difficulties, we restrict our attention to diagrams D such that $(b, d) \in [0, T] \times [0, T]$ for all $(b, d) \in D$, for some fixed $T > 0$. Note that, in our setting, $D_{\mathbb{X}}$ is in $\mathcal{D}_T$ as soon as T is larger than the diameter of $\mathbb{X}$. We denote by $\mathcal{D}_T$ the space of all such (restricted) persistence diagrams and we endow it with a metric called the *bottleneck distance* $\mathrm{d_b}$. Given two persistence diagrams, it is defined as the infimum of the δ for which we can find a matching between the diagrams, such that two points can only be matched if their distance is less than δ and all points at distance more than δ from the diagonal must be matched.

A fundamental property of persistence diagrams, proven in Chazal et al. (2012a), is their *stability*. Recall that the Hausdorff distance between two compact subsets X, Y of a metric space $(\mathbb{X}, \rho)$ is $H(X, Y) = \max\left\{ \max_{x \in X} \min_{y \in Y} \rho(x, y),\ \max_{y \in Y} \min_{x \in X} \rho(x, y) \right\}$. If $\mathbb{X}$ and $\widetilde{\mathbb{X}}$ are two compact metric spaces, then one has

$$\mathrm{d_b}(D_{\mathbb{X}}, D_{\widetilde{\mathbb{X}}}) \ \leq \ 2\mathrm{d}_{\mathrm{GH}}(\mathbb{X}, \widetilde{\mathbb{X}}), \tag{1}$$

where $\mathrm{d}_{\mathrm{GH}}(\mathbb{X}, \widetilde{\mathbb{X}})$ denotes the Gromov-Hausdorff distance, i.e., the infimum Hausdorff distance between $\mathbb{X}$ and $\widetilde{\mathbb{X}}$ over all possible isometric embeddings into a common metric space. If $\mathbb{X}$ and $\widetilde{\mathbb{X}}$ are already embedded in the same metric space then (1) holds for $H(\cdot, \cdot)$ in place of $2\mathrm{d}_{\mathrm{GH}}(\cdot, \cdot)$.

2.3 Persistence Landscapes

The persistence landscape, introduced in Bubenik (2012), is a collection of continuous, piecewise linear functions $\lambda \colon \mathbb{Z}^+ \times \mathbb{R} \to \mathbb{R}$ that summarizes a persistence diagram. To define the landscape, consider the set of functions created by tenting each each point $p = (x, y) = \left(\frac{b+d}{2}, \frac{d-b}{2}\right)$ representing a birth-death pair $(b, d) \in D$ as follows:

$$\Lambda_p(t) = \begin{cases} t - x + y & t \in [x - y, x] \\ x + y - t & t \in (x, x + y] \\ 0 & \text{otherwise} \end{cases} = \begin{cases} t - b & t \in [b, \frac{b+d}{2}] \\ d - t & t \in (\frac{b+d}{2}, d] \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We obtain an arrangement of piecewise linear curves by overlaying the graphs of the functions $\{\Lambda_p\}_p$; see Figure 2. The persistence landscape of D is a summary of this arrangement. Formally, the persistence landscape of D is the collection of functions

$$\lambda_D(k, t) = \underset{p}{\mathrm{kmax}}\, \Lambda_p(t), \quad t \in [0, T], k \in \mathbb{N}, \tag{3}$$

where kmax is the kth largest value in the set; in particular, 1max is the usual maximum function. We set $\lambda_D(k, t) = 0$ if the set $\{\Lambda_p(t)\}_p$ contains less than k points. For simplicity of exposition, if $D_{\mathbb{X}}$ is the persistence diagram of some metric space $\mathbb{X}$, then we use $\lambda_{\mathbb{X}}$ to denote $\lambda_{D_{\mathbb{X}}}$.

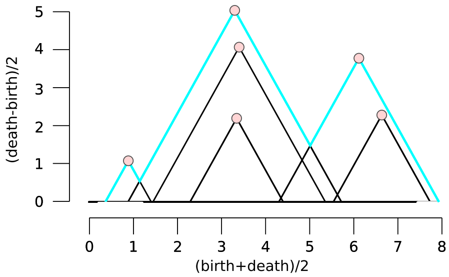

Figure 2: We use the rotated axes to represent a persistence diagram D. A feature $(b, d) \in D$ is represented by the point $(\frac{b+d}{2}, \frac{d-b}{2})$ (pink). In words, the x-coordinate is the average parameter value over which the feature exists, and the y-coordinate is the half-life of the feature. The cyan curve is the landscape $\lambda(1, \cdot)$.

We denote by $\mathcal{L}_T$ the space of persistence landscapes corresponding to $\mathcal{D}_T$. From the definition of persistence landscape, we immediately observe that $\lambda_D(k, \cdot)$ is one-Lipschitz. The following additional properties are proven in Bubenik (2012).

Lemma 1. *Let D, D' be persistence diagrams. We have the following for any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$:*
(i) $\lambda_D(k, t) \geq \lambda_D(k+1, t) \geq 0$.
(ii) $|\lambda_D(k, t) - \lambda_{D'}(k, t)| \leq \mathrm{d_b}(D, D')$.

For ease of exposition, we focus on the case $k = 1$, and set $\lambda_D(t) = \lambda_D(1, t)$. However, the results we present hold for $k > 1$. In fact, the results hold for more general summaries of persistence landscapes, including the silhouette defined in Chazal et al. (2014a).

3 The Multiple Samples Approach

Let $(\mathbb{X}, \rho)$ be a metric space of diameter at most $T/2$ and let $\mathcal{P}(\mathbb{X})$ be the space of probability measures on $\mathbb{X}$, such that, for any measure $\mu \in \mathcal{P}(\mathbb{X})$, its support $\mathbb{X}_\mu$ is a compact set. The space $\mathbb{X}_\mu$ is a natural object of interest in computational topology. Its persistent homology is usually approximated by the persistent homology of the distance function to a sample $X_N = \{x_1, \ldots, x_N\} \subset \mathbb{X}_\mu$. Fasy et al. (2013) propose several methods for the construction of confidence sets for the persistence diagram of $\mathbb{X}_\mu$, while Chazal et al. (2014b) establish optimal convergence rates for $d_b(D_{\mathbb{X}_\mu}, D_{X_N})$.

When N is too large, the computation of the persistent homology of X_N is prohibitive, due to the combinatorial complexity of the computation. Our aim is to study topological signatures of the data that can be efficiently computed in a reasonable time. We define such quantities by repeatedly sampling m points of $\mathbb{X}$ according to μ.

For any positive integer m, let $X = \{x_1, \cdots, x_m\} \subset \mathbb{X}$ be a sample of m points from the measure $\mu \in \mathcal{P}(\mathbb{X})$. The corresponding persistence landscape is λ_X and we denote by Ψ_μ^m the measure induced by $\mu^{\otimes m}$ on $\mathcal{L}_T$. Note that the persistence landscape λ_X can be seen as a single draw from the measure Ψ_μ^m. We consider the point-wise expectations of the (random) persistence landscape under this measure: $\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)], t \in [0, T]$. This quantity is relevant from a topological point of view, because it is stable under perturbation of the underlying measure μ. This stability result is the main result of the paper, presented in detail in the next section.

The average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ has a natural empirical counterpart, which can be used as its unbiased estimator. Let $S_1^m, \ldots, S_n^m$ be n independent samples of size m from μ. We define the empirical average landscape as

$$\overline{\lambda_n^m}(t) = \frac{1}{n} \sum_{i=1}^{n} \lambda_{S_i^m}(t), \qquad \text{for all } t \in [0, T], \tag{4}$$

and propose to use $\overline{\lambda_n^m}$ to estimate $\lambda_{\mathbb{X}_\mu}$. The variance of this estimator under the ℓ_∞-distance was studied in detail in Chazal et al. (2014a). Here instead we are concerned with the quantity $\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}_{\Psi_\mu^m}[\lambda_X]\|_\infty$, which can be seen as the bias component (see Section 5).

In addition to the average, we also consider using the *closest sample* to $\mathbb{X}_\mu$ in Hausdorff distance. The closest sample method consists in choosing a sample of m points of $\mathbb{X}$, as close as possible to $\mathbb{X}_\mu$, and then use this sample to build a landscape that approximates $\lambda_{\mathbb{X}_\mu}$. Let $S_1^m, \ldots, S_n^m$ be n independent samples of size m from $\mu^{\otimes m}$. The closest sample is

$$\widehat{C_n^m} = \arg \min_{S \in \{S_1^m, \ldots, S_n^m\}} H(S, X_\mu) \tag{5}$$

and the corresponding landscape function is $\widehat{\lambda_n^m} = \lambda_{\widehat{C_n^m}}$. Of course, the method requires the support of μ to be a known quantity.

Remark 2. *Computing the persistent homology of X_N is $O(\exp(N))$, whereas computing the average landscape is $O(n\exp(m))$ and the persistent homology of the closest sample is $O(nmN + \exp(m))$.*

Remark 3. *The general framework described above is valid for the case in which μ is a discrete measure with support $\mathbb{X}_\mu = \{x_1, \ldots, x_N\} \subset \mathbb{R}^D$. For example, the following situation is very common in practice. Let $X_N = \{x_1, \ldots, x_N\}$ be a given point cloud, for large but fixed $N \in \mathbb{R}$. When N is large, the computation of the persistent homology of X_N is unfeasible. Instead, we consider the discrete uniform measure μ that puts mass $1/N$ on each point of X_N, and we propose to estimate $\lambda_{X_N} = \lambda_{\mathbb{X}_u}$ by repeatedly subsampling $m \ll N$ points of X_N according to μ.*

We will study the ℓ_∞-risk of the proposed estimators, $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \overline{\lambda_n^m}\|_\infty\right]$ and $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right]$, under the following assumption on the underlying measure μ, which we will refer to as the (a, b, r_0)-*standard assumption*: there exist positive constants a, b and $r_0 \geq 0$ such that

$$\forall r > r_0, \, \forall x \in \mathbb{X}_\mu \, , \; \mu(B(x, r)) \geq 1 \wedge ar^b. \tag{6}$$

For $r_0 = 0$, this is known as the (a, b)-standard assumption and has been widely used in the literature of set estimation under Hausdorff distance (Cuevas and Rodríguez-Casal, 2004; Cuevas, 2009; Singh et al., 2009) and more recently in the statistical analysis of persistence diagrams (Chazal et al., 2014b; Fasy et al., 2013). We use the generalized version with $r_0 > 0$ to take into account the case in which μ is a discrete measure (in which case r_0 depends on N); see Appendix C for more details.

4 Stability of the Average Landscape

Consider the framework described in Section 3: m points are repeatedly sampled from the space $\mathbb{X}$ according to a measure $\mu \in \mathcal{P}(\mathbb{X})$. In this section, we show that the average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ is an interesting quantity on its own, since it carries some stable topological information about the underlying measure μ, from which the data are generated.

Chazal et al. (2014a) provide a way to construct confidence bands for $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$. Here, we compare the average landscapes corresponding to two measures that are close to each other in the Wasserstein metric.

Definition 4. Given a metric space $(\mathbb{X}, \rho)$, the pth Wasserstein distance between two measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$ is $W_{\rho,p}(\mu, \nu) = \left(\inf_\Pi \int_{\mathbb{X}\times\mathbb{X}} [\rho(x, y)]^p d\Pi(x, y)\right)^{\frac{1}{p}}$, where the infimum is taken over all measures on $\mathbb{X} \times \mathbb{X}$ with marginals μ and ν.

The Wasserstein distance is often colloquially referred to as the earth-movers distance, as Π can be seen as a transport plan. The following result shows that the average behavior of the landscapes of sets of m points sampled according to any measure μ is stable with respect to the Wasserstein distance.

Theorem 5. *Let $(\mathbb{X}, \rho)$ be a metric space of diameter bounded by $T/2$. Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where $\mu, \nu \in \mathcal{P}(\mathbb{X})$ are two probability measures. For any $p \geq 1$ we have*

$$\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq m^{\frac{1}{p}} W_{\rho,p}(\mu, \nu).$$

For measures that are not defined on the same metric space, the inequality of Theorem 5 can be extended to Gromov-Wasserstein metric: $\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2m^{\frac{1}{p}} GW_{\rho,p}(\mu, \nu)$.

The result of Theorem 5 is useful for two reasons. First, it tells us that for a fixed m, the expected "topological behavior" of a set of m points carries some stable information about the underlying measure from which the data are generated. Second, it provides a lower bound for the Wasserstein distance between two measures, based on the topological signature of samples of m points.

The dependence on m of the upper bound of Theorem 5 seems to be necessary in this setting: intuitively, when m grows, the samples of m points converge to the support of μ and ν w.r.t. the Hausdorff distance respectively. Therefore the expected landscapes should converge to the landscapes of the support of the measures. But, in general, two measures that are close in the Wasserstein metric can have support that have very different and unrelated topologies. Indeed, a similar dependence was also obtained in Blumberg et al. (2014) when analyzing the stability properties of persistent diagrams in the Gromov-Prohorov metric.

Note that in Theorem 5 we do not make any assumption on the measures μ and ν. If we assume that they both satisfy the (a, b, r_0)-standard assumption we can provide a different bound on the difference of the expected landscapes, based on the Hausdorff distance between the support of the two measures.

Theorem 6. *Let $(\mathbb{X}, \rho)$ be a metric space of diameter bounded by $T/2$. Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where $\mu, \nu \in \mathcal{P}(\mathbb{X})$ satisfy the (a, b, r_0)-standard assumption on $\mathbb{X}$. Define $r_m = 2\left(\frac{\log m}{am}\right)^{1/b}$. Then*

$$\|\mathbb{E}_{\Psi_\mu^m}(\lambda_X) - \mathbb{E}_{\Psi_\nu^m}(\lambda_Y)\|_\infty \le H(\mathbb{X}_\mu, \mathbb{X}_\nu) + 2r_0 \;+\; 2r_m \mathbb{1}_{(r_0,\infty)}(r_m) \;+\; 2\,C_1(a,b)\, r_m \,\frac{1}{(\log m)^2},$$

where $C_1(a, b)$ is a constant depending on a and b.

The following result follows by combining theorems 5 and 6.

Corollary 7. *Under the same assumptions of Theorem 6 we have that*

$$\Big\|\mathbb{E}_{\Psi_\mu^m}(\lambda_X) - \mathbb{E}_{\Psi_\nu^m}(\lambda_Y)\Big\|_\infty \le \min\Big\{m^{\frac{1}{p}} W_p(\mu, \nu),$$
$$H(\mathbb{X}_\mu, \mathbb{X}_\nu) + 2r_0 \;+\; 2r_m \mathbb{1}_{(r_0,\infty)}(r_m) \;+\; 2\,C_1(a,b)\, r_m \,\frac{1}{(\log m)^2}\Big\}.$$

5 Risk Analysis

In this section we study the performance of the average landscape $\overline{\lambda_n^m}$ and of the landscape of the closest sample $\widehat{\lambda_n^m}$, as estimators of $\lambda_{\mathbb{X}_\mu}$. We start by decomposing the ℓ_∞-risk of the average landscape as follows. Set $\lambda_1 = \lambda_{S_1^m}$, with S_1^m a sample of size m from μ. Then,

$$\mathbb{E}\left\|\lambda_{\mathbb{X}_\mu} - \overline{\lambda_n^m}\right\|_\infty \le \left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty + \mathbb{E}\left\|\overline{\lambda_n^m} - \mathbb{E}\lambda_1\right\|_\infty, \tag{7}$$

where the expectation of $\overline{\lambda_n^m}$ is wrt $(\Psi_\mu^m)^{\otimes n}$ and the expectation of λ_1 is wrt Ψ_μ^m.

For the bias term $\left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty$ we use the stability property to go back into $\mathbb{R}^d$:

$$\left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty \le \mathbb{E}_{\Psi_\mu^m}\left\|\lambda_{\mathbb{X}_\mu} - \lambda_1\right\|_\infty \le \mathbb{E}_{\mu^{\otimes m}} H(\mathbb{X}_\mu, X), \tag{8}$$

where X is a sample of size m from μ. Note that, if calculating $H(\mathbb{X}_\mu, X)$ is computationally feasible, then, in practice, $\mathbb{E}_{\mu^{\otimes m}} H(\mathbb{X}_\mu, X)$ can be approximated by the average of a large number B of values of $H(\mathbb{X}_\mu, X)$, for B different draws of subsamples $X \sim \mu^{\otimes m}$.

To give an explicit bound on the bias we assume that μ satisfies the (a, b, r_0)-standard assumption.

Theorem 8. *Let $r_m = 2\left(\frac{\log m}{am}\right)^{1/b}$. If μ satisfies the (a, b, r_0)-standard assumption, then*

$$\left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty \le r_0 \;+\; r_m\, \mathbb{1}_{(r_0,\infty)}(r_m) \;+\; C_1(a,b)\, r_m \,\frac{1}{(\log m)^2},$$

where $C_1(a, b)$ is a constant that depends on a and b.

Chazal et al. (2014a) controls the variance term, which is of the order of $1/\sqrt{n}$. Therefore, if r_0 is negligible, we see that n should be taken of the order of $(m/\log m)^{2/b}$.

We now turn to the closest sample estimator $\widehat{\lambda}_n$ and investigate its ℓ_∞ risk $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right]$, where the expectation is with respect to $(\Psi_\mu^m)^{\otimes n}$. As before, in our analysis we rely on the stability property $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right] \le \mathbb{E}\left[H(\mathbb{X}_\mu, \widehat{C_n^m})\right]$, where the second expectation is with respect to $(\mu^{\otimes m})^{\otimes n}$.

Theorem 9. *Let $r_m = 2\left(\frac{\log(2^b m)}{am}\right)^{\frac{1}{b}}$. If $\mu \in \mathcal{P}(\mathbb{X})$ satisfies the (a, b, r_0)-standard assumption, then*

$$\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right] \le r_0 + r_m \mathbb{1}_{(r_0,\infty)}(r_m) + C_2(a,b)\, r_m \frac{1}{n\,[\log(2^b m)]^{n+1}},$$

where $C_2(a, b)$ is a constant that depends on a and b.

Remark 10. *The risk of the closest subsample method can in principle be smaller than the average landscape method. In Appendix C we show that if μ is the discrete uniform measure on a point cloud of size N, sampled from a measure satisfying the $(a, b, 0)$-standard assumption, then r_0 is of the order of $(\frac{\log N}{N})^{1/b}$. When r_0 is negligible, the rates of theorems 8 and 9 are comparable, both of the order of $O(\frac{\log m}{m})^{1/b}$. However, the average method has another advantage: it is robust to outliers. This point is discussed in detail in Appendix D.*

6 Experiments

In this section, we illustrate our methods. Since computing the persistent homology of the Vietoris-Rips (VR) filtrations built on top of the large samples is infeasible, we resort to the subsampling strategy described in Section 3. More formally, let $X_N = \{x_1, \ldots, x_N\}$ and $Y_N = \{y_1, \ldots, y_N\}$ be two large point clouds. We draw n subsamples each of size $m \ll N$ points from μ and ν, the discrete uniform measures on X_N and Y_N, and we compare the corresponding average landscapes and closest subsample landscapes, induced by the persistent homology of the VR filtrations built on top of the subsamples. We apply this technique to two examples.

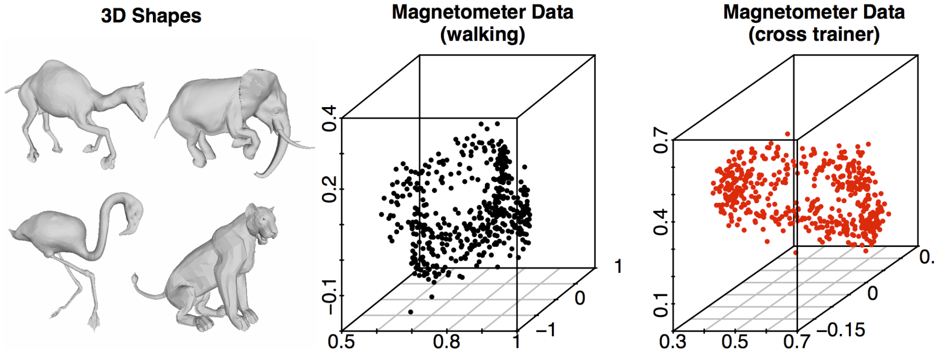

Figure 3: Left: 3D shapes of the first experiment. Middle and Left: 500 random points from the magnetometer data of the second experiment.

3D Shapes. We use the publicly available database of triangulated shapes (Sumner and Popović, 2004). We select a single *pose* (#2) of 4 different classes: *camel, elephant, flamingo, lion*. The 4 shapes are represented in Figure 3. In practice, each shape consists of a 3D point cloud embedded in the Euclidean space, with a number of vertices that ranges from 7K to 40K. The data are normalized, so that the diameter of each shape is 1. For $n = 100$ times we subsample $m = 300$ points from each shape; then we select the closest subsample to the corresponding original point cloud and compute $4 \times n$ persistence diagrams (dimension 1), one for each subsample. See Figure 4: the plot on the left shows the landscapes corresponding to the closest subsamples of m points among the n different subsamples from each shape; the plot in the middle shows the empirical average landscapes within each class, computed as the pointwise average of n landscapes, with a 95% uniform confidence band for the true average landscape, constructed using the method described in Chazal et al. (2014a); the dissimilarity matrix on the right shows the pairwise ℓ_∞ distances between the average landscapes (scale from yellow to red), which, according to Theorem 5, represent a lower bound for the pairwise Wasserstein distances of the discrete uniform measures on the 4 different shapes.

Magnetometer Data. For the second example, we consider the problem of distinguishing human activities performed while wearing inertial and magnetic sensor units. The dataset is publicly available at the UCI Machine Learning Repository[3] and is described in Barshan and Yüksek (2013),

[3] http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities

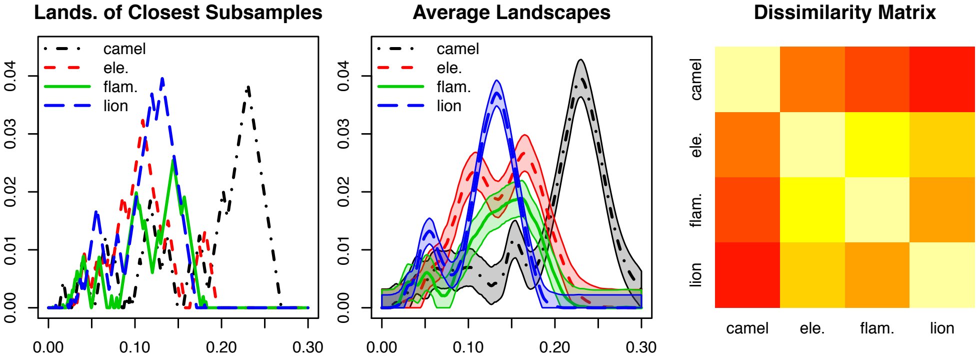

Figure 4: Subsampling methods applied to 3D shapes. For $n = 100$ subsamples of size $m = 300$, for each shape, we constructed the landscapes of the closest subsample (left), the average landscape with 95% confidence band (middle) and the dissimilarity matrix of the pairwise ℓ_∞ distance between average landscapes.

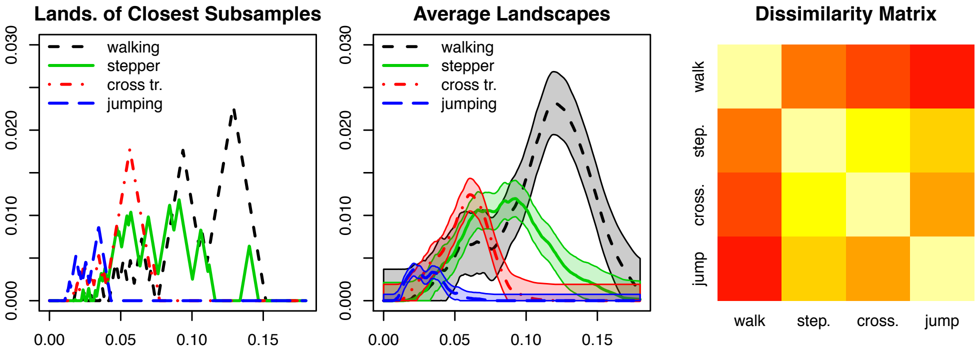

Figure 5: Subsampling methods applied to magnetometer data. For $n = 80$ subsamples of size $m = 200$, for each activity, we constructed the landscapes of the closest subsample (left), the average landscape with 95% confidence band (middle) and the dissimilarity matrix of the pairwise ℓ_∞ distance between average landscapes.

where it is used to classify 19 activities performed by 8 people that wear sensor units on the chest, arms and legs. For ease of illustration, we report here the results on 4 activities (walking, stepper, cross trainer, jumping) performed by a single person (#1). We use the data from the magnetometer of a single sensor (left leg), which measures the direction of the magnetic field in the space at a frequency of 25Hz. For each activity there are 7,500 consecutive measurements that we treat as a 3D point cloud in the Euclidean space. As an example, Figure 3 shows 500 points at random for 2 activities (walking and using a cross trainer). As in the previous example, for $n = 80$ times, we subsample $m = 200$ points from the point cloud of each activity, construct the landscapes of the closest subsamples, the average landscapes (dim 1), and the dissimilarity matrix based on the ℓ_∞ distances of the average landscapes. See Figure 5. To the best of our knowledge persistent homology has never been used to study data from accelerometers or magnetometers before. A remarkable advantage is that the methods of persistent homology are insensitive to the orientation of the input data, as opposed to other methods that require the exact calibration of the sensor units; see, for example, Altun et al. (2010) and Barshan and Yüksek (2013).

7 Conclusion

We have presented a framework for approximating the persistent homology of a set using subsamples. The method is simple and computationally fast. Moreover, we provided stability results for the new summaries and bounds on the risk of the proposed estimators. In the future we will release software for implementing the method. We plan to investigate other methods for speeding up the computations.

References

Kenneth S. Alexander. Rates of growth for weighted empirical processes. In *Proc. of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 475–493, 1985.

Kenneth S. Alexander. The central limit theorem for weighted empirical processes indexed by sets. *J. Multivar. Anal.*, 22(2):313–339, 1987a.

Kenneth S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probab. Theory Related Fields*, 75(3):379–423, 1987b.

Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

Billur Barshan and Murat Cihan Yüksek. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, page bxt075, 2013.

Andrew J. Blumberg, Itamar Gal, Michael A. Mandell, and Matthew Pancia. Persistent homology for metric measure spaces, and robust statistics for hypothesis testing and confidence intervals. *Found. Comput. Math.*, pages 1–45, May 2014.

Jean-Daniel Boissonnat, Tamal K. Dey, and Clément Maria. The compressed annotation matrix: An efficient data structure for computing persistent cohomology. In *Algorithms–ESA 2013*, volume 8125 of *LNCS*, pages 695–706. Springer, 2013.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *arXiv preprint 1207.6437*, 2012.

Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc.*, 46(2):255–308, 2009.

Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y Oudot. Proximity of persistence modules and their diagrams. In *Proc. 25th Annu. Sympos. Comput. Geom*, pages 237–246. ACM, 2009.

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

Frédéric Chazal, Vin De Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geom. Dedicata*, pages 1–22, 2012b.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Proc. 30th Annu. Sympos. Comput. Geom*, 2014a.

Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. In *Proc. 31st Int. Conf. Mach. Learn.*, volume 32, pages 10–18. JMLR W&CP, 2014b. arXiv preprint 1305.6239.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007.

Antonio Cuevas. Set estimation: another bridge between statistics and geometry. *Bol. Estad. Investig. Oper.*, 25(2):71–85, 2009. ISSN 1889-3805.

Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, pages 340–354, 2004.

Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. AMS, 2010.

Brittany Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Statistical inference for persistent homology: Confidence sets for persistence diagrams. *arXiv preprint 1303.7117*, 2013.

Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216, 2006.

Evarist Giné, Vladimir Koltchinskii, and Jon A. Wellner. Ratio limit theorems for empirical processes. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 249–278. Birkhäuser, Basel, 2003.

Aarti Singh, Clayton Scott, and Robert Nowak. Adaptive Hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B):2760–2782, 2009.

Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 399–405. ACM, 2004.

Appendix

A Technical results

In this section we present some technical results that will be used to prove the main theorems.
First, we expand the notation introduced in the body of the paper (Section 3). For any positive integer m, let $\phi_m : \mathbb{X}^m \to \mathcal{D}_T$ be the diagram and $\psi_m : \mathcal{D}_T \to \mathcal{L}_T$ the landscape, i.e. $\phi_m(X) = D_X$, for any $X = \{x_1, \cdots, x_m\} \subset \mathbb{X}$ and $\psi(D_X) = \lambda_{D_X} = \lambda_X$, for any $D_X \in \mathcal{D}_T$. Given $\mu \in \mathcal{P}(\mathbb{X})$, we denote by Φ_μ^m the push-forward measure of $\mu^{\otimes m}$ by ϕ_m, that is $\Phi_\mu^m = (\phi_m)_*\mu$. Similarly, we denote by Ψ_μ^m the push-forward (induced) measure of $\mu^{\otimes m}$ by $\psi \circ \phi_m$, that is $\Psi_\mu^m = (\psi \circ \phi_m)_*\mu$.

For a fixed integer $m > 0$, consider the metric space $(\mathbb{X}, \rho)$ and the space $\mathbb{X}^m$ endowed with a metric ρ_m. We impose two conditions on ρ_m:

(C1) Given a real number $p \geq 1$, for any $X = \{x_1, \ldots, x_m\} \subset \mathbb{X}$ and $Y = \{y_1, \ldots, y_m\} \subset \mathbb{X}$,

$$\rho_m(X, Y) \leq \left(\sum_{i=1}^m \rho(x_i, y_i)^p \right)^{\frac{1}{p}}, \tag{9}$$

(C2) For any $X = \{x_1, \ldots, x_m\} \subset \mathbb{X}$ and $Y = \{y_1, \ldots, y_m\} \subset \mathbb{X}$,

$$H(X, Y) \leq \rho_m(X, Y). \tag{10}$$

Two examples of distance that satisfy conditions (C1) and (C2) are the Hausdorff distance and the L_p-distance $\rho_m(X, Y) = \left(\sum_{i=1}^m \rho(x_i, y_i)^p\right)^{\frac{1}{p}}$.

Lemma 11. *For any probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$ and any metric $\rho_m : \mathbb{X}^m \times \mathbb{X}^m \to \mathbb{R}$ that satisfies (C1), we have*

$$W_{\rho_m, p}(\mu^{\otimes m}, \nu^{\otimes m}) \leq m^{\frac{1}{p}} W_{\rho, p}(\mu, \nu).$$

Remark: The bound of the above lemma is tight: it is an equality when μ is a Dirac measure and ν any other measure.

Proof. Let $\Pi \in \mathcal{P}(\mathbb{X} \times \mathbb{X})$ be a transport plan between μ and ν. Up to reordering the components of $\mathbb{X}^{2m}$, $\Pi^{\otimes m}$ is a transport plan between $\mu^{\otimes m}$ and $\nu^{\otimes m}$ whose p-cost is given by

$$\begin{aligned} \int_{\mathbb{X}^m \times \mathbb{X}^m} \rho_m(X, Y)^p d\Pi^{\otimes m}(X, Y) &\leq \int_{\mathbb{X}^m \times \mathbb{X}^m} \sum_{i=1}^m \rho(x_i, y_i)^p \; d\Pi(x_1, y_1) \cdots d\Pi(x_m, y_m) \\ &= m \int_{\mathbb{X} \times \mathbb{X}} \rho(x_1, y_1)^p d\Pi(x_1, y_1). \end{aligned}$$

The lemma follows by taking the minimum over all transport plans on both sides of this inequality. □

Lemma 12. *For any probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$ and any metric $\rho_m : \mathbb{X}^m \times \mathbb{X}^m \to \mathbb{R}$ that satisfies (C2), we have*

$$W_{d_b, p}\left(\Phi_\mu^m, \; \Phi_\nu^m\right) \leq W_{\rho_m, p}(\mu^{\otimes m}, \nu^{\otimes m}).$$

Proof. This is a consequence of the stability theorem for persistence diagrams. Given $X, Y \subset \mathbb{X}^m$, define

$$\Lambda_m(X, Y) = (D_X, D_Y).$$

If $\Pi \in \mathcal{P}(\mathbb{X}^m \times \mathbb{X}^m)$ is a transport plan between $\mu^{\otimes m}$ and $\nu^{\otimes m}$ then $\Lambda_{m,*}\Pi$ is a transport plan between Φ_μ^m and Φ_ν^m. Its p-cost is given by

$$\begin{aligned}\int_{\mathcal{D}_T \times \mathcal{D}_T} d_b(D_X, D_Y)^p d\Lambda_{m,*}\Pi(D_X, D_Y) &= \int_{\mathbb{X}^m \times \mathbb{X}^m} d_b(\phi_m(X), \phi_m(Y))^p d\Pi(X,Y) \\ &\leq \int_{\mathbb{X}^m \times \mathbb{X}^m} H(X,Y)^p d\Pi(X,Y) \ \text{(stability theorem)} \\ &\leq \int_{\mathbb{X}^m \times \mathbb{X}^m} \rho_m(X,Y)^p d\Pi(X,Y).\end{aligned}$$

The lemma follows by taking the minimum over all transport plans on both sides of this inequality. □

Lemma 13. *Let μ and ν be two probability measures on $\mathbb{X}$. Let $\lambda_X \sim \Psi_\mu^m$ and $\lambda_Y \sim \Psi_\nu^m$. Then*

$$\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq W_{d_b,p}\left(\Phi_\mu^m,\ \Phi_\nu^m\right).$$

Proof. Let Π be a transport plan between Φ_μ^m and Φ_ν^m. For any $t \in \mathbb{R}$ we have

$$\begin{aligned}\left| \mathbb{E}_{\Psi_\mu^m}[\lambda_X](t) - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y](t) \right|^p &= |\mathbb{E}[\lambda_X(t) - \lambda_Y(t)]|^p \\ &\leq \mathbb{E}\left[|\lambda_X(t) - \lambda_Y(t)|^p\right] \ \text{(Jensen inequality)} \\ &\leq \mathbb{E}\left[d_b(D_X, D_Y)^p\right] \ \text{(Stability of landscapes)} \\ &= \int_{\mathcal{D}_T \times \mathcal{D}_T} d_b(D_X, D_Y)^p d\Pi(D_X, D_Y) \\ &= C_p(\Pi)^p.\end{aligned}$$

□

The following lemma is similar to Theorem 2 in Chazal et al. (2014b).

Lemma 14. *Let X be a sample of size m from a measure $\mu \in \mathcal{P}(\mathbb{X})$ that satisfies the (a, b, r_0)-standard assumption. Let $r_m = 2\left(\frac{\log m}{am}\right)^{1/b}$. Then*

$$\mathbb{E}\left[H(X, \mathbb{X}_\mu)\right] \leq r_0 + 2\left(\frac{\log m}{am}\right)^{1/b} \mathbb{1}_{(r_0,\infty)}(r_m) + 2\, C_1(a,b)\left(\frac{\log m}{am}\right)^{1/b} \frac{1}{(\log m)^2},$$

where $C_1(a, b)$ is a constant depending on a and b.

Proof. Let $r > r_0$. It can be proven that $q := \mathrm{Cv}\left(\mathbb{X}_\mu, r/2\right) \leq \frac{4^b}{ar^b} \vee 1$, where Cv$(\mathbb{X}_\mu, 2r)$ denotes the number of balls of radius $r/2$ that are necessary to cover $\mathbb{X}_\mu$. Let $\mathcal{C} = \{x_1, \ldots, x_p\}$ be a set of centers such that $B(x_1, r/2), \ldots, B(x_p, r/2)$ is a covering of $\mathbb{X}_\mu$. Then,

$$\begin{aligned}\mathbb{P}\left(H(X, \mathbb{X}_\mu) > r\right) &\leq \mathbb{P}\left(H(X, \mathcal{C}) + H(\mathcal{C}, \mathbb{X}_\mu) > r\right) \\ &\leq \mathbb{P}\left(H(X, \mathcal{C}) > r/2\right) \\ &\leq \mathbb{P}\left(\exists i \in \{1, \cdots, p\} \text{ such that } X \cap B(x_i, r/2) = \emptyset\right) \\ &\leq \sum_{i=1}^p \mathbb{P}\left(X \cap B(x_i, r/2) = \emptyset\right) \\ &\leq \frac{4^b}{ar^b}\left[1 - \inf_{i=1\ldots p} \mathbb{P}(B(x_i, r/2))\right]^m \\ &\leq \frac{4^b}{ar^b}\left[1 - \frac{ar^b}{2^b}\right]^m \\ &\leq \frac{4^b}{ar^b} \exp\left(-m\frac{a}{2^b} r^b\right)\end{aligned}$$

Then

$$\mathbb{E}\left[H(X,\mathbb{X}_\mu)\right] = \int_{r>0} \mathbb{P}\left(H(X,\mathbb{X}_\mu) > r\right) dr$$
$$\leq r_0 + \int_{r>r_0} \mathbb{P}\left(H(X,\mathbb{X}_\mu) > r\right) dr. \tag{11}$$

If $r_m \leq r_0$ then the last quantity in (11) is bounded by

$$r_0 + \int_{r>r_m} \mathbb{P}\left(H(X,\mathbb{X}_\mu) > r\right) dr,$$

otherwise (11) is bounded by

$$r_0 + \int_{r>0} \mathbb{P}\left(H(X,\mathbb{X}_\mu) > r\right) dr \leq r_0 + r_m + \int_{r>r_m} \mathbb{P}\left(H(X,\mathbb{X}_\mu) > r\right) dr.$$

In either case, we follow the strategy in Chazal et al. (2014b) to obtain the following bound:

$$\int_{r>r_m} \mathbb{P}\left(H(X,\mathbb{X}_\mu) > r\right) dr \leq 2\,C(a,b)\left(\frac{\log m}{am}\right)^{1/b} \frac{1}{(\log m)^2},$$

which implies that

$$\mathbb{E}\left[H(X,\mathbb{X}_\mu)\right] \leq r_0 \,+\, r_m \mathbb{1}_{(r_0,\infty)}(r_m) \,+\, 2\,C(a,b)\left(\frac{\log m}{am}\right)^{1/b} \frac{1}{(\log m)^2}.$$

□

B Main Proofs

Proof of Theorem 5 It immediately follows from the three following inequalities of Lemmas 11, 12 and 13:

- upperbound on the Wasserstein distance between the tensor product of measures:
$$W_{\rho_m,p}(\mu^{\otimes m},\nu^{\otimes m}) \leq m^{\frac{1}{p}} W_{\rho,p}(\mu,\nu)$$
- from measures on $\mathbb{X}^m$ to measures on $\mathcal{D}$:
$$W_{d_b,p}\left(\Phi_\mu^m,\ \Phi_\nu^m\right) \leq W_{\rho_m,p}(\mu^{\otimes m},\nu^{\otimes m})$$
- from measures on $\mathcal{D}$ to difference of the expected landscapes:
$$\left\|\mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y]\right\|_\infty \leq W_{d_b,p}\left(\Phi_\mu^m,\ \Phi_\nu^m\right)$$

□

Proof of Theorem 6

$$\|\mathbb{E}_{\Psi_\mu^m}(\lambda_X) - \mathbb{E}_{\Psi_\nu^m}(\lambda_Y)\|_\infty = \int_{\varepsilon>0} \mathbb{P}_{\Psi_\mu^m \otimes \Psi_\nu^m}\left(\|\lambda_X - \lambda_Y\|_\infty > \varepsilon\right) d\varepsilon$$
$$= \varepsilon_0 + \int_{\varepsilon>\varepsilon_0} \mathbb{P}_{\Psi_\mu^m \otimes \Psi_\nu^m}\left(\|\lambda_X - \lambda_Y\|_\infty > \varepsilon\right) d\varepsilon. \tag{12}$$

The event $\{\|\lambda_X - \lambda_Y\|_\infty > \varepsilon\}$ inside the integral implies that

$$\varepsilon_0 \leq \varepsilon < H(X,Y) \leq H(X,\mathbb{X}_\mu) + H(\mathbb{X}_\mu,\mathbb{X}_\nu) + H(Y,\mathbb{X}_\nu), \tag{13}$$

where X and Y are two samples of m points from μ and ν, respectively. Let $\varepsilon_0 = H(\mathbb{X}_\mu,\mathbb{X}_\nu)$. By (13) it follows that at least one of the following conditions holds:

$$H(X,\mathbb{X}_\mu) \geq \frac{\varepsilon - \varepsilon_0}{2},$$
$$H(Y,\mathbb{X}_\nu) \geq \frac{\varepsilon - \varepsilon_0}{2}.$$

We assume that the first condition holds (the other case follows similarly). Then the last quantity in equation (12) can be bounded by

$$
\begin{aligned}
&\varepsilon_0 + \int_{\varepsilon > \varepsilon_0} \mathbb{P}\left(H(X, \mathbb{X}_\mu) \geq \frac{\varepsilon - \varepsilon_0}{2}\right) d\varepsilon \\
=&H(\mathbb{X}_\mu, \mathbb{X}_\nu) + 2\int_{u>0} \mathbb{P}\left(H(X, \mathbb{X}_\mu) \geq u\right) du \\
=&H(\mathbb{X}_\mu, \mathbb{X}_\nu) + 2\mathbb{E}\left[H(X, \mathbb{X}_\mu)\right] \\
\leq&H(\mathbb{X}_\mu, \mathbb{X}_\nu) + 2r_0 \; + \; 4\left(\frac{\log m}{am}\right)^{1/b} \mathbb{1}_{(r_0, \infty)}(r_m) \; + \; 4\, C_1(a,b)\left(\frac{\log m}{am}\right)^{1/b} \frac{1}{(\log m)^2},
\end{aligned}
$$

where the last inequality follows from Lemma 14. □

Proof of Theorem 8 It follows directly from (8) and Lemma 14 . □

Proof of Theorem 9

$$
\begin{aligned}
\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right] &\leq \mathbb{E}\left[H(\mathbb{X}_\mu, \widehat{C_n^m})\right] \\
&\leq \int_{r>0} \mathbb{P}\left(H(\mathbb{X}_\mu, \widehat{C_n^m}) > r\right) dr \\
&\leq r_0 + \int_{r>r_0} \left[\mathbb{P}\left(H(\mathbb{X}_\mu, S_1^m) > r\right)\right]^n dr \\
&\leq r_0 + \int_{r>r_0} \left[\frac{4^b}{ar^b} \exp\left(-m\frac{a}{2^b} r^b\right)\right]^n dr,
\end{aligned}
$$

where the last inequality follows from Lemma 14. If $r_m \leq r_0$ then the last term is upper bounded by

$$
r_0 + \int_{r>r_m} \left[\frac{4^b}{ar^b} \exp\left(-m\frac{a}{2^b} r^b\right)\right]^n dr,
$$

otherwise it is bounded by

$$
r_0 + r_m + \int_{r>r_m} \left[\frac{4^b}{ar^b} \exp\left(-m\frac{a}{2^b} r^b\right)\right]^n dr.
$$

In either case,

$$
\begin{aligned}
\int_{r>r_m} \left[\frac{4^b}{ar^b} \exp\left(-m\frac{a}{2^b} r^b\right)\right]^n dr &= 2\frac{2^{bn}}{b}(ma)^{-1/b} m^n \int_{u>\log m} u^{1/b-n-1} \exp(-nu) du \\
&\leq 2C_2(a,b)\left(\frac{\log(2^b m)}{am}\right)^{1/b} \frac{1}{n\left[\log(2^b m)\right]^{n+1}} \; ,
\end{aligned}
$$

where in the last inequality we applied the same strategy used to prove Theorem 2 in Chazal et al. (2014b). □

C About the (a, b, r_0)-standard assumption

The aim of this section is to explain why the (a, b, r_0)-standard assumption is relevant, in particular when μ is a discrete measure. Our argument is related to weighted empirical processes, which have been studied in details by Alexander; see Alexander (1985, 1987b,a). A new look on this problem has been proposed more recently in Giné and Koltchinskii (2006); Giné et al. (2003) by using Talagrand concentration inequalities. The following result from Alexander (1985) will be sufficient here. Let $(\mathbb{X}, \rho, \eta)$ be a measure metric space and let η_N be the empirical counterpart of η.

Proposition 15. *Let $\mathcal{C}$ be a VC class of measurable sets of index v of $\mathbb{X}$. Then for every δ, $\varepsilon > 0$ there exists K such that*

$$
\eta\left[\sup\left\{\left|\frac{\eta_N(C) - \eta(C)}{\eta(C)}\right| \; : \; \eta(C) \geq Kv\frac{\log N}{N}, \, C \in \mathcal{C}\right\} > \varepsilon\right] = O(N^{-(1+\delta)v}). \qquad (14)
$$

Assume that μ is the discrete uniform measure on a point cloud $X_N = \{x_1, \dots, x_N\}$ which has been sampled from η, thus $\mu = \eta_N$. Assume moreover that η satisfies an (a', b)-standard assumption ($r_0 = 0$). Let r_0 be a positive function of N chosen further. For any $r > r_0(N)$ and any $y \in \mathbb{X}_\mu$:

$$\begin{aligned}
\inf_{y \in \mathbb{X}_\mu} \mu(B(y,r)) &= \inf_{y \in \mathbb{X}_\mu} \eta_N(B(y,r)) \\
&= \inf_{y \in \mathbb{X}_\mu} \left\{ \eta(B(y,r)) \left[1 - \frac{\eta(B(y,r)) - \eta_N(B(y,r))}{\eta(B(y,r))} \right] \right\} \\
&\geq (1 \wedge a' r^b) \inf_{y \in \mathbb{X}_\mu} \left\{ 1 - \sup_{x \in \mathbb{X}} \left| \frac{\eta(B(x,r)) - \eta_N(B(x,r))}{\eta(B(x,r))} \right| \right\} \\
&\geq (1 \wedge a' r^b) \inf_{y \in \mathbb{X}_\mu} \left\{ 1 - \sup_{x \in \mathbb{X}, r' \geq r_0(N)} \left| \frac{\eta(B(x,r')) - \eta_N(B(x,r'))}{\eta(B(x,r'))} \right| \right\} (15)
\end{aligned}$$

Assume that the set of balls in $(\mathbb{X}, \rho)$ has a finite VC-dimension v. For instance, in $\mathbb{R}^d$, the VC-dimension of balls is d+1. Under this assumption we apply Alexander's Proposition with (for instance) $\delta = 1$ and $\varepsilon = 1/2$. Let $K > 0$ such that (14) is satisfied. Then, by setting

$$r_0(N) := \left(\frac{Kv}{a'} \frac{\log N}{N} \right)^{1/b},$$

we finally obtain using (14) and (15) that

$$\eta \left[\inf_{y \in \mathbb{X}_\mu,\, r \geq r_N} \mu(B(y,r)) \geq 1 \wedge \frac{a'}{2} r^b \right] = O(N^{-2v}).$$

In this quite general context, we see that by taking r_0 of the order of $\left(\frac{\log N}{N} \right)^{1/b}$, for large values of N the (a, b, r_0)-standard assumption is satisfied with high probability (in η).

D Robustness to Outliers

The average landscape method is insensitive to outliers, as can be seen by the stability result of Theorem 5. The probability mass of an outlier gives a minimal contribution to the Wasserstein distance on the right hand side of the inequality.

For example, suppose that $X_N = \{x_1, \dots, x_N\}$ is a random sample from the unit circle $\mathbb{S}^2$, and let $Y_N = X_N \backslash \{x_1\} \cup \{(0,0)\}$. See Figure 6. The landscapes λ_{X_N} and λ_{Y_N} are very different because of the presence of the outlier $(0,0)$. On the other hand, the average landscapes constructed by multiple subsamples of $m < N$ points from X_N and Y_N are close to each other. Formally, let μ be the discrete uniform measure that put mass $1/N$ on each point of X_N and similarly let ν be the discrete uniform measure on Y_N. The 1st Wasserstein distance between the two measure is 1/N and, according to Theorem 5, the difference between the average landscapes is $\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq \frac{m}{N}$.

More formally, we can show that the average landscape $\overline{\lambda_n^m}$ can be more accurate than the closest subsample method when there are outliers. In fact, $\overline{\lambda_n^m}$ can even be more accurate than the landscape corresponding to a large sample of N points.

Suppose that the large, given point cloud $X_N = \{x_1, \dots, x_N\}$ has a small fraction of outliers. Specifically, $X_N = \mathcal{G} \bigcup \mathcal{B}$ where $\mathcal{G} = \{x_1, \dots, x_G\}$ are the good observations and $\mathcal{B} = \{y_1, \dots, y_B\}$ are the outliers (bad observations). Let $G = |\mathcal{G}|$, $B = |\mathcal{B}|$ so that $N = G + B$ and let $\epsilon = B/N$ which we assume is small but positive. Our target is the landscape based on the non-outliers, namely, $\lambda_{\mathcal{G}}$. The presence of the outliers means that $\lambda_{X_N} \neq \lambda_{\mathcal{G}}$. Let $\beta = \inf_S ||\lambda_S - \lambda_{\mathcal{G}}||_\infty > \delta$, for some $\delta > 0$, where the infimum is over all subsets that contain at least one outlier. Thus, β denotes the minimal bias due to the outliers. We consider three estimators:

$$\begin{aligned}
\lambda_{X_N} &: \text{landscape from full given sample } X_N \\
\overline{\lambda_n^m} &: \text{average landscape from } n \text{ subsamples of size } m \\
\widehat{\lambda_n^m} &: \text{landscape of closest subsample, from } n \text{ subsamples of size } m.
\end{aligned}$$

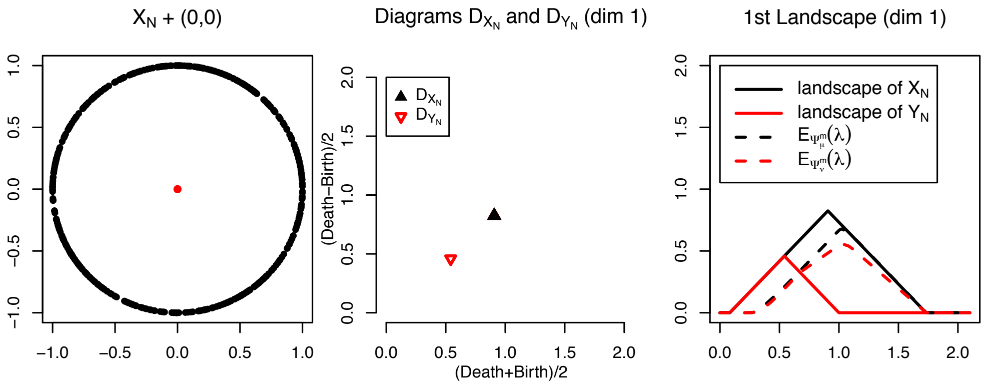

Figure 6: Left: X_N is the set of $N = 500$ points on the unit circle; $Y_N = X_N \backslash \{x_1\} \cup \{(0,0)\}$. Middle: persistence diagrams (dim 1) of the VR filtrations on X_N and Y_N, in the same plot, with different symbols. Right: landscapes of X_N, Y_N and the corresponding average landscapes constructed by subsampling $m = 100$ points from the two sets, for $n = 30$ times.

The last two estimators are defined in Section 3, and are constructed using n independent samples of size m from the discrete uniform measure that puts mass $1/N$ on each point of X_N.

Proposition 16. *If $\epsilon = o(1/n)$, then, for large enough n and m,*

$$\mathbb{E}\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty < \mathbb{E}\|\lambda_{X_N} - \lambda_{\mathcal{G}}\|_\infty. \tag{16}$$

In addition, if $nm\epsilon \to \infty$ then

$$\mathbb{P}\left[\mathbb{E}\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty < \|\widehat{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty\right] \to 1. \tag{17}$$

Proof. Say that a subsample is clean if it contains no outliers and that it is infected if it contains at least one outlier. Let $S_1, \ldots, S_n$ be the subsamples of X_N of size m. Let $I = \{i : \ S_i \text{ is infected}\}$ and $C = \{i : \ S_i \text{ is clean}\}$. Then

$$\overline{\lambda_n^m} = \frac{n_0}{n}\lambda_0 + \frac{n_1}{n}\lambda_1$$

where n_0 is the number of clean subsamples, $n_1 = n - n_0$ is the number of infected subsamples, $\lambda_0 = (1/n_0)\sum_{i \in C} \lambda_{S_i}$ and $\lambda_1 = (1/n_1)\sum_{i \in I} \lambda_{S_i}$. Hence,

$$\begin{aligned}\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty &\leq \frac{n_0}{n}\|\lambda_0 - \lambda_{\mathcal{G}}\|_\infty + \frac{n_1}{n}\|\lambda_1 - \lambda_{\mathcal{G}}\|_\infty \\ &\leq \frac{n_0}{n}\|\lambda_0 - \lambda_{\mathcal{G}}\|_\infty + \frac{Tn_1}{2n}.\end{aligned}$$

A subsample is clean with probability $(1-\epsilon)^m$. Thus, $n_0 \sim \text{Binomial}(n, (1-\epsilon)^m)$ and $n_1 \sim \text{Binomial}(n, 1-(1-\epsilon)^m)$. Let $\pi = 1-(1-\epsilon)^m$. By Hoeffding's inequality

$$\mathbb{P}\left(\frac{Tn_1}{2n} > \frac{\beta}{2}\right) = \mathbb{P}\left(\frac{Tn_1}{2n} - \frac{T\pi}{2} > \frac{\beta}{2} - \frac{T\pi}{2}\right) \leq \exp\left(-2n\left(\frac{\beta}{T} - \pi\right)^2\right).$$

Since $\epsilon = o(1/n)$, we eventually have that

$$\pi = 1 - (1-\epsilon)^m < \frac{\beta}{T} - \sqrt{\frac{\log n}{2n}},$$

which implies that $\mathbb{P}\left(\frac{Tn_1}{2n} > \frac{\beta}{2}\right) < 1/n$. So, except on a set of probability tending to 0,

$$\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty \leq \frac{n_0}{n}\|\lambda_0 - \lambda_{\mathcal{G}}\|_\infty + \frac{\beta}{2} \leq \|\lambda_0 - \lambda_{\mathcal{G}}\|_\infty + \frac{\beta}{2}$$

and thus, as soon as n, m and N are large enough,

$$\mathbb{E}\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty \leq \frac{\beta}{2} + \frac{\beta}{2} = \beta \leq \|\lambda_{X_N} - \lambda_{\mathcal{G}}\|_\infty.$$

This proves the first claim. To prove the second claim, note that the probability that at least one subsample is infected is $1-(1-\epsilon)^{nm} \sim 1-e^{-\epsilon nm} \to 1$. So with probabilty tending to 1, there will be an infected subsample. This subsample will minimize $H(X, S_j)$ and the landscape based on this selected subsample will have a bias of order β. □

In practice, we can increase the robustness further, by using filtered subsampling. This can be done using distance to k-th nearest neighbor or using a kernel density estimator. For example, let

$$\widehat{p}_h(x) = \frac{1}{N}\sum_{j=1}^{N} K\left(\frac{||x-X_i||}{h}\right)$$

be a kernel density estimator with bandwidth h and kernel K. Suppose that all subsamples are chosen from the filtered set $\mathcal{F} = \{X_i : \widehat{p}_h(X_i) > t\}$. Suppose that the good observations $\mathcal{G}$ are sampled from a distribution on a set $A \subset [0,1]^d$ satisfying the (a,b)-standard condition with $b < d$, $a > 0$ and that $\mathcal{B}$ consists of B observations sampled from a uniform on $[0,1]^d$. For any $x \in A$,

$$\mathbb{E}[\widehat{p}_h(x)] \approx \frac{ah^b}{h^d}$$

and for any outlier X_i we have (for h small enough) that $\widehat{p}_h(X_i) = 1/(nh^d)$. Hence, if we choose t such that

$$\frac{1}{nh^d} < t < \frac{a}{h^{d-b}}$$

then $\mathcal{F} = \mathcal{G}$ with high probability.