

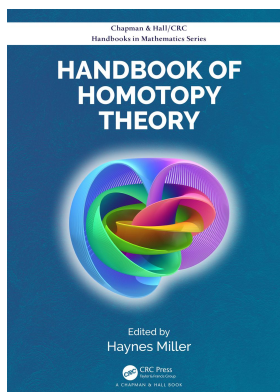
This article was downloaded by: 10.3.97.143

On: 10 Jun 2023

Access details: *subscription number*

Publisher: *CRC Press*

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: 5 Howick Place, London SW1P 1WG, UK



## Handbook of Homotopy Theory

Haynes Miller

### Persistent homology and applied homotopy theory

Publication details

<https://www.routledgehandbooks.com/doi/10.1201/9781351251624-8>

Gunnar Carlsson

**Published online on: 18 Dec 2019**

**How to cite :-** Gunnar Carlsson. 18 Dec 2019, *Persistent homology and applied homotopy theory* from: Handbook of Homotopy Theory CRC Press

Accessed on: 10 Jun 2023

<https://www.routledgehandbooks.com/doi/10.1201/9781351251624-8>

**PLEASE SCROLL DOWN FOR DOCUMENT**

Full terms and conditions of use: <https://www.routledgehandbooks.com/legal-notices/terms>

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

---

# Persistent homology and applied homotopy theory

---

Gunnar Carlsson

---

## 8.1 Introduction

Persistent homology is a technique that has been developed over the last 20 years. Initial ideas developed in the early 1990s [36], but the idea of persistence was introduced by Vanessa Robins in [60], rapidly followed by additional development ([34], [68]), and has been developing rapidly since that time. The original motivation for the method was to extend the ideas of algebraic topology from the category of spaces  $X$  to situations where we only have a sampling of the space  $X$ . Of course, a sample is a discrete space so there is nothing to be obtained unless one retains some additional information. One assumes the presence of a metric or a more relaxed “dissimilarity measure”, and uses this information restricted to the sample in constructing the algebraic invariant. Over time, persistent homology has been used in other situations, for example where one has a topological space with additional information, such as a continuous real valued function, and the sublevel sets of the function determine a filtration on the space. The output of standard persistent homology (we will discuss some generalizations) is represented in two ways, via *persistence barcodes* and *persistence diagrams*. Initially persistent homology was used, as homology is used for topological spaces, to obtain a large scale geometric understanding of complex data sets, encoded as finite metric spaces. Examples of this kind of application are [21], [39], [46], [59], and [25]. Another class of applications uses persistent homology to study data sets where the *points themselves* are metric spaces, such as databases of molecule structures or images. This second set of applications is developing very rapidly and is exemplified in [15], [16], [66], and [43]. Another direction in which persistent homology is being applied is in the study of coverage and evasion problems arising in sensor net technology [27]. Example of research in this direction are [31], [32], [1], and [38].

There are a number of different active research directions in this area.

- **Coordinatization of barcodes:** Barcodes in their native form, i.e. as a set of intervals, do not lend themselves to analysis by machine learning techniques. It is therefore important to represent them in a method more amenable to analysis. This can be achieved by appropriate coordinatizations of the space of barcodes. Several methods for this task are described in Section 8.7.

---

*Mathematics Subject Classification.* 55-04, 55N35, 55U05, 55U15, 55U99.

*Key words and phrases.* persistent homology, applied topology, topological data analysis, random complexes, persistence diagrams, persistence barcodes.

- **Generalized persistence:** Persistent homology has as its output a diagram of complexes, parametrized by the partially ordered set of real numbers, on which algebraic computations are performed so as to produce barcodes. There are other parameter categories that are useful in the study of data sets. We discuss two examples of this notion in Section 8.8, but we would expect there to be many different types of diagrams that will shed light on finite metric spaces.
- **Stability results:** Since noise and error are key elements in the study of data, it is important to develop methods that quantify the dependence of the barcode output on small perturbations of data. This requires the imposition of metrics on the set of barcodes, and proving theorems concerning the distances between barcodes that differ only by small perturbations. The progress that has been made in this direction is described in Section 8.5.
- **Probabilistic analysis and inference:** Because many of the applications of persistent homology occur in the study of data, it becomes important to study not only stability results under small perturbations, but also perturbations that are “probabilistically small”, i.e. which may be large, but where a large perturbation is a rare event. This means that one must study the distributions on barcode space that occur from applying persistent homology to complexes generated by various random models. This is a rapidly developing area within the subject.
- **Coverage and evasion problems:** This work centers around attempts to understand complements in Euclidean space of regions defined by collections of sensors. It has been approached using different methods, and appears to be a place where techniques such as Spanier-Whitehead duality and embedding analysis, applied in suitably generalized situations, should play a role. The general problem of understanding complements of objects embedded in Euclidean space of course also plays a role in robotics.
- **Symplectic geometry:** Although persistent homology is mainly used in situations where one is examining discrete approximations to continuous objects, it can be applied in any geometric situation where there is a metric, or where one is considering filtered objects. Such situations occur in symplectic geometry, and there is recent work applying the technique in studying, for example, Floer homology spectra. Examples of this kind of work are [13], [48], [56], [57], and [65].

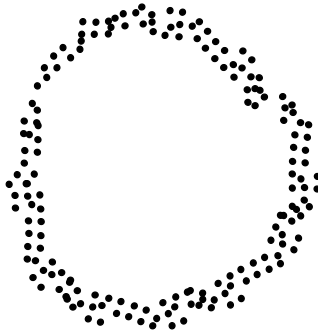
The goal of this paper is to discuss the different research directions and applications at a high level, so that the reader can orient him/herself in the techniques. We remark that there are a number of useful surveys on persistent homology and on topological data analysis more generally. The papers [17], [35], and [55] give different perspectives on this subject.

---

## 8.2 The motivating problem

Suppose that we are given a set of points  $X$  in the plane, and believe that it is reasonable to assume that the points are sampled (perhaps with error) from a geometric object. We

could ask for information about the homology of the underlying space from which  $X$  is sampled. Consider the finite set of points  $X$  in  $\mathbb{R}^2$  displayed in Figure 8.2.1 below.



**FIGURE 8.2.1.** Statistical circle

When we examine  $X$ , we observe that it appears to be sampled from a loop and would like to have algebraic tools that capture the “loopy” structure of the set. Note that we are only given a discrete set of points, so direct application of homological constructions will only produce the homology of a finite set of points. However, we could attempt to construct a space from the set  $X$ , which in a sense fills in the gaps between the points. We will need to use some additional information about the points, and that will in this case be restriction of the Euclidean metric to  $X$ . One distance based construction is the *Vietoris-Rips complex*.

**Definition 8.2.1.** For any finite metric space  $(X, d)$ , and every  $R \geq 0$ , let  $\mathcal{V}(X, R)$  denote the simplicial complex with vertex set equal to  $X$ , and such that  $\{x_0, \dots, x_k\}$  spans a  $k$ -simplex if and only if  $d(x_i, x_j) \leq R$  for all  $0 \leq i < j \leq k$ .

Notice that if  $R$  is smaller than the smallest interpoint distance, then  $\mathcal{V}(X, R)$  will be a discrete complex on the set  $X$ . On the other hand, if  $R$  is greater than the diameter of  $X$ , then  $\mathcal{V}(X, R)$  is a full simplex on  $X$ . For intermediate values, one obtains other complexes. In this case, there is a range of values of  $R$  in which  $\mathcal{V}(X, R)$  has the homotopy type of a circle. One could ask if there is a principled way to choose a threshold  $R$  based on only the distances between the points. After a great deal of experimentation, one finds that this is a very difficult, if not unsolvable problem. A question that one can ask is if there is a more structured object that one can study which incorporates all the thresholds in a single object, which can be analyzed in a number of different ways. Statisticians have confronted this problem in an analogous situation.

**Hierarchical Clustering:** The *clustering problem* in statistics is to determine ways to infer the set of connected components of a metric space  $X$  from finite samples  $Y$ . One of the approaches statisticians developed was to compute  $\pi_0 \mathcal{V}(Y, R)$  for a choice of threshold  $R$ , but they confronted the analogous problem to the one we discussed above, namely the selection of  $R$ . They came to the conclusion that the problem of selecting a threshold in a principled way is not a well posed problem, but managed to construct a structured output, called a *dendrogram*, which allowed them to study the behavior of all thresholds at once.

It is defined as follows. For each threshold  $R$ , we obtain a set of components  $\pi_0\mathcal{V}(Y, R)$ , which yields a partition  $\Pi_R$  of  $Y$ . If we have  $R \leq R'$ , then  $\Pi_R$  is a refinement of  $\Pi_{R'}$ . One definition of a dendrogram structure on a finite set  $Y$  is a parametrized family  $\{\Pi_R\}_{R \geq 0}$  of partitions of  $Y$ , with the property that  $\Pi_R$  refines  $\Pi_{R'}$  whenever  $R \leq R'$ , and so that for any partition  $\Pi$ , the intervals  $\{R \mid \Pi_R = \Pi\}$  are either empty or closed on the left and open on the right. We assume that there is an  $R_\infty$  so that  $\Pi_{R_\infty}$  is the partition with one block, namely  $Y$ . The reason for this terminology is that this information is equivalent to a tree  $\mathfrak{D}$  with a reference map to the non-negative real line. The tree  $\mathfrak{D}$  is defined as follows. The points of  $\mathfrak{D}$  are pairs  $(c, R)$ , where  $c$  is a block in the partition  $\Pi_R$ , and  $0 \leq R \leq R_\infty$ . We clearly have a reference map to  $[0, R_\infty]$  given by  $(c, R) \rightarrow R$ . To define a topology on this set, we construct an auxiliary space  $Z$ , defined as

$$Z = \coprod_c [0, R_\infty]_c$$

where  $c$  is a block in the partition  $\Pi_0$ , and  $[0, R_\infty]_c$  denotes a copy of the interval  $[0, R_\infty]$  labelled by  $c$ . There is a natural map  $\varphi$  from  $Z$  to  $\mathfrak{D}$  given by  $(c, t) \rightarrow (\rho_t(c), t)$ , where  $\rho_t$  denotes the projection  $Y/\Pi_0 \rightarrow Y/\Pi_t$ . It is clear that  $\varphi$  is a surjective map, and therefore that  $\mathfrak{D}$  is the quotient of  $Z$  by an equivalence relation. The topology on  $\mathfrak{D}$  is the quotient topology associated to the topology on  $Z$ . It is easy to check that this topology makes  $\mathfrak{D}$  into a rooted tree. A tree with a reference map can be laid out in the plane, as indicated in Figure 8.2.2, and one can recover directly the clustering at any given value of  $R$ .

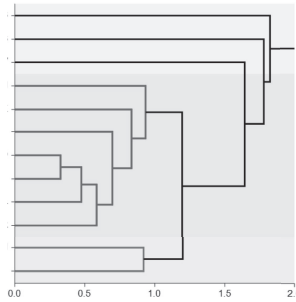


FIGURE 8.2.2. Dendrogram

The dendrogram can be regarded as the “right” version of the invariant  $\pi_0$  in the statistical world of finite metric spaces. The question now becomes if there are similar invariants that can capture the notions of higher homotopy groups or homology groups. In order to define them, we need a preliminary definition.

**Definition 8.2.2.** A submonoid  $\mathbb{A} \subseteq \mathbb{R}_+$  is said to be pure if given any  $r_0, r_1$ , and  $r_2$  in  $\mathbb{R}_+$ , with  $r_0$  and  $r_2$  in  $\mathbb{A}$ , so that  $r_0 + r_1 = r_2$ , then  $r_1 \in \mathbb{A}$ .  $\mathbb{A}$  is a totally ordered set in its own right by restriction of the total order on  $\mathbb{R}_+$ .

For example,  $\mathbb{N}$  (the non-negative integers) and  $\mathbb{Q}_+$  are both pure. We can now make a definition that includes the dendrogram as a special case.

**Definition 8.2.3.** Let  $\underline{C}$  denote any category. A persistence object in  $\underline{C}$  is a functor  $\mathbb{R}_+ \rightarrow \underline{C}$ , where  $\mathbb{R}_+$  denotes the ordered set of non-negative real numbers, regarded as a category, so that there is a unique morphism  $r_0 \rightarrow r_1$  whenever  $r_0 \leq r_1$ . More generally, if  $\mathbb{A} \subseteq \mathbb{R}_+$  is any pure submonoid, an  $\mathbb{A}$ -parametrized persistence object in  $\underline{C}$  we will mean a functor from the ordered set  $\mathbb{A}$  to  $\underline{C}$ . It is clear that the  $\mathbb{A}$ -parametrized persistence objects in  $\underline{C}$  form a category in their own right, where the morphisms are the natural transformations of functors. We will denote this category by  $\mathfrak{P}\mathbb{A}\underline{C}$ , with the special case of  $\mathbb{R}_+$  denoted by  $\mathfrak{P}\underline{C}$ .

It is readily observed that the functor  $\mathbb{R}_+ \rightarrow VR(-, R)$  is a persistence object in the category of simplicial complexes. It follows that that the functor  $\mathbb{R}_+ \rightarrow \underline{\text{Sets}}$  defined by

$$R \rightarrow \pi_0(\mathcal{V}(X, R))$$

where  $X$  is a finite metric space, is a persistence object in  $\underline{\text{Sets}}$ , or a *persistent set*. This is the case because  $\pi_0$  is a set valued functor. Similarly, if  $A$  is any abelian group, and  $i \geq 0$  is an integer, then  $R \rightarrow H_i(VR(-, R) : A)$  is a persistence object in the category of abelian groups, and it will be referred to as the *persistent homology with coefficients in  $A$* . More generally, we may assign persistent abelian groups to any persistent chain complex  $C_*$ , and we will refer to them as the persistent homology of  $C_*$ . The critical question becomes whether or not the isomorphism classes of these persistence objects are in any sense understandable, and useful for distinguishing or understanding the underlying metric spaces. We will see that in the case of coefficients in a field  $K$ , persistent homology is computable and interpretable.

---

### 8.3 The structure theorem for persistence vector spaces over a field

For the entirety of this section,  $K$  will denote a field, which will be fixed throughout. We are interested in the isomorphism classification of  $\mathbb{A}$ -parametrized persistence  $K$ -vector spaces, for pure submonoids  $\mathbb{A}$  of  $\mathbb{R}_+$ . This is very complicated in general, but is manageable for objects of  $\mathfrak{P}\text{Vect}(K)$  satisfying a finiteness condition, which is always satisfied for the Vietoris-Rips complexes associated with finite metric spaces. Let  $\mathbb{M}$  be any commutative monoid. By an  $\mathbb{M}$ -graded  $K$ -vector space, we will mean a  $K$ -vector space  $V$  equipped with a decomposition

$$V \cong \bigoplus_{\mu \in \mathbb{M}} V_\mu.$$

Given two  $\mathbb{M}$ -graded  $K$ -vector spaces  $V_*$  and  $W_*$ , by their tensor product we will mean the tensor product  $V \otimes_K W$  equipped with the  $\mathbb{M}$ -grading given by

$$(V \otimes_K W)_\mu = \bigoplus_{\mu_1 + \mu_2 = \mu} V_{\mu_1} \otimes W_{\mu_2}$$

and we write  $V_* \otimes_K W_*$  for this construction. An  $\mathbb{M}$ -graded  $K$ -algebra is then an  $\mathbb{M}$ -vector space  $R_*$  together with a homomorphism  $R_* \otimes_K R_* \rightarrow R_*$ , satisfying the associativity and

distributivity conditions. An important example is the monoid  $K$ -algebra  $K[\mathbb{M}]_*$ , for which the grading is given by

$$K[\mathbb{M}]_\mu = K \cdot \mu.$$

It will be convenient to write  $t^\mu$  for elements  $\mu \in K[\mathbb{M}]_*$ . We define the notion of a  $\mathbb{M}$ -graded  $R_*$ -module  $M_*$  in a similar way.

We specialize to the situation  $\mathbb{M} = \mathbb{A}$ , where  $\mathbb{A}$  is a pure submonoid  $\mathbb{A} \subseteq \mathbb{R}_+$ . We will demonstrate the classification of  $\mathbb{A}$ -parametrized persistence modules by using an equivalence of categories to the category of  $\mathbb{A}$ -graded  $K[\mathbb{A}]_*$ -modules.

**Proposition 8.3.1.** *Let  $K[\mathbb{A}]_*$  denote the monoid algebra of  $\mathbb{A}$  over  $K$ , for a pure submonoid  $\mathbb{A} \subseteq \mathbb{R}_+$ , regarded as a graded  $K$ -algebra. Let  $\underline{G}(\mathbb{A}, K)$  denote the category of  $\mathbb{A}$ -graded  $K[\mathbb{A}]_*$ -modules. Then there is an equivalence of categories*

$$\mathfrak{P}\mathbb{A}\text{Vect}(K) \cong \underline{G}(\mathbb{A}, K).$$

*Proof:* Given a functor  $\theta : \mathbb{A} \rightarrow \underline{\text{Vect}}(K)$ , we will denote by  $M(\theta)$  the graded  $K$ -vector space

$$M(\theta) = \bigoplus_{\alpha \in \mathbb{A}} \theta(\alpha)$$

where the elements of grading  $\alpha$  are precisely the elements of the summand  $\theta(\alpha)$ . We now extend the vector space structure to a graded  $K[\mathbb{A}]_*$ -module structure by defining the action  $t^\alpha \cdot : \theta(\alpha') \rightarrow \theta(\alpha + \alpha')$  to be equal to the linear transformation

$$\theta(\alpha' \rightarrow \alpha + \alpha') : \theta(\alpha') \rightarrow \theta(\alpha + \alpha')$$

where  $\alpha' \rightarrow \alpha + \alpha'$  denotes the unique morphism in  $\mathbb{A}$  from  $\alpha$  to  $\alpha + \alpha'$ . It is clear that this defines an  $\mathbb{A}$ -graded  $K[\mathbb{A}]_*$ -module structure on  $M(\theta)$ , and that  $M$  is a functor. We produce an inverse functor  $\eta : \underline{G}(\mathbb{A}, K) \rightarrow \mathfrak{P}\mathbb{A}\text{Vect}(K)$  on objects by setting

$$\eta(M_*)(\alpha) = M_\alpha$$

and on morphisms by

$$\eta(M_*)(\alpha' \rightarrow \alpha + \alpha') = t^\alpha \cdot : M_{\alpha'} \rightarrow M_{\alpha + \alpha'}.$$

The functors  $\theta$  and  $\eta$  are clearly inverse to each other.  $\square$

Let  $\mathbb{A}$  be any pure submonoid of  $\mathbb{R}_+$ . Then for any  $\alpha \in \mathbb{A}$ , we define  $F(\alpha)$  to be the free  $\mathbb{A}$ -graded  $K[\mathbb{A}]_*$ -module on a single generator in grading  $\alpha$ . For any pair  $\alpha, \alpha' \in \mathbb{A}$ , where  $\alpha' > \alpha$ , we define  $F(\alpha, \alpha')$  to be the quotient

$$F(\alpha) / (t^{\alpha' - \alpha} \cdot F(\alpha)).$$

The following result describes the isomorphism classification of finitely presented graded  $K[\mathbb{A}]_*$ -modules.

**Proposition 8.3.2.** Any finitely presented object of  $\underline{G}(\mathbb{A}, K)$  is isomorphic to a module of the form

$$\bigoplus_{s=1}^m F(\alpha_s) \oplus \bigoplus_{t=1}^n F(\alpha_t, \alpha'_t).$$

Moreover, the decomposition is unique up to reordering of summands. The kernel of any homomorphism between two finitely generated free  $\mathbb{A}$ -graded modules is itself a finitely generated free  $\mathbb{A}$ -graded module.

**Remark 8.3.3.** This result is formally very similar to the structure theorem for finitely generated modules over a principal ideal domain (PID). Indeed, for the case of  $\mathbb{A} = \mathbb{N}$ , where  $K[\mathbb{A}]$  is Noetherian, the result is exactly a structure theorem for finitely generated graded modules over the graded ring  $K[t]$ . For other choices of  $\mathbb{A}$ ,  $K[\mathbb{A}]_*$  is not necessarily Noetherian. However, it does turn out to be *coherent*, i.e. having the property that the kernel of any homomorphism between finitely generated free modules is finitely generated.

*Proof:* A proof is given in [18]. A proof in a different context was given earlier in [7].  $\square$

**Remark 8.3.4.** Notice that the proof also gives an algorithm for producing a matrix in the diagonal form given above.

The above theorem is summarized using the following definition.

**Definition 8.3.5.** By an  $\mathbb{A}$ -valued barcode, we will mean a finite set of elements

$$(\alpha, \alpha') \in \mathbb{A} \times (\mathbb{A} \cup \{+\infty\})$$

satisfying the condition  $\alpha < \alpha'$ . An  $\mathbb{A}$ -valued barcode is said to be finite if the right-hand endpoint  $+\infty$  does not occur. If  $\mathbb{A} = \mathbb{R}_+$ , we will simply refer to it as a barcode, without labeling by the monoid. We have shown that isomorphism classes of finitely presented  $K[\mathbb{A}]_*$ -modules are in bijective correspondence with  $\mathbb{A}$ -valued barcodes.

**Remark 8.3.6.** Barcodes are often represented visually as collections of intervals. The image in Figure 8.3.1 shows barcodes in dimensions zero and one.

Notice that the zero-dimensional barcode has one infinite interval, while the one-dimensional barcode is finite. There is an equivalent visual representation called the *persistence diagram* in which each interval is encoded as a point  $(x, y)$  in the plane, where  $x$  and  $y$  are the left and right hand endpoints respectively. An example is pictured in Figure 8.3.2.

An advantage of this representation made apparent in this image is that one can represent different homology groups in the same diagram. The dark gray dots are in this case the zero-dimensional persistence diagram and the light gray ones are the one-dimensional ones. When there are infinite intervals, one often selects an upper threshold  $\tau$  for the persistence parameter, and represents the infinite interval by one with left hand endpoint  $\tau$ .

We also have the following.

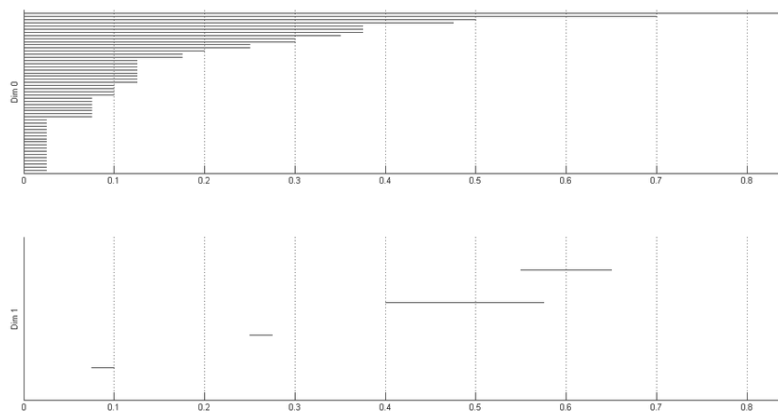
**Proposition 8.3.7.** The kernel of a homomorphism between finitely generated free  $K[\mathbb{A}]_*$ -modules is finitely generated free.



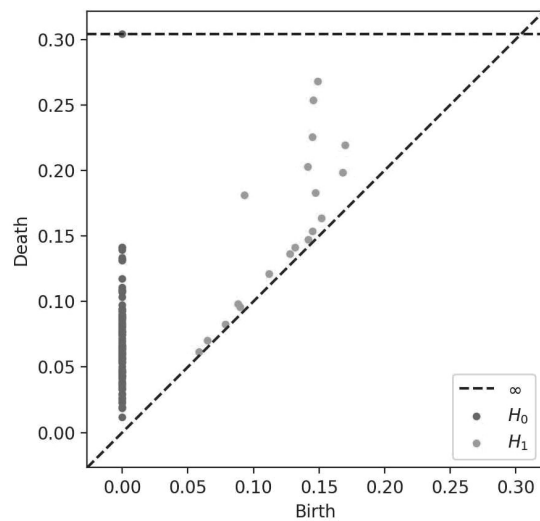
*Proof:* This follows immediately from the matrix analysis in the proof of Proposition 8.3.2.  $\square$

**Corollary 8.3.8.** *Given any chain complex of finitely generated free graded  $K[\mathbb{A}]_*$ -modules, the homology modules are finitely presented  $K[\mathbb{A}]_*$ -modules.*

We also interpret this result in terms of persistence vector spaces. The category  $\mathfrak{P}\mathbb{A}\text{Vect}(K)$  is clearly an abelian category. For any  $\alpha_0 \in \mathbb{A}$ , we define  $V(\alpha)$  to be the persistence vector space  $\{V_\alpha\}_{\alpha \in \mathbb{A}}$ , where  $V_\alpha = \{0\}$  for  $\alpha < \alpha_0$ ,  $V_\alpha = K$  for  $\alpha \geq \alpha_0$ , and where for any  $\alpha_0 \leq \alpha \leq \alpha'$  the linear transformation  $V_\alpha \rightarrow V_{\alpha'}$  is the identify on  $K$ . For any  $\alpha < \alpha'$ , we also define the persistence vector space  $V(\alpha, \alpha')$  to be the quotient of  $V(\alpha)$  by the image of the natural inclusion  $V(\alpha') \hookrightarrow V(\alpha)$ . Corollary 8.3.8 now has the following consequence.



**FIGURE 8.3.1.** Persistence barcodes



**FIGURE 8.3.2.** Persistence diagram

**Corollary 8.3.9.** *Let  $C_*$  be a chain complex of  $\mathbb{A}$ -persistence vector spaces, so that for every  $s$ ,  $C_s$  is a finite direct sum of persistence vector spaces, each of which is of the form  $V(\alpha)$  for some  $\alpha \in \mathbb{A}$ . Then for each  $s$ ,  $H_s(C_*)$  is isomorphic to a direct sum of finitely many persistence vector spaces, each of which is of the form  $V(\alpha)$  or  $V(\alpha, \alpha')$  for  $\alpha, \alpha' \in \mathbb{A}$ , and  $\alpha < \alpha'$ . In particular, the homology in each dimension  $s$  is uniquely described by an  $\mathbb{A}$ -valued barcode.*

## 8.4 Complex constructors

### 8.4.1 Introduction

All data that we consider consists of finite sets of points  $X$ . The space  $X$  itself is uninteresting topologically, since it is a discrete set of points. This means that we have to build a space using auxiliary information attached to the set of points. The auxiliary information we choose is a metric on the set  $X$ , so  $X$  is a finite metric space. In the context of data sets, metrics are often referred to as *dissimilarity measures*, since small distances between data points often reflect notions of similarity between data points. Often the metric chosen is the restriction of a well known and analyzed metric on an ambient space containing  $X$ , such as  $n$ -dimensional Euclidean space. Other choices that are often appropriate are Hamming distance, correlation distances, and normalized variants (mean centering of coordinate functions, normalizing variance to 1) of Euclidean distance. One method for constructing spaces based on metrics is the Vietoris-Rips complex that we have seen above. It is actually a persistence object in the category of simplicial complexes. All the constructions we will look at except the Mapper construction are naturally persistence objects in the category of simplicial complexes. The Mapper construction can also be equipped with many such structures, but they are not canonical. Because of the presence of a persistence structure, the Čech, Vietoris-Rips, alpha, and witness complexes all have persistence barcodes associated to them in all non-negative dimensions. It is obvious from the constructions that the zero-dimensional barcodes have a single infinite interval and that all higher dimensional barcodes are finite.

### 8.4.2 Čech construction

Let  $(X, d)$  denote a finite metric space. Given a threshold parameter value  $R$ , let  $\mathcal{U}_R$  denote the covering of  $X$  by all balls  $B_R(x) = \{x' \mid d(x, x') \leq R\}$ . The *Čech complex at scale  $R$*  is the nerve of the covering  $\mathcal{U}_R$ , and we denote it by  $\check{C}(X; R)$ . It is clear that for  $R \leq R'$ , there is an inclusion  $\check{C}(X; R) \hookrightarrow \check{C}(X; R')$ , and that therefore  $\{\check{C}(X; R)\}_R$  is a persistence object in the category of simplicial complexes. From the theoretical point of view, it has the advantage that given a covering of a topological space (with suitable point set hypotheses) by open sets, the nerve lemma (see [47], Thm. 15.21) gives a criterion that guarantees that the nerve of the covering is homotopy equivalent to the original space.

### 8.4.3 Vietoris-Rips complex

The Čech construction from the previous section is computationally expensive, because it involves computing simplices individually at every level. We can create another construction that has a strong relationship with the Čech construction. Recall that a simplicial complex  $Z$  is said to be a *flag complex* if for any collection  $\sigma$  of vertices  $\{z_0, \dots, z_k\}$  for which each pair  $(z_i, z_j)$  is an edge,  $\sigma$  is a  $k$ -simplex of  $X$ . From the computational point of view, flag complexes are attractive because one needs only enumerate all the edges in the complex, rather than all the higher order simplices. The Vietoris-Rips complex which was defined in Definition 8.2.1 is by definition a flag complex for every parameter value  $R$ . There is a relationship between the persistent Čech and Vietoris-Rips complexes.

**Proposition 8.4.1.** *There are inclusions*

$$\check{C}(X, R) \hookrightarrow \mathcal{V}(X, 2R)$$

and

$$\mathcal{V}(X, R) \hookrightarrow \check{C}(X, R).$$

For specific situations, this bound can be improved. For example, it is shown in [31] that if  $X \subseteq \mathbb{R}^d$  is equipped with the restricted metric, then

$$\mathcal{V}(X, R') \subseteq \check{C}(X, R/2) \subseteq \mathcal{V}(X, \varepsilon) \text{ if } \frac{R}{R'} \geq \sqrt{\frac{2d}{d+1}}.$$

This result allows us to compare homology computed using the Čech and Vietoris-Rips methods.

### 8.4.4 Alpha complex

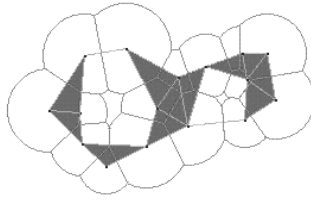
There is another kind of complex whose dimension is low and which generally has a moderate number of simplices. It is called the *alpha complex*, or the *alpha shapes complex*, and was introduced in [33], with a thorough description in [5]. It applies to data  $X$  that is embedded Euclidean space  $\mathbb{R}^n$ , and so that the metric on  $X$  is the restriction of the Euclidean metric to  $X$ . Typically the number  $n$  is relatively small, say  $\leq 5$ , because the construction becomes quite expensive in higher dimensions. Also, the complex is generically of dimension  $\leq n$ . The notion of *generic* is the following. Given any set of points  $X \subseteq \mathbb{R}^n$ , it is possible for the alpha complex to have dimension higher than  $n$ , but it is possible to perturb all the points by an arbitrarily small amount and obtain a complex that is  $n$ -dimensional.

For any point  $x \in X$ , we define the *Voronoi cell* of  $x$ , denoted by  $V(x)$ , by

$$V(x) = \{y \in Y \mid d(x, y) \leq d(x', y) \text{ for all } x' \in X\}.$$

The collection of all Voronoi cells for a finite subset of Euclidean space is called its *Voronoi diagram*.

For each  $x \in X$ , we also denote by  $B_\varepsilon(x)$  the set  $\{y \in Y \mid d(x, y) \leq \varepsilon\}$ . By the  *$\alpha$ -cell* of  $x \in V(x)$  with scale parameter  $\varepsilon$ , we will mean the set  $A_\varepsilon(x) = B_\varepsilon(x) \cap V(x)$ . The  *$\alpha$ -complex with scale parameter  $\varepsilon$*  of a subset  $x \in X$ , denoted by  $\alpha_\varepsilon(X)$  will be the abstract simplicial complex with vertex set  $X$ , and where the set  $\{x_0, \dots, x_k\}$  spans a  $k$ -simplex if



**FIGURE 8.4.1.** Alpha complex in  $\mathbb{R}^2$

and only if

$$\bigcap_{i=0}^k A_\varepsilon(x_i) \neq \emptyset.$$

It is of course the nerve of the covering of  $\mathbb{R}^n$  by the sets  $A_\varepsilon(x_i)$ . See Figure 8.4.1 for an example.

### 8.4.5 Witness complex

The *witness complex* was introduced in [30]. It can be thought of as an analogue of the alpha complex for non-Euclidean data. The construction of the Voronoi cells is not dependent on the fact that the embedding space is Euclidean. It can be constructed for any embedding of a data set in a larger metric space. Given a data set  $X$ , we therefore select a subset of *landmarks*  $\mathcal{L} \subseteq X$ . We can now build the analogues of the Voronoi cells for each of the landmark points within  $X$ , and construct the nerve of the covering. In this case, both the ambient space and the landmark set are usually taken to be finite. We will also need to introduce persistence into this picture. The construction is as follows.

**Definition 8.4.2.** *Given a metric space  $(X, d)$ , a finite subset  $\mathcal{L} \subseteq X$ , called the landmark set, and a persistence parameter  $R$ , and for every  $x \in X$  we denote by  $m_x$  the distance from  $x$  to the set  $\mathcal{L}$ . We define a simplicial complex  $W(X, \mathcal{L}, R)$  as follows. The vertex set of  $W(X, \mathcal{L}, R)$  is the set  $\mathcal{L}$ , and  $\{l_0, l_1, \dots, l_k\}$  spans a  $k$ -simplex if and only if there is a point  $x \in X$  (the witness) so that  $d(x, l_i) \leq m_x + R$  for all  $i$ . The family of complexes  $\{W(X, \mathcal{L}, R)\}_R$  form a persistence simplicial complex.*

There are several variants on this construction. For example, there is the “lazy” version in which the 1-simplices are identical to the 1-simplices of the witness construction, but in which we declare that any higher dimensional simplex is an element of the complex if and only if all its one dimensional faces are. Each of the lazy complexes is a flag complex. The lazy witness complex bears the same relationship to the witness complex as the Vietoris-Rips complex bears to the Čech complex. There is also the *weak witness complex*,  $\{W^{weak}(X, \mathcal{L}, R)\}_R$ , which is defined as follows. For each point  $x \in X$ , we let  $\delta_x$  denote the distance to the *second* closest element of  $\mathcal{L}$  to  $x$ . We then declare that a pair  $(\lambda_1, \lambda_2)$  spans an edge of  $W^{weak}(X, \mathcal{L}, R)$  if and only if there is an  $x \in X$  so that

$$\max(d(\lambda_1, x), d(\lambda_2, x)) \leq \delta_x + R.$$

A higher dimensional simplex  $\{\lambda_0, \dots, \lambda_n\}$  is contained in  $W^{weak}(X, \mathcal{L}, R)$  if and only if all of its edges are contained in it. This is a very useful construction because the persistence “starts faster” than the standard complex. It is often the case that one obtains the ultimate result even at  $R = 0$ , or for very small values of  $R$ .

### 8.4.6 Mapper

Another construction is based on Morse theoretic ideas. We motivate it by considering a space level construction. Suppose that we have a continuous map  $r : X \rightarrow B$  of spaces, and suppose further that  $B$  is equipped with an open covering  $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ . We obtain the open covering  $r^{-1}\mathcal{U} = \{r^{-1}U_\alpha\}_{\alpha \in A}$  which can be refined into a new covering  $r^{-1}\mathcal{U}^*$  by decomposing each set  $r^{-1}U_\alpha$  into its connected components. We note that the dimension of the nerve of  $r^{-1}\mathcal{U}^*$  is less than or equal to the dimension of the nerve of  $\mathcal{U}$ , so this construction, like the alpha complex, produces complexes of bounded dimension. The analogue of this construction for finite metric spaces is obtained by assuming that the finite metric space is equipped with a map  $\rho$  to a reference space  $B$ , and replacing the connected component construction by the output of a clustering algorithm. A simple algorithm to use is single linkage hierarchical clustering, where one makes a choice of threshold based on a simple heuristic, such as the one found in [63]. Once this is done, one obtains a covering of the finite set  $X$  by the collection of all clusters constructed in each of the sets  $\rho^{-1}(U_\alpha)$ , and constructs the nerve complex. This construction is referred to as *Mapper*. Usually,  $B$  is chosen to be  $\mathbb{R}^n$ , for a small positive integer  $n$ , and therefore the reference map is determined by an  $n$ -tuple of real valued functions on the metric space  $X$ . The reference maps can be chosen in many ways, giving different views of the data. Some standard choices are density estimators, measures of centrality, coordinates in linear algebraic algorithms such as principal component analysis and multidimensional scaling ([42]), or individual coordinates when a metric space is obtained as a subspace of  $\mathbb{R}^N$  for some  $N$ . The method has been used extensively in work on life sciences data sets, see for example [51], [53], [54], and [61].

---

## 8.5 Metrics on barcode space, and stability theorems

Since persistent homology is used to analyze data sets, and data sets are often noisy in the sense that one does not want to assign meaning to small changes, it is important to analyze the stability of persistent homology outputs to small changes in the underlying data. In order to do this, it is very useful to construct metrics on the output barcodes so that one can assert the continuity of the assignment of a barcode to a finite metric space or to the graph of a function. Informally, one wants to prove that small changes in the data give rise to small changes in the barcodes. Since small changes will often result in a change in the number of bars, it will be important to construct a set of all barcodes, on which we can impose a metric. The following is a natural construction. Let  $n$  be a positive integer, and let  $B_n$  denote the set of unordered  $n$ -tuples of closed intervals  $[x, y]$ , where we permit  $x = y$ , and require  $x, y \geq 0$ . It is understood that  $B_0$  consists of a single point, namely the empty set of intervals. The set  $B_n$  can be described as the orbit space of the  $\Sigma_n$ -action on the set  $\mathcal{I}^n$  that permutes coordinates, where  $\mathcal{I}$  denotes the set of closed intervals. To consider all barcodes, we form

$$\mathfrak{B}_+ = \coprod_{n \geq 0} B_n$$

and define the full barcode space  $\mathfrak{B}$  to be the quotient  $\mathfrak{B}_+ / \sim$ , where  $\sim$  is the equivalence relation generated by the relations

$$\{I_1, \dots, I_{k-1}, [x_i, x_i], I_{k+1}, \dots, I_n\} \sim \{I_1, \dots, I_{k-1}, I_{k+1}, \dots, I_n\}.$$

The idea is that intervals of length zero, which do not represent nonzero vector spaces, should be ignored. We will need to construct metrics on  $\mathfrak{B}$  and to prove continuity results for these metrics.

### 8.5.1 Metrics on barcode space

The general idea for the construction of metrics on barcode space is to consider the set of all partial matchings between the intervals in the barcodes, assign a penalty to each such matching, and finally to minimize this penalty over the set of all matchings. Partial matchings are a little awkward, so for a pair of barcodes  $B_1$  and  $B_2$  one instead considers actual bijections  $B'_1 = B_1 \cup Z \rightarrow B_2 \cup Z = B'_2$ , where  $Z$  is the set consisting of a countable number of copies of the interval of length zero  $[x, x]$  for each  $x \geq 0$ . To start, one assigns a penalty  $\pi([x_1, x_2], [y_1, y_2]) = \|(y_1 - x_1, y_2 - x_2)\|_\infty$  for every pair of intervals including those of length zero. Next, given two barcodes  $B_1$  and  $B_2$ , let  $\mathcal{D}(B_1, B_2)$  denote the set of all bijections  $\theta : B'_1 \rightarrow B'_2$  for which  $\pi(I, \theta(I)) \neq 0$  for only finitely many  $I \in B'_1$ . Given a positive number  $p$ , we extend the penalty function  $\pi$  from individual intervals to barcodes by forming

$$W_p(B_1, B_2) = \inf_{\theta \in \mathcal{D}(B_1, B_2)} \left( \sum_{I \in B'_1} \pi(I, \theta(I))^p \right)^{\frac{1}{p}}.$$

As usual,  $p = \infty$  is interpreted as the  $L_\infty$  norm.

**Definition 8.5.1.** For any  $p > 0$  and also for  $p = \infty$ , we refer to  $W_p(B_1, B_2)$  as the  $p$ -Wasserstein distance. For  $p = \infty$ , this distance is often referred to as the bottleneck distance.

It is readily verified that under this definition,  $W_p$  defines a metric on  $\mathfrak{B}$ .

### 8.5.2 Stability theorems

The metrics defined in the previous section have stability theorems associated to them, that assert that the assignment of barcodes is a continuous process. There are two situations of interest.

1. Gromov has defined (see [41]) a metric  $d_{GH}$  on the set of all isometry classes of compact metric spaces, called the Gromov-Hausdorff metric. For any integer  $k \geq 0$ , one can view the assignment to any finite metric space its barcode as a map from the set of isometry classes of finite metric spaces to  $\mathfrak{B}$ , and one can ask about its continuity properties.
2. Let  $f : X \rightarrow \mathbb{R}_+$  denote a continuous function. One can assign to each such  $f$  and each integer  $i \geq 0$  the persistence vector space  $\{H_i(f^{-1}([0, r]))\}_{r \geq 0}$ . Under suitable situations (e.g. where  $X$  is a finite simplicial complex and  $f$  is linear on simplices, or where  $M$  is a compact manifold and  $f$  is smooth), one can show that the associated barcode is of

finite type. It is then an interesting question to ask about the continuity properties of this assignment, where one assigns various metrics to the set of functions on  $X$ .

There are theorems in both these cases. The following theorem demonstrates the continuity of the assignment of a  $k$ -dimensional barcode with coefficients in a field to a finite metric space, when the metric on the set of isometry classes of finite metric spaces is the Gromov-Hausdorff distance.

**Theorem 8.5.2.** [26] *For any two finite metric spaces  $X$  and  $Y$  and integer  $k \geq 0$ , let  $B(X)$  and  $B(Y)$  denote the  $k$ -dimensional barcodes for the Vietoris-Rips complexes of  $X$  and  $Y$  in a field  $K$ . Then*

$$W_\infty(B(X), B(Y)) \leq d_{GH}(X, Y).$$

There is a direct analogue for continuous real valued functions on a topological space. In order to state it, we need a pair of definitions.

**Definition 8.5.3.** *Let  $X$  be a topological space and  $f$  a continuous real valued function on  $X$ . A real number  $a$  is said to be a homological critical value of  $f$  if for some  $k$  and all sufficiently small  $\varepsilon > 0$  the inclusion*

$$H_k(f^{-1}(-\infty, a - \varepsilon]) \rightarrow H_k(f^{-1}(-\infty, a + \varepsilon])$$

*is not an isomorphism.*

**Definition 8.5.4.** *Let  $X$  be a topological space and  $f$  be a continuous real-valued function. We say  $f$  is tame if there are finitely many homological critical values and all the homology groups  $H_k(f^{-1}(-\infty, a])$  with coefficients in a field  $K$  are finite dimensional.*

**Remark 8.5.5.** Tameness holds in a number of familiar situations.

1. Morse functions on compact smooth manifolds.
2. Functions on finite simplicial complexes that are linear on each simplex.
3. Morse functions on compact Whitney-stratified spaces.

The theorem is as follows.

**Theorem 8.5.6.** [28] *Let  $X$  be triangulable space, and suppose  $f, g : X \rightarrow \mathbb{R}$  are tame continuous functions. Let  $B(f)$  and  $B(g)$  denote the barcodes attached to  $\{H_k(f^{-1}(-\infty, r]; K)\}_r$  and  $\{H_k(g^{-1}(-\infty, r]; K)\}_r$ , respectively. Then*

$$W_\infty(B(f), B(g)) \leq \|f - g\|_\infty.$$

The situation for the  $p$ -Wasserstein distances where  $p \neq \infty$  is more complex. We will need constraints on the metric space as well as on the functions. For the space  $X$ , it is required to be a triangulable compact metric space, so  $X$  is homeomorphic to a finite simplicial complex. In addition, though, there is a requirement that the number of simplices required to construct a triangulation where the diameter of the simplices is less than a threshold  $r$ .

Specifically, for a given  $r > 0$ , we define  $N(r)$  to be the minimal number of simplices in a triangulation of  $X$  for which each simplex has diameter  $\leq r$ . We will assume that  $N(r)$  grows polynomially with  $r^{-1}$ , i.e. that there are constants  $C$  and  $m$  so that  $N(r) \leq \frac{C}{r^m}$ . It is easy to observe that this result holds for a finite simplicial complex  $X$  equipped with the Euclidean metric obtained by restricting along a piecewise linear embedding  $X \hookrightarrow \mathbb{R}^n$ , as well as for a compact Riemannian manifold. In [29], it is proved that any metric space satisfying this condition also satisfies a homological condition. To state the homological condition, given a barcode  $B = \{[x_1, y_1], \dots, [x_n, y_n]\}$ , we define  $P_k(B)$  to be the sum

$$P_k(B) = \sum_i (y_i - x_i)^k.$$

**Lemma 8.5.7.** [29] *Suppose that  $X$  is as above, and that  $k$  is a nonnegative real number. Then there is a constant  $C_X$  so that  $P_k(B(f)) \leq C_X$  for every tame function  $f : X \rightarrow \mathbb{R}$  with Lipschitz constant  $L(f) \leq 1$ , where  $B(f)$  is defined as in Theorem 8.5.6.*

**Definition 8.5.8.** *When the conclusion of Lemma 8.5.7 holds for a metric space  $X$  and real number  $k \geq 1$ , we say that  $X$  implies bounded degree- $k$  persistence.*

The theorem is now as follows.

**Theorem 8.5.9.** [29] *Let  $X$  be a triangulable metric space that implies bounded degree- $k$  persistence for some real number  $k \geq 1$ , and let  $f, g : X \rightarrow \mathbb{R}$  be two tame Lipschitz functions. Let  $C_X$  be the constant in Definition 8.5.7, and let  $L(f)$  and  $L(g)$  denote the Lipschitz constants for  $f$  and  $g$  respectively. Then*

$$W_p(f, g) \leq C^{\frac{1}{p}} \cdot \|f - g\|_{\infty}^{1 - \frac{k}{p}}$$

for all  $p \geq k$ , where  $C = C_X \max\{L(f)^k, L(g)^k\}$ .

---

## 8.6 Tree-like metric spaces

Persistent homology gives ways of assessing the shape of a finite metric space. One situation where this is very useful is in problems in evolution. The notion that there is a “tree of life” is a very old one which actually predates Darwin. Different organisms of the same type have attached to them sequences of the same length in a genetic alphabet  $A$ . Therefore any set of organisms produces a subset of the set of sequences of fixed length in  $A$ . One can assign a metric to the space of all such sequences using Hamming distances or variants thereof. The notion that there is a tree of the various organisms within a fixed type can be restated in mathematical terms as stating that the space  $\mathcal{S}(A)$  corresponding to the organisms in the family is well modeled by a *tree-like metric space*, i.e. a metric space that is isometric to the set of nodes in a tree, possibly with weighted edges, equipped with the distance function that assigns to each pair of vertices of the tree the length of the shortest edge path between them. This approximability could be called the *phylogenetic hypothesis*



for the particular class of organisms. Testing this hypothesis for particular genetic data sets has usually been done by attempting to fit trees to a given metric space and attempting to assess how well the approximation fits. Given the persistent homology construction, one is tempted to develop criteria attached to the barcodes that can distinguish between tree-like and non-tree-like metric spaces. The following theorem, proved in [25] gives such a criterion.

**Theorem 8.6.1.** *Let  $X$  be a finite tree-like metric space. Then the  $k$ -dimensional persistent homology of  $X$  vanishes for  $k > 0$ .*

**Remark 8.6.2.** This theorem was proved in the context of a study of data sets of viral sequences. In that paper it was also shown that representative cycles for generators of persistent homology in positive degrees gave important clues to the mechanism of the failure of the phylogenetic hypothesis.

---

## 8.7 Persistence and feature generation

### 8.7.1 Introduction

The output of persistent homology is an interesting data type, consisting as it does of finite collections of intervals. When humans are directly interpreting barcodes, they are typically able to interpret barcodes directly from this description. However, there is a whole class of problems where computers are used to “interpret” the barcodes. For example, suppose that we have a database of complex molecules. Each molecule is given as a collection of atoms and bonds, and the bonds may be equipped with lengths. The set of atoms in a molecule can be endowed with a metric by considering the edge-path distance using the lengths of the bonds as the lengths of the edges. What we have is now a data set in which each of the *data points* is a finite metric space, and therefore possesses a barcode. If there are many molecules, we cannot hope to interpret these barcodes by eye, and must therefore allow a computer to deal with them. Machine learning algorithms are generally not well equipped to deal with data points described as sets, and it is therefore important to encode them somehow as vectors, which are the natural input to such algorithms. In this section we will describe three distinct methods for “vectorizing” barcodes, i.e. for creating coordinates on the set of barcodes. Specifically, we will define coordinates on the space  $\mathfrak{B}$  constructed in Section 8.5.

### 8.7.2 Algebraic functions

This method proceeds from the observation that the sets  $B_n$  can be viewed as subsets of a real algebraic variety. The set  $\mathcal{I}$  embeds as a subset of the two-dimensional affine space  $\mathbb{A}^2(\mathbb{R})$ , and is defined by the inequalities  $x, y \geq 0$  and  $y \geq x$  for an interval coordinatized by the pair  $(x, y)$ . Consequently, we have an embedding

$$\mathcal{I}^n \hookrightarrow \mathbb{A}^2(\mathbb{R})^n \cong \mathbb{A}^{2n}(\mathbb{R})$$

and it is equivariant with respect to the permutation actions on  $\mathcal{I}^n$  and  $\mathbb{A}^2(\mathbb{R})^n$ . Under the identification  $\mathbb{A}^2(\mathbb{R})^n \cong \mathbb{A}^{2n}(\mathbb{R})$ , with  $\mathbb{A}^{2n}(\mathbb{R})$  coordinatized using coordinates  $(x_1, \dots, x_n, y_1, \dots, y_n)$ , the corresponding action simply permutes the  $x_i$ 's and  $y_i$ 's among themselves. It is a standard result in algebraic geometry (see [52]) that for any action of a finite group on an affine algebraic variety (over  $\mathbb{R}$  in this case), there is an orbit variety, whose affine coordinate ring is the ring of invariants of the group action. It is easy to verify that in this case, the orbit set of the action on the closed real points of  $\mathbb{A}^{2n}(\mathbb{R})$  is exactly the symmetric product  $Sp^n(\mathbb{R}^2)$ . Since  $B_n \subseteq Sp^n(\mathbb{R}^2)$ , elements in the affine coordinate ring of the orbit variety can be regarded as functions on  $B_n$ , so we now have an algebra of functions  $\mathcal{A}_n$  on  $B_n$ . This means that we can describe functions on the sets of barcodes with exactly  $n$  intervals. What one really wants is a ring of functions on all of  $\mathfrak{B}$ . In order to construct such a ring, we observe that  $\mathfrak{B}$  can be described as a quotient of the direct limit of the system

$$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow \dots \tag{8.7.1}$$

where the inclusion  $B_n \rightarrow B_{n+1}$  is given by

$$\{[x_1, y_1], [x_2, y_2], \dots, [x_n, y_n]\} \rightarrow \{[x_1, y_1], [x_2, y_2], \dots, [x_n, y_n], [0, 0]\}.$$

There is a corresponding direct system of affine schemes

$$Spec(\mathcal{A}_0) \rightarrow Spec(\mathcal{A}_1) \rightarrow Spec(\mathcal{A}_2) \rightarrow \dots \tag{8.7.2}$$

which is compatible with the system (8.7.1) above. The colimit of the system (8.7.2) is an affine scheme, whose affine coordinate ring is the inverse limit of the system

$$\mathcal{A}_0 \leftarrow \mathcal{A}_1 \leftarrow \mathcal{A}_2 \leftarrow \dots$$

which we will denote  $\overline{\mathcal{A}}$ . The ring  $\overline{\mathcal{A}}$  can be analyzed, but is a bit too complicated to be used in applications. To define a smaller subring, we note that  $Spec(\overline{\mathcal{A}})$  is equipped with an action by the algebraic group  $\mathbb{G}_m$ , and we can define a subring  $\overline{\mathcal{A}}^{fin}$  to consist of all those functions  $f$  so that all the translates of  $f$  under the  $\mathbb{G}_m$ -action span a finite dimensional vector subspace within  $\overline{\mathcal{A}}$ . The ring  $\overline{\mathcal{A}}^{fin}$  is actually a graded ring, since the  $\mathbb{G}_m$ -action determines a grading on it. Within  $\overline{\mathcal{A}}^{fin}$  we define a subring  $\mathcal{A}^{fin}$  that consists of all elements of  $\overline{\mathcal{A}}^{fin}$  that respect the equivalence relation  $\sim$ . The main result of [3] is the following.

**Theorem 8.7.1.** *The rings in question have the following properties.*

1. The ring  $\overline{\mathcal{A}}^{fin}$  has the structure

$$\overline{\mathcal{A}}^{fin} \cong \mathbb{R}[x_{i,j}; 0 \leq i, 0 \leq j, \text{ and } i + j > 0].$$

2. The subring  $\mathcal{A}^{fin}$  is identified with

$$\mathbb{R}[x_{i,j}; 0 < i, 0 \leq j, \text{ and } i + j > 0].$$

3. The element  $x_{i,j}$  is the function given on a barcode  $\{[x_1, y_1], \dots, [x_n, y_n]\}$  by

$$\sum_{s=1}^n (y_s - x_s)^i (y_s + x_s)^j.$$

4. The elements of  $\mathcal{A}^{fin}$  separate points in  $\mathfrak{B}$ .

5. The ring  $\mathcal{A}^{fin}$  injects into the ring of functions on  $\mathfrak{B}$ .

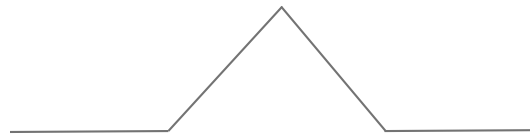
We remark that these functions are not continuous for the bottleneck distance on  $\mathfrak{B}$ . In [45], a tropical version of this work is presented, which gives functions that are continuous for the bottleneck distance. For the  $p < \infty$  situation, the functions are continuous for the  $p$ -Wasserstein distance if one assigns  $\mathfrak{B}$  the direct limit topology associated to the filtration of  $\mathfrak{B}$  by the images of the spaces  $B_n$ , defined in Section 8.5.1.

### 8.7.3 Persistence landscapes

*Persistence landscapes* were introduced in [12] as another vectorization of barcodes. The vectorization in this case consists of an embedding of the set  $\mathfrak{B}$  in a set of sequences of functions on the real line. Let  $(a, b)$  denote a pair of elements of  $\mathbb{R}$  with  $a \leq b$ . Then we define a function  $f_{a,b}(t)$  on the real line by setting

$$f_{(a,b)}(t) = \min(t - a, b - t)_+$$

where  $c_+ = \max(c, 0)$ . A quick analysis of  $f_{(a,b)}$  shows that it is zero for  $t \leq a$  and  $t \geq b$ , that on the interval  $[a, \frac{a+b}{2}]$  it is equal to the graph of a line of slope 1 including the point  $(a, 0)$ , and on the interval  $[\frac{a+b}{2}, b]$  it is the graph of a line of slope  $-1$  including the point  $(b, 0)$ . The shape of the graph is that of a pyramid, as in Figure 8.7.1.



**FIGURE 8.7.1.** Graph of  $f_{a,b}$

**Remark 8.7.2.** Note that for  $a = b$ ,  $f_{(a,b)}(t) \equiv 0$ .

Given a persistence barcode  $B = \{(a_1, b_1), \dots, (a_n, b_n)\}$ , we define a family of functions  $\lambda_k(t)$  parametrized by a positive integer  $k$ . For  $k = 1$ , we let  $\lambda_1(B)(t)$  denote the maximum of all the values  $f_{(a_i, b_i)}(t)$  over all  $i$ . For  $k > 1$ , we set  $\lambda_k(B)(t)$  equal to the  $k$ -th largest value occurring in the set  $\{f_{(a_i, b_i)}(t)\}_i$ . The family of functions  $\{\lambda_k(B)(t)\}_{k>0}$  is defined to be the persistence landscape of the barcode  $B$ . It is clear that

$$\lambda_k(B)(t) \geq 0$$

that

$$\lambda_k(B)(t) \geq \lambda_{k+1}(B)(t)$$

and that

$$\lambda_k(B)(l) = 0 \text{ for } k > n.$$

In [12] it is also proved that each function  $\lambda_k(B)$  is 1-Lipschitz, i.e. that

$$|\lambda_k(B)(t) - \lambda_k(B)(t')| \leq |t - t'|.$$

To summarize, the persistence landscape lies in the vector space  $\mathfrak{F}$  of real valued functions on  $\mathbb{N} \times \mathbb{R}$ , and it follows directly from Remark 8.7.2 that the definition gives us a function  $PL : \mathfrak{B} \rightarrow \mathfrak{F}$ .

One extremely useful fact about the persistence landscape is that it is compatible with the various distances assigned to barcode spaces. We let  $\mathfrak{F}_p \subset \mathfrak{F}$  denote the space of functions with finite  $L_p$ -norm  $\| \cdot \|_p$ . It is clear that the function  $PL$  takes values in  $\mathfrak{F}_p$  for all  $p$ . Recall the definition of the  $p$ -Wasserstein distance  $W_p$  between barcodes from Section 8.5.1. Bubenik now proves the following in [12].

**Theorem 8.7.3.** *The function  $d_p(B, B') = \|PL(B) - PL(B')\|_p$  is a metric on  $\mathfrak{B}$ . The two metrics  $W_{p+1}$  and  $d_p$  generate the same topology on  $\mathfrak{B}$ . It follows that the map  $PL$  is continuous when  $\mathfrak{B}$  is equipped with the metric  $W_{p+1}$  and  $\mathfrak{F}_p$  is equipped with the metric associated with the  $L_p$  norm.*

**Remark 8.7.4.** Bubenik also provides explicit inequalities involving the two metrics in [12]

Bubenik also proves the following continuity theorem.

**Theorem 8.7.5.** *The map  $PL$  is 1-Lipschitz from  $\mathfrak{B}$  equipped with the bottleneck distance to the space of persistent landscapes equipped with the sup norm distance. This is equivalent to the algebraic statement*

$$|\lambda_k(B)(t) - \lambda_k(B')(t)| \leq W_\infty(B, B').$$

**Remark 8.7.6.** The map  $PL$  separates points.

### 8.7.4 Persistence images

There is another approach that proceeds by treating a barcode, recoded as a persistence diagram, as a collection of point masses and then smoothing the corresponding measure to produce an image, which is finally discretized by selecting pixels and assigning each pixel the average value of the function on a box surrounding that pixel. It is described in [2]. The detailed description is given in several steps. The input is a persistence diagram (it is more natural to use the persistence diagram view in this case), a collection of points  $B = \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ . We assume we are given a function  $\phi : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$  so that (a)  $\phi_u = \phi(u, -)$  is a probability distribution on  $\mathbb{R}^2$  for each  $u \in \mathbb{R}^2$  and (b) the mean of  $\phi_u$  is  $u$ . A standard choice is that of a spherically symmetric Gaussian with mean  $u$  and a fixed variance  $\sigma$ . We also assume we are given a continuous and piecewise differentiable nonnegative weighting function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  that is zero along the  $x$ -axis.

- Apply the coordinate change  $(x, y) \rightarrow (x, y - x) = (\xi, \eta)$  to  $\mathbb{R}^2$ , to obtain the new set of points  $\mathcal{B} = \{(\xi_1, \eta_1) \dots, (\xi_n, \eta_n)\}$  of the same cardinality in  $\mathbb{R}^2$ . The points that correspond to short intervals are now all located near the  $\xi$ -axis. The  $\xi$ -axis itself corresponds to intervals of length zero.
- Construct a new function  $\rho_{\mathcal{B}}(z) : \mathbb{R}^2 \rightarrow \mathbb{R}$ , called the *persistence surface of  $\mathcal{B}$* , defined by

$$\rho(z) = \sum_{i=1}^n f(\xi_i, \eta_i) \phi((\xi_i, \eta_i), z).$$

Notice that  $\rho$  vanishes on the  $x$ -axis.

- To construct a finite dimensional representation, we first assume that the persistence diagrams we will be dealing with will always lie in a bounded region in  $\mathbb{R}^2$ , and divide a box containing that region into a square grid. Construct the vector with coordinates in one to one correspondence with the squares of the grid, and assign the entry corresponding to a square to be the integral over that square of  $\rho_{\mathcal{B}}$ .

The following is proved in [2].

**Theorem 8.7.7.** *The map  $PI : \mathfrak{B} \rightarrow \mathbb{R}^N$  that assigns to a persistence diagram a vector using the above procedure is continuous when the metric on  $\mathfrak{B}$  is the 1-Wasserstein distance.*

**Remark 8.7.8.** In [2], estimates proving this result are given both in the case of a general choice of  $\phi$  and the special case where  $\phi$  is given by Gaussian distributions with fixed variance. Of course, the estimates in the latter case are stronger.

---

## 8.8 Generalized persistence

### 8.8.1 Introduction

Persistent homology operates on functors  $F$  from the category  $\mathbb{R}_+$  to simplicial complexes, by composing them with the homology functor to obtain a persistence vector space. It is useful to consider other parameter categories for the diagrams, which can also help clarify the structure of data sets. There are at least two such constructions that have been discussed. The first is *zig-zag* persistence, introduced in [19], and the second is *multidimensional persistence*, discussed for example in [22]. The first is designed to study the relationship between homology of constructions that are not nested within each other, such as distinct samples from a given space, and the second provides invariants of situations where it is natural to study filtrations of spaces involving more than one parameter, such as filtering by both the scale parameter  $R$  and a measure of density. We describe both extensions of the standard persistent homology methods.

### 8.8.2 Zig-zag persistence

Consider the triangulation of  $\mathbb{R}$  whose vertices are the integers. The set of vertices of the barycentric subdivision of this simplicial complex is equipped with a partial ordering, by recognizing that its elements are in one-to-one correspondence with the simplices in the original triangulation, and that that set is equipped with a partial ordering by treating it as a subset of the power set of  $\mathbb{R}$ . In concrete terms, we may view it as in one to one correspondence with the set all integers and half integers, with every integer  $n$  being less than or equal to the elements  $n \pm \frac{1}{2}$ . We'll refer to this partially ordered set as  $\mathfrak{Z}$ . A partially ordered set (and its corresponding category) is said to be connected if any two objects can be connected by a zig-zag sequence of morphisms. Connected partially ordered subsets of  $\mathfrak{Z}$  are always determined by a pair  $(x, y)$ , where  $x$  and  $y$  are both integers or half integers, via the rule that assigns to the pair  $(x, y)$  the collection of objects  $z$  for which  $x \leq z \leq y$  (in the total ordering on  $\mathbb{R}$ ).

**Definition 8.8.1.** For any category  $\underline{C}$ , a zig-zag persistence object in  $\underline{C}$  is a functor from a connected subcategory of  $\mathfrak{Z}$  to  $\underline{C}$ . Suppose further that  $\underline{C}$  is equipped with an object  $c_0$  that is both initial and terminal. Then for any connected subcategory  $\mathfrak{Z}_0 \subseteq \mathfrak{Z}$  and object  $c \in \underline{C}$ , we define the interval object for  $\mathfrak{Z}_0$  and  $c$  to be the functor  $F = F_{\mathfrak{Z}_0, c} : \mathfrak{Z} \rightarrow \underline{C}$  defined on objects by  $F(x) = c$  for all  $x \in \mathfrak{Z}_0$  and  $F(x) = c_0$  for any  $x \notin \mathfrak{Z}_0$ , and on morphisms by  $F(x \leq y) = id_c$  whenever  $x, y \in \mathfrak{Z}_0$ . The behavior on morphisms into or out of  $\mathfrak{Z}_0$  is uniquely determined by the fact that  $c_0$  is both initial and terminal. When  $\underline{C}$  is the category of vector spaces over a field  $K$ , then it is understood that  $c$  will be chosen to be a one-dimensional vector space over  $K$ .

Given a zig-zag persistence object in the category of simplicial complexes, we may apply a homology functor  $H_i(-; A)$  for an abelian group  $A$  to obtain a zig-zag persistence object in the category of abelian groups. The category of abelian groups has the zero object as an object that is both initial and terminal, and so the notion of interval objects makes sense. Of course if  $A = K$ , where  $K$  is a field, then we obtain a zig-zag persistence vector space. It turns out that they can be classified up to isomorphism.

**Theorem 8.8.2.** (P. Gabriel, [37]) Let  $\mathfrak{Z}' \subseteq \mathfrak{Z}$  be any finite connected subcategory, and let  $F$  denote any zig-zag persistence object in the category of finite dimensional vector spaces over a field  $K$  defined on  $\mathfrak{Z}'$ . Then there is a finite direct sum decomposition

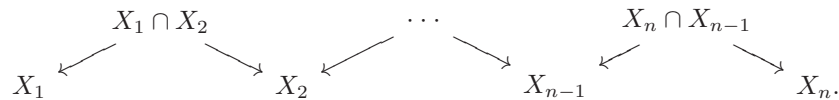
$$F \cong \bigoplus_i F_i$$

where each  $F_i$  is an interval object for  $\mathfrak{Z}_0$  and  $K$ , where  $\mathfrak{Z}_0 \subseteq \mathfrak{Z}'$  is a connected subcategory of  $\mathfrak{Z}'$ . Moreover, the sum is unique up to isomorphism and reordering of the sum.

**Corollary 8.8.3.** The classification of zig-zag persistence vector spaces based on a connected subcategory  $\mathfrak{Z}$  is given by barcodes where the intervals have endpoints integers or half integers. We'll refer to these barcodes as the zig-zag persistence barcodes of the zig-zag persistence vector space.

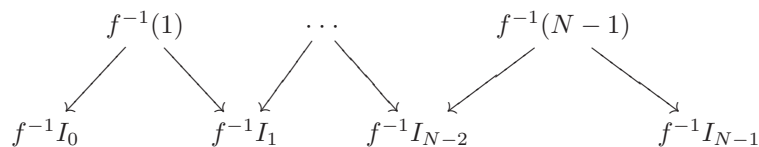
This theory with applications is discussed in [19] and [20]. Here are some particular situations in which this construction can be used.

1. **Samples:** Given a finite metric space, one can ask to what extent the persistent homology is captured on smaller samples from the data set. For example, suppose that we have taken a very large uniform sample  $X$  from the circle, and equip them with a metric by restricting the intrinsic metric (say) on the circle to these points. We will then with high probability find that the one-dimensional persistence barcode for the Vietoris-Rips construction on  $X$  will contain one long bar and many much shorter ones. Supposing that we do not actually know that the sample is coming from a circle, but simply observe that we obtain a one-dimensional barcode with one long bar and many shorter ones. A hypothesis suggested by this observation is that the data is obtained by sampling from a space with the homotopy type of the circle, but we may wonder if instead it has somehow appeared “by accident”. One way to provide confirmation of our hypothesis would be to observe that we obtain the same result for various subsamples of our space, and that they are compatible in an appropriate sense. Zig-zag persistence provides a way for carrying this out. We suppose that we have chosen samples  $X_1, \dots, X_n \subseteq X$ , and create a persistence object in the category of finite metric spaces and distance non-increasing maps



If we apply  $\mathcal{V}(-, r)$  for a fixed choice of  $r$ , guided by the beginning and endpoints of the observed long bar in the barcode for  $X$ , we obtain a persistence object in the category of simplicial complexes. If we further apply  $H_1(-, K)$  for a field  $K$ , we obtain a zig-zag persistence  $K$ -vector space. By the classification Theorem 8.8.2 above, we obtain a decomposition of the resulting zig-zag persistence  $K$ -vector space. The interpretation of the informal idea that each sample has a one-dimensional homology class and that they are consistent is the presence of an interval  $K$ -vector space for a relatively long interval within the set  $\{1, \frac{3}{2}, 2, \dots, n - \frac{1}{2}, n\}$ , or equivalently a relatively long bar in the zig-zag persistence barcode.

2. **Functions on spaces:** Suppose that we have a topological space  $X$  equipped with a continuous map  $f : X \rightarrow \mathbb{R}_+$ . Then we have the ordinary persistence  $K$ -vector spaces  $\{H_i(f^{-1}([0, r], K))\}_r$ , which encode information about the evolution of the homology of the sublevel sets of  $f$  as  $r$  increases. However, one might be interested in gaining information about approximations to *level sets* instead. They can provide more useful invariants in a number of cases, and are approachable through zig-zag persistent homology. We construct a zig-zag diagram of topological spaces as follows.



where  $I_k = [k, k+1]$  for all  $k$ . Again we can apply  $H_i(-, K)$  to this diagram, and obtain a

zig-zag persistence barcode. It contains information about how the spaces  $f^{-1}[k, k + 1]$  change as  $k$  changes, and how they assemble together. This situation is studied in [20].

3. **Witness complexes:** One of the problems with the witness complex is that we have very little theory about the extent to which it reflects accurately the persistent homology of the underlying metric space. A related problem is that there is no direct relationship between the construction for two different landmark sets. Even if  $\mathcal{L}_1 \subseteq \mathcal{L}_2$ , there are no maps directly relating  $W(X, \mathcal{L}_1, R)$  and  $W(X, \mathcal{L}_2, R)$  for two different landmark sets. One approach is to attempt to assess in some manner how consistent the results of the constructions based on  $\mathcal{L}_1$  and  $\mathcal{L}_2$  are. It turns out that given two landmark sets  $\mathcal{L}_1$  and  $\mathcal{L}_2$ , it is possible to construct an intermediate bivariate construction  $W(X, \{\mathcal{L}_1, \mathcal{L}_2\}, R)$  for which there is an evident diagram

$$W(X, \mathcal{L}_1, R) \longleftarrow W(X, \{\mathcal{L}_1, \mathcal{L}_2\}, R) \longrightarrow W(X, \mathcal{L}_2, R).$$

This is itself a short zig-zag diagram of length three, but if we have landmark sets  $\{\mathcal{L}_i\}_{i=0}^N$  we can clearly construct a longer diagram that includes the bivariate constructions  $W(X, \{\mathcal{L}_i, \mathcal{L}_{i+1}\}, R)$  for  $i = 0, \dots, N - 1$ . The construction is quite simple. Its vertex set is  $\mathcal{L}_1 \times \mathcal{L}_2$ , and a subset

$$\{l_1^1 \times l_2^1, l_1^2 \times l_2^2, \dots, l_1^k \times l_2^k\} \subseteq \mathcal{L}_1 \times \mathcal{L}_2$$

spans a simplex in  $W(X, \{\mathcal{L}_i, \mathcal{L}_{i+1}\}, R)$  if and only if there is a point  $x \in X$  so that  $x$  is a witness for  $\{l_1^1, l_1^2, \dots, l_1^k\}$  and  $\{l_2^1, l_2^2, \dots, l_2^k\}$  in the complexes  $W(X, \mathcal{L}_1, R)$  and  $W(X, \mathcal{L}_2, R)$ , respectively. The projections to  $W(X, \mathcal{L}_1, R)$  and  $W(X, \mathcal{L}_2, R)$  are defined in the evident way.

### 8.8.3 Multidimensional persistence

There are many situations where it can be useful to introduce families of spaces varying with more than one real parameter. For example, in any kind of topological analysis of data sets, it usually is the case that if one considers the persistent homology of the entire data set, the presence of outliers means that we do not typically obtain the “right homology”. For example, if we have data sampled from the unit circle, but a small number of points sprinkled throughout the unit disc, then persistent homology will end up reflecting the homology of the disc rather than that of the circle. This is often circumvented by selecting only the points of sufficient density, as measured by a density estimator, since outliers will typically have very low density. The question then becomes, though, how to choose the threshold for density. Also, it turns out that in general, there will be variation through different topologies as one changes the threshold. The solution to this problem is to attempt to study all thresholds at once, just as we do when considering the scale parameter in the Vietoris-Rips construction. This leads us to the following definition.

**Definition 8.8.4.** *Let  $\underline{C}$  be a category. Then by an  $n$ -dimensional persistence object in  $\underline{C}$  we mean a functor  $\mathbb{R}_+^n \rightarrow \underline{C}$ , where  $\mathbb{R}_+^n$  denotes the  $n$ -fold product of copies of the category  $\mathbb{R}_+$ .*



**Example 8.8.5.** Let  $X$  be any metric space, and suppose that  $X$  is equipped with a function  $f : X \rightarrow \mathbb{R}$ . It might be a density estimator, but it might also be a measure of centrality. Then if  $X[s]$  denotes the subset  $\{x \in X \mid f(x) \leq s\}$ , we obtain family of spaces  $\mathcal{V}(X[s], r)$ , and by applying homology with coefficients in a field  $K$ , we obtain a 2-dimensional persistence  $K$ -vector space parametrized by the pair  $(r, s)$ . While density is used as described above to remove outliers or noise, the case of a centrality measure allows one to capture the presence of the analogues of ends in a finite metric space.

**Example 8.8.6.** Multidimensional persistent homology can also be used to capture geometric information that is not strictly topological (see [23]). For example, given a Riemannian manifold, one can compute Gaussian curvature at each point and filter by that quantity. Considering the entire manifold, one can obtain a one-dimensional persistence vector space by applying homology over a field  $K$ . For computational purposes, though, we would need to deal with a sample and use a second parameter, namely the scale parameter in a Vietoris-Rips complex. This kind of analysis can for example be used to distinguish between various ellipsoids.

The equivalence of categories in Proposition 8.3.1 has the following straightforward analogue.

**Proposition 8.8.7.** *The category of  $n$ -dimensional persistence vector spaces over  $K$  is equivalent to the category of  $\mathbb{R}_+^n$ -graded modules over the  $\mathbb{R}_+^n$ -graded ring  $K[\mathbb{R}_+^n]$ . The analogous result with  $\mathbb{R}_+^n$  replaced by  $\mathbb{N}^n$  also holds.*

Although useful, this result does not give us a classification of multidimensional persistence  $K$ -vector spaces analogous to the barcode classification that works in the one-dimensional case. The reason can be understood as analogous to the commutative algebraic situation, where finitely generated  $K[x]$ -modules can be classified because  $K[x]$  is a principal ideal domain, but  $K[x_1, \dots, x_n]$ -modules cannot be classified for  $n \geq 2$ . In a sense it is provable that there is no classification strictly analogous to the  $n = 1$  case, because the classification in the case  $n \geq 2$  depends on the structure of the field  $K$ , as is demonstrated in [22]. This is not the case when  $n = 1$ , because the classification is always by barcodes, and that is independent of the structure of  $K$ .

Although there is no complete classification of multidimensional persistence vector spaces, there exist interesting invariants. For any  $k$ -dimensional persistence  $K$ -vector space  $\{V_x\}_{x \in \mathbb{R}^k}$ , and pair of points  $x, y \in \mathbb{R}^k$ , with  $x \leq y$  in the natural partial ordering on  $\mathbb{R}^k$ , we can define  $r(x, y)$  to be the rank of the linear transformation  $V_x \rightarrow V_y$ . We extend the definition to all pairs  $x, y \in \mathbb{R}^k$  by setting  $r(x, y) = 0$  when  $x$  is not less than or equal to  $y$ . The function  $r$  is therefore defined on  $\mathbb{R}^k \times \mathbb{R}^k$ , and we refer to it as the *rank invariant*. In the case  $k = 1$ , the rank invariant is complete, in that it differentiates between distinct barcodes. There is an analogue for multidimensional persistence  $K$ -vector spaces of the results described in Section 8.7.2. To state it, we first define an analogue of barcodes. By a *cube* in  $\mathbb{R}^k$ , we mean a set of the form  $I_1 \times I_2 \times \dots \times I_k$ , where each  $I_s$  is an interval  $[a_s, b_s]$ , and write  $C(a_1, \dots, a_k, b_1, \dots, b_k)$  for this cube. We write  $\mathcal{C}_k$  for the set of all  $k$ -dimensional cubes. There is a straightforward analogue to the space  $\mathfrak{B}$  defined in Section 8.7.1. We first define  $B_n = Sp^n(\mathcal{C}_k)$ , and define  $\mathfrak{B}(k)_+ = \coprod Sp^n(\mathcal{C}_k)$ . Next, we say a cube  $C(a_1, \dots, a_k, b_1, \dots, b_k)$

is negligible if  $a_i = b_i$  for some  $i$ , and define an equivalence relation  $\sim$  on  $\mathfrak{B}(k)_+$  to be the equivalence relation generated by the relation

$$\{C_1, C_2, \dots, C_n\} \sim \{C_1, \dots, C_{n-1}\} \text{ for all negligible } C_n.$$

We define  $\mathfrak{B}(k)$  to be  $\mathfrak{B}(k)_+ / \sim$ . Every equivalence class under  $\sim$  has a unique minimal representative consisting entirely of non-negligible cubes. Let  $\mathcal{M}(k)$  denote the set of isomorphism classes of  $k$ -dimensional persistence  $K$ -vector spaces. For each cube  $C = C(a_1, \dots, a_k, b_1, \dots, b_k)$  we let  $\mu(C)$  denote the isomorphism class of the  $k$ -dimensional persistence  $K$ -vector space  $\{V_{\vec{x}}\}_{\vec{x} \in \mathbb{R}_+^k}$  for which  $V_{\vec{x}} = K$  whenever  $\vec{x} \in C$ ,  $V_{\vec{x}} = \{0\}$  whenever  $\vec{x} \notin C$ , and for which all induced morphisms  $V_{\vec{x}} \rightarrow V_{\vec{y}}$  for  $\vec{x} \leq \vec{y}$  and  $\vec{x}, \vec{y} \in C$  are equal to the identity. There is an obvious map  $\theta : \mathfrak{B}(k) \rightarrow \mathcal{M}(k)$  which assigns to each minimal representative  $\{C_1, \dots, C_n\}$  the direct sum  $\oplus_i \mu(C_i)$ .

**Theorem 8.8.8.** [64] *The constructions above satisfy the following properties.*

1. *The set  $\mathfrak{B}(k)$  is a subset of the set of real points of an affine scheme  $\text{Spec}(A)$ .*
2. *The ring  $A$  is complicated, but there is a  $\mathbb{G}_m$ -action on  $\text{Spec}(A)$  that allows us to define a more manageable subring  $A^{fin} \subseteq A$ .*
3.  *$A^{fin}$  is isomorphic to the polynomial ring  $\mathbb{R}[x_{\vec{a}, \vec{b}}]$  where  $\vec{a}$  and  $\vec{b}$  are  $k$ -vectors of integers for which  $a_i \geq 1$  and  $b_i \geq 0$  for all  $i$ .*
4. *The ring  $A^{fin}$  separates points in  $\mathfrak{B}(k)$ , and maps injectively to the ring of all real valued functions on  $\mathfrak{B}(k)$ .*
5. *There is a natural lift of the ring homomorphism  $A^{fin} \rightarrow F(\mathfrak{B}(k), \mathbb{R})$  along  $\theta$  to a ring homomorphism  $j : A^{fin} \rightarrow F(\mathcal{M}(k), \mathbb{R})$ .  $F(X, \mathbb{R})$  denotes the ring of real-valued functions on a set  $X$ .*
6. *For any  $\alpha \in A^{fin}$ , the function  $j(\alpha) : \mathcal{M}(k) \rightarrow \mathbb{R}$  factors through the rank invariant. For any two elements  $X, Y \in \mathcal{M}(k)$ , if for all  $\alpha \in A^{fin}$ ,  $j(\alpha)(X) = j(\alpha)(Y)$ , then the rank invariants of  $X$  and  $Y$  are equal.*

This result gives one approach to the study of multidimensional persistence. There is a great deal of other work on this topic. The paper [24] deals with computing persistent homology using commutative algebra techniques, via the equivalence of categories from Proposition 8.8.7. They demonstrate that the multigrading yields significant simplification. In [62] and [24], an algebraic framework is constructed for dealing with the fact that there is always noise in the applications to data analysis. In [14], multidimensional persistence is studied by examining the family of all one dimensional persistence modules obtained by considering lines with varying angles in the persistence domain. M. Lesnick in [49] has defined metric properties of the set of isomorphism classes of multidimensional persistence vector spaces, and proven uniqueness results for them. Finally, in [50], software is developed for visualization and interrogation of two-dimensional persistence vector spaces.

One generalization that has not been studied yet is to multidimensional persistence where some of the persistence directions might be “zig-zag” directions rather than ordinary

persistence directions. Formally, this would mean functors from the categories of the form  $\mathbb{R}_+^m \times \mathfrak{Z}^n$ . This would be very useful in a number of situations. For example, in the zig-zag constructions discussed in Section 8.8 for samples and for witness complexes, we were forced to choose a threshold for the scale parameter. If we had a way of representing functors from  $\mathbb{R}_+ \times \mathfrak{Z}$  to vector spaces, we would not be forced to make this selection.

## 8.9 Coverage and evasion problems

There is an interesting set of technologies used for sensing of various kinds called *sensor nets*. A sensor net consists of a collection of sensors distributed throughout a domain. The sensors are very primitive in the sense that they are only capable of sensing the presence of an intruder or of another sensor within a fixed detection radius  $R$ . We also assume each sensor is given an identifying label or number, and that other sensors can sense that identifier when they are within the radius  $R$  of each other. One problem of interest is whether or not the balls of radius  $R$  cover the domain, and it does not have an immediate solution due to the fact that the positions of the sensors are not available. V. de Silva, and R. Ghrist have developed a very interesting method for addressing this problem based on persistent homology (see [31] and [32]). The rough idea is as follows. Suppose that one has a domain  $D$  in the plane, with a connected and compact curve boundary  $\partial D$ , that one has a collection of points  $\{v_i\}_{i \in I}$  in the region, one for each sensor, and that one knows in some way that  $\partial D$  is covered by the open balls  $B_R(v_i)$ . Let  $\mathcal{U}$  denote the family of open subsets  $\{B_R(v_i) \cap D\}_{i \in I}$  of  $D$ , and let  $\mathcal{U}^\partial$  denote the covering  $\{B_R(v_i) \cap \partial D\}_{i \in I}$  of  $\partial D$ . Suppose further that one knows that spaces

$$B_R(v_{i_1}) \cap \cdots \cap B_R(v_{i_s}) \cap D \quad \text{and} \quad B_R(v_{i_1}) \cap \cdots \cap B_R(v_{i_s}) \cap \partial D$$

are all either contractible or empty, for all choices of subsets  $\{i_1, \dots, i_s\} \subseteq I$ . The conditions ensure that  $D$  is homotopy equivalent to the nerve of  $\mathcal{U}$ , and that  $\partial D$  is homotopy equivalent to the nerve of the covering  $\mathcal{U}^\partial$ , as a consequence of the nerve theorem. It further ensures that the pair  $(N\mathcal{U}, N\mathcal{U}^\partial)$  is equivalent to the pair  $(D, \partial D)$ . For any field, the relative group  $H_2(D, \partial D; K) \cong K$ , since  $D$  is a connected orientable manifold with boundary  $\partial D$ , and it follows that the relative group  $H_2(N\mathcal{U}, N\mathcal{U}^\partial; K) \cong K$ . On the other hand, suppose that the sets  $B_R(v_i)$  do not cover all of  $D$ , and let  $D_0 \subseteq D$  denote the union

$$\bigcup_i B_r(v_i) \cap D.$$

The space  $D_0$  is a non-compact manifold with boundary, and consequently the relative group  $H_2(D_0, \partial D; K) \cong 0$ . As above, it will follow that

$$H_2(N\mathcal{U}, N\mathcal{U}^\partial; K) \cong 0.$$

Consequently, the simplicial complex of the nerve of the covering, which can be computed using the information available from the sensors, determines whether or not we have a covering based on its simplicial homology. The conditions on the coverings given above are of course impossible to verify, but de Silva and Ghrist are able to formulate a persistent homology condition that is a reasonable substitute, and which gives a homological criterion in terms of a 2-dimensional persistent homology group which is sufficient to guarantee coverage.

In order to understand the result in [32], we first observe that the information from the sensors do not give us access to the Čech complex, since we have no way of determining the intersection of balls without precise knowledge of the distances between their centers. However, we do have access to the Vietoris-Rips complex, since for any pair of points, we can tell whether or not they are within a distance  $R$ , where  $R$  is the detection radius. We also have the comparison results for the Vietoris-Rips complex and the Čech complex given in Proposition 8.4.1. The overall idea in [31] and [32] is to leverage the relationship between the complex we have access to (Vietoris-Rips) and the complex from which we can deduce coverage (Čech). In order to formulate such a result, we assume that we are attaching a second number to each sensor, namely its *covering radius*  $R_c$ . It is understood that each sensor covers a disc of radius  $R_c$  around it, and that it can detect other sensors at the detection radius  $R$  given above. We further assume that  $R_c \geq R/\sqrt{3}$ . This allows us to guarantee that if  $\{x_0, x_1, x_2\}$  forms a two simplex in the Vietoris-Rips complex  $\mathcal{V}(X, R)$ , then they span a two simplex in  $\check{C}(X, R_c)$ , by the second statement in Proposition 8.4.1. There are now the following assumptions made in [32].

1. The sensors lie in a compact connected domain  $D \subseteq \mathbb{R}^2$ , whose boundary  $\partial D$  is connected and piecewise linear with vertices called *fence nodes*.
2. Each fence node  $v$  is within  $R$  of its fence node neighbors on  $\partial D$ .

The following theorem is proved in [31] and [32].

**Theorem 8.9.1.** *Let  $\mathcal{R}$  denote the Vietoris-Rips complex of the set of all sensors, and let  $\mathcal{F}$  denote the subcomplex on the fence vertices. If the sensors satisfy the conditions above, and if there exists  $[\alpha] \in H_2(\mathcal{R}, \mathcal{F})$  so that  $\partial([\alpha]) \neq 0$ , where  $\partial : H_2(\mathcal{R}, \mathcal{F}) \rightarrow H_1(\mathcal{F})$  is the connecting homomorphism, then the balls of radius  $R_c$  around the sensors cover  $U$ .*

This theorem is in some situations not ideal, due to the strong assumptions on the boundary. In [32], it is shown that the use of the persistent homology of the pair  $(\mathcal{R}, \mathcal{F})$  can be used to obtain coverage results with much weaker hypotheses.

Another interesting direction is the study of time varying situations, where the sensors move in time. In this case, there are situations where the balls around the sensors do not cover the region at any fixed time, but that no “evader” can avoid being sensed at some time. This kind of problem is referred to as an *evasion problem*, and has been studied in [1] and [38]. The two approaches are quite distinct, the approach in [1] using zig-zag persistence, and the approach in [38] develops a new kind of cohomology with semigroup coefficients. The approach in [38] yields “if and only if” results.

## 8.10 Probabilistic analysis

### 8.10.1 Random complexes

A very interesting direction of research is the study of the distributions on the space of barcodes  $\mathfrak{B}$  (defined in Section 8.5) obtained by sampling points from Euclidean space using various models of randomness, i.e. sampling from various distributions on  $\mathbb{R}^n$ , or using the theory of random graphs [11]. Since we do not have a library of well understood distributions on  $\mathfrak{B}$ , one can instead study the distributions on the real line obtained by pushing forward a distribution on  $\mathfrak{B}$  along a map from  $\mathfrak{B}$  to  $\mathbb{R}$ . An interesting such map is

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} = B \rightarrow \max\left\{\frac{y_1}{x_1}, \dots, \frac{y_n}{x_n}\right\} = \lambda(B).$$

In fact, this map is only defined for barcodes in  $\mathfrak{B}$  for which the endpoints of the intervals in the barcode all lie in  $\{x \in \mathbb{R} \mid x > 0\}$ . When one is computing homology in dimensions  $\geq 1$  of a Čech or Vietoris-Rips complex of a set of points in Euclidean space, the barcode satisfies this property. A very interesting result in this direction is proved in [10]. They proceed by sampling from  $[0, 1]^d$  using a uniform Poisson process of intensity  $n$ . This means that the sampling is done from a uniform distribution on  $[0, 1]^d$ , but that the number of points sampled is governed by a Poisson distribution. The description of all these notions is beyond the scope of the present paper, but we refer the reader to [44]. The main result of [10] is the following.

**Theorem 8.10.1.** *We suppose that  $d$  and  $n$  are as above, that  $d \geq 2$ , that we are computing the  $k$ -dimensional persistence barcode  $B$ , and that  $1 \leq k \leq d - 1$ . Let*

$$\Delta_k(n) = \left(\frac{\log n}{\log \log n}\right)^{\frac{1}{k}}.$$

*Then there exist constants  $A_k$  and  $B_k$  so that*

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(A_k \leq \frac{\Pi_k(n)}{\Delta_k(n)} \leq B_k\right) = 1$$

*where  $\mathbb{P}$  denotes probability, and where  $\Pi_k(n)$  denotes the value of  $\lambda(B)$  for a barcode generated as above.*

### 8.10.2 Robust estimators

The stability theorem in Section 8.5.2 deals with the effect on persistence barcodes of small perturbations in the metric space, where perturbations are small in the sense of the Gromov-Hausdorff distance. In reality, though, one expects that in a perturbation of a metric space, a small number of distances may undergo relatively large perturbations. However, one believes that the number will be small, and that the points involved will be of small measure in an underlying measure. In order to deal with this problem, one incorporates a measure-theoretic component in one's definitions.

**Definition 8.10.2.** *By a metric measure space we will mean a complete separable metric space  $M$  equipped with a Borel measure  $\mu$ .*

In [40], a metric  $d_{GPr}$  called the *Gromov-Prokhorov* metric is introduced on the measure preserving isometry classes of compact metric measure spaces. It is constructed by combining the Gromov-Hausdorff metric on the isometry classes of compact metric spaces with the Prokhorov metric  $d_{Pr}$  on measures on a fixed metric space (see [58]) by methods which we will not discuss here.

The paper [8] studies the distributions on the space of persistence barcodes arising from the persistence barcodes obtained by sampling from a fixed metric measure spaces. More precisely, they study distributions on the completion of the metric space  $\overline{\mathfrak{B}}$  of  $\mathfrak{B}$  equipped with the bottleneck distance. Let  $\mu(n, k, X)$  denote the distribution on  $\overline{\mathfrak{B}}$  that arises from sampling a set  $S$  of  $n$  points on a metric measure space  $X$ , and computing the  $k$ -dimensional barcode on  $S$ . We have the following.

**Theorem 8.10.3.** *The inequality*

$$d_{Pr}(\mu(n, k, X), \mu(n, k, X')) \leq nd_{GPr}(X, X')$$

*holds for all compact metric measure spaces  $X$  and  $X'$ .*

This result is the used in [8] to develop *robust statistics* for distinguishing the results of sampling from a fixed metric space. Robust statistics are computable quantities attached to samples from distributions that are relatively insensitive to small changes in parameter values in the distribution from which the samples are gathered, and also to the presence of outliers. An elementary example of this idea is the median, which is relatively insensitive to outliers and is considered a robust statistic, while the mean is not. For the problem at hand, [8] defines a precise notion of robustness.

**Definition 8.10.4.** *Let  $f$  be a function from the set of isomorphism classes of finite metric spaces to a metric space  $(W, d_W)$ . We say the  $f$  is robust with robustness coefficient  $r > 0$  if for any nonempty finite metric space  $(X, d_X)$ , if for any nonempty finite metric space  $(X', d_{X'})$ , there exists a bound  $\delta$  such that for all isometric embeddings of  $X$  in a finite metric space  $(X', d_{X'})$  for which  $\#(X')/\#(X) < 1+r$ , it is the case that  $d(f(X), f(X')) < \delta$ . There is a corresponding uniform notion that states that the bound  $\delta$  may be chosen universally, for all  $X$ . There is a corresponding uniform notion that states that the bound  $\delta$  may be chosen universally, for all  $X$ .*

We now obtain the following result for finite metric spaces, which are being regarded as metric measure spaces by assigning to each metric space the uniform measure.

**Theorem 8.10.5.** [8] *For fixed  $n$  and  $k$ ,  $\mu(n, k, -)$  is uniformly robust with robustness coefficient  $r$  and estimate bound  $\delta = nr/(1+r)$  for any  $r$ .*

Since the space  $\overline{\mathfrak{B}}$  is relatively inaccessible, it is useful to use this result to construct real valued statistics that also satisfy the robustness property. One way to do this is to consider a fixed reference distribution  $\mathcal{P}$  on  $\overline{\mathfrak{B}}$ , and define  $\Delta_{\mathcal{P}}(n, k, X)$  to be the number  $d_{Pr}(\mu(n, k, X), \mathcal{P})$ .

**Theorem 8.10.6.** [8] For fixed  $n$  and  $k$ , and  $\mathcal{P}$ ,  $\Delta_{\mathcal{P}}(n, k, -)$  is uniformly robust with robustness coefficient  $r$  and estimate bound  $nr/(1+r)$  for any  $r$ .

It is also possible to obtain a somewhat simpler result, which does not require calculation of the full distribution  $\mu(n, k, X)$ . Instead of fixing a reference distribution  $\mathcal{P}$ , we choose a reference barcode  $B \in \overline{\mathfrak{B}}$ , and define  $\Delta_B^{med}(n, k, X)$  to be the median of the distribution of  $d_B(B, -)$  applied to samples of  $k$ -dimensional barcodes attached to samples of size  $n$  taken from the metric measure space  $X$ .

**Theorem 8.10.7.** For fixed  $n$ ,  $k$ , and  $B$ , the function  $\Delta_B^{med}(n, k, -)$  from finite metric spaces (with uniform probability measure) is robust with robustness coefficient greater than  $\ln 2/n$ .

### 8.10.3 Random fields

Suppose that we have a function  $f$  on a manifold. We have seen in Section 8.5 that we can associate to  $f$  the persistence vector spaces  $\{H_i(f^{-1}((-\infty, r], K))\}_r$  for  $i$  a nonnegative integer and  $K$  a field, and further that there are stability results that show that small changes in the function  $f$  lead to small changes, as measured by the bottleneck distance, in the corresponding sublevel set barcode. One can also ask, though, what the expected behavior of various of the features attached to barcode space is for a class of functions chosen at random. An initial question is what one means by a function chosen at random. The notion of a *random field* is defined in [4] as follows.

**Definition 8.10.8.** By a real-valued random field on a topological space  $T$  we mean a measurable mapping

$$F : \Omega \rightarrow \mathbb{R}^T$$

where  $\mathbb{R}^T$  denotes the set of all real-valued functions on  $T$  and  $(\Omega, \mathcal{F}, \mathbb{P})$  is a complete probability space. The set  $\mathbb{R}^T$  is equipped with the product  $\sigma$ -algebra over the set  $T$ .  $F$  creates a probability measure on  $\mathbb{R}^T$ , from which one can sample. Similarly, one can define an  $s$ -dimensional vector-valued random field as a measurable mapping

$$F : \Omega \rightarrow (\mathbb{R}^T)^s.$$

**Remark 8.10.9.** Note that the definition does not use the topology of  $T$ . However, in much of the work done in this area, one studies additional hypotheses on the random fields in which the topology of  $T$  plays a role.

The idea here is that rather than being a function, a random field is an assignment to each  $t \in T$  a distribution on  $\mathbb{R}$ , rather than a fixed value. Each of the restrictions  $\mathbb{R}^T \rightarrow \mathbb{R}^{\{t\}} \cong \mathbb{R}$  produces a random variable, and therefore the corresponding distribution, which we denote  $F_t$ . In fact, for any finite set of points  $t_1, \dots, t_n \in T$ , we obtain a distribution on  $\mathbb{R}^n$  which we denote  $F_{t_1, \dots, t_n}$ . There is a particular class of random fields called the *Gaussian random fields* that is particularly amenable to analysis.

**Definition 8.10.10.** A real-valued random field  $F$  is a Gaussian random field if for all  $n$ , the  $n$ -dimensional distributions  $F_{t_1, \dots, t_n}$  are multivariate Gaussian distributions on  $\mathbb{R}^n$ .

An  $s$ -dimensional random field is Gaussian, if all the distributions  $F_{t_1, \dots, t_n}$  are multivariate Gaussian distributions on  $\mathbb{R}^{n \times s}$ . Note that in Gaussian fields, the behavior of the random function is completely determined by the expectation function  $m(t) = E(F_{t_1, \dots, t_n})$ , and the covariance function  $C$ .

The first example of this construction comes out of work of Wiener (see [67] and [6]), using analysis of Brownian motion. Wiener studied the case  $T = \mathbb{R}_+$ , and produced a Gaussian random field  $W$ , where the expected value of  $W_t$  is always  $= 0$  and the variance is given by  $C(s, t) = \min(s, t)$ . He also showed that when one samples from the associated measure on  $\mathbb{R}^{\mathbb{R}_+}$ , one obtains continuous functions with probability 1, and so one calls  $W$  a *continuous Gaussian field*. Given any analytic property of functions on manifolds, such as  $k$ -th order differentiability, smoothness, or the property of being a Morse function, one can create and study Gaussian random fields whose samples have the given property with probability 1. Further, there are frequently a priori conditions on the covariance function of the random field that can be readily verified and guarantee the satisfaction of such properties.

The paper [9] proves a result concerning the persistent homology of the sublevel sets of functions sampled from Gaussian random fields. We consider the real-valued function  $\sigma$  on barcodes given by

$$\sigma\{[a_1, b_1], \dots, [a_n, b_n]\} = \sum_i (b_i - a_i).$$

For any fixed  $x \in \mathbb{R}$  and barcode  $\beta = \{[a_1, b_1], \dots, [a_n, b_n]\}$ , we define the  $x$ -truncation of  $\beta$ ,  $\beta[x]$ , to be the barcode

$$\{[a_1, \min(b_1, x)], \dots, [a_n, \min(b_n, x)]\}$$

where it is understood that for any  $i$  such that  $x \leq a_i$ , the interval  $[a_i, b_i]$  is simply deleted. Finally, we define

$$\chi^{pers}(M, f, x) = \sum_{i=0}^{\infty} (-1)^i \sigma(\{H_i(f^{-1}((-\infty, r], K))\}_r[x]).$$

In [9], the following result is proved concerning the distribution of  $\chi^{pers}(M, f, x)$  for Gaussian random fields on Riemannian manifolds which produce Morse functions with probability one.

**Theorem 8.10.11.** *Let  $M$  be a closed  $d$ -dimensional Riemannian manifold, and let  $F$  be a smooth real valued Gaussian random field so that  $F$  is Morse with probability 1, and so that the mean is identically zero and the variance is identically equal to 1. Then for any  $x \in \mathbb{R}$ , we have*

$$\begin{aligned} \mathbb{E}\{\chi^{pers}(M, f, x)\} &= \chi(M)(\varphi(x) + x\Phi(x)) \\ &+ \varphi(x) \sum_{j=1}^d (2\pi)^{-j/2} \mathcal{L}_j(M) H_{j-2}(-x) \end{aligned}$$

where

1.  $H_n$  denotes the  $n$ -th Hermite polynomial.
2.  $\varphi(x) = (2\pi)^{-1/2} e^{-x^2/2}$  is the density function for the standard Gaussian distribution.



3.  $\Phi(x) = \int_{-\infty}^x \varphi(u) du$ .
4.  $\mathcal{L}_j(M)$  denotes the  $j$ -th Lipschitz-Killing curvature of  $M$  (defined for example in [4], Section 7.6) with respect to a metric constructed from the covariance metric attached to  $F$ .

**Remark 8.10.12.** The point of this result is that it gives a theoretical estimate for the persistent Euler characteristic of sublevel sets in terms of classical invariants of the manifolds. Also, the result in [9] is actually proved in a much more general context, that of regular stratified spaces and stratified Morse theory, which in particular permits the study of manifolds with boundary. It also includes the study of random fields that are of the form  $G \circ (F_1, \dots, F_k)$ , where  $G$  is a deterministic function from  $\mathbb{R}^k$  to  $\mathbb{R}$ , and  $(F_1, \dots, F_k)$  is a vector-valued Gaussian random field.

---

## Bibliography

- [1] H. Adams and G. Carlsson, *Evasion paths in mobile sensor networks*, The International Journal of Robotics Research, 34 (1), 2015, 90–104.
- [2] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, *Persistence images: a stable vector representation of persistent homology*, J. Mach. Learn. Res., 18, 2017, 1–35.
- [3] A. Adcock, E. Carlsson, and G. Carlsson, *The ring of algebraic functions on persistence barcodes*, Homology, Homotopy, and Appl., 18, 2016, 381–402.
- [4] R. Adler and J. Taylor, *Random fields and geometry*, Springer, 2009.
- [5] N. Akkiraju, H. Edelsbrunner, M. Facello, P. Fu, E. Mucke, and C. Varela, *Alpha shapes: definition and software*, in Proc. Internat. Comput. Geom. Software Workshop 1995.
- [6] P. Baldi, *Stochastic calculus, an introduction through theory and exercises*, Springer Universitext, 2017.
- [7] S. Barannikov, *The framed Morse complex and its invariants* Adv. Soviet Math., 21, 1994, 93–115.
- [8] A. Blumberg, I. Gal, M. Mandell, and M. Pancia, *Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces*, Found. Comput. Math., 14, 2014, 745–789.
- [9] O. Bobrowski and M. Borman, *Euler integration of Gaussian random fields and persistent homology*, J. Topol. and Anal., 4 (1), 2012, 49–70.
- [10] O. Bobrowski, M. Kahle, and P. Skraba, *Maximally persistent cycles in random geometric complexes*, Ann. Appl. Probab., 27 (4), 2017, 2032–2060.
- [11] B. Bollobás, *Random graphs*, second edition, Cambridge University Press, 2011.
- [12] P. Bubenik, *Statistical topological data analysis using persistence landscapes*, The Journal of Machine Learning Research 16 (1), 2015, 77–102.
- [13] L. Buhovsky, V. Humilière, S. Seyfaddini. *The action spectrum and  $C^0$  symplectic topology*, arXiv:1808.09790, 2018.
- [14] F. Cagliari, B. Di Fabio, and M. Ferri, *One-dimensional reduction of multidimensional persistent homology*, Proc. Amer. Mat. Soc., 138 (8), 2010, 3003–3017.
- [15] Z. X. Cang, Lin Mu and G. Wei, *Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening*, PLOS Computational Biology, 14 (1), 2018, e100592.
- [16] Z. X. Cang and G. Wei, *TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions*, PLOS Computational Biology, 13 (7), 2017, e1005690.
- [17] G. Carlsson, *Topology and data*, Bull. Amer. Math. Soc., 46 (2), 2009, 255–308.
- [18] G. Carlsson, *Topological pattern recognition for point cloud data*, Acta Numer., 23, 2014, 289–368
- [19] G. Carlsson and V. de Silva, *Zigzag persistence*, Found. Comput. Math., 10 (4), 2010, 367–405

- [20] G. Carlsson, V. de Silva, and D. Morozov, *Zigzag persistent homology and real-valued functions*, Proceedings of the twenty-fifth symposium on computational geometry, ACM, 2009, 247–256.
- [21] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, *On the local behavior of spaces of natural images*, Int. J. Comput. Vis., 76 (1) 2008, 1–12.
- [22] G. Carlsson and A. Zomorodian, *The theory of multidimensional persistence*, Discrete Comput. Geometry 42 (1), 2009,71–93.
- [23] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, *Persistence barcodes for shapes*, International Journal of Shape Modeling, 11 (02), 2005, 149–187.
- [24] W. Chacholski, M. Scalamiero, and F. Vaccarino, *Combinatorial presentation of multidimensional persistent homology*, J. Pure Appl. Alg., 221 (5), 2017, 1055–1075.
- [25] J. Chan, G. Carlsson, and R. Rabadan, *Topology of viral evolution*, Proc. Nat. Acad. Sci. USA, 110 (46), 2013, 18566–18571.
- [26] F. Chazal, D. Cohen-Steiner, L. Guibas, F. Mémoli, and S. Oudot, *Gromov-Hausdorff stable signatures for shapes using persistence*, Eurographics Symposium on Geometry Processing, 28 (5), 2009.
- [27] C. Chong and S. Kumar, *Sensor networks: Evolution, opportunities, and challenges*, Proceedings of the IEEE, 91,8, 2003, 1247–1256.
- [28] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, *Stability of persistence diagrams*, Discrete and Computational Geometry, 37 (1), 2007, 103–120.
- [29] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko, *Lipschitz functions have  $L_p$ -stable persistence*, Found. Comput. Math., 10 (2), 2010, 127–139.
- [30] V. de Silva and G. Carlsson, *Topological estimation using witness complexes*, Symposium on Point Based Graphics, ETH, Zürich, Switzerland, 2004.
- [31] V. de Silva and R. Ghrist, *Coverage in sensor networks via persistent homology*, Alg. Geom. Topol. 7, 2007, 339–358.
- [32] V. de Silva and R. Ghrist, *Homological sensor networks*, Notices Amer. Math. Soc., 54 (1), 2007, 10–17.
- [33] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, *On the shape of a set of points in the plane*, IEEE Transactions on Information Theory, 29 (4), 1983, 551–559.
- [34] H. Edelsbrunner, D. Letscher, A. Zomorodian, *Topological persistence and simplification*, Discrete and Computational Geometry, 28 (4), 2002, 511–533.
- [35] H. Edelsbrunner and J. Harer, *Persistent homology - a survey*, Contemp. Math. 453, Amer. Math. Soc., 2008, 257–282.
- [36] P. Frosini, *A distance for similarity classes of submanifolds of a Euclidean space*, Bull. Aust. Math. Soc. 42 (3), 1990, 407–415.
- [37] P. Gabriel, *Unzerlegbare Darstellungen I*, Manuscripta Math 6, 1972, 71–103.
- [38] R. Ghrist and S. Krishnan, *Positive Alexander duality for pursuit and evasion*, SIAM Journal on Applied Algebra and Geometry, 1 (1), 2017, 308–327.
- [39] C. Giusti, E. Pastalkova, C. Curto, and V. Itskov, *Clique topology reveals intrinsic geometric structure in neural correlations*, Proc. Nat. Acad. Sci. USA, 112 (44), 2015, 13455–13460.
- [40] A. Greven, A. Pfaffelhuber, and A. Winter, *Convergence in distribution of random metric measure spaces*, Probab. Theory Related Fields, 145, 2009, 285–322.
- [41] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*, Birkhäuser, 2007.
- [42] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning. Data mining, inference, and prediction*, Springer Series in Statistics, Springer, 2009.
- [43] Y. Hiraoka, T. Nakamura, A. Hirata, E.G. Excolar, K. Matsue, and Y. Nishiura, *Hierarchical structures of amorphous solids characterized by persistent homology*, Proc. Nat. Acad. Sci. USA, 113 (26), 2016, 7035–7040,
- [44] P. Jones and P. Smith, *Stochastic processes. An introduction*, CRC Press, 2018.
- [45] S. Kalisnik, *Tropical coordinates on the space of persistence barcodes*, Found. Comput. Math., 19 (1), 2019, 101–129,
- [46] L. Kanari, P. Dlotko, M. Scalamiero, R. Levi, J. C. Shillcock, K. Hess, and H. Markram, *A topological representation of branching morphologies*, Neuroinformatics, 16 (1), 2018, 3–13.
- [47] D. Kozlov, *Combinatorial Algebraic Topology*, Algorithms and Computation in Mathematics, 21, Springer, 2008.
- [48] F. Le Roux, S. Seyfaddini, and C. Viterbo, *Barcodes and area-preserving homeomorphisms*, arXiv:1810.03039, 2018.

- [49] M. Lesnick, *The theory of the interleaving distance on multidimensional persistence modules*, Found. Comput. Math., 15 (3), 2015, 613–650.
- [50] M. Lesnick and M. Wright, *Interactive visualization of 2-D persistence modules*, arXiv:1512.00180, 2015.
- [51] L. Li, W. Cheng, G. Glicksberg, O. Gottesman, R. Tarnier, R. Chen, E. Bottinger, and J. Dudley, *Identification of type 2 diabetes subgroups through topological analysis of patient similarity*, Science Translational Medicine, 7(311), doi: 10.1126/scitranslmed.aaa9364, 2015.
- [52] D. Mumford, J. Fogarty, and F. Kirwan, *Geometric invariant theory*, Springer Verlag, 2002.
- [53] M. Nicolau, A. Levine, and G. Carlsson, *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival*, Proc. Nat. Acad. Sci. USA, 108 (17), 2011, 7265–7270. doi: 10.1073/pnas.1102826108.
- [54] A. Olin, E. Henckel, Y. Chen, T. Lakshmikanth, C. Pou, J. Mikes, A. Gustafsson, A. Bernhardsson, C. Zhang, K. Bohlin, and P. Brodin, *Stereotypic immune system development in newborn children*, Cell, 174 (5), 2018, 1277–1292.e14. doi: 10.1016/j.cell.2018.06.045.
- [55] N. Otter, M. Porter, U. Tillmann, P. Grindrod, and H. Harrington, *A roadmap for the computation of persistent homology*, EPJ Data Science, 6 (17), 2017.
- [56] L. Polterovich and E. Shelukhin, *Autonomous Hamiltonian flows, Hofer’s geometry and persistence modules*, Selecta Math., 22, 2016, 227–296.
- [57] L. Polterovich, E. Shelukhin, and V. Stojisavljević, *Persistence modules with operators in Morse and Floer theory*, Moscow Math. J. 17 (4), 2017, 757–786.
- [58] Y. Prokhorov, *Convergence of random processes and limit theorems in probability theory*, Theory Probab. Appl., 1, 1956, 157–214.
- [59] M. W. Reimann, M. Nolte, M. Scolamiero, K. Turner, R. Perin, G. Chindemi, P. Dlotko, R. Levi, K. Hess, and H. Markram, *Cliques of neurons bound into cavities provide a missing link between structure and function*, Front. Comput. Neurosci., 12 June 2017, <https://doi.org/10.3389/fncom.2017.00048>.
- [60] V. Robins, *Towards computing homology from finite approximations*, Proceedings of the 14th Summer Conference on General Topology and its Applications (Brookville, NY, 1999), Topology Proc. 24, 1999, 503–532.
- [61] M. Saggari, O. Sporns, J. Gonzalez-Castillo, P. Bandettini, G. Carlsson, G. Glover, and A. Reiss, *Towards a new approach to reveal dynamical organization of the brain using topological data analysis*, Nature Communications, 11 (9), 1399, 2018, doi: 10.1038/s41467-018-03664-4.
- [62] M. Scolamiero, W. Chacholski, A. Lundman, R. Ramanujam, and S. Öberg, *Multidimensional persistence and noise*, Foundations of Computational Mathematics, 17 (6), 2017, 1367–1406.
- [63] G. Singh, F. Memoli, and G. Carlsson, *Topological methods for the analysis of high dimensional data sets and 3D object recognition*, Eurographics Symposium on Point-based Graphics, 2007, 91–100.
- [64] J. Skryzalin and G. Carlsson, *Numeric invariants from multidimensional persistence*, J. Appl. Comput. Topol., 1, 2017, 89–119.
- [65] M. Usher and J. Zhang, *Persistent homology and Floer-Novikov theory*, Geom. Topol., 6, 2016, 3333–3430.
- [66] K. Xia and G. Wei, *Persistent homology analysis of protein structure, flexibility and folding*, Int. J. Numer. Methods Biomed. Eng., 30, 2014, 814–844.
- [67] N. Wiener, *Nonlinear problems in random theory*, Technology Press Research Monographs, The Technology Press of the Massachusetts Institute of Technology and John Wiley & Sons, Chapman & Hall, Ltd., 1958.
- [68] A. Zomorodian and G. Carlsson, *Computing persistent homology*, Discrete Comput. Geom., 33 (2), 2005, 249–274.

DEPARTMENT OF MATHEMATICS, STANFORD UNIVERSITY, STANFORD, CA 94305,  
U.S.A.

*E-mail address:* carlsson@stanford.edu