GEOMETRIC AND PROBABILISTIC LIMIT THEOREMS IN TOPOLOGICAL DATA ANALYSIS

SARA KALIŠNIK, CHRISTIAN LEHN, AND VLADA LIMIC

ABSTRACT. We develop a general framework for the probabilistic analysis of random finite point clouds in the context of topological data analysis. We extend the notion of a barcode of a finite point cloud to compact metric spaces. Such a barcode lives in the completion of the space of barcodes with respect to the bottleneck distance, which is quite natural from an analytic point of view. As an application we prove that the barcodes of i.i.d. random variables sampled from a compact metric space converge to the barcode of the support of their distribution when the number of points goes to infinity. We also examine more quantitative convergence questions for uniform sampling from compact manifolds, including expectations of transforms of barcode valued random variables in Banach spaces. We believe that the methods developed here will serve as useful tools in studying more sophisticated questions in topological data analysis and related fields.

CONTENTS

1. INTRODUCTION

Topological Data Analysis (TDA) is a fast growing field whose aim is to provide a set of new topological and geometric tools for analyzing data. One of the most widely used tools is persistent homology. The ideas behind persistent homology can be traced back to the works of Patrizio Frosini [Fro92] on size functions, and of Vanessa Robins [Rob99] on using experimental data to infer the topology of attractors in dynamical systems, though the method only gained prominence with the pioneering works of Edelsbrunner, Letscher and

Zomorodian [ELZ02] and Carlsson and Zomorodian [CZ05]. Persistent homology has been used to address problems in fields ranging from sensor networks [GdS06, AC15], medicine [FS10, ARC14], neuroscience [CBK09, CIVCY13, GPCI15], as well as imaging analysis [PC14].

The input of persistent homology is usually a point cloud, i.e. a finite metric space. Since finitely many points do not carry any nontrivial topological information, the idea is to consider the homology of thickenings of this point cloud in order to deduce information about the data or the distribution it is sampled from. The output is a barcode, i.e. a multiset of intervals, where each interval ("bar") represents a topological feature present at parameter values specified by the interval. This space of barcodes $\mathcal{B}$ comes equipped with natural metrics, for example the Wasserstein and the Bottleneck distance.

The present paper grew out of an attempt to understand how some of the fundamental aspects of persistent homology *and probability theory* could interact in order to allow for further statistical applications. Here and in the rest of the introduction we will present some of our key results.

Firstly, we wish to extend the notion of a barcode from finite sets to compact sets. This is done in

Proposition 2.17. *Let $k \in \mathbb{N}_0$ be a nonnegative integer, and let M be a metric space. Then for every compact set $K \subset M$ there is a barcode $\beta_k(K) \in \hat{\mathcal{B}}_\infty$ such that $K \mapsto \beta_k(K)$ is a 1-Lipschitz map from the space of compact subsets of M, equipped with the Hausdorff distance, to the completion $\hat{\mathcal{B}}_\infty$ of the barcode space $\mathcal{B}$, with respect to the bottleneck distance.*

This result can also be obtained from the main theorem of [CSEH07]. It was later explicitly stated and proved in [CdSO14, Proposition 5.1] and relied on a measure theoretic approach to persistent homology introduced in [COGDS16]. For completeness, we include a simple, conceptually clear, and self-contained proof. See Remark 2.19 for an extension to totally bounded spaces.

Since we use a limiting procedure to define β_k on compact subsets of M, the barcode $\beta_k(K)$ has to live in the completion $\hat{\mathcal{B}}_\infty$, which is a natural space for doing analysis with barcodes.

Suppose now that the point cloud is obtained by sampling independent and identically distributed (i.i.d.) points from an unknown distribution with compact support C. The following question seems very natural, and it is somewhat surprising that it has not yet been answered:

What happens to the barcode as we sample more and more such points?

In Section 3, we provide the following intuitive answer.

Theorem 3.1. *Let M be a metric space, $X_1, X_2, \ldots$ be i.i.d. M-valued random variables, and $k \in \mathbb{N}_0$. Define $P_n = \{X_1, X_2, \ldots, X_n\}$. If the distribution of the X_i has support equal to a compact subset $C \subset M$, then*

$$\beta_k(C) = \lim_{n \longrightarrow \infty} \beta_k(P_n) \quad \text{almost surely.}$$

In the stochastic setting we also address questions about the mean and deviation from (or concentration about) the mean. For this discussion we consider random variables taking values in some Banach space. Starting from a barcode valued random variable β (e.g. $\beta = \beta_k(P_n)$ as above), one can obtain a Banach space valued random variable, as in [ACC16, Bub15, DP19, Kal18, RHBK15] and many others. In this paper we use the functions proposed by [Kal18]

as a primary example, though the results are stated in full generality for Lipschitz continuous maps

$$T : \hat{\mathcal{B}}_\infty \longrightarrow V$$

from the completed space of barcodes to some space V. Mapping to a Banach space is necessary to be able to talk about stochastic quantities such as expectation. But even more importantly, in order to use the information contained in the barcode using machine learning algorithms, one needs to produce a vector valued output. As mentioned above, there is a plethora of methods to produce such an output and our setup covers all of them, the only hypothesis being that the map T is Lipschitz continuous. Note that it has been shown in [CB19] that one cannot embed the space of barcodes with finitely many functions which is why we stress the (infinite dimensional) Banach setup.

The study of probabilistic properties of $T \circ \beta$ naturally leads to the law of large numbers and a central limit theorem, as Bubenik first observed in [Bub15]. They can be formulated as follows.

Theorems 4.1 and 4.2. —

- *Let $T : \hat{\mathcal{B}}_\infty \longrightarrow V$ be a continuous map from the (bottleneck completed) space of barcodes to a separable Banach space V and let $\{X_i\}_{i\in\mathbb{N}}$ be an i.i.d. sequence of $\hat{\mathcal{B}}_\infty$-valued random variables such that $\mathbb{E}[\|T(X_1)\|] < \infty$. Then the sequence $(S_n)_n$ of empirical means*

$$S_n := \frac{T(X_1) + \ldots + T(X_n)}{n}$$

  converges almost surely to $\mathbb{E}[T(X_1)]$.
- *Suppose that in addition $\mathbb{E}[T(X_1)] = 0$ and $\mathbb{E}[\|T(X_1)\|^2] < \infty$, and let S_n be as above. If V is of type 2, then $(\sqrt{n}S_n)_n$ converges weakly to a Gaussian random variable with the covariance structure of $T(X_1)$.*

The reader may be concerned about the vacuousness of the just stated result due to its rather abstract setting. We respond by addressing the important situation of barcodes of compact metric spaces (in particular including that of point clouds sampled from a distribution with compact support).

Theorem 4.3. *Let M be a metric space, $K(M)$ the complete metric space of all compact subsets of M with the Hausdorff distance, and X a random variable taking values in $K \in K(M)$. Consider the k-th barcode map $\beta_k : K(M) \longrightarrow \hat{\mathcal{B}}_\infty$ and let $T : \hat{\mathcal{B}}_\infty \longrightarrow V$ be a continuous map, where V is a separable Banach space of type 2. Then $\|T(\beta_k(X))\|$ has finite n-th moments for all $n \geq 0$.*

Once the existence of barcode expectations is settled, it is important to know how to calculate them for random point clouds of bigger and bigger samples, drawn from an unknown distribution. The TDA pipeline is too complicated for permitting one to find an explicit symbolic way for such calculations in general. The only reasonable way of doing so is to make an *educated guess*! We infer directly from Theorem 3.1

Corollary 1.1. Let M be a metric space, and $X_1, X_2, \ldots$ an i.i.d. sequence of M-valued random variables. Set $k \in \mathbb{N}_0$, and put $P_n = \{X_1, X_2, \ldots, X_n\}$. If the distribution of the X_i has support equal to a compact subset $C \subset M$, and if $T : \hat{\mathcal{B}}_\infty \to V$ is a continuous map to a Banach space of type 2, then

$$T\left(\beta_k(C)\right) = \lim_{n \to \infty} \mathbb{E}[T\left(\beta_k(P_n)\right)].$$

For some specific underlying probability distributions, explicit calculations and more careful asymptotic estimates may be possible. We consider the simplest (and paradigm) example of the circle $\mathbb{S}^1 \subset \mathbb{R}^2$, and i.i.d. points sampled from it. The interesting barcode here is the β_1-barcode, and it is uniquely determined by its length. In Theorem 5.5 we give an explicit formula for the expectation of the length.

The principal contribution of this work is that we devise a new concrete general framework for analysis of random finite point clouds and their corresponding barcodes. The fact that the proofs of our main results are not technically involved is in our opinion a firm indication that the framework here proposed is natural and potentially very useful in studying more sophisticated TDA questions.

Related work. There are a number of related approaches to studying the statistical properties of persistent homological estimators (see [CM17] for an overview).

A closely related work is by Bubenik [Bub15], who develops statistical inference via an embedding with "persistence landscapes", which is further studied in [CFL+15b] and [CFL+15a]. Like Bubenik, we use CLT and LLN theory in Banach spaces, but on an object different from his. Unlike him, we study natural geometric and probabilistic limits directly on the barcodes of large point clouds (Theorem 3.1 and Theorem 6.5). In particular, in Theorem 6.5, we establish a connection with the work of [NSW08]. The just mentioned theorems are linked in spirit to [BM15], who also study homology approximations based on large point clouds drawn from a compact manifold, and their analysis is based on [NSW08] as well. Unlike us, for large n, Bobrowski and Mukherjee [BM15] approximate simultaneously the homology of the manifold in a large range m_n of degrees, with the homology in the corresponding degrees of the point cloud inflated by r_n, where r_n is a power of $\frac{\log(n)}{n}$. In our context of persistent homology, we look at a continuum (a segment) of radii (away from zero) and aim to match, for large n, the homotopy type of the manifold with that of the inflated point cloud, simultaneously for all radii in thus fixed interval.

Chazal et al. [CGLM15] establish convergence rates for metric spaces endowed with a probability measure that satisfies the (a, b)-standard assumption, see section 2.2 of that paper. In our study of almost sure convergence, we do not impose any conditions on the measure (except for compact support), see Theorem 3.1. Hiraoka, Shirai, and Trinh [HST18], Owada [Owa18], Adler and Owada [OA17] also study limit theorems for persistence diagrams; but in their case, the point clouds are stationary point processes on $\mathbb{R}^n$. Similar results also appear in [CGLM15].

The foundational work of Mileyko, Mukherjee and Harer in [MMH11] introduces probability measures on barcode space, and these ideas are developed further (with Turner) in the context of Frechét means as ways of summarizing barcode distributions in [TMMH14]. Since we work

with embeddings into Banach spaces, we do not need to rely on the theory developed in these papers.

Another active area of research in TDA deals with topological features of random simplicial complexes and noise [BK18, ABW14]. The present paper has a different focus, but it would be interesting to incorporate noise into our framework. This will be the subject of a forthcoming work.

Outline of the article. In Section 2, we recall the definition of persistent homology, barcodes, and the space of barcode representations. This is the basis for what follows. Most results presented in this section are not new, nevertheless we occasionally included arguments (such as for Lemma 2.16) to make the text more self-contained. As explained above, equivalent statements to Proposition 2.17, where the definition of barcode representations is extended from finite point clouds to compact subsets of a given metric space (with the induced metric), already appeared in the literature. As our approach through completions is conceptually very clear, we nevertheless chose to include it. This definition is fundamental for the rest of the paper. The generalized barcode representations live in the completion $\hat{\mathcal{B}}_\infty$ of the space of barcodes with respect to the bottleneck distance, thus they can be thought of as representing barcodes consisting of countably many intervals with a finite metric distance to a given barcode. In the same spirit as Proposition 2.17 we show in Theorem 2.8 that the filtration associated to a limit of tame functions also has an associated barcode representation.

Following Bubenik [Bub15, Section 3.2], we study a Law of Large Numbers and a Central Limit Theorem in Section 4. The main new contribution here is Theorem 4.3 as explained above. Section 2.4 contains the probabilistic limit theorem 3.1 for barcodes which are probably the most fundamental contribution of this article. It heavily relies on Lemma 3.2 which is a more geometric limit theorem for random point clouds in the space of compact subspaces of $\mathbb{R}^d$.

Section 5 is independent of the preceding two sections and gives a hint at a more quantitative version of a limit theorem. For this we consider the simplest nontrivial example of a compact metric space in $\mathbb{R}^2$ – the circle $\mathbb{S}^1$. We fix the number n of points and we would like to determine the expected barcode of a random n-point cloud (consisting of independent uniform samples from $\mathbb{S}^1$). To give meaning to this idea, we need to find some quantity which determines the barcode and over which we can average (in order to speak of expectations). In this simple case, the elementary geometry of the circle (and of point clouds on it) only allow for a restricted barcode which is entirely determined by its length, see Corollary 5.3 and the discussion thereafter. We give explicit formulas for the *expected length* for $n = 3$ in Proposition 5.4 and arbitrary n in Theorem 5.5.

Such quantities as the length which determine the barcode completely can of course no longer be explicitly given for arbitrary compact submanifolds of $\mathbb{R}^d$ which is why in order to talk about expectation we consider embeddings (or more generally continuous maps) $T : \hat{\mathcal{B}}_\infty \to V$ to some Banach space $(V, \|\cdot\|_V)$. Building on the work of Niyogi–Smale–Weinberger [NSW08], who investigated when an ε-neighborhood of a random point sample on a compact submanifold of $\mathbb{R}^d$ is homotopy equivalent to that manifold, we give an estimate for the distance in V of the

expectation of the transform (under T) of the barcode for a random n-point cloud for fixed n from the transform of the barcode for the manifold M from which the point cloud is sampled.

Section 7 shows that our hypotheses on the existence of a (Lipschitz continuous) map from the barcode space to a Banach space can be fulfilled using functions introduced in Kalisnik's work [Kal18]. Finally, Section 8 gives a glimpse at open problems in this context.

Notation and Conventions. Let (M, d) be a metric space. For $x \in M$ and $t \in \mathbb{R}_{\geq 0}$, let $B_t(x) = \{y \in M \mid d(x, y) < t\}$ be the open t-ball of x and $\overline{B}_t(x) = \{y \in M \mid d(x, y) \leq t\}$ the closed t-ball around x. For a subset $P \subset M$ we will denote $P_t := \{x \in M \mid d(x, P) \leq t\}$ the t-neighborhood of P which is closed if P is.

We denote by $\mathcal{P}(X)$ the power set of X and by $F(X) \subset \mathcal{P}(X)$ the set of finite nonempty subsets of X. Throughout this paper, we take homology groups with coefficients in a field $\mathbf{k}$. For $n \in \mathbb{N}$ denote by $[n]$ the set $\{1, \ldots, n\}$.

Recall that a *multiset* is a set A together with multiplicities, i.e., a map $A \to \mathbb{N}_0$. We will usually suppress the map in the notation and just speak of a multiset A. Also, we will use set notation such as $A = \{x_1, x_2, x_3, \ldots\}$.

We use Θ for asymptotically comparable in the Big O notation. For example, $f = \Theta(\frac{\log(n)}{n})$ if and only if there exist positive constants C_1 and C_2 such that $C_1 \frac{\log(n)}{n} \leq f(n) \leq C_2 \frac{\log(n)}{n}$ for all large n.

2. FROM PERSISTENT HOMOLOGY TO BARCODES

2.1. **Persistence.** In many applications data lies in a metric space, for example, in a subset of a Euclidean space with an inherited distance function. From this (necessarily finite, and often large) sample one wishes to learn some basic characteristics, such as the number of components or the existence of holes and voids, of the underlying space from which we sampled. Finite metric spaces are discrete spaces, and as such do not per se have interesting topological structure in their own right. The philosophy of topological data analysis is that data does have an inherent topology and in order to uncover it, one assigns a 1-parameter family of topological spaces or a filtration to a finite metric space M [Car09, Car14, ELZ02]. Applying the degree-k homology functor H_k to this filtration yields what is called a persistence module [COGDS16].

Definition 2.1. Let $\mathbf{k}$ be a field. A **persistence module** (over $\mathbf{k}$) is an indexed family of vector spaces

$$V = \left(\{V_t\}_{t \in \mathbb{R}}, \{\phi_s^t\}_{s \leq t \in \mathbb{R}} \right)$$

of $\mathbf{k}$-vector spaces V_t and linear maps $\phi_s^t : V_s \longrightarrow V_t$ for every $s \leq t$ such that $\phi_t^t = \mathrm{id}_{V_t}$ and $\phi_r^t = \phi_s^t \circ \phi_r^s$ for all $r \leq s \leq t$.

One could also replace the field $\mathbf{k}$ by a ring R (e.g. $R = \mathbb{Z}$ is a natural choice) and define an R-persistence module by replacing $\mathbf{k}$-vector spaces by R-modules in the above definition. This might give finer information about the topology of the point clouds, but is also much more complicated from the representation theoretic point of view, see e.g. the discussion in [Car09] before Theorem 2.10 (p. 267). For example, analogs of essential results like Gabriel's theorem (stated here as Theorem 2.8) are not available for $R = \mathbb{Z}$. As our work builds on that in an essential way, we work with fields and vector spaces throughout the paper.

Recall that if we work with field coefficients, homology is a collection of functors $(H_n)_{n \in \mathbb{N}_0}$ from the category of topological spaces to the category of $\mathbf{k}$-vector spaces. We refer the reader to standard textbooks such as Bredon [Bre97] or Hatcher [Hat02]. It is sometimes useful to consider *reduced homology* whose definition we briefly recall: denote by pt the one point space. Then for every topological space X there is a unique continuous map $p_X : X \longrightarrow \mathrm{pt}$. One defines the reduced degree k homology of X as

$$\tilde{H}_k(X) := \ker\left(H_k(X) \longrightarrow H_k(\mathrm{pt})\right)$$

where $H_k(X) \longrightarrow H_k(\mathrm{pt})$ is the map in homology induced by p_X. As $H_k(\mathrm{pt}) = \mathbf{k}$ if $k = 0$ and is trivial otherwise, we have $H_k(X) = \tilde{H}_k(X)$ for every $k \neq 0$. Reduced homology is also a functor on the category of topological spaces.

Definition 2.2. Let X be a topological space and let $f : X \longrightarrow \mathbb{R}$ be a continuous function. This function defines a filtration, called the **sublevelset filtration** of (X, f), by setting $X_t = f^{-1}\left((-\infty, t]\right)$. For $k \in \mathbb{N}_0$ the sublevel set filtration of (X, f) defines a persistence module $(\mathrm{PH}_k(X, f), \phi)$ by $\mathrm{PH}_k(X, f)_t = \tilde{H}_k(X_t)$ and $\phi_s^t : \tilde{H}_k(X_s) \longrightarrow \tilde{H}_k(X_t)$ induced by the inclusion $X_s \hookrightarrow X_t$. For $X \subset \mathbb{R}^d$ we will simply write $\mathrm{PH}_k(X)$ instead of $\mathrm{PH}_k(X, f)$ if $X \subset \mathbb{R}^d$ and $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is the distance–to–X function. We refer to $\mathrm{PH}_k(X)$ (respectively $\mathrm{PH}_k(X, f)$) as the **persistent homology** in degree k of X (respectively of (X, f)).

Definition 2.3. A persistence module V is called **tame** if all V_t have finite dimension and there exist finitely many $t_1 < \ldots < t_m \in \mathbb{R}$ such that ϕ_s^t is an isomorphism whenever $s, t \in (t_i, t_{i+1})$ for some i (where we set $t_0 = -\infty$, $t_{m+1} = \infty$). The function f is called **tame** if the module $\mathrm{PH}_k(X, f)$ is tame for all k.

Example 2.4. It is clear that for an arbitrary smooth manifold $M \subset \mathbb{R}^d$ the k-th persistence module $\mathrm{PH}_k(M)$ is not necessarily tame. Take for example a strictly decreasing sequence $(r_n)_{n \in \mathbb{N}}$ of positive rational numbers such that $\sum_{n \in \mathbb{N}} r_n < \infty$ and put $R_n := \sum_{m=1}^{n} r_n$. If M is the union over all $n \in \mathbb{N}$ of circles K_n with radius R_n centered at the origin, then the persistent homology $\mathrm{PH}_1(M)$ will decompose as a direct sum of interval modules and this decomposition will give rise to an element $b \in \hat{\mathcal{B}}_\infty \setminus \mathcal{B}$.

In certain cases a persistence module can be expressed as a direct sum of "interval modules", which can be thought of as the building blocks of the theory. Here we have four types of intervals and recall the representation from [COGDS16]:

$$
\begin{array}{cc}
\text{interval} & \text{decorated pair} \\
(p, q) & (p^+, q^-) \\
(p, q] & (p^+, q^+) \\
[p, q) & (p^-, q^-) \\
[p, q] & (p^-, q^+)
\end{array}
$$

Definition 2.5. For an interval (p^*, q^*), where * is either $+$ or $-$, denote by $\mathbb{I}(p^*, q^*)$ the persistence module

$$
(\mathbb{I}(p^*, q^*))_t = \begin{cases} \mathbf{k}, & \text{for } t \in (p^*, q^*) \\ 0, & \text{otherwise} \end{cases} \text{ and } \phi_s^t = \begin{cases} \text{id}_\mathbf{k}, & \text{for } s \le t, \text{ and } s, t \in (p^*, q^*) \\ 0, & \text{otherwise} \end{cases}.
$$

Definition 2.6. A persistence module V over $\mathbf{k}$ is called *decomposable* if it can be decomposed as a direct sum

$$
V \cong \bigoplus_{m \in \Lambda} \mathbb{I}(p_m^*, q_m^*),
$$

where Λ is some index set and $* \in \{+, -\}$. If V is decomposable, then the **barcode** associated to V is the **multiset**

$$
\{(p_m^*, q_m^*) \mid m \in \Lambda\}.
$$

We call a decomposable persistence module V *of finite type* if Λ is a finite set.

Remark 2.7. The barcode of V is also called the **persistence** of V.

Not all persistence modules decompose in this way [COGDS16], and there is a considerable body of literature trying to ascertain under which conditions persistence modules are decomposable [Gab72, CCSG+09a, COGDS16, CB15]. We will restrict to the case of most interest to us.

Theorem 2.8 (Gabriel [Gab72])**.** *Let X be a topological space and let $f : X \longrightarrow \mathbb{R}$ be a tame function in the sense of Definition 2.3. Then $\text{PH}_k(X, f)$ is decomposable and of finite type.*

Example 2.9. Examples of (X, f) with a tame function f include:
- X a compact manifold and f a Morse function (where tameness is the result of Morse theory, see [Mil63] for a general reference to this classical field).
- X a compact polyhedron and f a piecewise linear function, see Theorem 2.2 in [COGDS16].
- $X = \mathbb{R}^d$ and f the distance to P function for a finite set $P \subset \mathbb{R}^d$. In this case, tameness is a direct consequence of the nerve theorem. A textbook reference for the general nerve theorem is e.g. [Hat02, Corollary 4G.3].

Let $P \subset \mathbb{R}^d$ be a finite set and $f : \mathbb{R}^d \to \mathbb{R}$ the distance to P function. Then $P_t = f^{-1}((-\infty, t])$ is just the closed t-neighborhood of P and $\text{PH}_k(P)_t = \tilde{H}_k(P_t)$ is decomposable by Theorem 2.8 for $k \in \mathbb{N}_0$. Furthermore, all non-zero intervals appearing in the barcode are closed on the left and open on the right (also known as *closed-open type*), or equivalently of the third type in the above table describing the decorated pair notation.

We can of course define $\text{PH}_k(P)$ even when P is not finite as it may still be decomposable. For example, $\text{PH}_k(P)$ is decomposable and of finite type for a semi-algebraic set P as a consequence of Hardt's theorem, see the discussion in section 3.2 of [HW19].

2.2. **Persistent Homology of Finite Subsets of Metric Spaces.** As mentioned in Example 2.9, the persistent homology $\mathrm{PH}_k(P)$ of a finite point cloud $P \subset \mathbb{R}^d$ can be calculated using the nerve theorem. It tells us in particular, that the homology of P_t is the same as the homology of a simplicial complex, the so-called *Čech complex* with parameter t.

Recall that given a metric space M, a finite set $P \subset M$, and a parameter $t \geq 0$, the Čech complex $\check{C}_t(P)$ is the abstract simplicial complex whose vertex set is P, and where $\{x_0, x_1, \ldots, x_k\}$ spans a k-simplex if and only if $\bigcap_{i=0}^{k} \overline{B}_t(x_i) \neq \emptyset$. The *Čech filtration* of P is the nested family of Čech complexes obtained by varying parameter t from 0 to ∞. This can be used as a definition.

Definition 2.10. Let M be a metric space and let $P \subset M$ be a finite subset. For $k \in \mathbb{N}_0$ we define the persistent homology $\mathrm{PH}_k(P)$ in degree k of P to be the persistence module obtained from taking the homology of the nested family of Čech complexes associated to P. In formulas:

$$\mathrm{PH}_k(P)_t := \tilde{H}_k(\check{C}_t(P)) \quad \text{for } t \in \mathbb{R}_{\geq 0}.$$

From the construction and Theorem 2.8, we immediately deduce

Corollary 2.11. Let M be a metric space and let $P \subset M$ be a finite subset. Then for every $k \in \mathbb{N}_0$, the persistence module $\mathrm{PH}_k(P)$ is tame and decomposable. $\qquad \square$

Note that the two definitions (Definition 2.2 and Definition 2.10) of $\mathrm{PH}_k(P)$ for a finite subset $P \subset \mathbb{R}^d$ coincide.

2.3. **Barcode Space.** In this subsection we describe a useful way of representing barcodes. Given an interval $I \subset \mathbb{R}_{\geq 0}$ of finite length, we encode it as a point $(x, d) \in \mathbb{R}_{\geq 0}^2$ where x is the left endpoint of I and d is its length. The price we pay with this simplified representation is the loss of information about the inclusion of endpoints of the intervals. However, restricted to only one single interval type, this representation map is injective. In the cases we are mainly interested in, this is indeed the case. We are led to the following

Definition 2.12. Let us denote $A := \coprod_{n \in \mathbb{N}_0} \mathbb{R}_{\geq 0}^{2n}$. Let $\sim$ on A be an equivalence relation generated by the relations

$$(x_1, d_1, \ldots, x_n, d_n) \sim (y_1, e_1, \ldots, y_m, e_m) \Longleftrightarrow$$
$$(x_{\sigma(1)}, d_{\sigma(1)}, \ldots, x_{\sigma(n)}, d_{\sigma(n)}) = (y_1, e_1, \ldots, y_n, e_n) \text{ and}$$
$$e_{n+1} = \ldots = e_m = 0 \text{ for some } \sigma \in S_n, \ n \leq m \in \mathbb{N}_0$$

where S_n denotes the symmetric group on n elements. A **barcode representation** is an equivalence class of $(x_1, d_1, \ldots, x_n, d_n)$ with respect to $\sim$. The **space of barcode representations** is the quotient of the disjoint union A by the equivalence relation defined above:

$$\mathcal{B} := A/\sim .$$

For simplicity, we will sometimes also refer to $\mathcal{B}$ as the **barcode space**. We denote by $\mathcal{B}_n \subset \mathcal{B}$ the image of $\coprod_{m \leq n} \mathbb{R}_{\geq 0}^{2m}$ under the canonical map $A \to \mathcal{B}$.

We adopt the notation of Definition 2.5. Let $b = \{(x_1^*, (x_1 + d_1)^*), \ldots, (x_n^*, (x_n + d_n)^*)\}$ with $* \in \{+, -\}$ be a barcode such that all intervals have non-negative left endpoint x_i and

finite length d_i. Then we call $(x_1, d_1, \ldots, x_n, d_n) \in \mathcal{B}$ the **barcode representation of the barcode** b.

The equivalence relation $\sim$ defined above says that two barcode representations are equivalent if they coincide up to permutation of intervals and after deleting zero length intervals (i.e. (x_i, d_i) with $d_i = 0$).

As already pointed out, given a finite subset P of a metric space M, the persistence module $\mathrm{PH}_k(P)$ is decomposable and of finite type by Theorem 2.8, Example 2.9, and Corollary 2.11. Therefore, there is an associated barcode all of whose intervals have finite length. This – and in fact only for $k = 0$ – is where we need to use reduced instead of ordinary homology. We can define the following barcode map from the set of finite nonempty subsets of a metric space M to the barcode space.

Definition 2.13. Let us fix $k \in \mathbb{N}_0$. Given a finite subset P of some metric space M, we denote by $\beta_k(P)$ the barcode representation of the barcode associated to the persistence module $\mathrm{PH}_k(P)$. This defines a map

$$\beta_k : F(M) \longrightarrow \mathcal{B}$$

where $F(M)$ is the set of finite nonempty subsets of M. We will refer to this map as the k-th barcode map.

The barcode space comes equipped with natural metrics. In order to define them, we first specify the distance between any pair of intervals, as well as the distance between any interval and the equivalence class of the zero length interval which for this purpose is represented by the set $\Delta = \{(x, 0) \,|\, -\infty < x < \infty\}$. We put

$$\mathrm{d}_\infty \left((x_1, d_1), (x_2, d_2) \right) = \max \left(|x_1 - x_2|, |(x_1 + d_1) - (x_2 + d_2)| \right).$$

The distance between (the representation of) an interval and the set Δ is

$$\mathrm{d}_\infty((x, d), \Delta) = \frac{d}{2}.$$

Recall that $[n] = \{1, 2, \ldots, n\}$. Let $b_1 = \{I_i\}_{i \in [n]}$ and $b_2 = \{J_j\}_{j \in [m]}$ be barcodes. For any bijection θ from a subset $A \subseteq [n]$ to $B \subseteq [m]$, the *penalty* $P_\infty(\theta)$ of θ is

$$(1) \qquad P_\infty(\theta) = \max \left(\max_{a \in A} \left(\mathrm{d}_\infty(I_a, J_{\theta(a)}) \right), \max_{a \in [n] \setminus A} \mathrm{d}_\infty(I_a, \Delta), \max_{b \in [m] \setminus B} \mathrm{d}_\infty(I_b, \Delta) \right).$$

Definition 2.14. The *bottleneck distance* is defined by

$$\mathrm{d}_\infty(b_1, b_2) = \min_\theta P_\infty(\theta),$$

where with the notation above the minimum is over all possible bijections θ from subsets $A \subset [n]$ to subsets $B \subset [m]$.

There are other metrics also commonly used for barcode spaces. Keeping the notation and changing the penalty (1) for the bottleneck distance to

$$(2) \qquad P_p(\theta) = \sum_{a \in A} \mathrm{d}_\infty(I_a, J_{\theta(a)})^p + \sum_{a \in [n] \setminus A} \mathrm{d}_\infty(I_a, \Delta)^p + \sum_{b \in [m] \setminus B} \mathrm{d}_\infty(I_b, \Delta)^p$$

yields the *pth-Wasserstein distance* ($p \geq 1$) between $b_1, b_2 \in \mathcal{B}$:

$$d_p(b_1, b_2) = \left(\min_\theta P_p(\theta) \right)^{\frac{1}{p}}.$$

Let us consider an example in order to get acquainted with these metrics.

Example 2.15. Let $\mathcal{B}_1 \subset \mathcal{B}$ consist of barcodes containing a single interval (bar). We set $b_1 = (x_1, d_1), b_2 = (x_2, d_2) \in \mathcal{B}_1$ and calculate

$$d_\infty(b_1, b_2) = \min \left(\max \left(|x_1 - x_2|, |x_1 + d_1 - (x_2 + d_2)| \right), \max \left(\frac{d_1}{2}, \frac{d_2}{2} \right) \right).$$

Then we see that if for arbitrary fixed $x_1, x_2 \in \mathbb{R}_{\geq 0}$ the length of both intervals is small, the bottleneck distance between b_1 and b_2 is equally small, even if the intervals are far away from each other. The pth-Wasserstein distance behaves similarly.

The barcode space $\mathcal{B}$ is not a complete metric space, neither with respect to the bottleneck, nor with respect to any of the Wasserstein distances [MMH11]. This is a consequence of the fact that appending bars of smaller and smaller but nonzero length to any given barcode can easily yield a Cauchy sequence of barcodes, with respect to any of the above metrics, and clearly without a limit in $\mathcal{B}$. For the sake of concreteness, let $x > 0$ be fixed, and consider the barcode b_n consisting of all intervals $I_k := (x, \frac{1}{k})$ for all $1 \leq k \leq n$ (so that $b_n \in B_n$). In this case, we have for $n < m$

$$d(b_n, b_m) \leq \max_{n+1 \leq k \leq m} d_\infty(\{I_k\}, \Delta) = \frac{1}{2(n+1)}.$$

A limit could only be a barcode consisting of infinitely many bars, which is impossible.

In order to overcome this problem, we shall consider the completions

(3) $$\left(\hat{\mathcal{B}}_p, d_p \right) \quad \text{and} \quad \left(\hat{\mathcal{B}}_\infty, d_\infty \right)$$

of $\mathcal{B}$ with respect to the Wasserstein and bottleneck distances.

2.4. **Limits of Barcodes.** In subsection 2.1 we recalled the classical construction of barcodes from finite point clouds. Here we present a generalization, which is natural in the context of our probabilistic investigations. Let (M, d) be a metric space and consider the family

$$K(M) := \{Y \subset M \mid Y \text{ compact, non-empty}\}$$

of all compact subsets of M. Together with the Hausdorff distance

(4) $$d_H(A, B) := \max \left(\inf \{t \in \mathbb{R}_{\geq 0} \mid A \subset B_t\}, \inf \{t \in \mathbb{R}_{\geq 0} \mid B \subset A_t\} \right),$$

the set $K(M)$ becomes a metric space. It is well known that $(K(M), d_H)$ is complete whenever (M, d) is complete, and compact whenever (M, d) is compact. Given a bounded subset $A \subset M$, we consider the continuous function, the "distance from A", defined by

$$d_A : M \to \mathbb{R}_{\geq 0}, \quad d_A(x) := \inf \{d(x, y) \mid y \in A\}.$$

We can also describe compact metric spaces in terms of functions. The following result should be rather standard, but it turned out to be easier to give a proof than to find an exact reference.

Lemma 2.16. Let M be a metric space, and denote by $(L_\infty(M), \|\cdot\|_\infty)$ the Banach space of bounded functions $f : M \to \mathbb{R}$, equipped with the supremum norm.

    (1) For $A, B \in K(M)$ the function $d_A - d_B$ is bounded on M.

    (2) The function $n_\infty : K(M) \times K(M) \to \mathbb{R}_{\geq 0}$, $n_\infty(A, B) := \|d_A - d_B\|_\infty$ defines a metric on $K(M)$ such that

$$(K(M), d_H) \to (K(M), n_\infty), \quad A \mapsto A$$

    is an isometry.

    (3) If M is compact, then the function d_A for $A \subset M$ is bounded and $A \mapsto d_A$ defines a continuous injective map

$$(K(M), d_H) \hookrightarrow (L_\infty(M), \|\cdot\|_\infty),$$

    which is an isometry of metric spaces onto its image.

Proof. For (1) let us denote $R := \sup_{a \in A, b \in B} d(a, b)$ which is $< \infty$ by compactness. For a given $x \in M$, we choose $a \in A, b \in B$ such that $d_A(x) = d(a, x)$, $d_B(x) = d(b, x)$ which is again possible by compactness. Without loss of generality $d_A(x) \geq d_B(x)$. The triangle inequality gives

$$|d_A(x) - d_B(x)| = d(a, x) - d(b, x) \leq d(a, b) \leq R$$

and the claim follows.

For (2) let $A, B \in K(M)$. We will first prove that $d_H(A, B) \leq \|d_A - d_B\|_\infty$. Suppose that $|d_A(x) - d_B(x)| \leq t$ for some $t \in \mathbb{R}_{\geq 0}$ and for all $x \in M$. Then in particular for $a \in A$ we deduce $d_B(a) \leq t$ so that $A \subset B_t$. By symmetry the other inclusion follows and therefore $d_H(A, B) \leq t$.

For the inequality in the other direction, let us now assume that for some $t \in \mathbb{R}_{\geq 0}$ we find $A \subset B_t$ and $B \subset A_t$. Let $x \in M$ be given. It suffices to show that $|d_A(x) - d_B(x)| \leq t$. We may assume $d_A(x) - d_B(x) > 0$. By compactness, the infimum is a minimum so that $d_A(x) = d(a, x)$ and $d_B(x) = d(b, x)$ for some $a \in A$, $b \in B$. As $B \subset A_t$ there is $a' \in A$ such that $d(a', b) \leq t$. From $d_A(x) = d(a, x)$ it follows that $d(a', x) \geq d(a, x)$ and we infer

$$|d_A(x) - d_B(x)| = d(a, x) - d(b, x) \leq d(a', x) - d(b, x) \leq d(a', b) \leq t.$$

Thus $\|d_A - d_B\|_\infty \leq t$.

Let us now prove (3). Every compact metric space has a finite radius $R := \sup_{x, y \in M} d(x, y)$. Obviously $\|d_A\|_\infty \leq R$. The rest of the claim follows from (2). $\qquad\square$

Proposition 2.17. Let $k \in \mathbb{N}_0$ be a nonnegative integer and M be a metric space.

    (1) The map $\beta_k : F(M) \to \hat{\mathcal{B}}_\infty$ is Lipschitz continuous with Lipschitz constant equal to 1.

    (2) There is a unique continuous extension $K(M) \to \hat{\mathcal{B}}_\infty$ of $\beta_k : F(M) \to \mathcal{B} \subset \hat{\mathcal{B}}_\infty$. We will denote it by the same symbol β_k. The extended map is also Lipschitz continuous with Lipschitz constant 1.

Proof. The claim in (1) was proved in [CCSG$^+$09b].

For (2), we first show that $F(M) \subset K(M)$ is dense. Given a compact subset $K \subset M$ and $\varepsilon > 0$ we will show that $B_\varepsilon(K) := \{A \in K(M) \mid d_H(A, K) < \varepsilon\} \subset K(M)$ intersects $F(M)$

nontrivially. Since K is compact, there is $P = \{x_1, \ldots, x_n\} \subset K$ such that $K \subset B_\varepsilon(P)$. On the other hand, $P \subset K \subset K_\varepsilon$ so that $d_H(P, K) < \varepsilon$. As Lipschitz functions are in particular uniformly continuous, $\beta_k : F(M) \to \hat{\mathcal{B}}_\infty$ extends to $K(M)$. The fact that the extension is again Lipschitz with the same Lipschitz constant is also standard. $\qquad\square$

Definition 2.18. We call the map $\beta_k : K(M) \to \hat{\mathcal{B}}_\infty$ the barcode map and for a compact set $K \subset M$ we refer to $\beta_k(K)$ as the k-th barcode of K.

This definition and the definition of [CCSG$^+$09b, COGDS16] are equivalent concepts as both produce barcode maps which are continuous functions from the space $K(M)$ to some space of barcodes and they coincide on the dense subset $F(M) \subset K(M)$ of finite subsets of M.

Remark 2.19. Note that the barcode map β_k can easily be extended to a map on totally bounded sets. Since in the proof of Proposition 2.17 we reduced to the case where M is complete, then a totally bounded subset is compact if and only if it is closed. Therefore, for every totally bounded set there is a compact set (its closure) at Hausdorff distance zero (see (4), although totally bounded spaces only form a pseudo metric space for the Hausdorff spaces). One way to define a barcode for a totally bounded set is therefore to define it via Proposition 2.17 as the barcode of its completion which is compact.

One can naturally generalize Proposition 2.17 to the setting of tame functions.

Definition 2.20. Let M be a metric space. We denote by $C(M, \mathbb{R})$ the set of continuous functions with values in $\mathbb{R}$ and endow it with the metric

$$d : C(M, \mathbb{R}) \times C(M, \mathbb{R}) \to [0, \infty], \quad d(f, g) = \|f - g\|_\infty .$$

We will denote by $T(M) \subset C(M, \mathbb{R})$ the subset of tame functions and by $\widehat{T}(M)$ its completion.

There is no harm in allowing the metric to take value ∞. The induced topology is the same as the one induced by the metric

$$(f, g) \mapsto d'(f, g) := \frac{d(f, g)}{1 + d(f, g)} \in [0, 1].$$

The metrics d, d' also feature the same notion of Cauchy sequences. Working with d is however more appropriate for the inequalities we need.

Theorem 2.21. *Let $k \in \mathbb{N}_0$ be a nonnegative integer and let M be a metric space.*
  (1) *The map $\beta_k : T(M) \to \mathcal{B}$ is Lipschitz continuous with Lipschitz constant equal to 1.*
  (2) *There is a unique continuous extension $\widehat{T}(M) \to \hat{\mathcal{B}}_\infty$ of $\beta_k : T(M) \to \mathcal{B} \subset \hat{\mathcal{B}}_\infty$. As before we denote it by the same letter, and note that the extension is also 1-Lipschitz continuous.*

Proof. As in Proposition 2.17, the first part follows from [CCSG$^+$09b]. The second part is implied by the same extension argument for uniformly continuous maps. Note that Lipschitz continuity implies uniform continuity. $\qquad\square$

3. Barcodes of compact sets as almost sure limits

In this section, we will address a very natural convergence problem for stochastic barcodes. It is somewhat surprising that this question has never been addressed before, at least not in full generality.

Let M be a metric space. We consider i.i.d. M-valued random variables $X_1, X_2, \ldots$ whose distribution has support equal to a compact subset $C \subset M$. Recall that the *support* of a measure μ on a σ-algebra containing the Borel σ-algebra $B(M)$ is defined to be the closed subset

$$\operatorname{supp}(\mu) := \{x \in M \mid \forall \varepsilon > 0 : \mu(B_\varepsilon(x)) > 0\}.$$

Let us consider the finite random set $P_n = \{X_1, X_2, \ldots, X_n\}$ and for a fixed k the sequence of barcodes $(\beta_k(P_n))_{n \in \mathbb{N}}$. We would like to describe the limit of this sequence for $n \to \infty$. If P_n were a deterministic sequence approaching C in the Hausdorff distance, then the limit of $\beta_k(P_n)$ would be $\beta_k(C)$ by definition of the latter, see Proposition 2.17. Now, the P_n are random variables, and we prove the following

Theorem 3.1. *Let M be a metric space, let $X_1, X_2, \ldots$ be i.i.d. M-valued random variables, let $k \in \mathbb{N}_0$, and put $P_n = \{X_1, X_2, \ldots, X_n\}$. If the distribution of the X_i has support equal to a compact subset $C \subset M$, then*

$$\beta_k(C) = \lim_{n \to \infty} \beta_k(P_n) \quad \text{almost surely.}$$

This theorem immediately results from the following lemma by continuity of the barcode map, see Proposition 2.17. This statement is a 'folk theorem', and a variation of it with extra assumptions appears in the work of Cuevas and Fraiman [CF97]. We include it here for completeness because we could not find this precise statement in the literature.

Lemma 3.2. *Let M be a metric space, let $X_1, X_2, \ldots$ be i.i.d. M-valued random variables, and put $P_n = \{X_1, X_2, \ldots, X_n\}$. If the distribution of the X_i has support equal to a compact subset $C \subset M$, then*

$$\lim_{n \to \infty} d_H(C, P_n) = 0 \text{ almost surely.}$$

Proof. As $\operatorname{supp}(X_i) = C$ we have $P_n \subset C$ with probability 1. Thus,

$$d_H(C, P_n) = \inf \{\varepsilon > 0 \mid C \subset B_\varepsilon(P_n)\}.$$

By construction, $P_n \subset P_{n+1}$ almost surely for all n so that

$$d_H(C, P_{n+1}) \leq d_H(C, P_n) \text{ almost surely,}$$

and $0 \leq \lim_{n \to \infty} d_H(C, P_n)$ exists almost surely due to monotonicity. It thus suffices to show that $d_H(C, P_n) \to 0$ in probability. Here we use the property that if $Z_n \to Z$ in probability and $Z_n \to Y$ almost surely, then $Z = Y$ almost surely. For $\gamma > 0$ let us denote the event

$$A_\gamma^n = \{d_H(C, P_n) > \gamma\}.$$

We need to show that $\mathbb{P}(A_\gamma^n) \xrightarrow{n \to \infty} 0$ for all $\gamma > 0$. Let us fix some $\gamma > 0$. We have

$$(5) \qquad A_\gamma^n = \{C \not\subset (P_n)_\gamma\} = \{\exists y \in C : y \notin (P_n)_\gamma\} = \{\exists y \in C : B_\gamma(y) \cap P_n = \emptyset\}.$$

Since C is compact, it is totally bounded, i.e., for each $\varepsilon > 0$ we can find $c_1, \ldots, c_{N(\varepsilon)} \in C$ such that $C \subset \bigcup_{i=1}^{N(\varepsilon)} B_\varepsilon(c_i)$. For $\varepsilon = \frac{\gamma}{2}$ it must be that

$$A_\gamma^n \subset \bigcup_{i=1}^{N\left(\frac{\gamma}{2}\right)} \left\{ B_{\frac{\gamma}{2}}(c_i) \cap P_n = \emptyset \right\} \text{ almost surely}$$

from (5). Indeed, if $\xi \in C$ is a random point satisfying $B_\gamma(\xi) \cap P_n = \emptyset$, then for $i \leq N\left(\frac{\gamma}{2}\right)$ such that $\xi \in B_{\frac{\gamma}{2}}(c_i)$ we must have $B_{\frac{\gamma}{2}}(c_i) \cap P_n = \emptyset$ (otherwise we could find a point in P_n at distance smaller than γ from ξ by the triangle inequality). Since the random points X_j are i.i.d., we have for each i

$$\mathbb{P}\left(\left\{ B_{\frac{\gamma}{2}}(c_i) \cap P_n = \emptyset \right\}\right) = \prod_{j=1}^n \left(1 - \mathbb{P}\left(X_j \in B_{\frac{\gamma}{2}}(c_i)\right)\right) = \left(1 - \mathbb{P}(X_1 \in B_{\frac{\gamma}{2}}(c_i)\right)^n.$$

Due to subadditivity of $\mathbb{P}$ we conclude

$$\mathbb{P}\left(A_\gamma^n\right) \leq \mathbb{P}\left(\bigcup_{i=1}^{N\left(\frac{\gamma}{2}\right)} \left\{ B_{\frac{\gamma}{2}}(c_i) \cap P_n = \emptyset \right\}\right) \leq \sum_{i=1}^{N\left(\frac{\gamma}{2}\right)} \left(1 - \mathbb{P}(X_1 \in B_{\frac{\gamma}{2}}(c_i))\right)^n.$$

Each term in the finite sum on the right-hand-side goes to zero as $n \to \infty$, since all the c_i were chosen in the support of the distribution of X_1. Since $\gamma > 0$ is arbitrary, the claim follows as noted above. $\qquad\square$

It is worthwhile emphasizing that there is no condition on the distribution of the random variables such as absolute continuity, the above result is completely general and vaguely reminiscent of the Glivenko-Cantelli theorem.

4. LLN and CLT for barcodes

We deduce a law of large numbers (LLN) and a central limit theorem (CLT) for $\hat{\mathcal{B}}_\infty$-valued random variables. This becomes meaningful via Theorem 7.1 in Section 7. In the context of *persistence landscapes*, Bubenik [Bub15] observed that LLN and CLT can be deduced from general probability theory in Banach spaces. In this section we mirror his approach in the present (barcode representation) context. For a general reference on probability theory in Banach spaces we refer to the monographs by Vakhania, Tarieladze, and Chobanyan [VTC87] respectively by Ledoux and Talagrand [LT91].

Let us recall the definition of the Pettis integral. Let $(V, \|\cdot\|_V)$ be a Banach space. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random vector $\xi : \Omega \to V$, an element $v \in V$ is called the *Pettis integral* of ξ if for each continuous linear functional $\varphi : V \to \mathbb{R}$ we have

$$\varphi(v) = \int_\Omega \varphi\left(\xi(\omega)\right) d\mathbb{P}(\omega).$$

The vector v is also called the *expectation* of ξ and is denoted by $\mathbb{E}[\xi]$ or $\int_\Omega \xi(\omega) d\mathbb{P}(\omega)$. If $\mathbb{E}[\|\xi\|_V] < \infty$, then $\mathbb{E}[\xi]$ exists and satisfies $\|\mathbb{E}[\xi]\|_V \leq \mathbb{E}[\|\xi\|_V]$, see [VTC87, II.3.1 (c)].

Theorem 4.1 (LLN for barcodes). *Let $T : \hat{\mathcal{B}}_\infty \to V$ be a continuous map from the space of barcodes to a separable Banach space V. Let $\{X_i\}_{i \in \mathbb{N}}$ be an i.i.d. sequence of $\hat{\mathcal{B}}_\infty$-valued random barcodes such that $\mathbb{E}[\|T(X_1)\|] < \infty$. Then the sequence of random variables $(S_n)_n$ where*

$$(6) \qquad S_n := \frac{T(X_1) + \ldots + T(X_n)}{n}$$

converges almost surely to $\mathbb{E}[T(X_1)]$.

Proof. As the $\{X_n\}_n$ are i.i.d., so are the random variables $\{T(X_n)\}_n$. Thus, the theorem follows from the general theory of Banach space valued probability, see [LT91, Corollary 7.10].
□

Let us recall the concept of *type* and *cotype* of a Banach space, see e.g. [LT91, II.9.2]. A *Rademacher* (or *Bernoulli*) *sequence* is a sequence of independent random variables with values ± 1 both taken with probability $1/2$. For $1 \le p \le 2$ a Banach space $(V, \|\cdot\|)$ is said to be *of type p* if for every Rademacher sequence $(\varepsilon_i)_{i \in \mathbb{N}}$ there exists a constant C such that for all finite sequences (x_i) the inequality

$$\left\| \sum_i \varepsilon_i x_i \right\|_p \le C \cdot \left(\sum_i \|x_i\|^p \right)^{\frac{1}{p}}$$

holds. Here, $\|\cdot\|_p$ is defined as follows:

$$\|X\|_p = \left(\int_\Omega \|X\|^p d\mathbb{P} \right)^{\frac{1}{p}},$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is the underlying probability space, and the norm $\|\cdot\|$ is the norm of the Banach space V. Similarly, $(V, \|\cdot\|)$ is said to be *of cotype q* for $1 \le q \le \infty$ if instead there is a constant D such that

$$\left(\sum_i \|x_i\|^q \right)^{\frac{1}{q}} \le D \cdot \left\| \sum_i \varepsilon_i x_i \right\|_q.$$

By [HJP76, Theorem 2.1], being of type p is equivalent the existence of a constant $C > 0$ such that

$$\mathbb{E}\left[\left\| \sum_{j=1}^n X_j \right\|^p \right] \le C \cdot \sum_{j=1}^n E\left[\|X_j\|^p \right]$$

for all independent $X_1, \ldots, X_n$ with mean 0 and finite p-th moment.

Note that every Banach space is of type 1 and that a Hilbert space is of type 2 and cotype 2. It can be shown that even the converse is true, i.e. a Banach space of type 2 and cotype 2 is a Hilbert space, see [Kwa72, Theorem 1.1].

Theorem 4.2 (CLT for barcodes). *Let $T : \hat{\mathcal{B}}_\infty \to V$ be a continuous map from the space of barcodes to a separable Banach space V of type 2. Let $\{X_i\}_{i \in \mathbb{N}}$ be an i.i.d. sequence of $\hat{\mathcal{B}}_\infty$-valued random barcodes such that $\mathbb{E}[T(X_1)] = 0$ and $\mathbb{E}[\|T(X_1)\|^2] < \infty$ and let S_n be the V-valued random variable from (6). Then $(\sqrt{n}S_n)_n$ converges weakly to a Gaussian random variable with the covariance structure of $T(X_1)$.*

Proof. Separability of V implies that any probability measure on V is Radon. Thus, the claim follows from [HJP76, Theorem 3.6]. □

We will show next that for important classes of examples the hypotheses of Theorem 4.1 and Theorem 4.2 are fulfilled. Let M be a metric space. For a finite set $P \subset M$ recall that $\beta_k(P)$ is its k-th barcode, see Definition 2.13. For a compact set $K \subset M$, the barcode $\beta_k(K)$ is defined in Proposition 2.17.

Theorem 4.3. *Let M be a metric space and let X be a random variable with values in a compact set $\mathcal{K} \subset K(M)$. Let $T : \hat{\mathcal{B}}_\infty \longrightarrow V$ be a continuous map to a separable Banach space V of type 2. Then $\|T(\beta_k(X))\|$ has finite n-th moments for all $n \geq 0$ where β_k denotes the k-th barcode.*

Proof. The map β_k is continuous with respect to the bottleneck (in the codomain) and the Hausdorff (in the domain) distances. Thus, the image

$$C = \{\|T(\beta_k(K))\| \mid K \in \mathcal{K}\} \subset \mathbb{R}_{\geq 0}$$

is compact. Let $R := \sup C < \infty$. If $(\Omega, \mathcal{F}, \mathbb{P})$ is the underlying probability space on which X is defined, then clearly $\|T(\beta_k(X))\| \leq R$ holds $\mathbb{P}$-almost surely, and in particular

$$\mathbb{E}[\|T(\beta_k(X))\|^n] \leq R^n \int_\Omega d\mathbb{P} = R^n.$$

□

The following is our main application.

Corollary 4.4. Let $M = \mathbb{R}^d$ and let $X_1, \ldots, X_n$ be random variables with values in a compact subset $W \subset \mathbb{R}^d$. Then $P_n = \{X_1, \ldots, X_n\}$ is a random variable with values in the compact subset $\mathcal{K} = K(W) \subset K(\mathbb{R}^d)$. Thus, for every continuous map $T : \hat{\mathcal{B}}_\infty \to V$ as in Theorem 4.3 the LLN and CLT (Theorems 4.1 and 4.2) apply to a sequence of i.i.d. copies of P_n for fixed n.

5. Sampling from the circle: expected barcode lengths

We wish to consider the question of approximation by expectations (of transformations) of random barcodes, where the barcodes are obtained from i.i.d. samples with a fixed (large) sample size.

We first compute expectations in the context of i.i.d. sampling in the simplest example at work - the circle

$$\mathbb{S}^1 = \{x = (x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 = 1\}$$

with uniform samples. Recall that the uniform distribution on an m-dimensional manifold $M \subset \mathbb{R}^d$ of finite volume is defined by

$$\mathbb{P}(A) := \frac{\text{vol}(A)}{\text{vol}(M)} \qquad \forall A \subset M \text{ measurable.}$$

Here, vol is the m-dimensional volume of measurable subsets of M.

In our study, we will more precisely focus on the length of the β_1-barcode for the unit circle[1], and approach the question more generally in Section 6. In order to get these more precise results, we need to be more concrete on the distribution.

Recall that for a finite set $P \subset \mathbb{S}^1$ and $t \geq 0$ we denoted by P_t the closed t-neighborhood of P. Before allowing P to be random, we deduce some general properties of deterministic P_t.

Lemma 5.1. If $t \in [0, 1)$, the projection $\pi : P_t \to \mathbb{S}^1$, $v \mapsto \frac{v}{\|v\|}$ is a homotopy equivalence onto its image $\pi(P_t) \subset \mathbb{S}^1$. If $t \geq 1$, then P_t is star-shaped for $0 \in P_t \subset \mathbb{R}^2$. In particular, P_t is contractible in that case.

Example 5.2. Before we proceed to the proof of the lemma, let us illustrate what happens in two simple examples.

FIGURE 1. Three points whose t-neighborhood has a cycle.

FIGURE 2. Six points whose t-neighborhood has no cycle.

The first example is $P = \{x \in \mathbb{C} \mid x^3 = 1\}$ and $t = \frac{\sqrt{3}}{2}$ as depicted in Figure 1. Even though P_t contains a nontrivial 1-cycle (the triangle between the three points), it does not contain $\mathbb{S}^1$. However, its image $\pi(P_t)$ is the full circle and indeed, P_t and $\mathbb{S}^1$ are homotopy equivalent. The homotopy equivalence is realized by exhibiting a subspace of P_t that maps homeomorphically to the sphere, namely the orange triangle.

The second example is depicted in Figure 2. In this case both P_t and $\pi(P_t)$ have three connected components and each of them is contractible. As in the previous example, the homotopy equivalence is shown by noting that the orange polygonal chain inside P_t maps homeomorphically to $\pi(P_t)$. This chain is obtained by considering each connected component of P_t separately and within such a component connecting every point of P through a straight line segment with its left and right neighbor (if existent) and furthermore connecting the "leftmost" and the "rightmost" point (call them x_ℓ and x_r) via a straight line segment to the unique leftmost point on the boundary of the t-ball around x_ℓ respectively to the unique rightmost point on the boundary of the t-ball around x_r.

As explained in Example 2.9 and Section 2.2, the homotopy type of an "inflated point cloud" $P_t \subset \mathbb{R}^d$ can be calculated using the nerve theorem. The homology of P_t is the same as the

---

[1]The β_1-barcode of $\mathbb{S}^1$ is shown to consist of at most one interval in Corollary 5.3, thus we may speak of its length by which we just mean the length of that interval.

homology of the Čech complex $\check{C}_t(P)$ with parameter $t \geq 0$. The Čech filtration also gives a computational tool to get one's hand on the persistent homology of a finite point cloud. However, it turns out that there is no actual homology computation to be done in this section, because by Corollary 5.3 below the persistent homology of a finite point cloud on the circle will be rather simple.

Proof of Lemma 5.1. The statement for $t \geq 1$ is clear because every point in P_t is contained in a convex ball containing $0 \in P_t$ with center on the circle. We will therefore assume that $t < 1$ from now on. Let us construct a homotopy inverse to π.

As was anticipated in the examples, the homotopy equivalence will be obtained by exhibiting a subspace $G \subset P_t$ which under π maps homeomorphically onto $\pi(P_t)$. The homotopy inverse to π will then be $\iota := (\pi|_G)^{-1} : \pi(P_t) \to G \subset P_t$ to the effect that $\pi \circ \iota = \mathrm{id}_{\pi(P_t)}$ and $\iota \circ \pi$ will be homotopic to the identity on P_t via the homotopy $(x, t) \mapsto tx + (1-t)\iota(\pi(x))$. Observe that with a point x also the line segment between x and $\pi(x)$ is in P_t so that the homotopy is well-defined.

First note that every connected component of P_t is closed and maps onto a closed interval $I \subset \mathbb{S}^1$ where a closed interval on the circle is just the image of a closed interval in $\mathbb{R}$ under the parametrization $t \mapsto (\cos(t), \sin(t))$. Thus, it is sufficient to treat each connected component separately. Moreover, connected components of P_t are again of the form P'_t for a subset $P' \subset P$ because balls $\bar{B}_t(x)$ are connected. In other words, we may assume that P_t is connected.

If $\pi(P_t) = \mathbb{S}^1$, we put $n = \#P$ and let G be the n-gon connecting the centers of the circles in circular order by line segments. This is the triangle in the first example from Example 5.2 above.

Suppose now that $\pi(P_t) \neq \mathbb{S}^1$. Without loss of generality 1 is not in the image of π. We write $P = \{p_1, \ldots, p_n\}$ such that $\arg(p_i) < \arg(p_{i+1})$ for all $i = 1, \ldots, n-1$ where for all $z \neq 1$ we denote $\arg(z) \in (0, 2\pi)$ the unique point such that $e^{i \arg(z)} = z$. Moreover, there are unique points $p_0 \in \bar{B}_t(p_1)$ and $p_{n+1} \in \bar{B}_t(p_n)$ such that

$$\arg(p_0) = \min\{\arg(z) \mid z \in \pi(P_t)\} \quad \text{and} \quad \arg(p_{n+1}) = \max\{\arg(z) \mid z \in \pi(P_t)\}.$$

Then, we define G to be the polygonal chain which is the union of the line segments connecting p_i and p_{i+1} for all $i = 0, \ldots, n$. We leave it to the reader to verify that $\pi|_G$ is a homeomorphism onto $\pi(P_t)$. $\qquad\square$

Corollary 5.3. *For every* $t \in [0, 1)$ *and* $P \subset \mathbb{S}^1$ *finite we have*

$$H_1(P_t, \mathbf{k}) = 0 \text{ or } H_1(P_t, \mathbf{k}) = \mathbf{k}.$$

Proof. By Lemma 5.1 (whose notation we use) it suffices to show that $H_1(\pi(P_t), \mathbf{k}) = 0$ or $H_1(\pi(P_t), \mathbf{k}) = \mathbf{k}$. We have seen in the proof of the preceding lemma that the connected components of $\pi(P_t)$ are either all homeomorphic to closed intervals in $\mathbb{R}$ or $\pi(P_t) = \mathbb{S}^1$, whence the two cases. $\qquad\square$

As usual, we denote by $\beta_k(P)$ the barcode obtained from the k-th persistent homology of a finite set $P \subset \mathbb{R}^d$. By Corollary 5.3 we know that the β_1-barcode of a point cloud $P \subset \mathbb{S}^1$

consists of at most one interval. We denote the **length** of this interval by

$$\ell(\beta_1(P)) \in [0, 1]$$

and also sometimes refer to it as the **length of the barcode**. Before stating the main result of this section, Theorem 5.5, in its most general form, it might be instructive to consider the following special case.

Proposition 5.4. Suppose that $P_3 = \{X_1, X_2, X_3\} \subset \mathbb{S}^1$ is composed of three independent uniformly distributed points on the circle $\mathbb{S}^1$. Then

$$\mathbb{E}[\ell(\beta_1(P_3))] = \frac{9(\sqrt{3} - 2)}{\pi^2} + 1/4.$$

Proof. We parametrize the circle by the interval $I = (-\pi, \pi]$. Using the rotational symmetry we may assume that $X_1 = \pi$ and that $X_2 = \vartheta$, $X_3 = \varphi$ where ϑ, φ are uniformly distributed random angles. It follows from Lemma 5.1 that the time of death of the β_1-barcode is $t_d = 1$. Its time of birth is

$$(7) \qquad t_b = \begin{cases} 1 & \text{if } X_1, X_2, X_3 \text{ lie on a half circle} \\ \max\left(\frac{|X_1 - X_2|}{2}, \frac{|X_1 - X_3|}{2}, \frac{|X_2 - X_3|}{2}\right) \end{cases}$$

where $|\cdot|$ denotes the Euclidean norm. We have

$$|X_1 - X_2| = \sqrt{(1 + \cos(\vartheta))^2 + \sin(\vartheta)^2} = 2\cos\left(\frac{\vartheta}{2}\right),$$

$$|X_1 - X_3| = 2\cos\left(\frac{\varphi}{2}\right),$$

$$|X_2 - X_3| = 2\sin\left(\frac{\vartheta - \varphi}{2}\right).$$

Now we wish to calculate

$$\mathbb{E}[\ell] = \int_{I \times I} (t_d - t_b) \, d\mathbb{P}$$

where $\mathbb{P} = \frac{1}{4\pi^2}\mu$ is the uniform measure on $I \times I$ and μ is the Lebesgue measure. We observe that $\ell = t_d - t_b = 0$ whenever X_1, X_2, X_3 lie on a half circle in $\mathbb{S}^1$ by (7). Let $G \subset I \times I$ be the event that X_1, X_2, X_3 do not lie on a half circle. We have

$$G = G_0 \cup (-G_0) \quad \text{where } G_0 = \{(\vartheta, \varphi) \in I \times I \mid \vartheta \geq 0, \vartheta - \pi < \varphi < 0\}.$$

This event, as well as the function ℓ, are invariant under $(\vartheta, \varphi) \mapsto (-\vartheta, -\varphi)$. Thus

$$\mathbb{E}[\ell] = 2 \int_{G_0} (1 - t_b) \, d\mathbb{P}$$

$$= \frac{1}{2\pi^2} \int_0^\pi \int_{\vartheta - \pi}^0 (1 - t_b(\vartheta, \varphi)) \, d\varphi d\vartheta.$$

Next we divide $G_0 = G_{12} \cup G_{13} \cup G_{23}$ into three subevents corresponding to whether $|X_1 - X_2|$, $|X_1 - X_3|$, or $|X_2 - X_3|$ is maximal. For example, $|X_1 - X_2|$ is maximal on $G_{12} = \{(\vartheta, \varphi) \mid 0 < \vartheta < \frac{\pi}{3}, -\pi + 2\vartheta < \varphi < -\vartheta\}$. Again by symmetry considerations, these events have the

same probabilities, and the integrals (expectations) restricted to them have equal values, so that

$$
\begin{aligned}
\mathbb{E}[\ell] &= \frac{3}{2\pi^2} \int_{G_{12}} (1 - t_b)\, d\mu \\
&= \frac{3}{2\pi^2} \int_0^{\frac{\pi}{3}} \int_{2\vartheta - \pi}^{-\vartheta} \left(1 - \cos\left(\frac{\vartheta}{2}\right)\right) d\varphi d\vartheta \\
&= \frac{9(\sqrt{3} - 2)}{\pi^2} + \frac{1}{4}
\end{aligned}
$$

as claimed. $\qquad\square$

We note $\frac{9(\sqrt{3}-2)}{\pi^2} + \frac{1}{4} \approx 0,00565963600183$. The just made calculation can be generalized as follows.

Theorem 5.5. *Suppose that $P_n = \{X_1, \ldots, X_n\} \subset \mathbb{S}^1$ is a random point set on the circle, i.e., $X_1, \ldots, X_n$ are independent, uniformly distributed $\mathbb{S}^1$–valued random variables. Then*

$$
\mathbb{E}[\ell\left(\beta_1(P_n)\right)] = 1 - \left(\sum_{k \geq 1} (-1)^{k-1} \binom{n}{k} \int_0^{\min\left(\frac{1}{2}, \frac{1}{k}\right)} \pi \cos(\pi t)(1 - kt)^{n-1}\, dt \right).
$$

Proof. This time we parametrize the circle by the interval $[0, 2\pi]$, modulo 2π. Let Θ_i with values in $(0, 1]$ be specified through the identity $X_i = (\cos(2\pi\Theta_i), \sin(2\pi\Theta_i)) = \exp\{2\pi i\, \Theta_i\}$. It is again natural to identify one of the points (for example the last one) with the angle $0 = 2\pi$. Let $\Theta_{(i)}$ be the i-th order statistic of $(\Theta_1, \ldots, \Theta_{n-1})$, i.e. the i-th smallest value among $(\Theta_1, \ldots, \Theta_{n-1})$, and let us set in addition $\Theta_{(0)} := 0$ and $\Theta_{(n)} := 1$. The *normalized (angular) spacings* between the points are defined as follows: $S_i := \Theta_{(i)} - \Theta_{(i-1)}$ for $i = 1, \ldots, n$. We also define

$$
X_{(i)} := \exp\{2\pi i\, \Theta_{(i)}\}, \quad i = 0, 1, \ldots, n,
$$

so that the 2-dimensional random points are ordered via their respective angles (similarly to the proof of Lemma 5.1).

It is easy to check by induction (or alternatively look in [BK07] or [Dev81]) that the joint distribution of the spacings vector $(S_1, \ldots, S_n)$ is uniform on the unit $n - 1$-simplex, as given by,

$$
\mathbb{P}(S_1 > a_1, S_2 > a_2, \ldots, S_n > a_n) = \begin{cases} (1 - \sum_j a_j)^{n-1}, & \sum_j a_j < 1 \\ 0, & \sum_j a_j \geq 1 \end{cases}
$$

As in the case of three random points above, from Lemma 5.1 we know that the β_1-barcode dies at time $t_d = 1$ and is born at time

$$
(8) \qquad t_b = \begin{cases} 1, & \text{if } X_1, X_2, \ldots, X_n \text{ lie on a half circle} \\ \max_{i=1}^n |X_{(i)} - X_{(i-1)}|/2 & \text{otherwise} \end{cases}.
$$

The first condition in (8) is equivalent to the *maximal spacing* $M_n := \max_{i=1}^n S_i$ being $\geq 1/2$. However on $\{M_n < 1/2\}$ we have

$$
\max_{i=1}^n \frac{|X_{(i)} - X_{(i-1)}|}{2} = \sin(\pi M_n).
$$

For the remainder of the calculation let us abbreviate $\ell(\beta_1(P_n))$ by ℓ. Due to the just made observations we conclude that $\mathbb{E}[\ell] = \mathbb{E}\left[(1 - \sin(\pi M_n))\mathbb{1}_{\{M_n < 1/2\}}\right]$. From the above given expression for the joint residual distribution of spacings and the inclusion-exclusion formula, one deduces the following expression for the residual distribution of M_n:

$$\mathbb{P}(M_n > x) = \mathbb{P}(M_n \geq x) = \sum_{k \geq 1:\, kx < 1} (-1)^{k-1} \binom{n}{k} (1 - kx)^{n-1}.$$

This formula is attributed to Whitworth [Whi97]. Let us define $g : [0, 1] \mapsto [0, 1]$ as

$$g(t) := \begin{cases} \sin(\pi x), & x < 1/2 \\ 1, & x \geq 1/2. \end{cases}$$

Now $\mathbb{E}[\ell] = 1 - \mathbb{E}[g(\pi M_n)]$. Since g is non-negative and differentiable (of class C^1), we can apply a well-known change of (order of) integration formula

$$\mathbb{E}[g(\pi M_n)] = \int_{t \geq 0} g'(t)\mathbb{P}(M_n \geq t)\, dt = \int_0^{\frac{1}{2}} \pi \cos(\pi t)\mathbb{P}(M_n \geq t)\, dt,$$

which equals

$$\sum_{k \geq 1} (-1)^{k-1} \binom{n}{k} \int_0^{\min\left(\frac{1}{2}, \frac{1}{k}\right)} \pi \cos(\pi t)(1 - kt)^{n-1}\, dt.$$

$\square$

Remark 5.6 (Related work). Similar computations to ours were made in Bubenik and Kim [BK07] in the setting of Vietoris-Rips filtration (as opposed to Čech filtration), and with respect to the angular (unlike Euclidean taken here) metric on points.

6. Approximation by expected transformations of random barcodes

The calculations made in the previous section demonstrate that expected functionals of barcodes can be quite difficult (and, for more complicated examples, impossible) to obtain explicitly. Theorem 3.1 applied to $\mathbb{S}^1$ on the other hand tells us that as n gets large, in the notation of the previous section, the length $\ell(\beta_1(P_n))$ of the single bar comprising $\beta_1(\mathbb{S}^1)$ must converge to 1. If interested in the asymptotics of $\ell(\beta_1(P_n))$ and $\mathbb{E}[\ell(\beta_1(P_n))]$, we refer the reader to Devroye [Dev81]. In particular, since $\sin(x) \sim x$ for small x, one can apply [Dev81], Lemma 2.5 saying $\frac{nM_n}{\log n} \to 1$ in probability, whereas in the last section M_n denotes the maximal spacing. Therefore, $M_n \to 0$ almost surely, and $1 - \ell(\beta_1(P_n))$ is of order $\frac{\log n}{n}$ with an overwhelming probability as $n \to \infty$. Similar considerations based on [Dev81], Lemma 2.6 lead to $\mathbb{E}[\ell(\beta_1(P_n))] = 1 - \Theta(\frac{\log n}{n})$ as $n \to \infty$.

This is an interesting example that motivates the study of the quality of such an approximation in general.

Similarly to Section 3, one could consider, for a fixed (and relatively large) $n \in \mathbb{N}$, i.i.d. $\mathbb{R}^d$-valued random variables $X_1, \ldots, X_n$, where the joint distribution has support on some compact subset $M \subset \mathbb{R}^d$. The k-th barcode of the resulting random finite set $P_n = \{X_1, \ldots, X_n\}$ yields a random barcode $\beta_k(P_n)$ for each k. Suppose that $T : \mathcal{B} \to V$ is a continuous function from the barcode space to some Banach space. By Theorem 4.1 and Theorem 4.3, the expected value

$\mathbb{E}[T(\beta_k(P_n))]$, can be well approximated by the empirical means (taken over many i.i.d. samples of point clouds of size n).

We restrict our hypotheses somewhat with respect to those of Section 3, in assuming in addition that M is a compact m-dimensional manifold in $\mathbb{R}^d$, and the distribution of X_1 above is uniform on M. We are working on relaxing these hypotheses in a forthcoming project. Let us first introduce some notation. Recall that the *medial axis* of M is defined as the closure of the set of points in $\mathbb{R}^d$ that do not have a unique nearest point on M. We denote by $\tau = \inf_{p \in M} \sigma(p)$ the infimum of the distances $\sigma(p)$ of $p \in M$ from the medial axis of M, i.e., every point in the open τ-neighborhood has a unique nearest point on M. It follows from compactness that τ is positive. The quantity τ is referred to as the *reach* of M.

Under the above assumptions, we can rely on the work by Niyogi et al. [NSW08]. The result [NSW08], Theorem 3.1 is not sufficient for our purposes, therefore we prove a stronger statement in Theorem 6.1 and explain how this also follows from the analysis in [NSW08], see also Remark 6.2. Let

$$c_1(\varepsilon) := \frac{\text{vol}(M)}{\cos\left(\arcsin\left(\frac{\varepsilon}{8\tau}\right)\right)^m \text{vol}\left(B^m_{\varepsilon/4}(0)\right)},$$

(9)

$$c_2(\varepsilon) := \frac{\text{vol}(M)}{\cos\left(\arcsin\left(\frac{\varepsilon}{16\tau}\right)\right)^m \text{vol}\left(B^m_{\varepsilon/8}(0)\right)},$$

where the superscript m indicates that the balls of radii $\varepsilon/4$ and $\varepsilon/8$, respectively, are taken in $\mathbb{R}^m$ (and not necessarily in the ambient space $\mathbb{R}^d$). In particular, the smaller the ε, the larger are $c_{1,2}$, and they are of order $1/\varepsilon^m$. We will use these constants throughout this section.

Let $A \subset \mathbb{R}^d$ be a set and $t \geq 0$. As in Section 2.3 we denote by A_t the closed t-neighborhood of A. For every manifold M with reach τ as above and for every $0 \leq t < \tau$ the inclusion $\iota_t : M \hookrightarrow M_t$ is a homotopy equivalence. This is almost by definition of the reach: a homotopy inverse is given by the projection $\pi : M_t \to M$ to the nearest point on M. Note that for any $p \in M_t$ the line segment connecting p to $\pi(p)$ is entirely contained in M_t (even in the fiber of π over $\pi(p)$) so that a simple convex combination between $\iota \circ \pi$ and the identity gives a homotopy equivalence. For $A \subset M$ and $t \in [0, \tau)$ we denote

$$(10) \qquad \qquad \chi_{A,t} : A_t \hookrightarrow M_t \xrightarrow{\pi} M$$

the composition of the inclusion with the projection.

Theorem 6.1. *Let $M \subset \mathbb{R}^d$ be a smooth compact submanifold of dimension m and let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from M for the uniform distribution. Denoting $P_n := \{X_1, \ldots, X_n\}$ we have that if $\varepsilon \in (0, \sqrt{\frac{3}{5}}\tau)$, then for each $\delta > 0$ and each*

$$(11) \qquad \qquad n > c_1(\varepsilon)\left(\log(c_2(\varepsilon)) + \log\frac{1}{\delta}\right),$$

the map $\chi_{P_n,t} : (P_n)_t \to M$ from (10) is a homotopy equivalence for all $t \in \left[\varepsilon, \sqrt{\frac{3}{5}}\tau\right)$ with probability at least $1 - \delta$.

Remark 6.2. We could have restricted δ to $(0, 1]$, but prefer this statement (trivially true if $\delta > 1$ since any probability is non-negative) in view of applications below. A careful comparison with [NSW08], Theorem 3.1, reveals several differences, but only one is responsible for the fact that the just stated result is non-trivially stronger in the stochastic sense. The claim in Theorem 6.1 is that for any $0 \leq \varepsilon < \sqrt{\frac{3}{5}}\tau$ the map $\chi_{P_n,t} : P_{nt} \to M$ from (10) is a homotopy equivalence *on the whole interval* of parameters $t \in [\varepsilon, \sqrt{\frac{3}{5}}\tau)$ on one and the same event of a sufficiently large probability. The claim in [NSW08] is only that $\chi_{P_n,\varepsilon}$ is a homotopy equivalence at the given parameter ε on an event of a sufficiently large probability. However, an intersection of many (let alone, infinitely many) highly probable events may have a drastically smaller probability. This does however not happen here, for the reasons we give next. We do not contribute any new argument for this, the stronger formulation stated in Theorem 6.1 is merely a consequence of ordering the arguments of [NSW08] accordingly.

Proof of Theorem 6.1. Recall that P_n is called ε-*dense* if the open ε-neighborhood of P_n covers M. For a given $\varepsilon \in (0, \sqrt{\frac{3}{5}}\tau)$, $\delta > 0$, and n satisfying (11), the event A_ε defined by the random point cloud $P_n \subset M$ being $\frac{\varepsilon}{2}$-*dense* in M, has probability at least $1 - \delta$ by Lemma 5.1 in [NSW08]. Therefore, on the same event A_ε the same point cloud is $\frac{t}{2}$-dense for every $t \in [\varepsilon, \sqrt{\frac{3}{5}}\tau)$.

Now we infer from Proposition 3.1 in [NSW08] the deterministic statement that whenever a subset $P \subset M$ is $\frac{t}{2}$-dense, the map $\chi_{P,t} : P_t \to M$ is a homotopy equivalence. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the corresponding probability space. Then we apply the above reasoning and the just mentioned proposition to obtain

$$
\begin{aligned}
A_\varepsilon &= \left\{ \omega \in \Omega \,\middle|\, P_n(\omega) \text{ is } \frac{\varepsilon}{2}\text{-dense} \right\} \\
&= \left\{ \omega \in \Omega \,\middle|\, P_n(\omega) \text{ is } \frac{t}{2}\text{-dense for all } t \in \left[\varepsilon, \sqrt{\frac{3}{5}}\tau\right) \right\} \\
&= \left\{ \omega \in \Omega \,\middle|\, \chi_{P_n(\omega),t} \text{ is a homotopy equivalence for all } t \in \left[\varepsilon, \sqrt{\frac{3}{5}}\tau\right) \right\},
\end{aligned}
$$

Together with Lemma 5.1 from [NSW08] for these sets A_ε, the claim follows. Note that the quantities from that Lemma 5.1 are bounded according to the analysis in section 5 of [NSW08] in such a way that (11) holds. $\square$

We will make essential use of the following easy but important observation.

Lemma 6.3. Let $M \subset \mathbb{R}^d$ be a smooth compact submanifold of dimension m and reach τ, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from M for the uniform distribution, and put $P_n := \{X_1, \ldots, X_n\}$. Then for each $\varepsilon \in (0, \sqrt{\frac{3}{5}}\tau)$, each $\delta > 0$, and each

$$
(12) \qquad\qquad n > c_1(\varepsilon) \left(\log(c_2(\varepsilon)) + \log\frac{1}{\delta} \right),
$$

we have that

$$
d_\infty\left(\beta_k(M), \beta_k(P_n)\right) \leq \frac{\varepsilon}{2}
$$

with probability at least $1 - \delta$.

Proof. By [NSW08, Proposition 3.2] for every $\varepsilon \in (0, \sqrt{\frac{3}{5}}\tau)$ the sample P_n is $\frac{\varepsilon}{2}$-dense in M. Because of $P_n \subset M$ this just means that $d_H(P_n, M) \leq \frac{\varepsilon}{2}$ for the Hausdorff distance d_H. Therefore, the claim follows by (1) of Proposition 2.17. □

To formulate our next result, we introduce an operator on barcodes. For any two a, b such that $0 < a \leq b < \infty$, let $R_{[a,b]}$ denote the restriction map $R_{[a,b]} : \hat{\mathcal{B}}_\infty \to \hat{\mathcal{B}}_\infty$ defined as follows: for each finite barcode representation $b = \{I_i\}_{i=1}^n \in \mathcal{B}$ with $I_i = (x_i, d_i) \in \mathbb{R}^2_{\geq 0}$ we first define

$$I_i^{|(a,b)} := (\max(x_i, a), \min(x_i + d_i, b) - \max(x_i, a))$$

if $\min(x_i + d_i, b) \geq \max(x_i, a)$ and $I_i^{|(a,b)} := (x_i, 0)$ otherwise. Finally, we put $R_{[a,b]}(b) = \{I_i^{|(a,b)}\}_{i=1}^n$. Since the thus defined $R_{[a,b]} : \mathcal{B} \to \mathcal{B}$ is clearly a 1-Lipshitz map, we can extend it as usual to $R_{[a,b]} : \hat{\mathcal{B}}_\infty \to \hat{\mathcal{B}}_\infty$. Note that the coordinates of $I_i^{|(a,b)}$ are just the starting point and the length of the interval $[x_i, x_i + d_i] \cap [a, b]$ if nonempty.

For further use we also record that for a given barcode $\beta \in \hat{\mathcal{B}}_\infty$ the barcode $R_{[a,b]}(\beta)$ depends continuously on a and b.

Setup 6.4. We fix a Lipschitz continuous map $T : \hat{\mathcal{B}}_\infty \to V$ to some Banach space $(V, \|\cdot\|)$ with Lipschitz constant $L(T) > 0$. Due to compactness and continuity, the transformed barcodes $T(\beta_k(M))$ and $T(\beta_k(P_n))$ are uniformly bounded over n by some finite number, which we denote by $C(M; T)$. We also know that, for large n, both $T(\beta_k(P_n))$ and $\mathbb{E}[T(\beta_k(P_n))]$ (due to the dominated convergence theorem) approximate $T(\beta_k(M))$. The question is how large can the difference of $T(\beta_k(M))$ and $\mathbb{E}[T(\beta_k(P_n))]$ be? By interpreting Theorem 6.1 we arrive to the following conclusion.

Theorem 6.5. *Let $M \subset \mathbb{R}^d$ be a smooth compact submanifold of dimension m and reach τ, let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from M for the uniform distribution, and denote $P_n := \{X_1, \ldots, X_n\}$. Let $\varepsilon \in \left[0, \sqrt{\frac{3}{5}}\tau\right)$, and put $I_\varepsilon := \left[\varepsilon, \sqrt{\frac{3}{5}}\tau\right)$. Then for all $k \in \mathbb{N}_0$ the following hold:*

(1) *Let $\bar{I}_\varepsilon = \left[\varepsilon, \sqrt{\frac{3}{5}}\tau\right]$ denote the closure of the interval I_ε and $c_1(\varepsilon) > 0$ and $c_2(\varepsilon) > 0$ be as in (9). Then:*

$$\mathbb{E}\left[\left\|T \circ R_{\bar{I}_\varepsilon}(\beta_k(P_n)) - T \circ R_{\bar{I}_\varepsilon}(\beta_k(M))\right\|_V\right] \leq 3c_2(\varepsilon) \exp\left(\frac{-n}{c_1(\varepsilon)}\right) C(M; T).$$

(2) *For the unrestricted barcodes we have:*

$$\mathbb{E}\left[\|T(\beta_k(P_n)) - T(\beta_k(M))\|_V\right] \leq 3c_2(\varepsilon) \exp\left(\frac{-n}{c_1(\varepsilon)}\right) C(M; T) + \frac{L(T) \cdot \varepsilon}{2}.$$

Here, T, $C(M; T)$ and $L(T)$ are as in Setup 6.4.

Proof. Let us prove (1). By continuity of the projection as a function of the (endpoints of) the interval, it suffices to prove the inequality for every closed interval contained in I_ε. Let $I \subset I_\varepsilon$ be such an interval.

Due to Theorem 6.1, with our choice of n we have that for all $s \in I_\varepsilon$ the homology of M equals that of the point cloud thickened by s, except on an event E_ε of probability at most δ. Condition (11) is equivalent to

$$\delta > c_2(\varepsilon) \exp\left(-\frac{n}{c_1(\varepsilon)}\right).$$

In particular, we could take $\delta(n) = 3c_2(\varepsilon) \exp\left(-\frac{n}{c_1(\varepsilon)}\right)/2$. Therefore, we find $\mathbb{P}(E_\varepsilon) \leq \frac{3}{2}c_2(\varepsilon) \exp\left(\frac{-n}{c_1(\varepsilon)}\right)$, and on the complement of E_ε we know that the homology of the inflated point cloud P_s does not change when $s \in I$ varies, and is equal to that of M and hence to that of M_s.

In particular, $T \circ R_I(\beta_k(P_n)) = T \circ R_I(\beta_k(M))$ for all $k \in \mathbb{N}_0$ on the complement E_ε^c. To arrive at the above stated bound, for each given k, we apply the trivial upper bound $\|T \circ R_I(\beta_k(P_n)) - T \circ R_I(\beta_k(M))\|_V \leq 2C(M;T)$ on E_ε, and take expectation.

For the proof of (2), we just have to note that on E_ε^c we have $d_\infty(\beta_k(P_n), \beta_k(M)) \leq \frac{\varepsilon}{2}$ by Lemma 6.3. The claim follows as T is $L(T)$-Lipschitz and $\mathbb{P}(E_\varepsilon^c) \leq 1$. □

In particular, the theorem implies that the quantity $\|\mathbb{E}\left[T(\beta_k(P_n))\right] - T(\beta_k(M))\|_V$ satisfies the same inequalities as in Theorem 6.5 thanks to the basic properties of the Pettis integral, see Section 4.

7. Embedding the space of barcodes

In Section 4 we have deduced LLN and CLT for random variables induced from random barcodes. We have been working with Lipschitz continuous maps from $\hat{\mathcal{B}}_\infty$ to some Banach space. In this section we will take a look at one such example by building on work of the first named author [Kal18]. Let $\mathcal{B}$ denote the space of barcode representations. Our goal is to describe a Lipschitz continuous embedding $\hat{\mathcal{B}}_\infty \hookrightarrow \ell_1$.

Let us consider the operations $\boxplus, \oplus, \odot$ on $\mathbb{R}$ defined as

$$a \oplus b := \min(a, b), \quad a \boxplus b := \max(a, b), \quad a \odot b := a + b.$$

We call $(\mathbb{R}, \boxplus, \odot)$ the max-plus semiring and $(\mathbb{R}, \oplus, \odot)$ the tropical semiring.

Just as ordinary polynomials are formed by multiplying and adding real variables, max-plus polynomials can be formed by multiplying and adding variables in the max-plus semiring. Let $x_1, x_2, \ldots, x_N$ be variables that represent elements in the max-plus semiring. A *max-plus monomial expression* is any product of these variables, where repetition is permitted. By commutativity, we can sort the product and write monomial expressions with the variables raised to exponents:

$$p(x_1, x_2, \ldots, x_N) = a_1 \odot x_1^{a_1^1} x_2^{a_2^1} \ldots x_N^{a_N^1} \boxplus a_2 \odot x_1^{a_1^2} x_2^{a_2^2} \ldots x_N^{a_N^2} \boxplus \ldots \boxplus a_m \odot x_1^{a_1^m} x_2^{a_2^m} \ldots x_N^{a_N^m}.$$

Here the coefficients $a_1, a_2, \ldots a_m$ are in $\mathbb{R}$, and the exponents a_j^i for $1 \leq j \leq N$ and $1 \leq i \leq m$ are in $\mathbb{N}_0$.

Different max-plus polynomial expressions may happen to define the same functions. Thus, if p and q are max-plus polynomial expressions and

$$p(x_1, x_2, \ldots, x_N) = q(x_1, x_2, \ldots, x_N)$$

for all $(x_1, x_2, \ldots, x_N) \in \mathbb{R}^N$, then p and q are said to be *functionally equivalent*, and we write $p \sim q$. Max-plus polynomials are the semiring of equivalence classes of max-plus polynomial expressions with respect to $\sim$.

The goal of [Kal18] was to identify sufficiently many max-plus polynomials on $\mathcal{B}$ to separate points. This involves finding functions invariant under the action of the symmetric group. To be able to list these functions, consider the set $\mathscr{E}_N$ of $(N \times 2)$-matrices with entries in $\{0, 1\}$. The symmetric group S_N acts on $\mathscr{E}_N$ by permuting the rows. To a matrix $E = (e_{i,j})_{i,j} \in \mathscr{E}_N$ we associate the max-plus monomial $P(E) = x_{1,1}^{e_{1,1}} x_{1,2}^{e_{1,2}} \ldots x_{N,1}^{e_{N,1}} x_{N,2}^{e_{N,2}}$. Suppose that the S_N-orbit of E is $[E] = \{E_1, E_2, \ldots, E_m\}$. Then $P_E = P(E_1) \boxplus P(E_2) \boxplus \ldots \boxplus P(E_m)$ is a 2-symmetric max-plus polynomial and a we can define a function $P_{k,E}$ on $\mathcal{B}_n$ as

$$(13) \qquad P_{k,E}(x_1, d_1, \ldots, x_N, d_N) := P_E(x_1 \oplus d_1^k, d_1, \ldots, x_N \oplus d_N^k, d_N).$$

For $m, n \in \mathbb{N}_0$ with $m + n \geq 1$ we denote by $E_{m,n}$ the matrix

$$\left. \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{array}{l} \left.\rule{0pt}{22pt}\right\} m \text{ times} \\ \left.\rule{0pt}{22pt}\right\} n \text{ times} \end{array} \right.$$

and write $P_{k,m,n}$ for the polynomial $P_{k,E_{m,n}}$. This is a function on $\mathcal{B}$; if b is a barcode with N bars, then

- if $m + n = N$, we use Equation (13);
- if $m + n > N$, then we add $N - m - n$ 0 length bars to b and then use Equation (13);
- if $m + n < N$, then we add $N - m - n$ 0 length rows to the $E_{m,n}$ matrix and then use Equation (13) for this matrix.

It was shown in [MKnGC17, Theorem 6.7] that the set of functions $\{P_{k,m,n}\}_{k,m,n \in \mathbb{N}_0^3}$ separates points from $\mathcal{B}$. Furthermore, all of these functions are Lipschitz [Kal18], i.e. for $C(k, m, n) = 2(2m \max(k, 1) + 2m + 2n)$, the estimate

$$(14) \qquad |P_{k,m,n}(b) - P_{k,m,n}(b')| \leq C(k, m, n) \, d_B(b, b')$$

holds for $b, b' \in \mathcal{B}$.

We fix once and for all an enumeration $(k_1, m_1, n_1), (k_2, m_2, n_2), \ldots$ and consider the corresponding coordinates on the barcode space. We obtain:

Theorem 7.1. *The sequence* $(\frac{1}{C(k_t, m_t, n_t) t^2} P_{k_t, m_t, n_t})_{t \in \mathbb{N}}$ *of functions* $\mathcal{B} \to \mathbb{R}$ *defines an injective map* $\iota : \mathcal{B} \hookrightarrow \ell_1$. *This map is Lipschitz continuous.*

Proof. Let $b \in \mathcal{B}$ be a barcode. We will first prove that $\iota(b)$ is well-defined, i.e., lies in ℓ_1. Let us write $b = (x_1, d_1, \ldots, x_N, d_N)$ for $x_i, d_i \in \mathbb{R}_{\geq 0}^2$, and let $M := \max_{i=1}^N \max(x_i, d_i)$. For any $k, m, n \in \mathbb{N}_0$ we claim that $\frac{1}{C(k_t, m_t, n_t)} P_{k,m,n}(b) \leq 2MN$. Since $P_{k,m,n}(b)$ is the maximum of $P(E)(b)$ where E runs through the orbit of $E_{m,n}$ and the monomials $P(E)$ have degree $2N$,

$$P(E)(x_1 \oplus d_1^k, d_1, \ldots, x_N \oplus d_N^k, d_N) \leq P(E)(x_1, d_1, \ldots, x_N, d_N) \leq P(E)(M, \ldots, M) \leq 2NM.$$

Since $C(k, m, n) \geq 1$, $\frac{1}{C(k_t, m_t, n_t)} P_{k,m,n}(b) \leq 2MN$. Consequently,

$$\sum_{t \in \mathbb{N}} \frac{1}{C(k_t, m_t, n_t) t^2} |P_{k_t, m_t, n_t}(b)| \leq \sum_{t \in \mathbb{N}} \frac{2MN}{t^2} < \infty.$$

As mentioned above the functions $P_{k,m,n}$ separate points on $\mathcal{B}$ so that ι is indeed injective.

This embedding is Lipschitz since it follows from Equation (14) that

$$\sum_{t=1}^{\infty} \left| \left(\frac{1}{C(k_t, m_t, n_t) t^2} P_{k_t, m_t, n_t} \right)(b) - \left(\frac{1}{C(k_t, m_t, n_t) t^2} P_{k_t, m_t, n_t} \right)(b') \right| \leq \sum_{t=1}^{\infty} \frac{1}{t^2} d_B(b, b')$$

$$= \frac{\pi^2}{6} d_B(b, b').$$

$\square$

Example 7.2. It is easy to see that the scaling by $\frac{1}{t^2}$ in the definition of ι is necessary. Consider e.g. $b = (1, 1, \ldots, 1, 1) \in \mathcal{B}_N \subset \mathcal{B}$. Then the sequence $a_\ell = P_{0,\ell,0}(b) = 2\ell$ if $\ell \leq N$ and $a_\ell = P_{0,\ell,0}(b) = 2N$ otherwise. In particular, $\sum_\ell a_\ell$ diverges.

Remark 7.3. Note that for $1 \leq p \leq q \leq \infty$ we have $\ell_p \subset \ell_q$ and the inclusion is Lipschitz continuous. In particular, we have a Lipschitz continuous embedding $\hat{\mathcal{B}}_\infty$ into the separable Hilbert space ℓ_2, thus into a separable Banach space of type 2 as in the assumptions of several results in Section 4.

8. Discussion

The focus in the present paper is on perfect data, sampled without noise. It seems important to allow for noise, and therefore for data issued from distributions with unbounded support. Once we allow for noise (potentially with unbounded support), with the number of points n being large, the maximal error will typically also be large with high probability. To overcome this problem, it seems reasonable to assume that, for each n, the random points are sampled independently from a distribution indexed by n, in such a way that the maximal error stays bounded in n with high probability. We postpone this study to a future work.

REFERENCES

[ABW14]   Robert J. Adler, Omer Bobrowski, and Shmuel Weinberger. Crackle: The homology of noise. *Discrete & Computational Geometry*, 52(4):680–704, Dec 2014. – cited on p. 5

[AC15]   Henry Adams and Gunnar Carlsson. Evasion paths in mobile sensor networks. *The International Journal of Robotics Research*, 34(1):90–104, 2015. – cited on p. 2

[ACC16]   A. Adcock, E. Carlsson, and G. Carlsson. The ring of algebraic functions on persistence bar codes. *Homology, Homotopy and Applications*, 18:381–402, 2016. – cited on p. 2

[ARC14]   Aaron Adcock, Daniel Rubin, and Gunnar Carlsson. Classification of hepatic lesions using the matching metric. *Computer Vision and Image Understanding*, 121:36–42, 2014. – cited on p. 2

[BK07]   Peter Bubenik and Peter T. Kim. A statistical approach to persistent homology. *Homology Homotopy Appl.*, 9(2):337–362, 2007. – cited on p. 21, 22

[BK18]   Omer Bobrowski and Matthew Kahle. Topology of random geometric complexes: a survey. *Journal of Applied and Computational Topology*, 1(3):331–364, Jun 2018. – cited on p. 5

[BM15]   Omer Bobrowski and Sayan Mukherjee. The topology of probability distributions on manifolds. *Probability Theory and Related Fields*, 161(3):651–686, Apr 2015. – cited on p. 4

[Bre97]   Glen E. Bredon. *Topology and geometry*, volume 139 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997. Corrected third printing of the 1993 original. – cited on p. 7

[Bub15]   Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16:77–102, 2015. – cited on p. 2, 3, 4, 5, 15

[Car09]   G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009. – cited on p. 6, 7

[Car14]   Gunnar Carlsson. Topological pattern recognition for point cloud data. *Acta Numerica*, 23:289–368, 2014. – cited on p. 6

[CB15]   William Crawley-Boevey. Decomposition of pointwise finite-dimensional persistence modules. *J. Algebra Appl.*, 14(5):1550066, 8, 2015. – cited on p. 8

[CB19]   Mathieu Carrière and Ulrich Bauer. On the metric distortion of embedding persistence diagrams into separable hilbert spaces. In *35th International Symposium on Computational Geometry, SoCG 2019, June 18-21, 2019, Portland, Oregon, USA.*, pages 21:1–21:15, 2019. – cited on p. 3

[CBK09]   Moo K. Chung, Peter Bubenik, and Peter T. Kim. Persistence diagrams of cortical surface data. In Jerry L. Prince, Dzung L. Pham, and Kyle J. Myers, editors, *Information Processing in Medical Imaging*, pages 386–397, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. – cited on p. 2

[CCSG$^+$09a] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry*, SCG '09, pages 237–246, New York, NY, USA, 2009. ACM. – cited on p. 8

[CCSG$^+$09b] Frédéric Chazal, David Cohen-Steiner, Leonidas J. Guibas, Facundo Mémoli, and Steve Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. In *Proceedings of the Symposium on Geometry Processing*, SGP '09, pages 1393–1403, Aire-la-Ville, Switzerland, Switzerland, 2009. Eurographics Association. – cited on p. 12, 13

[CdSO14]   Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, Dec 2014. – cited on p. 2

[CF97]   Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6):2300–2312, 1997. – cited on p. 14

[CFL$^+$15a]   Frederic Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Subsampling methods for persistent homology. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2143–2151, Lille, France, 07–09 Jul 2015. PMLR. – cited on p. 4

[CFL$^+$15b]   Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. *Journal of Computational Geometry*, 6(2):583–595, 2015. – cited on p. 4

[CGLM15]   Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.*, 16(1):3603–3635, January 2015. – cited on p. 4

[CIVCY13]   Carina Curto, Vladimir Itskov, Alan Veliz-Cuba, and Nora Youngs. The neural ring: An algebraic tool for analyzing the intrinsic structure of neural codes. *Bulletin of Mathematical Biology*, 75(9):1571–1611, 2013. – cited on p. 2

[CM17]   Frédéric Chazal and Bertrand Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *ArXiv e-prints*, page arXiv:1710.04019, October 2017. – cited on p. 4

[COGDS16]   Frédéric Chazal, Steve Y. Oudot, Marc Glisse, and Vin De Silva. *The Structure and Stability of Persistence Modules*. SpringerBriefs in Mathematics. Springer Verlag, 2016. – cited on p. 2, 6, 7, 8, 13

[CSEH07]   David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007. – cited on p. 2

[CZ05]   Gunnar Carlsson and Afra J. Zomorodian. Computing persistent homology. *Discrete and Computational Geometry*, 33:249–274, 2005. – cited on p. 2

[Dev81]   Luc Devroye. Laws of the iterated logarithm for order statistics of uniform spacings. *Ann. Probab.*, 9(5):860–867, 1981. – cited on p. 21, 22

[DP19]   Vincent Divol and Wolfgang Polonik. On the choice of weight functions for linear representations of persistence diagrams. *Journal of Applied and Computational Topology*, 3(3):249–283, 2019. – cited on p. 2

[ELZ02]   H. Edelsbrunner, D. Letscher, and A. J. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002. – cited on p. 2, 6

[Fro92]   Patrizio Frosini. Measuring shapes by size functions. *Intelligent Robots and Computer Vision X: Algorithms and Techniques*, pages 122–133, 1992. – cited on p. 1

[FS10]   Massimo Ferri and Ignazio Stanganelli. Size Functions for the Morphological Analysis of Melanocytic Lesions. *Journal of Biomedical Imaging*, 2010:5:1–5:5, 2010. – cited on p. 2

[Gab72]   P. Gabriel. Unzerlegbare Darstellungen I. *Manuscripta Mathematica*, 6:71–103, 1972. – cited on p. 8

[GdS06]   Robert Ghrist and Vin de Silva. Coordinate-free coverage in sensor networks with controlled boundaries via homology. *International Journal of Robotics Research*, 25:1205–1222, 2006. – cited on p. 2

[GPCI15]   Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015. – cited on p. 2

[Hat02]   Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002. – cited on p. 7, 8

[HJP76]   J. Hoffmann-Jørgensen and G. Pisier. The law of large numbers and the central limit theorem in Banach spaces. *Ann. Probability*, 4(4):587–599, 1976. – cited on p. 16, 17

[HST18]   Yasuaki Hiraoka, Tomoyuki Shirai, and Khanh Duy Trinh. Limit theorems for persistence diagrams. *Annals of Applied Probability*, 28(5):2740–2780, 10 2018. – cited on p. 4

[HW19]   Emil Horobeţ and Madeleine Weinstein. Offset hypersurfaces and persistent homology of algebraic varieties. *Comput. Aided Geom. Design*, 74:101767, 14, 2019. – cited on p. 8

[Kal18]   Sara Kališnik. Tropical coordinates on the space of persistence barcodes. *Foundations of Computational Mathematics*, Jan 2018. – cited on p. 2, 6, 26, 27

[Kwa72]   S. Kwapień. Isomorphic characterizations of inner product spaces by orthogonal series with vector valued coefficients. *Studia Math.*, 44:583–595, 1972. Collection of articles honoring the completion by Antoni Zygmund of 50 years of scientific activity, VI. – cited on p. 16

[LT91]   Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes. – cited on p. 15, 16

[Mil63]   J. Milnor. *Morse theory*. Based on lecture notes by M. Spivak and R. Wells. Annals of Mathematics Studies, No. 51. Princeton University Press, Princeton, N.J., 1963. – cited on p. 8

[MKnGC17]   A. Monod, S. Kališnik, J.A. Pati no Galindo, and L. Crawford. Tropical sufficient statistics for persistent homology. *submitted*, 2017. – cited on p. 27

[MMH11]   Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011. – cited on p. 4, 11

[NSW08]   Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008. – cited on p. 4, 5, 23, 24, 25

[OA17]   Takashi Owada and Robert J. Adler. Limit theorems for point processes under geometric con-
            straints (and topological crackle). *Ann. Probab.*, 45(3):2004–2055, 05 2017. – cited on p. 4

[Owa18]   Takashi Owada. Limit theorems for betti numbers of extreme sample clouds with application to
            persistence barcodes. *Ann. Appl. Probab.*, 28(5):2814–2854, 10 2018. – cited on p. 4

[PC14]   Jose A. Perea and Gunnar Carlsson. A Klein-Bottle-Based Dictionary for Texture Representation.
            International Journal of Computer Vision, 107(1):75–97, 2014. – cited on p. 2

[RHBK15]   Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for
            topological machine learning. In *The IEEE Conference on Computer Vision and Pattern Recog-
            nition (CVPR)*, 2015. – cited on p. 2

[Rob99]   Vanessa Robins. Towards computing homology from finite approximations. *Topology proceedings*,
            24:503–532, 1999. – cited on p. 1

[TMMH14]   Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distribu-
            tions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, Jul 2014. – cited
            on p. 4

[VTC87]   N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability distributions on Banach
            spaces*, volume 14 of *Mathematics and its Applications (Soviet Series)*. D. Reidel Publishing Co.,
            Dordrecht, 1987. Translated from the Russian and with a preface by Wojbor A. Woyczynski. –
            cited on p. 15

[Whi97]   William Allen Whitworth. *Choice and chance*. Cambridge Univ. Press, 1897. – cited on p. 22