A FRACTAL DIMENSION FOR MEASURES VIA PERSISTENT HOMOLOGY

HENRY ADAMS, MANUCHEHR AMINIAN, ELIN FARNELL, MICHAEL KIRBY, JOSHUA MIRTH, RACHEL NEVILLE, CHRIS PETERSON, AND CLAYTON SHONKWILER

Abstract. We use persistent homology in order to define a family of fractal dimensions, denoted $\dim^i_{\mathrm{PH}}(\mu)$ for each homological dimension $i \geq 0$, assigned to a probability measure μ on a metric space. The case of 0-dimensional homology ($i = 0$) relates to work by Michael J Steele (1988) studying the total length of a minimal spanning tree on a random sampling of points. Indeed, if μ is supported on a compact subset of Euclidean space $\mathbb{R}^m$ for $m \geq 2$, then Steele's work implies that $\dim^0_{\mathrm{PH}}(\mu) = m$ if the absolutely continuous part of μ has positive mass, and otherwise $\dim^0_{\mathrm{PH}}(\mu) < m$. Experiments suggest that similar results may be true for higher-dimensional homology $0 < i < m$, though this is an open question. Our fractal dimension is defined by considering a limit, as the number of points n goes to infinity, of the total sum of the i-dimensional persistent homology interval lengths for n random points selected from μ in an i.i.d. fashion. To some measures μ, we are able to assign a finer invariant, a curve measuring the limiting distribution of persistent homology interval lengths as the number of points goes to infinity. We prove this limiting curve exists in the case of 0-dimensional homology when μ is the uniform distribution over the unit interval, and conjecture that it exists when μ is the rescaled probability measure for a compact set in Euclidean space with positive Lebesgue measure.

1. Introduction

Let X be a metric space equipped with a probability measure μ. While fractal dimensions are most classically defined for a space, there are a variety of fractal dimension definitions for a measure, including the Hausdorff or packing dimension of a measure [32, 61, 25]. In this paper we use persistent homology to define a fractal dimension $\dim^i_{\mathrm{PH}}(\mu)$ associated to a measure μ for each homological dimension $i \geq 0$. Roughly speaking, $\dim^i_{\mathrm{PH}}(\mu)$ is determined by how the lengths of the persistent homology intervals for a random sample, X_n, of n points from X vary as n tends to infinity.

Our definition should be thought of as a generalization, to higher homological dimensions, of fractal dimensions related to minimal spanning trees, as studied, for example, in [70]. Indeed, the lengths of the 0-dimensional (reduced) persistent homology intervals corresponding to the Vietoris–Rips complex of a sample X_n are equal to the lengths of the edges in a minimal spanning tree with X_n as the set of vertices. In particular, if X is a subset of Euclidean space $\mathbb{R}^m$ with $m \geq 2$, then [70, Theorem 1] by Steele implies that $\dim^0_{\mathrm{PH}}(\mu) \leq m$, with equality when the absolutely continuous part of μ has positive mass (Proposition 4.2). Theoretical extensions of our work are considered in [69, 68], and an independent generalization of Steele's work to higher homological dimensions is considered in [27].

To some metric spaces X equipped with a measure μ we are able to assign a finer invariant that contains more information than just the fractal dimension. Consider the set of the lengths of all intervals in the i-dimensional persistent homology for X_n. Experiments suggest that when probability measure μ is absolutely continuous with respect to the Lebesgue measure on $X \subseteq \mathbb{R}^m$, the scaled set of interval lengths in each homological dimension i converges point-wise to some fixed probability distribution (depending on μ and i). It is easy to prove the weaker notion of convergence distribution-wise in the simple case of 0-dimensional homology when μ is the uniform distribution over the unit interval, in which case we can also derive a formula for the limiting distribution. Experiments suggest that when μ is the rescaled probability measure corresponding to a compact set $X \subseteq \mathbb{R}^m$ of positive Lebesgue measure, then a limiting rescaled distribution exists that depends only on m, i, and the volume of μ (see Conjecture 6.1). We would be interested to know the formulas for the limiting distributions with higher Euclidean and homological dimensions.

Whereas Steele in [70] studies minimal spanning trees on random subsets of a space, Kozma, Lotker, and Stupp in [46] study minimal spanning trees built on extremal subsets. Indeed, they define a fractal dimension for a metric space X as the infimum, over all powers d, such that for any minimal spanning tree T on a finite

number of points in X, the sum of the edge lengths in T each raised to the power d is bounded. They relate this extremal minimal spanning tree dimension to the box counting dimension. Their work is generalized to higher homological dimensions by Schweinhart [67]. By contrast, we instead generalize Steele's work [70] on measures to higher homological dimensions. Three differences between [46, 67] and our work are the following.

- The former references define a fractal dimension for metric spaces, whereas we define a fractal dimension for measures.
- The fractal dimension in [46, 67] is defined using extremal subsets, whereas we define our fractal dimension using random subsets.
- We can estimate our fractal dimension computationally using log-log plots as in Section 5, whereas we do not know a computational technique for estimating the fractal dimensions in [46, 67].

After describing related work in Section 2, we give preliminaries on fractal dimensions and on persistent homology in Section 3. We present the definition of our fractal dimension and prove some basic properties in Section 4. We demonstrate example experimental computations in Section 5; our code is publicly available at `https://github.com/CSU-PHdimension/PHdimension`. Section 6 describes how limiting distributions, when they exist, form a finer invariant. Sections 7 and 8 discuss the computational details involved in sampling from certain fractals and estimating asymptotic behavior, respectively. Finally we present our conclusion in Section 9. One of the main goals of this paper is to pose questions and conjectures, which are shared throughout.

2. Related work

2.1. **Minimal spanning trees.** The paper [70] studies the total length of a minimal spanning tree for random subsets of Euclidean space. Let X_n be a random sample of points from a compact subset of $\mathbb{R}^d$ according to some probability distribution. Let M_n be the sum of all the edge lengths of a minimal spanning tree on vertex set X_n. Then for $d \geq 2$, Theorem 1 of [70] says that

$$M_n \sim C n^{(d-1)/d} \quad \text{as} \quad n \to \infty, \tag{1}$$

where the relation $\sim$ denotes asymptotic convergence, with the ratio of the terms approaching one in the specified limit. Here, C is a constant depending on d and on the integral $\int f^{(d-1)/d}$, where f is the density of the absolutely continuous part of the probability distribution[1]. There has been a wide variety of related work, including for example [5, 6, 7, 42, 71, 72, 73, 74]. See [45] for a version of the central limit theorem in this context. The papers [58, 59] study the length of the longest edge in the minimal spanning tree for points sampled uniformly at random from the unit square, or from a torus of dimension at least two, and [47] extends this to any Ahlfors regular measure with connected support (i.e., to any connected semi-uniform metric measure space). By contrast, [46] studies Euclidean minimal spanning trees built on extremal finite subsets, as opposed to random subsets.

2.2. **Umbrella theorems for Euclidean functionals.** As Yukich explains in his book [79], there are a wide variety of Euclidean functionals, such as the length of the minimal spanning tree, the length of the traveling salesperson tour, and the length of the minimal matching, which all have scaling asymptotics analogous to (1). To prove such results, one needs to show that the Euclidean functional of interest satisfies translation invariance, subadditivity, superadditivity, and continuity, as in [22, Page 4]. Superadditivity does not always hold, for example it does not hold for the minimal spanning tree length functional, but there is a related "boundary minimal spanning tree functional" that does satisfy superadditivity. Furthermore, the boundary functional has the same asymptotics as the original functional, which is enough to prove scaling results. It is intriguing to ask if these techniques will work for functionals defined using higher-dimensional homology.

[1] If the compact subset has Hausdorff dimension less than d, then [70] implies $C = 0$.

2.3. **Random geometric graphs.** In this paper we consider simplicial complexes (say Vietoris–Rips or Čech) with randomly sampled points as the vertex set. The 1-skeleta of these simplicial complexes are random geometric graphs. We recommend the book [57] by Penrose as an introduction to random geometric graphs; related families of random graphs are also considered in [60]. Random geometric graphs are often studied when the scale parameter $r(n)$ is a function of the number of vertices n, with $r(n)$ tending to zero as n goes to infinity. Instead, in this paper we are more interested in the behavior over all scale parameters simultaneously. From a slightly different perspective, the paper [44] studies the expected Euler characteristic of the union of randomly sampled balls (potentially of varying radii) in the plane.

2.4. **Persistent homology.** Vanessa Robins' thesis [65] contains many related ideas; we describe one such example here. Given a set $X \subseteq \mathbb{R}^m$ and a scale parameter $\varepsilon \geq 0$, let

$$X_\varepsilon = \{y \in \mathbb{R}^m \mid \text{there exists some } x \in X \text{ with } d(y,x) \leq \varepsilon\}$$

denote the *ε-offset of X*. The ε-offset of X is equivalently the union of all closed ε balls centered at points in X. Furthermore, let $C(X_\varepsilon) \in \mathbb{N}$ denote the number of connected components of X_ε. In Chapter 5, Robins shows that for a generalized Cantor set X in $\mathbb{R}$ with Lebesgue measure 0, the box-counting dimension of X is equal to the limit

$$\lim_{\varepsilon \to 0} \frac{\log(C(X_\varepsilon))}{\log(1/\varepsilon)}.$$

Here Robins considers the entire Cantor set, whereas we study random subsets thereof.

The paper [51], which heavily influenced our work, introduces a fractal dimension defined using persistent homology. This fractal dimension depends on thickenings of the entire metric space X, as opposed to random or extremal subsets thereof. As a consequence, the computed dimension of some fractal shapes (such as the Cantor set cross the interval) disagrees significantly with the Hausdorff or box-counting dimension.

Schweinhart's paper [67] takes a slightly different approach from ours, considering extremal (as opposed to random) subsets. After fixing a homological dimension i, Schweinhart assigns a fractal dimension to each metric space X equal to the infimum over all powers d such that for any finite subset $X' \subseteq X$, the sum of the i-dimensional persistent homology bar lengths for X', each raised to the power d, is bounded. For low-dimensional metric spaces Schweinhart relates this dimension to the box counting dimension.

More recently, Divol and Polonik [27] independently obtain generalizations of [70, 79] to higher homological dimensions. In particular, they prove our Conjecture 4.3 in the case when X is a cube, and remark that a similar construction holds when the cube is replaced by any convex body. Related results are obtained in two papers by Schweinhart, which are in part inspired by our work: in [69] when X is a ball or sphere, and afterwards in [68] when points are sampled from a fractal according to an Ahlfors regular measure.

There is a growing literature on the topology of random geometric simplicial complexes, including in particular the homology of Vietoris–Rips and Čech complexes built on top of random points in Euclidean space [13, 43, 3]. The paper [14] shows that for n points sampled from the unit cube $[0,1]^d$ with $d \geq 2$, the maximally persistent cycle in dimension $1 \leq k \leq d-1$ has persistence of order $\Theta((\frac{\log n}{\log \log n})^{1/k})$, where the asymptotic notation big Theta means both big O and big Omega. The homology of Gaussian random fields is studied in [4], which gives the expected k-dimensional Betti numbers in the limit as the number of points increases to infinity, and also in [12]. The paper [30] studies the number of simplices and critical simplices in the alpha and Delaunay complexes of Euclidean point sets sampled according to a Poisson process. An open problem about the birth and death times of the points in a persistence diagram coming from sublevelsets of a Gaussian random field is stated in Problem 1 of [29]. The paper [19] shows that the expected persistence diagram, from a wide class of random point clouds, has a density with respect to the Lebesgue measure. We refer the reader also to [41, 55], which are related to our Conjecture 6.2 in the setting of point processes.

The paper [16] explores what attributes of an algebraic variety can be estimated from a random sample, such as the variety's dimension, degree, number of irreducible components, and defining polynomials; one of their estimates of dimension is inspired by our work.

In an experiment in [1], persistence diagrams are produced from random subsets of a variety of synthetic metric space classes. Machine learning tools, with these persistence diagrams as input, are then used to classify the metric spaces corresponding to each random subset. The authors obtain high classification rates between the different metric spaces. It is likely that the discriminating power is based not only on the

underlying homotopy types of the shape classes, but also on the shapes' dimensions as detected by persistent homology.

3. Preliminaries

This section contains background material and notation on fractal dimensions and persistent homology.

3.1. **Fractal dimensions.** The concept of fractal dimension was introduced by Hausdorff and others [40, 15, 31] to describe spaces like the Cantor set. It was later popularized by Mandelbrot [52], and found extensive application in the study of dynamical systems. The attracting sets of a simple dynamical system is often a submanifold, with an obvious dimension, but in non-linear and chaotic dynamical systems the attracting set may not be a manifold. The Cantor set, defined by removing the middle third from the interval $[0, 1]$, and then recursing on the remaining pieces, is a typical example. It has the same cardinality as $\mathbb{R}$, but it is nowhere-dense, meaning it at no point resembles a line. The typical fractal dimension of the Cantor set is $\log_3(2)$. Intuitively, the Cantor set has "too many" points to have dimension zero, but also should not have dimension one.

We speak of fractal dimensions in the plural because there are many different definitions. In particular, fractal dimensions can be divided into two classes, which have been called "metric" and "probabilistic" [33]. The former describe only the geometry of a metric space. Two widely-known definitions of this type, which often agree on well-behaved fractals, but are not in general equal, are the box-counting and Hausdorff dimensions. For an inviting introduction to fractal dimensions see [32]. Dimensions of the latter type take into account both the geometry of a given set and a probability distribution supported on that set—originally the "natural measure" of the attractor given by the associated dynamical system, but in principle any probability distribution can be used. The information dimension is the best known example of this type. For detailed comparisons, see [34]. Our persistent homology fractal dimension, Definition 4.1, is of the latter type.

For completeness, we exhibit some of the common definitions of fractal dimension. The primary definition for sets is given by the Hausdorff dimension [35].

Definition 3.1. Let S be a subset of a metric space X, let $d \in [0, \infty)$, and let $\delta > 0$. The *Hausdorff measure* of S is

$$H_d(S) = \inf_{\delta} \left(\inf \left\{ \sum_{j=1}^{\infty} \operatorname{diam}(B_j)^d \mid S \subseteq \bigcup_{j=1}^{\infty} B_j \text{ and } \operatorname{diam}(B_j) \leq \delta \right\} \right),$$

where the inner infimum is over all coverings of S by balls B_j of diameter at most δ. The *Hausdorff dimension* of S is

$$\dim_H(S) = \inf_d \{H_d(S) = 0.\}$$

The Hausdorff dimension of the Cantor set, for example, is $\log_3(2)$.

In practice it is difficult to compute the Hausdorff dimension of an arbitrary set, which has led to a number of alternative fractal dimension definitions in the literature. These dimensions tend to agree on well-behaved fractals, such as the Cantor set, but they need not coincide in general. Two worth mentioning are the box-counting dimension, which is relatively simple to define, and the correlation dimension.

Definition 3.2. Let $S \subseteq X$ a metric space, and let N_ε denote the infimum of the number of closed balls of radius ϵ required to cover S. Then the *box-counting dimension* of S is

$$\dim_B(S) = \lim_{\varepsilon \to 0} \frac{\log(N_\varepsilon)}{\log(1/\varepsilon)},$$

provided this limit exists. Replacing the limit with a lim sup gives the *upper* box-counting dimension, and a lim inf gives the *lower* box-counting dimension.

The box-counting definition is unchanged if N_ϵ is instead defined by taking the number of open balls of radius ε, or the number of sets of diameter at most ε, or (for S a subset of $\mathbb{R}^n$) the number of cubes of side-length ε [77, Definition 7.8], [32, Equivalent Definitions 2.1]. It can be shown that $\dim_B(S) \geq \dim_H(S)$. This inequality can be strict; for example if $S = \mathbb{Q} \cap [0, 1]$ is the set of all rational numbers between zero and one, then $\dim_H(S) = 0 < 1 = \dim_B(S)$ [32, Chapter 3]. If S is a self-similar shape that is nice enough,

i.e. satisfies an "open set" condition, then [32, Theorem 9.3] (for example) shows that the box-counting and Hausdorff dimensions agree: $\dim_B(S) = \dim_H(S)$.

In Section 4 we introduce a fractal dimension based on persistent homology which shares key similarities with the Hausdorff and box-counting dimensions. It can also be easily estimated via log-log plots, and it is defined for arbitrary metric spaces (though our examples will tend to be subsets of Euclidean space). A key difference, however, will be that ours is a fractal dimension for measures, rather than for subsets.

There are a variety of classical notions of a fractal dimension for a measure, including the Hausdorff, packing, and correlation dimensions of a measure [32, 61, 25]. We give the definitions of two of these.

Definition 3.3 ((13.16) of [32])**.** The *Hausdorff dimension* of a measure μ with total mass one is defined as

$$\dim_H(\mu) = \inf\{\dim_H(S) \mid S \text{ is a Borel subset with } \mu(S) > 0\}.$$

We have $\dim_H(\mu) \leq \dim_H(\operatorname{supp}(\mu))$, and it is possible for this inequality to be strict [32, Exercise 3.10][2]. We also give the definition of the correlation dimension of a measure.

Definition 3.4. Let X be a subset of $\mathbb{R}^m$ equipped with a measure μ, and let X_n be a random sample of n points from X. Let $\theta\colon \mathbb{R} \to \mathbb{R}$ denote the Heaviside step function, meaning $\theta(x) = 0$ for $x < 0$ and $\theta(x) = 1$ for $x \geq 0$. The *correlation integral* of μ is defined (for example in [37, 76]) to be

$$C(r) = \lim_{n\to\infty} \frac{1}{n^2} \sum_{\substack{x,x' \in X_n \\ x \neq x'}} \theta\left(r - \|x - x'\|\right).$$

It can be shown that $C(r) \propto r^\nu$, and the exponent ν is defined to be the *correlation dimension* of μ.

In [37, 38] it is shown that the correlation dimension gives a lower bound on the Hausdorff dimension of a measure. The correlation dimension can be easily estimated from a log-log plot, similar to the methods we use in Section 5. A different definition of the correlation dimension is given and studied in [24, 53]. The correlation dimension is a particular example of the family of *Rènyi dimensions*, which also includes the *information dimension* as a particular case [63, 64]. A collection of possible axioms that one might like to have such a fractal dimension satisfy is given in [53].

3.2. **Persistent homology.** The field of applied and computational topology has grown rapidly in recent years, with the topic of persistent homology gaining particular prominence. Persistent homology has enjoyed a wealth of meaningful applications to areas such as image analyis, chemistry, natural language processing, and neuroscience, to name just a few examples [2, 10, 21, 26, 49, 50, 78, 80]. The strength of persistent homology lies in its ability to characterize important features in data across multiple scales. Roughly speaking, homology provides the ability to count the number of independent k-dimensional holes in a space, and persistent homology provides a means of tracking such features as the scale increases. We provide a brief introduction to persistent homology in this preliminaries section, but we point the interested reader to [8, 28, 39] for thorough introductions to homology, and to [17, 23, 36] for excellent expository articles on persistent homology.

Geometric complexes, which are at the heart of the work in this paper, associate to a set of data points a simplicial complex—a combinatorial space that serves as a model for an underlying topological space from which the data has been sampled. The building blocks of simplicial complexes are called simplices, which include vertices as 0-simplices, edges as 1-simplices, triangles as 2-simplices, tetrahedra as 3-simplices, and their higher-dimensional analogues as k-simplices for larger values of k. An important example of a simplicial complex is the Vietoris–Rips complex.

Definition 3.5. Let X be a set of points in a metric space and let $r \geq 0$ be a scale parameter. We define the Vietoris–Rips simplicial complex $\mathrm{VR}(X; r)$ to have as its k-simplices those collections of $k+1$ points in X that have diameter at most r.

In constructing the Vietoris–Rips simplicial complex we translate our collection of points in X into a higher-dimensional complex that models topological features of the data. See Figure 1 for an example of a Vietoris–Rips complex constructed from a set of data points, and see [28] for an extended discussion.

[2]See also [33] for an example of a measure whose *information dimension* is less than the Hausdorff dimension of its support.

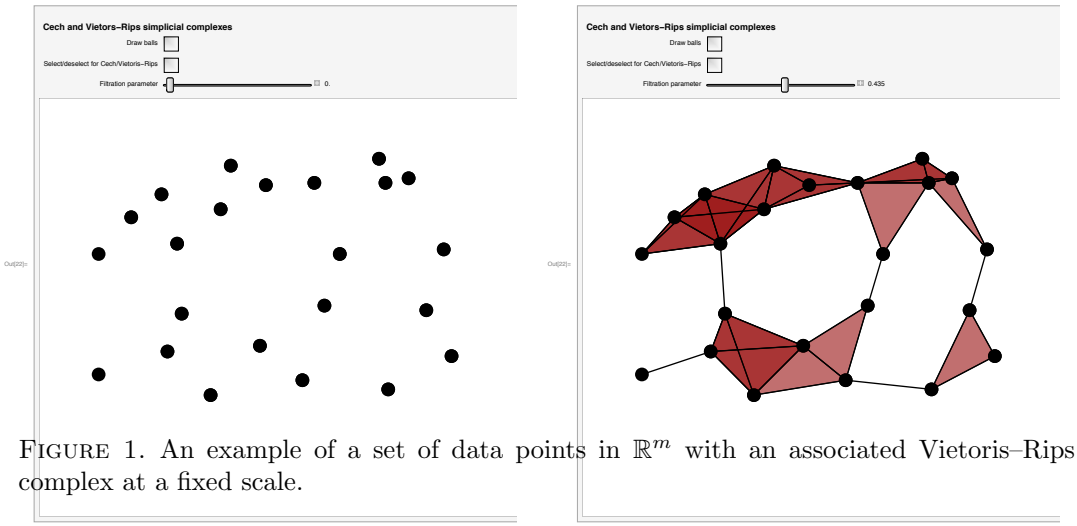

FIGURE 1. An example of a set of data points in $\mathbb{R}^m$ with an associated Vietoris–Rips complex at a fixed scale.

It is readily observed that for various data sets, there is not necessarily an ideal choice of the scale parameter so that the associated Vietoris–Rips complex captures the desired features in the data. The perspective behind persistence is to instead allow the scale parameter to increase and to observe the corresponding appearance and disappearance of topological features. To be more precise, each hole appears at a certain scale and disappears at a larger scale. Those holes that persist across a wide range of scales often reflect topological features in the shape underlying the data, whereas the holes that do not persist for long are often considered to be noise. However, in the context of this paper (estimating fractal dimensions), the holes that do not persist are perhaps better described as measuring the local geometry present in a random finite sample.

For a fixed set of points, we note that as scale increases, simplices can only be added and cannot be removed. Thus, for $r_0 < r_1 < r_2 < \cdots$, we obtain a filtration of Vietoris–Rips complexes

$$\mathrm{VR}(X; r_0) \subseteq \mathrm{VR}(X; r_1) \subseteq \mathrm{VR}(X; r_2) \subseteq \cdots .$$

The associated inclusion maps induce linear maps between the corresponding homology groups $H_k(\mathrm{VR}(X; r_i))$, which are algebraic structures whose ranks (roughly speaking) count the number of independent k-dimensional holes in the Vietoris–Rips complex. A technical remark is that homology depends on the choice of a group of coefficients; it is simplest to use field coefficients (for example $\mathbb{R}$, $\mathbb{Q}$, or $\mathbb{Z}/p\mathbb{Z}$ for p prime), in which case the homology groups are furthermore vector spaces. The corresponding collection of vector spaces and linear maps is called a *persistent homology module*.

A useful tool for visualizing and extracting meaning from persistent homology is a barcode. The basic idea is that each generator of persistent homology can be represented by an interval, whose start and end times are the *birth* and *death* scales of a homological feature in the data. These intervals can be arranged as a barcode graph in which the x-axis corresponds to the scale parameter. See Figure 2 for an example. If Y is a finite metric space, then we let $\mathrm{PH}^i(Y)$ denote the corresponding collection of i-dimensional persistent homology intervals. Indeed, any persistent homology module decomposes uniquely as a direct sum of interval summands.

Zero-dimensional barcodes always produce one infinite interval, as in Figure 2, which are problematic for our purposes. Therefore, in the remainder of this paper we will always use reduced homology, which has the effect of simply eliminating the infinite interval from the 0-dimensional barcode while leaving everything else unchanged. As a consequence, there will never be any infinite intervals in the persistent homology of a Vietoris–Rips simplicial complex, even in homological dimension zero.

Remark 1. It is well-known (see for example [65]) and easy to verify that for any finite metric space X, the lengths of the 0-dimensional (reduced) persistent homology intervals of the Vietoris–Rips complex of X correspond exactly to the lengths of the edges in a minimal spanning tree with vertex set X.

4. Definition of the persistent homology fractal dimension for measures

Let X be a metric space equipped with a probability measure μ, and let $X_n \subseteq X$ be a random sample of n points from X distributed independently and identically according to μ. Build a filtered simplicial complex K on top of vertex set X_n, for example a Vietoris–Rips complex $\mathrm{VR}(X; r)$ (Definition 3.5), an intrinsic Čech complex $\check{\mathrm{C}}(X, X; r)$, or an ambient Čech complex $\check{\mathrm{C}}(X, \mathbb{R}^m; r)$ if X is a subset of $\mathbb{R}^m$ [18]. Recall that the

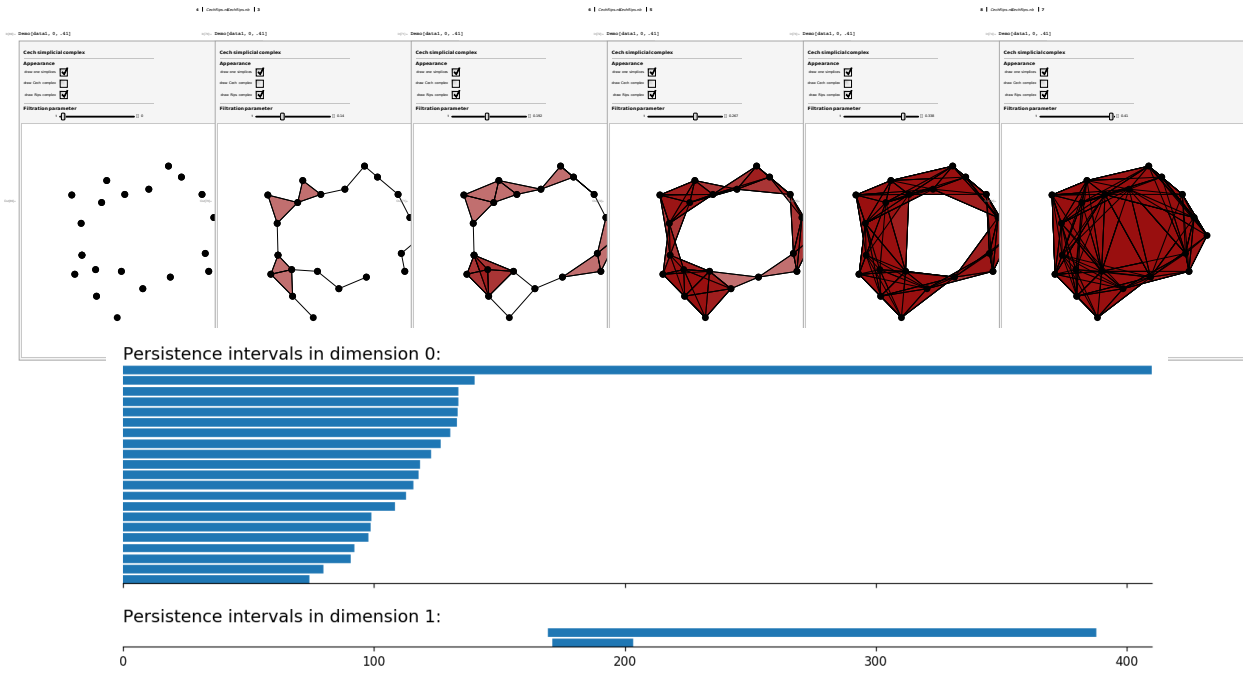

FIGURE 2. An example of Vietoris–Rips complexes at increasing scales, along with associated persistent homology intervals. The 0-dimensional persistent homology intervals shows how 21 connected components merge into a single connected component as the scale increases. The 1-dimensional persistent homology intervals show two 1-dimensional holes, one short-lived and the other long-lived.

i-dimensional persistent homology of this filtered simplicial complex, which decomposes as a direct sum of interval summands, is denoted by $\mathrm{PH}^i(X_n)$. We let $L^i(X_n)$ be the sum of the lengths of the intervals in $\mathrm{PH}^i(X_n)$. In the case of homological dimension zero, the sum $L^0(X_n)$ is simply the sum of all the edge lengths in a minimal spanning tree with X_n as its vertex set (since we are using reduced homology).

Definition 4.1 (Persistent homology fractal dimension)**.** Let X be a metric space equipped with a probability measure μ, let $X_n \subseteq X$ be a random sample of n points from X distributed according to μ, and let $L^i(X_n)$ be the sum of the lengths of the intervals in the i-dimensional persistent homology for X_n. We define the *i-dimensional persistent homology fractal dimension of μ* to be

$$\dim^i_{\mathrm{PH}}(\mu) = \inf_{d>0} \left\{ d \;\middle|\; \exists \text{ constant } C(i,\mu,d) \text{ such that } L^i(X_n) \le C n^{(d-1)/d} \text{ with probability one as } n \to \infty \right\}.$$

The constant C can depend on i, μ, and d. Here "$L^i(X_n) \le Cn^{(d-1)/d}$ with probability one as $n \to \infty$" means that we have $\lim_{n\to\infty} \mathbb{P}[L^i(X_n) \le Cn^{(d-1)/d}] = 1$. This dimension may depend on the choices of filtered simplicial complex (say Vietoris–Rips or Čech), and on the choice of field coefficients for homology computations; for now those choices are suppressed from the definition.

A measure μ on $X \subseteq \mathbb{R}^m$ is *nonsingular* if the absolutely continuous part of μ has positive mass.

Proposition 4.2. *Let μ be a measure on $X \subseteq \mathbb{R}^m$ with $m \ge 2$. Then $\dim^0_{\mathrm{PH}}(\mu) \le m$, with equality if μ is nonsingular.*

Proof. By Theorem 2 of [70], we have that $\lim_{n\to\infty} n^{-(m-1)/m} L^0(X_n) = c \int_{\mathbb{R}^m} f(x)^{(m-1)/m}\, dx$, where c is a constant depending on m, and where f is the absolutely continuous part of μ. To see that $\dim^0_{\mathrm{PH}}(\mu) \le m$, note that

$$L^0(X_n) \le \left(c \int_{\mathbb{R}^m} f(x)^{(m-1)/m}\, dx + \varepsilon \right) n^{(m-1)/m}$$

with probability one as $n \to \infty$ for any $\varepsilon > 0$. □

We conjecture that the i-dimensional persistent homology of compact subsets of $\mathbb{R}^m$ have the same scaling properties as the functionals in [70, 79].

Conjecture 4.3. *Let μ be a probability measure on a compact set $X \subseteq \mathbb{R}^m$ with $m \geq 2$, and let μ be nonsingular. Then for all $0 \leq i < m$, there is a constant $C \geq 0$ (depending on μ, m, and i) such that $L^i(X_n) = Cn^{(m-1)/m}$ with probability one as $n \to \infty$.*

Let μ be a probability measure with compact support that is absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^m$ for $m \geq 2$. Note that Conjecture 4.3 would imply that the persistent homology fractal dimension of μ is equal to m. The tools of subadditivity and superadditivity behind the umbrella theorems for Euclidean functionals, as described in [79] and Section 2.2, may be helpful towards proving this conjecture. In some cases, for example when X is a cube or ball (or more generally convex), then versions of Conjecture 4.3 are proven in [27, 69].

One could alternatively define birth-time (for $i > 0$) or death-time fractal dimensions by replacing $L^i(X_n)$ with the sum of the birth times, or alternatively the sum of the death times, in the persistent homology barcodes $\mathrm{PH}^i(X_n)$.

5. Experiments

A feature of Definition 4.1 is that we can use it to estimate the persistent homology fractal dimension of a measure μ. Indeed, suppose we can sample from X according to the probability distribution μ. We can therefore sample collections of points X_n of size n, compute the statistic $L^i(X_n)$, and then plot the results in a log-log fashion as n increases. In the limit as n goes to infinity, we expect the plotted points to be well-modeled by a line of slope $\frac{d-1}{d}$, where d is the i-dimensional persistent homology fractal dimension of μ. In many of the experiments in this section, the measures μ are simple enough (or self-similar enough) that we would expect the persistent homology fractal dimension of μ to be equal to the Hausdorff dimension of μ.

In our computational experiments, we have used the persistent homology software packages Ripser [9], Javaplex [75], and code from Duke (see the acknowledgements in Section 10). For the case of 0-dimensional homology, we can alternatively use well-known algorithms for computing minimal spanning trees, such as Kruskal's algorithm or Prim's algorithm [48, 62]. We estimate the slope of our log-log plots (of $L^i(X_n)$ as a function of n) using both a line of best fit, and alternatively a technique designed to approximate the asymptotic scaling described in Section 8. Our code is publicly available at `https://github.com/CSU-PHdimension/PHdimension`.

5.1. Estimates of persistent homology fractal dimensions.

We display several experimental results, for shapes of both integral and non-integral fractal dimension. In Figure 3, we show the log-log plots of $L^i(X_n)$ as a function of n, where X_n is sampled uniformly at random from a disk, a square, and an equilateral triangle, each of unit area in the plane $\mathbb{R}^2$. Each of these spaces constitutes a manifold of dimension two, and we thus expect these shapes to have persistent homology fractal dimension $d = 2$ as well. Experimentally, this appears to be the case, both for homological dimensions $i = 0$ and $i = 1$. Indeed, our asymptotically estimated slopes lie in the range 0.49 to 0.54, which is fairly close to the expected slope of $\frac{d-1}{d} = \frac{1}{2}$.

In Figure 4 we perform a similar experiment for the cube in $\mathbb{R}^3$ of unit volume. We expect the cube to have persistent homology fractal dimension $d = 3$, corresponding to a slope in the log-log plot of $\frac{d-1}{d} = \frac{2}{3}$. This appears to be the case for homological dimension $i = 0$, where the slope is approximately 0.65. However, for $i = 1$ and $i = 2$, our estimated slope is far from $\frac{2}{3}$, perhaps because our computational limits do not allow us to take n, the number of randomly chosen points, to be sufficiently large.

In Figure 5 we use log-log plots to estimate some persistent homology fractal dimensions of the Cantor set cross the interval (expected dimension $d = 1 + \log_3(2)$), of the Sierpiński triangle (expected dimension $d = \log_2(3)$), of Cantor dust in $\mathbb{R}^2$ (expected dimension $d = \log_3(4)$), and of Cantor dust in $\mathbb{R}^3$ (expected dimension $d = \log_3(8)$). As noted in Section 3, various notions of fractal dimension tend to agree for well-behaved fractals. Thus, in each case above, we provide the Hausdorff dimension d in order to define an expected persistent homology fractal dimension. The Hausdorff dimension is well-known for the Sierpiński triangle, Cantor dust in $\mathbb{R}^2$, and Cantor dust in $\mathbb{R}^3$. The Hausdorff dimension for the Cantor set cross the interval can be shown to be $1 + \log_3(2)$, which follows from [32, Theorem 9.3] or [54, Theorem III]. In Section 5.2 we define these fractal shapes in detail, and we also explain our computational technique for sampling points from them at random.

Summarizing the experimental results for self-similar fractals, we find reasonably good estimates of fractal dimension for homological dimension $i = 0$. More specifically, for the Cantor set cross the interval, we expect $\frac{d-1}{d} \approx 0.3869$, and we find slope estimates from a linear fit of all data and an asymptotic fit to be 0.3799 and 0.36488, respectively. In the case of the Sierpiński triangle, the estimate is quite good: we expect $\frac{d-1}{d} \approx 0.3691$, and the slope estimates from both a linear fit and an asymptotic fit are approximately 0.37. Similarly, the estimates for Cantor dust in $\mathbb{R}^2$ and $\mathbb{R}^3$ are close to the expected values: (1) For Cantor dust in $\mathbb{R}^2$, we expect $\frac{d-1}{d} \approx 0.2075$ and estimate $\frac{d-1}{d} \approx 0.25$. (2) For Cantor dust in $\mathbb{R}^3$, we expect $\frac{d-1}{d} \approx 0.4717$ and estimate $\frac{d-1}{d} \approx 0.49$. For $i > 0$ many of these estimates of the persistent homology fractal dimension are not close to the expected (Hausdorff) dimensions, perhaps because the number of points n is not large enough. The theory behind these experiments has now been verified in [68].

It is worth commenting on the Cantor set, which is a self-similar fractal in $\mathbb{R}$. Even though the Hausdorff dimension of the Cantor set is $\log_3(2)$, it is not hard to see that the 0-dimensional persistent homology fractal dimension of the Cantor set is 1. This is because as $n \to \infty$ a random sample of points from the Cantor set will contain points in $\mathbb{R}$ arbitrarily close to 0 and to 1, and hence $L_0(X_n) \to 1$ as $n \to \infty$. This is not surprising—we do not necessarily expect to be able to detect a fractional dimension less than one by using minimal spanning trees (which are 1-dimensional graphs). For this reason, if a measure μ is defined on a subset of $\mathbb{R}^m$, we sometimes restrict attention to the case $m \geq 2$. See Figure 6 for our experimental computations on the Cantor set.

Finally, we include one example with data drawn from a two-dimensional manifold in $\mathbb{R}^3$. We sample points from a torus with major radius 5 and minor radius 3. We expect the persistent homology fractal dimensions to be 2, and this is supported in the experimental evidence for 0-dimensional homology shown in Figure 7 with approximate slope $\frac{d-1}{d} = \frac{1}{2}$.

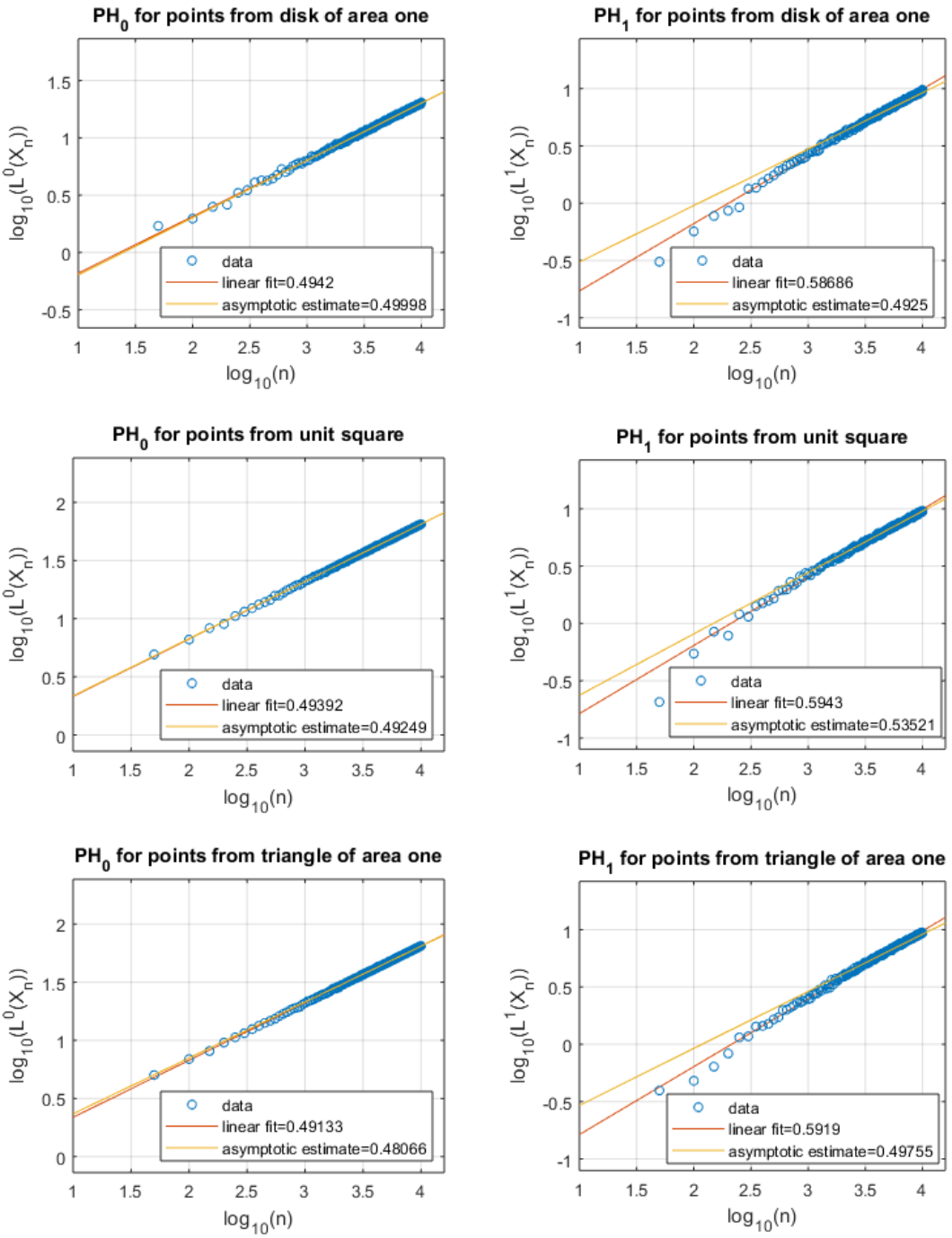

FIGURE 3. Log scale plots and slope estimates of the number n of sampled points versus $L^0(X_n)$ (left) or $L^1(X_n)$ (right). Subsets X_n are drawn uniformly at random from (top) the unit disk in $\mathbb{R}^2$, (middle) the unit square, and (bottom) the unit triangle. All cases have slope estimates close to $\frac{d-1}{d} = \frac{1}{2}$, which is consistent with the expected dimension. The asymptotic scaling estimates of the slope are computed as described in Section 8.

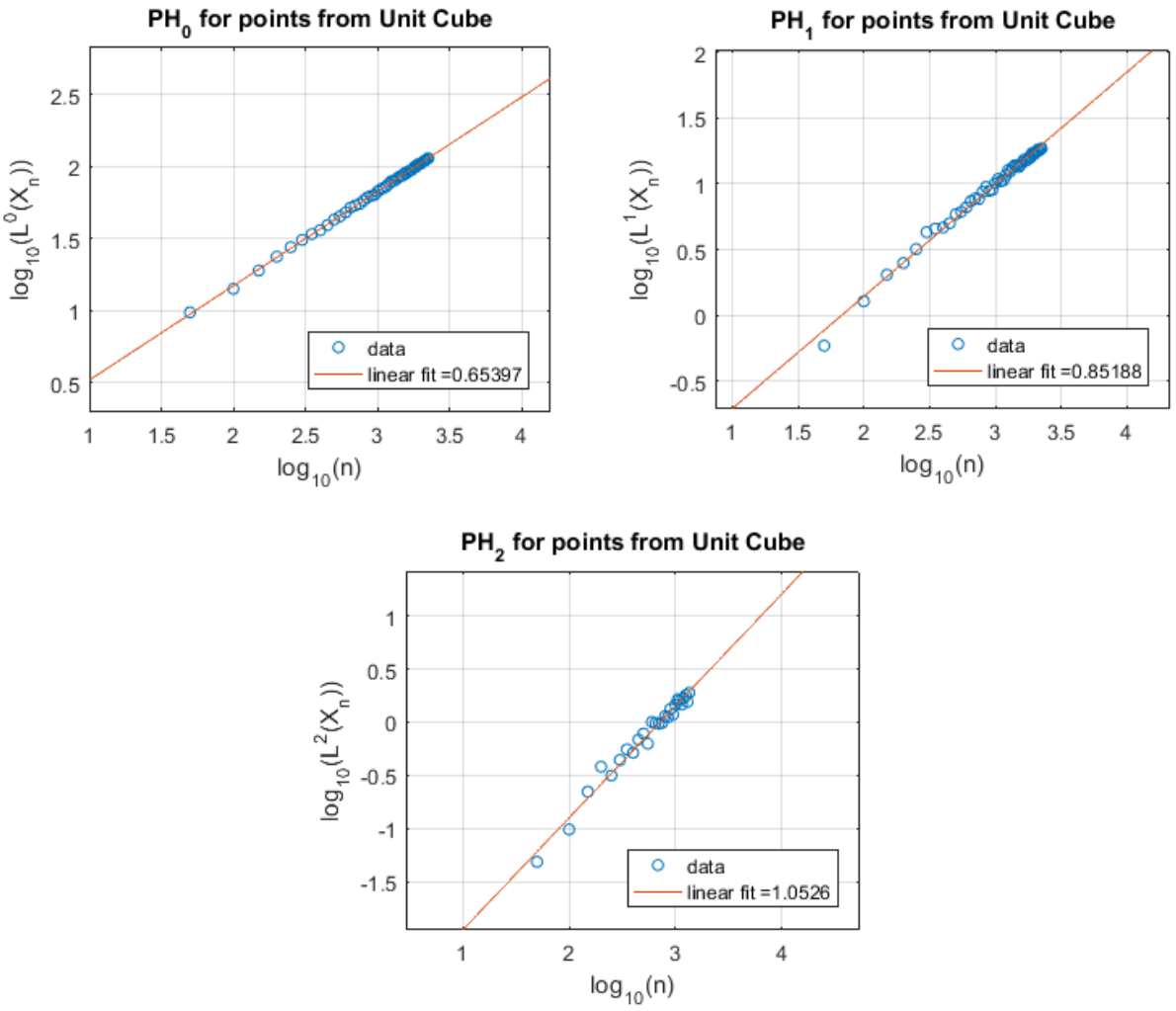

FIGURE 4. Log scale plots of the number n of sampled points from the cube versus $L^0(X_n)$ (left), $L^1(X_n)$ (right), and $L^2(X_n)$ (bottom). The dimension estimate from 0-dimensional persistent homology is reasonably close to the expected slope $\frac{d-1}{d} = \frac{2}{3}$, while the 1- and 2-dimensional cases are less accurate, likely due to computational limitations.

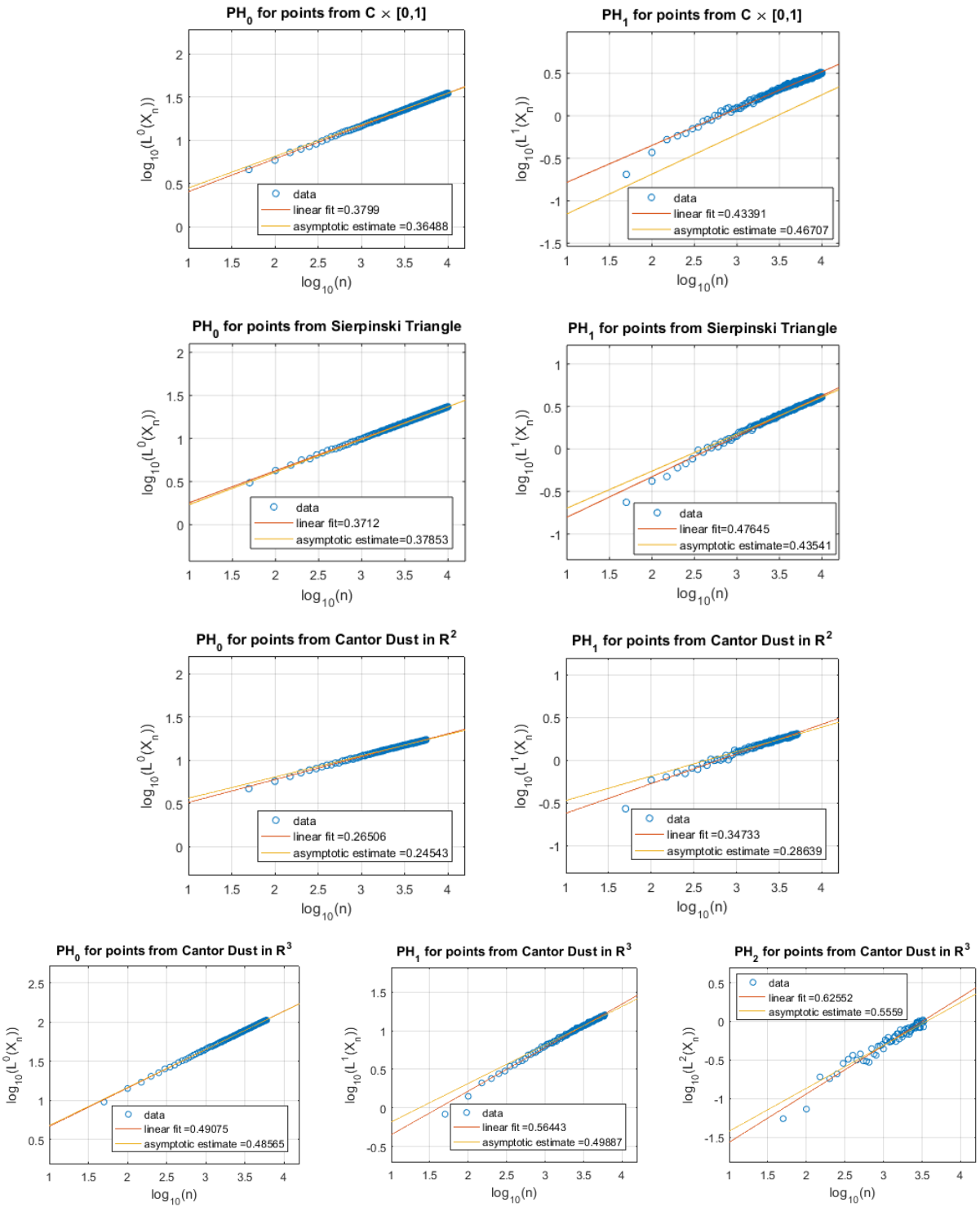

FIGURE 5. (Top row) Cantor set cross unit interval; expected slope $\frac{d-1}{d} \approx 0.3869$. (Second row) Sierpiński triangle; expected slope $\frac{d-1}{d} \approx 0.3691$. (Third row) Cantor dust in $\mathbb{R}^2$; expected slope $\frac{d-1}{d} \approx 0.2075$. (Bottom row) Cantor dust in $\mathbb{R}^3$; expected slope $\frac{d-1}{d} \approx 0.4717$. The 0-dimensional estimates are close to the expected dimensions. The higher-dimensional estimates are not as accurate, perhaps due to computational limitations.

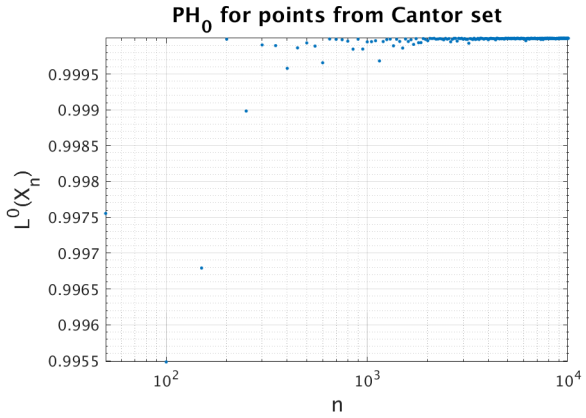

FIGURE 6. Log scale plot of the number n of sampled points from the Cantor set versus $L^0(X_n)$. Note that $L^0(X_n)$ approaches one, as expected.

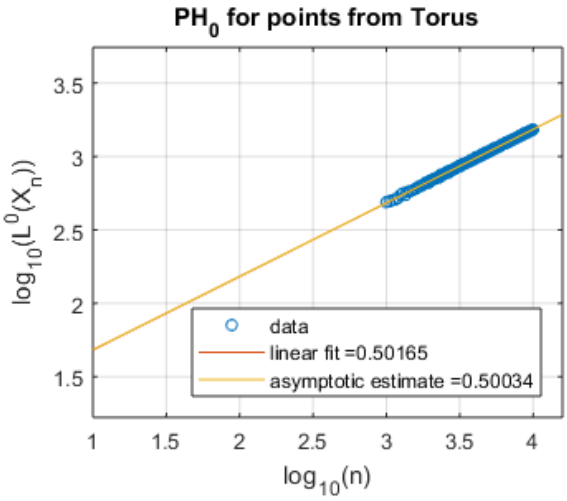

FIGURE 7. Log scale plot of the number n of sampled points from a torus with major radius 5 and minor radius 3 versus $L^0(X_n)$. Estimated lines of best fit from $L^0(X_n)$ have slope approximately equal to $\frac{1}{2}$, recovering $\frac{d-1}{d}$ for a dimension estimate of $d = 2$. We restrict to 0-dimensional homology in this setting due to computational limitations.

5.2. **Randomly sampling from self-similar fractals.** The Cantor set $C = \cap_{l=0}^{\infty} C_l$ is a countable intersection of nested sets $C_0 \supseteq C_1 \supseteq C_2 \supseteq \cdots$, where the set C_l at level l is a union of 2^l closed intervals, each of length $\frac{1}{3^l}$. More precisely, $C_0 = [0,1]$ is the closed unit interval, and C_l is defined recursively via

$$C_l = \frac{C_{l-1}}{3} \cup \left(\frac{2}{3} + \frac{C_{l-1}}{3} \right) \quad \text{for } l \geq 1.$$

In our experiment for the Cantor set (Figure 6), we do not sample from the Cantor distribution on the entire Cantor set C, but instead from the left endpoints of level C_l of the Cantor set, where l is chosen to be very large (we use $l = 100{,}000$). More precisely, in order to sample points, we choose a binary sequence $\{a_i\}_{i=1}^{l}$ uniformly at random, meaning that each term a_i is equal to either 0 or 1 with probability $\frac{1}{2}$, and furthermore the value a_i is independent from the value of a_j for $i \neq j$. The corresponding random point in the Cantor set is $\sum_{i=1}^{l} \frac{2a_i}{3^i}$. Note that this point is in C and furthermore is the left endpoint of some interval in C_l. So we are selecting left endpoints of intervals in C_l uniformly at random, but since l is large this is a good approximation to sampling from the entire Cantor set according to the Cantor distribution.

We use a similar procedure to sample at random for our experiments on the Cantor set cross the interval, on Cantor dust in $\mathbb{R}^2$, on Cantor dust in $\mathbb{R}^3$, and on the Sierpiński triangle (Figure 5). The Cantor set cross the interval is $C \times [0,1] \subseteq \mathbb{R}^2$, equipped with the Euclidean metric. We computationally sample by choosing a point from C_l as described in the paragraph above for $l = 100{,}000$, and by also sampling a point from the unit interval $[0,1]$ uniformly at random. Cantor dust is the subset $C \times C$ of $\mathbb{R}^2$, which we sample by choosing two points from C_l as described previously. The same procedure is done for the Cantor dust $C \times C \times C$ in $\mathbb{R}^3$. The Sierpiński triangle $S \subseteq \mathbb{R}^2$ is defined in a similar way to the Cantor set, with $S = \cap_{l=0}^{\infty} S_l$ a countable intersection of nested sets $S_0 \supseteq S_1 \supseteq S_2 \supseteq \cdots$. Here each S_l is a union of 3^l triangles. We choose $l = 100{,}000$ to be large, and then sample points uniformly at random from the bottom left endpoints of the triangles in S_l. More precisely, we choose a ternary sequence $\{a_i\}_{i=1}^{l}$ uniformly at random, meaning that each term a_i is equal to either 0, 1, or 2 with probability $\frac{1}{3}$. The corresponding random point in the Sierpiński triangle is $\sum_{i=1}^{l} \frac{1}{2^i} \vec{v}_i \in \mathbb{R}^2$, where vector $\vec{v}_i$ is given by

$$\vec{v}_i = \begin{cases} (0,0)^T & \text{if } a_i = 0 \\ (1,0)^T & \text{if } a_i = 1 \\ (\frac{1}{2}, \frac{\sqrt{3}}{2})^T & \text{if } a_i = 2. \end{cases}$$

Note this point is in S and furthermore is the bottom left endpoint of some triangle in S_l.

6. Limiting distributions

To some metric measure spaces, (X, μ), we are able to assign a finer invariant that contains more information than just the persistent homology fractal dimension. Consider the set of the lengths of all intervals in $\mathrm{PH}^i(X_n)$, for each homological dimension i. Experiments suggest that for some $X \subseteq \mathbb{R}^m$, the scaled set of interval lengths in each homological dimension converges point-wise to some fixed probability distribution which depends on μ and on i.

More precisely, for a fixed probability measure μ, let $\hat{F}_n^{(i)}$ be the empirical cumulative distribution function of the i-dimensional persistent homology interval lengths in $\mathrm{PH}^i(X_n)$, where X_n is a fixed sample of n points from X drawn in an i.i.d. fashion according to μ. If μ is absolutely continuous with respect to the Lebesgue measure on some compact set, then the function $\hat{F}_n^{(i)}(t)$ converges point-wise to the Heaviside step function as $n \to \infty$, since the fraction of interval lengths less than any fixed $\varepsilon > 0$ is converging to one as $n \to \infty$. More interestingly, for μ a sufficiently nice measure on $X \subseteq \mathbb{R}^m$, the rescaled empirical cumulative distribution function $\hat{F}_n^{(i)}(n^{-1/m}t)$ may converge to a non-constant curve. A back-of-the-envelope motivation for this rescaling is that if $L^i(X_n) = Cn^{(m-1)/m}$ with probability one as $n \to \infty$ (Conjecture 4.3), then the average length of a persistent homology interval length is

$$\frac{L^i(X_n)}{\#\text{ intervals}} = \frac{Cn^{(m-1)/m}}{\#\text{ intervals}},$$

which is proportional to $n^{-1/m}$ if the number of intervals is proportional to n. We make this precise in the following conjectures.

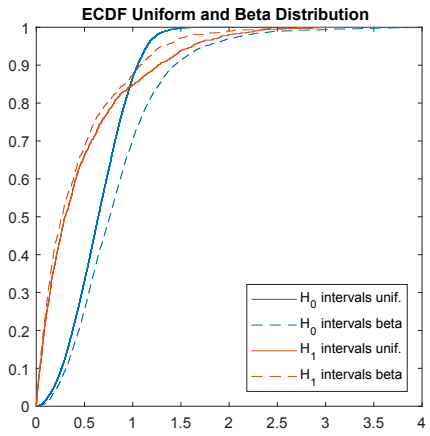

FIGURE 8. Empirical CDF's for the H_0 and H_1 interval lengths computed from 10,000 points sampled from the unit square according to the uniform distribution and beta distribution with shape and size parameter both set to 2. The limiting distributions appear to be different.

Conjecture 6.1. *Let μ be a probability measure on a compact set $X \subseteq \mathbb{R}^m$, and let μ be absolutely continuous with respect to the Lebesgue measure. Then the limiting distribution $\hat{F}^{(i)}(t) = \lim_{n\to\infty} \hat{F}_n^{(i)}(n^{-1/m}t)$, which depends on μ and i, exists.*

In Section 6.1 we show that Conjecture 6.1 holds when μ is the uniform distribution on an interval, and in Section 6.2 we perform experiments in higher dimensions.

Question 1. Assuming Conjecture 6.1 is true, what is the limiting rescaled distribution when μ is the uniform distribution on an m-dimensional ball, or alternatively an m-dimensional cube?

Conjecture 6.2. *Let the compact set $X \subseteq \mathbb{R}^m$ have positive Lebesgue measure, and let μ be the corresponding probability measure (i.e., μ is the restriction of the Lebesgue measure to X, rescaled to have mass one). Then the limiting distribution $\hat{F}^{(i)}(t) = \lim_{n\to\infty} \hat{F}_n^{(i)}(n^{-1/m}t)$ exists and depends only on m, i, and the volume of X.*

Question 2. Assuming Conjecture 6.2 is true, what is the limiting rescaled distribution when X has unit volume?

Remark 2. Conjecture 6.2 is false if μ is not a uniform measure (i.e. a rescaled Lebesgue measure). Indeed, the uniform measure on a square (experimentally) has a different limiting rescaled distribution than a (nonconstant) beta distribution on the same unit square, as seen in Figure 8.

Remark 3. Conjecture 6.2 is related to [19], and in the case of stationary point processes, to [41, Theorem 1.11] and [55].

6.1. **The uniform distribution on the interval.** In the case where μ is the uniform distribution on the unit interval $[0, 1]$, then a weaker version of Conjecture 6.1 (convergence distribution-wise) is known to be true, and furthermore a formula for the limiting rescaled distribution is known. If X_n is a subset of $[0, 1]$ drawn uniformly at random, then (with probability one) the points in X_n divide $[0, 1]$ into $n+1$ pieces. The joint probability distribution function for the lengths of these pieces is given by the flat Dirichlet distribution, which can be thought of as the uniform distribution on the n-simplex (the set of all $(t_0, \ldots, t_n)$ with $t_i \geq 0$ for all i, such that $\sum_{i=0}^{n} t_i = 1$). Note that the intervals in $\mathrm{PH}^0(X_n)$ have lengths $t_1, \ldots, t_{n-1}$, omitting t_0 and t_n which correspond to the two subintervals on the boundary of the interval.

The probability distribution function of each t_i, and therefore of each interval length in $\mathrm{PH}^0(X_n)$, is the marginal of the Dirichlet distribution, which is given by the Beta distribution $B(1, n)$ [11]. After simplifying,

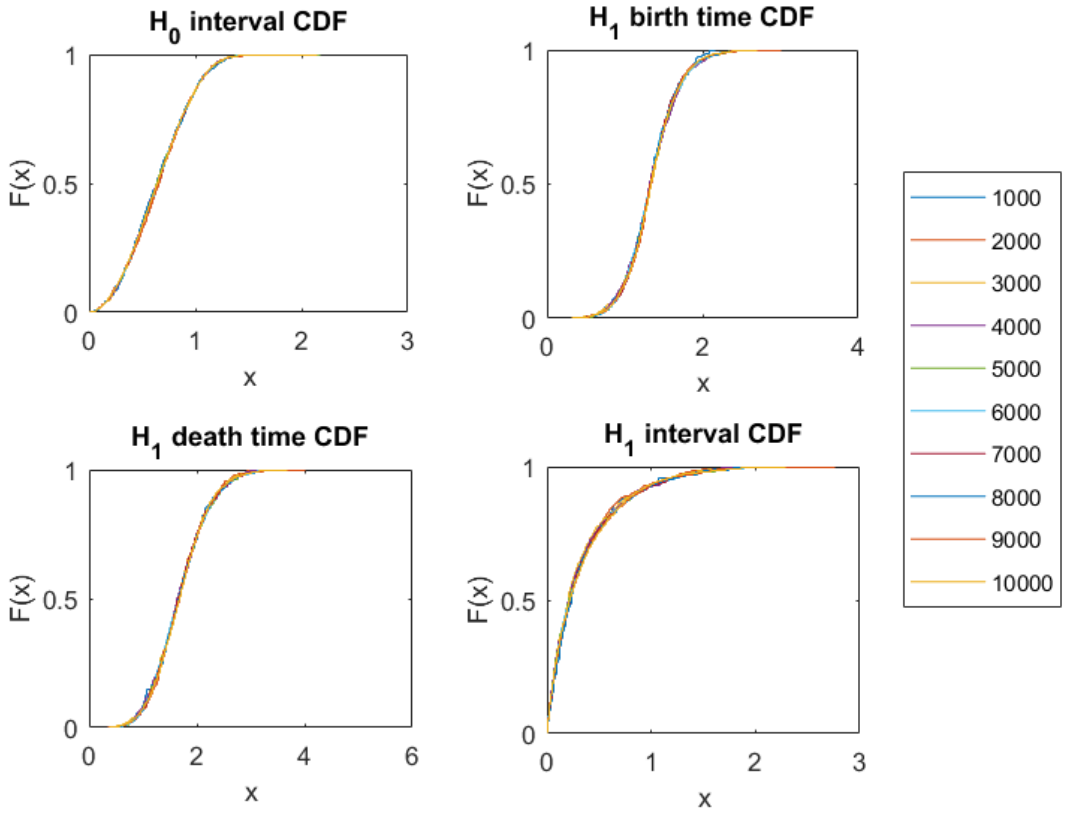

FIGURE 9. Empirical CDF's for H_0 interval lengths, H_1 birth times, H_1 death times, and H_1 interval lengths computed from an increasing number of n points drawn uniformly from the 2-dimensional unit square, and rescaled by $n^{1/2}$. It is plausible that both $F_n^{(0)}(n^{-1/2}t)$ and $F_n^{(1)}(n^{-1/2}t)$ converge point-wise to a limiting probability distribution.

the true cumulative distribution function (which we denote by $F_n^{(0)}$ instead of the empirical cumulative distribution function $\hat{F}_n^{(0)}$) of $B(1,n)$ is given by [66]

$$F_n^{(0)}(t) = \frac{B(t;1,n)}{B(1,n)} = \frac{\int_0^t s^0(1-s)^{n-1}ds}{\frac{\Gamma(1)\Gamma(n)}{\Gamma(n+1)}} = 1-(1-t)^n.$$

As n goes to infinity, $F_n^{(0)}(t)$ converges pointwise to the constant function 1. However, after rescaling, $F_n^{(0)}(n^{-1}t)$ converges to a more interesting distribution independent of n. Indeed, we have $F_n^{(0)}\left(\frac{t}{n}\right) = 1-(1-\frac{t}{n})^n$, and the limit as $n \to \infty$ is

$$\lim_{n\to\infty} F_n^{(0)}\left(\tfrac{t}{n}\right) = 1-e^{-t}.$$

This is the cumulative distribution function of the exponential distribution with rate parameter one. Therefore, the rescaled interval lengths in the limit as $n \to \infty$ are distributed according to the exponential distribution Exp(1).

6.2. **Experimental evidence for Conjecture 6.1 in $\mathbb{R}^2$.** We now move to the case where μ is the uniform distribution on the unit square in $\mathbb{R}^2$. It is known that the sum of the edge lengths of the minimal spanning tree, given by $L^0(X_n)$ where X_n is a random sample of n points from the unit square, converges as $n \to \infty$ to $Cn^{1/2}$, for a constant C [70]. However, to our knowledge the limiting distribution of all (rescaled) edge lengths

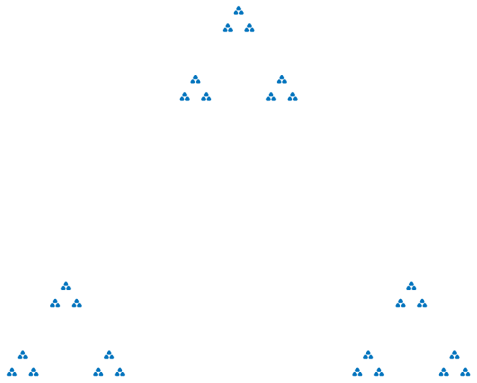

FIGURE 10. Plot of 20,000 points sampled at random from the Sierpiński triangle of separation $\delta = 2$.

is not known. We instead analyze this example empirically. The experiments in Figure 9 suggest that as n increases, it is plausible that both $F_n^{(0)}(n^{-1/2}t)$ and $F_n^{(1)}(n^{-1/2}t)$ converge point-wise to a limiting probability distribution. We have tried to fit these limiting probability distributions to standard distributions, without yet having found obvious candidates.

6.3. **Examples where a limiting distribution does not exist.** In this section we give experimental evidence that the assumption of being a rescaled Lebesgue measure in Conjecture 6.1 is necessary. Our example computation is done on a separated Sierpiński triangle.

For a given separation value $\delta \geq 0$, the *separated Sierpiński triangle* can be defined as the set of all points in $\mathbb{R}^2$ of the form $\sum_{i=1}^{\infty} \frac{1}{(2+\delta)^i}\vec{v}_i$, where each vector $\vec{v}_i \in \mathbb{R}^2$ is either $(0,0)$, $(1,0)$, or $(\frac{1}{2}, \frac{\sqrt{3}}{2})$. The Hausdorff dimension of this self-similar fractal shape is $\log_{2+\delta}(3)$ ([32, Theorem 9.3] or [54, Theorem III]), and note that when $\delta = 0$, we recover the standard (non-separated) Sierpiński triangle. See Figure 10 for a picture when $\delta = 2$. Computationally, when we sample a point from the separated Sierpiński triangle, we sample a point of the form $\sum_{i=1}^{l} \frac{1}{(2+\delta)^i}\vec{v}_i$, where in our experiments we use $l = 100{,}000$.

In the following experiment we sample random points from the separated Sierpiński triangle with $\delta = 2$. As the number of random points n goes to infinity, it appears that the rescaled[3] CDF of H_0 interval lengths are not converging to a fixed probability distribution, but instead to a periodic family of distributions, in the following sense. If you fix $k \in \mathbb{N}$ then the distributions on $n = k, 3k, 9k, 27k, \ldots, 3^j k, \ldots$ points appear to converge as $j \to \infty$ to a fixed distribution. Indeed, see Figure 11 for the limiting distribution on 3^j points, and for the limiting distribution on $3^j \cdot 2$ points. However, the limiting distribution for $3^j k$ points and the limiting distribution for $3^j k'$ points appear to be the same if and only if k and k' differ by a power of 3. See Figure 12, which shows four snapshots from one full periodic orbit.

Here is an intuitively plausible explanation for why the rescaled CDFs for the separated Sierpiński triangle converge to a periodic family of distributions, rather than a fixed distribution: Imagine focusing a camera at the origin of the Sierpiński triangle and zooming in. Once you get to $(2+\delta)\times$ magnification, you see the same image again. This is one full period. However, for magnifications between $1\times$ and $(2+\delta)\times$ you see a different image. In our experiments sampling random points, zooming in by a factor of $(2+\delta)\times$ is the same thing as sampling three times as many points (indeed, the Hausdorff dimension is $\log_{2+\delta}(3)$). When zooming in you see the same image only when the magnification is at a multiple of $2+\delta$, and analogously

[3]Since the separated Sierpiński triangle has Hausdorff dimension $\log_{2+\delta}(3)$, the rescaled distributions we plot are $F_n^{(0)}(n^{-1/m}t)$ with $m = \log_{2+\delta}(3)$.

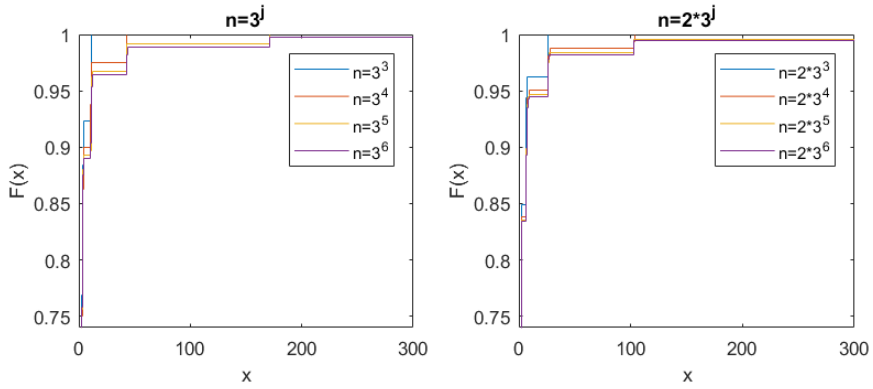

FIGURE 11. This figure shows the empirical rescaled CDFs of H_0 interval lengths for $n = 3^j$ points (left) and for $n = 3^j \cdot 2$ points (right) sampled from the separated Sierpiński triangle with $\delta = 2$. Each figure appears to converge to a fixed limiting distribution as $j \to \infty$, but the two limiting distributions are not equal.

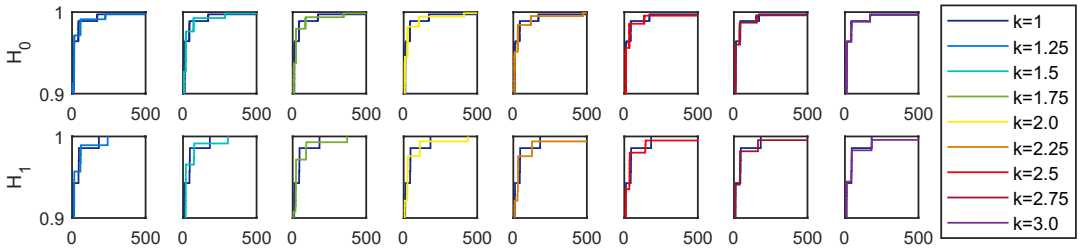

FIGURE 12. Empirical rescaled CDF's for H_0 interval lengths, and H_1 interval lengths computed from an increasing number of $n = k \cdot 3^6$ points from the separated Sierpiński triangle with $\delta = 2$, moving left to right. Note that as k increases between adjacent powers of three, the "bumps" in the distribution shift to the right, until the starting distribution reappears.

when sampling random points perhaps we should expect to see the same probability distribution of interval lengths only when the number of points is multiplied by a power of 3.

7. Another way to randomly sample from the Sierpiński triangle

An alternate approach to constructing a sequence of measures converging to the Sierpiński triangle is using a particular Lindenmayer system, which generates a sequence of instructions in a recursive fashion [56, Figure 7.16]. Halting the recursion at any particular level l will give a (non-fractal) approximation to the Sierpiński triangle as a piecewise linear curve with a finite number of segments; see Figure 13.

Let μ_l be the uniform measure on the piecewise linear curve at level l. In Figure 14 we sample n points from μ_l and compute $L^i(X_n)$, displayed in a log-log plot, for $i = 0$ and 1. Since each μ_l for l fixed is non-fractal (and 1-dimensional) in nature, the ultimate asymptotic behavior will be $d = 1$ once the number of points n is sufficiently large (depending on the level l). However, for level l sufficiently large (depending on the number of points n) we see that there is an intermediate regime in the log-log plots which scale with the expected fractal dimension near $\log_2(3)$. As pointed out by an anonymous reviewer, one could potentially prove that the scaling in the intermediate regime is indeed $\log_2(3)$, as follows. The 0 or 1-dimensional persistent homology of the entire Sierpiński curve at level l could likely be computed, for example using ideas similar to [51, Proposition 3.2]. Then, the difference between the persistent homology of the entire curve and a random sample X_n of n points could perhaps be controlled by using the stability of persistent homology [18]

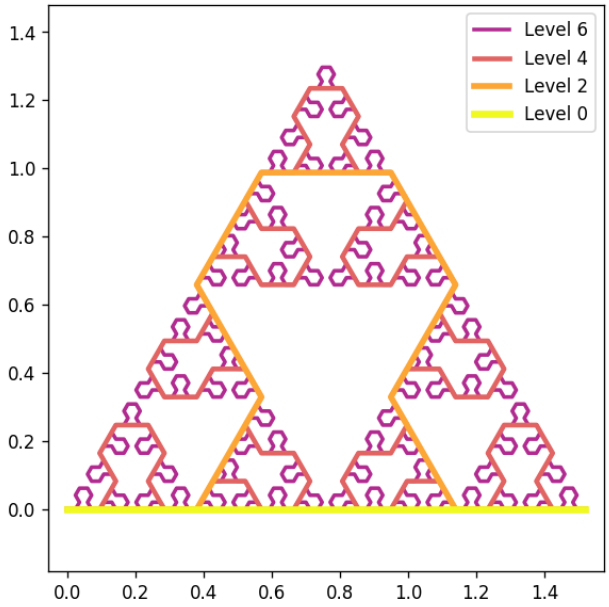

FIGURE 13. The Sierpiński triangle as the limit of a sequence of curves. We can uniformly randomly sample from the curve at level l to generate a sequence of measures μ_l converging to the Sierpiński triangle measure as $l \to \infty$.

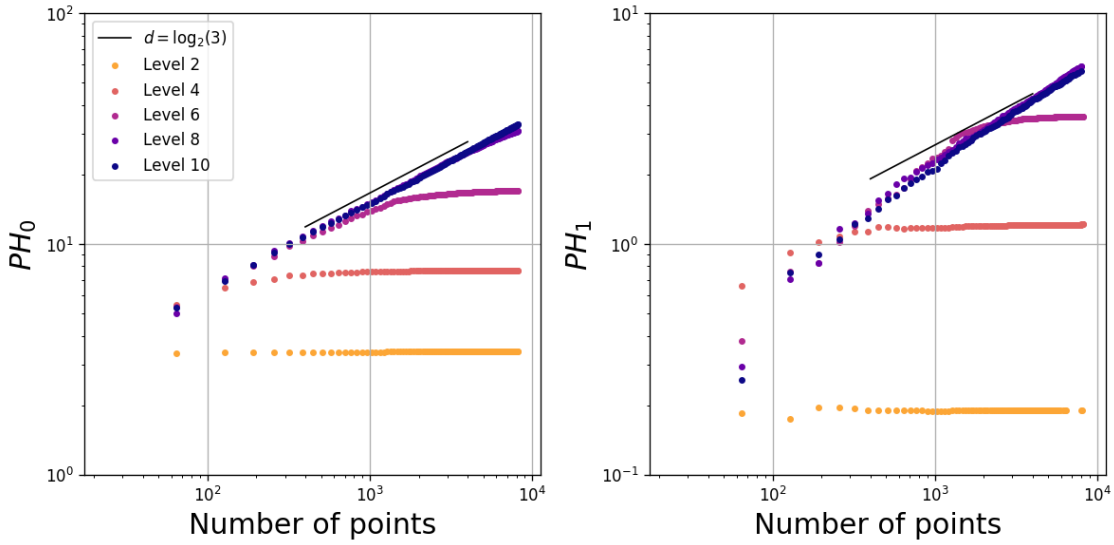

FIGURE 14. Scaling behaviors for various "depths" of the Sierpinski arrowhead curves visualized in Figure 13, in homological dimensions 0 and 1.

and ideas analogous to those in [68, Lemma 9 and Proposition 5], although rigorously controlling the effects of noise in all homological dimensions may not be easy. We expect a similar relationship between the number of points n and the level l to hold for many types of self-similar fractals.

We also give intuition why, for any fixed level l, the 0-dimensional persistent homology dimension of the curve μ_l is one. Note that μ_l consists of 3^l line segments (see Figure 13). Suppose X_N is a sample of N points from μ_l that is dense enough so that the minimal spanning tree with vertex set X_N consists exclusively of edges between two vertices that are either on the same line segment of μ_l or on adjacent line segments of μ_l.

If we then consider a nested sequence $X_N \subseteq X_{N+1} \subseteq X_{N+2} \subseteq \ldots$ of increasing finite subsets of μ_l, it follows that $L^0(X_n)$ for $n \geq N$ is a monotonically increasing sequence bounded above by the length of the curve μ_l. In this setting we have $L^0(X_n) \leq C$ where C is the length of μ_l; note that $C = Cn^{(d-1)/d}$ when $d = 1$.

8. Asymptotic approximation of the scaling exponent

From Definition 4.1 we consider how to estimate the exponent $(d-1)/d$ numerically for a given metric measure space (X, μ). For a fixed number of points n, a pair of values (n, ℓ_n) is produced, where $\ell_n = L^i(X_n)$ for a sampling X_n from (X, μ) of cardinality n. If the scaling holds asymptotically for n sampled past a sufficiently large point, then we can approximate the exponent by sampling for a range of n values and observing the rate of growth of ℓ_n. A common technique used to estimate power law behavior (see for example [20]) is to fit a linear function to the log-transformed data. The reason for doing this is a hypothesized asymptotic scaling $y \sim e^C x^\alpha$ as $x \to \infty$ becomes a linear function after taking the logarithm: $\log(y) \sim C + \alpha \log(x)$.

However, the expected power law in the data only holds asymptotically for $n \to \infty$. We observe in practice that the trend for small n is subdominant to its asymptotic scaling. Intuitively we would like to throw out the non-asymptotic portion of the sequence, but deciding where to threshold depends on the sequence. We propose the following approach to address this issue.

Suppose in general we have a countable set of measurements (n, ℓ_n), with n ranging over some subset of the positive integers. Create a sequence in monotone increasing order of n so that we have a $(n_k, \ell_{n_k})_{k=1}^\infty$ with $n_k > n_j$ for $k > j$. For any pairs of integers p, q with $1 \leq p < q$, we denote the log-transformed data of the corresponding terms in the sequence as

$$S_{pq} = \left\{ \left(\log(n_k), \log(\ell_{n_k}) \right) \mid p \leq k \leq q \right\} \subseteq \mathbb{R}^2.$$

Each finite collection of points S_{pq} has an associated pair of linear least-squares coefficients (C_{pq}, α_{pq}), where the line of best fit to the set S_{pq} is given by $y = C_{pq} + \alpha_{pq} x$. For our purposes we are more interested in the slope α_{pq} than the intercept C_{pq}. We expect that we can obtain the fractal dimension by considering the joint limits in p and q: if we define α as

$$\alpha = \lim_{p,q \to \infty} \alpha_{pq},$$

then we can recover the dimension by solving $\alpha = \frac{d-1}{d}$. A possibly overly restrictive assumption is that the asymptotic behavior of ℓ_{n_k} is monotone. If this is the case, we may expect *any* valid joint limit $p, q \to \infty$ will be defined and produce the same value. For example, setting $q = p + r$ we expect the following to hold:

$$\alpha = \lim_{p \to \infty} \lim_{r \to \infty} \alpha_{p,p+r}.$$

In general, the joint limit may exist under a wider variety of ways in which one allows q to grow relative to p.

Now define a function $A : \mathbb{R}^2 \to \mathbb{R}$, which takes on values $A(\frac{1}{p}, \frac{1}{q}) = \alpha_{pq}$, and define $A(0,0)$ so that A is continuous at the origin. Assuming $\alpha_{pq} \to \alpha$ as above, then any sequence $(x_k, y_k)_k \to (0,0)$ will produce the same limiting value $A(0,0)$ and the limit $\lim_{(x,y)\to(0,0)} A(x,y)$ is well-defined. This suggests an algorithm for finite data:

(1) Obtain a collection of estimates α_{pq} for various values of p, q, and then
(2) use the data $\{(\frac{1}{p}, \frac{1}{q}, A(\frac{1}{p}, \frac{1}{q}))\}$ to extrapolate an estimate for $A(0,0) = \alpha$, from which we can solve for the fractal dimension d.

For simplicity, we currently fix $q = n_{\max}$ and collect estimates varying only p; i.e., we only collect estimates of the form $\alpha_{p\, n_{\max}}$. In practice it is safest to use a low-order estimator to limit the risks of extrapolation. We use linear fit for the two-dimensional data $A(\frac{1}{p}, \frac{1}{q})$ to produce a linear approximation $\hat{A}(\xi, \eta) = a + b\xi + c\eta$, giving an approximation $\alpha = A(0,0) \approx \hat{A}(0,0) = a$.

Shown in Figure 15 is an example applied to the function

$$f(x) = \left(100x + \frac{1}{10} x^2 \right) (1 + 0.1\varepsilon(x)) \tag{2}$$

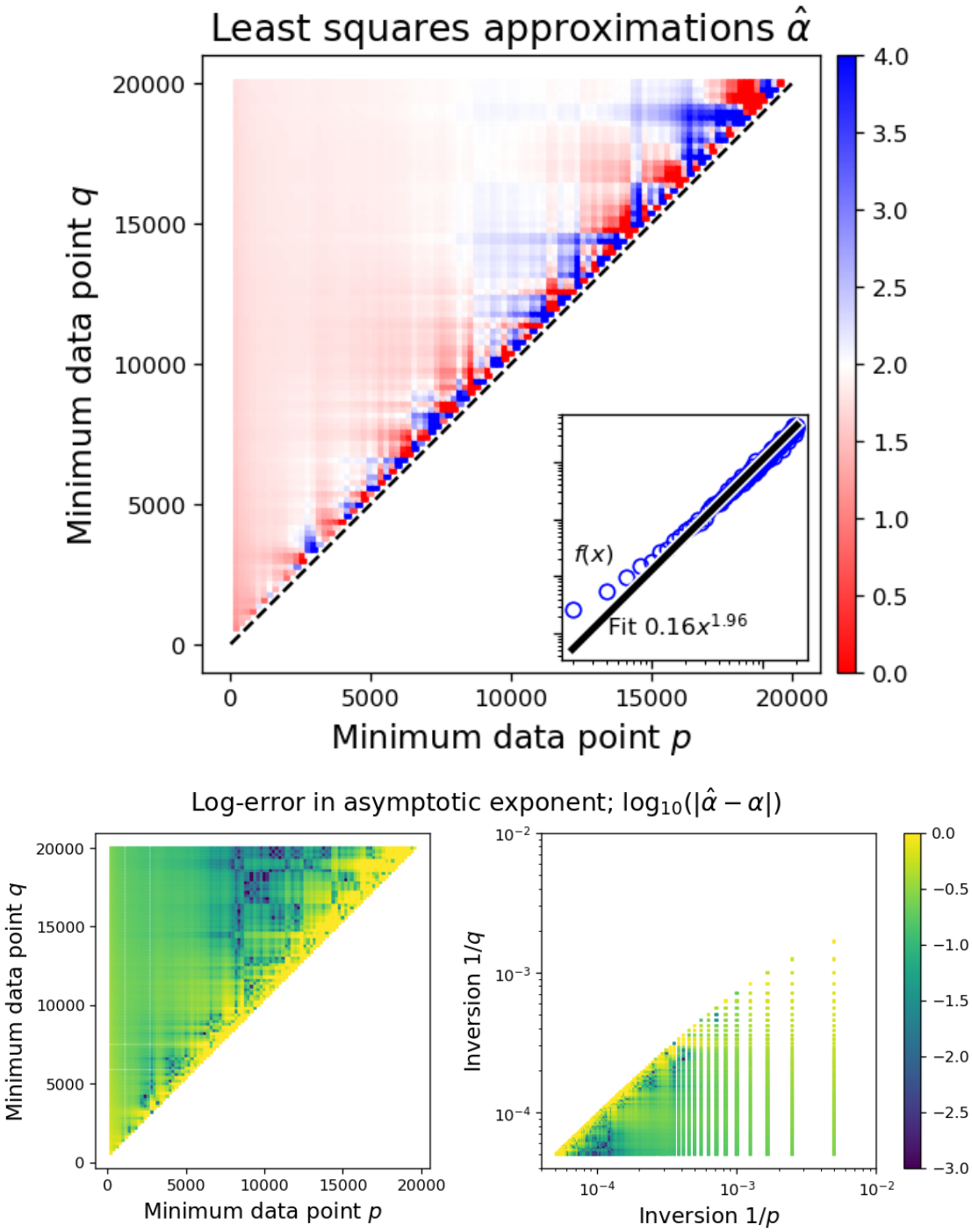

FIGURE 15. Top: Approximations α_{pq} for selections of (p, q) in the function $f(x)$ in (2). Top inset: sampling of $f(x)$ (blue circles) and the corresponding asymptotic best fit (black). (Bottom left) Log-absolute-error of the coefficients. Note that the approximation is generally poor for $|p - q|$ small, due to a small number of sample points. (Bottom right) Same values, with the coordinates mapped as $\xi = 1/p$, $\eta = 1/q$. The value to be extrapolated is at $(\xi, \eta) = (0, 0)$.

with $\varepsilon = dW(x)$, with $W(x)$ a sampling of standard Brownian noise, and x regularly sampled in $[400, 20000]$. The theoretical asymptotic is $\alpha = 2$ and should be attainable for sufficiently large x and enough sample points to overcome noise. Note that there is a balance needed to both keep a sufficient number of points to have a robust estimation (we want $q-p$ to be large) and to avoid including data in the pre-asymptotic regime (thus p must be relatively large). Visually, this is seen near the top side of the triangular region, where the error drops to roughly the order of 10^{-3}. The challenge for an arbitrary function is not knowing precisely where this balance is; see [20, Sections 1, 3.3-3.4] in the context of estimating $x_{\min}$ (in their language) for the tails of probability density functions.

It is important to note that the effects of noise and pre-asymptotic data in estimation of α can be non-negligible even for what are seemingly sufficiently large values of x. For example, we observe that even when removing noise ($\varepsilon(x) \to 0$) and performing a similar power fit on the restriction of the data to $x \in [19000, 20000]$ we obtain an estimated exponent $\hat{\alpha} \approx 1.9393$. Note the transition from first to second order behavior begins at $x = 10^3$, which is an order of magnitude earlier. Given this, we expect a rule of thumb recovering more than one significant digit reliably when performing random sampling requires sampling at least two orders of magnitude beyond when a transition in power law behavior occurs (this can certainly be made precise if one has a formula for the function in advance).

We note that the asymptotic estimates of slope in Figures 3, 5, and 7 often perform better than the lines of best fit, especially in Figure 3 for 1-dimensional homology. This improved performance is likely because whereas a linear fit places all random samples X_n of n data points (for varying values of n) on an equal footing, an asymptotic estimate weights more heavily the random samples X_n in which n is large.

9. Conclusion

When points are sampled at random from a subset of Euclidean space, there are a wide variety of Euclidean functionals (such as the minimal spanning tree, the traveling salesperson tour, the optimal matching) which scale according to the dimension of Euclidean space [79]. In this paper we explore whether similar properties are true for persistent homology, and how one might use these scalings in order to define a persistent homology fractal dimension for measures. We provide experimental evidence for some of our conjectures, though that evidence is limited by the sample sizes on which we are able to compute. Our hope is that our experiments are only a first step toward inspiring researchers to further develop the theory underlying the scaling properties of persistent homology.

10. Acknowledgements

We would like to thank Visar Berisha, Vincent Divol, Al Hero, Sara Kališnik, Louis Scharf, and Benjamin Schweinhart for their helpful conversations. We would like to acknowledge the research group of Paul Bendich at Duke University for allowing us access their persistent homology package. The ideas within this paper were greatly improved by the comments and suggestions from a very helpful anonymous referee. The first author would like to thank the organizers of the *2018 Abel Symposium on Topological Data Analysis* in Geiranger, Norway, for hosting a fantastic conference which was the inspiration for these proceedings. This work was supported by a grant from the Simons Foundation/SFARI (#354225, CS).

References

[1] Henry Adams, Sofya Chepushtanova, Tegan Emerson, Eric Hanson, Michael Kirby, Francis Motta, Rachel Neville, Chris Peterson, Patrick Shipman, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252, 2017.

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

[3] Robert J Adler, Omer Bobrowski, Matthew S Borman, Eliran Subag, and Shmuel Weinberger. Persistent homology for random fields and complexes. In *Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown*, pages 124–143. Institute of Mathematical Statistics, 2010.

[4] Robert J Adler, Omer Bobrowski, and Shmuel Weinberger. Crackle: The persistent homology of noise. *arXiv preprint arXiv:1301.1466*, 2013.

[5] David Aldous and J Michael Steele. Asymptotics for Euclidean minimal spanning trees on random points. *Probability Theory and Related Fields*, 92(2):247–258, 1992.

[6] David Aldous and J Michael Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on discrete structures*, pages 1–72. Springer, 2004.

[7] Kenneth S Alexander. The RSW theorem for continuum percolation and the CLT for Euclidean minimal spanning trees. *The Annals of Applied Probability*, 6(2):466–494, 1996.

[8] Mark A Armstrong. *Basic topology*. Springer Science & Business Media, 2013.

[9] Ulrich Bauer. Ripser: A lean C++ code for the computation of Vietoris–Rips persistence barcodes. *Software available at https://github.com/Ripser/ripser*, 2017.

[10] Paul Bendich, J S Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *The Annals of Applied Statistics*, 10(1):198 – 218, 2016.

[11] Martin Bilodeau and David Brenner. *Theory of multivariate statistics*. Springer Science & Business Media, 2008.

[12] Omer Bobrowski and Matthew Strom Borman. Euler integration of Gaussian random fields and persistent homology. *Journal of Topology and Analysis*, 4(01):49–70, 2012.

[13] Omer Bobrowski and Matthew Kahle. Topology of random geometric complexes: A survey. *Journal of Applied and Computational Topology*, 2018.

[14] Omer Bobrowski, Matthew Kahle, and Primoz Skraba. Maximally persistent cycles in random geometric complexes. *arXiv preprint arXiv:1509.04347*, 2015.

[15] Georges Bouligand. Ensembles impropres et nombre dimensionnel. *Bull. Sci. Math.*, 52:361–376, 1928.

[16] Paul Breiding, Sara Kalisnik Verovsek, Bernd Sturmfels, and Madeleine Weinstein. Learning algebraic varieties from samples. *arXiv preprint arXiv:1802.09436*, 2018.

[17] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[18] Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, pages 1–22, 2013.

[19] Frédéric Chazal and Vincent Divol. The density of expected persistence diagrams and its kernel based estimation. *arXiv preprint arXiv:1802.10457*, 2018.

[20] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[21] Anne Collins, Afra Zomorodian, Gunnar Carlsson, and Leonidas J. Guibas. A barcode shape descriptor for curve point cloud data. *Computers & Graphics*, 28(6):881 – 894, 2004.

[22] Jose A Costa and Alfred O Hero. Determining intrinsic dimension and entropy of high-dimensional shape spaces. In *Statistics and Analysis of Shapes*, pages 231–252. Springer, 2006.

[23] Justin Michael Curry. Topological data analysis and cosheaves. *Japan Journal of Industrial and Applied Mathematics*, 32(2):333–371, 2015.

[24] Colleen D Cutler. Some results on the behavior and estimation of the fractal dimensions of distributions on attractors. *Journal of Statistical Physics*, 62(3-4):651–708, 1991.

[25] Colleen D Cutler. A review of the theory and estimation of fractal dimension. In *Dimension estimation and models*, pages 1–107. World Scientific, 1993.

[26] Yuri Dabaghian, Facundo Mémoli, Loren Frank, and Gunnar Carlsson. A topological paradigm for hippocampal spatial map formation using persistent homology. *PLoS computational biology*, 8(8):e1002581, 2012.

[27] Vincent Divol and Wolfgang Polonik. On the choice of weight functions for linear representations of persistence diagrams. *arXiv preprint arXiv: arXiv:1807.03678*, 2018.

[28] Herbert Edelsbrunner and John L Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, 2010.

[29] Herbert Edelsbrunner, A Ivanov, and R Karasev. Current open problems in discrete and computational geometry. *Modelirovanie i Analiz Informats. Sistem*, 19(5):5–17, 2012.

[30] Herbert Edelsbrunner, Anton Nikitenko, and Matthias Reitzner. Expected sizes of Poisson–Delaunay mosaics and their discrete Morse functions. *Advances in Applied Probability*, 49(3):745–767, 2017.

[31] Gerald A Edgar. *Classics on fractals*. Addison–Wesley, 1993.

[32] Kenneth Falconer. *Fractal geometry: mathematical foundations and applications; 3rd ed.* Wiley, Hoboken, NJ, 2013.

[33] J.D. Farmer. Information dimension and the probabilistic structure of chaos. *Zeitschrift für Naturforschung A*, 37(11):1304–1326, 1982.

[34] J.D. Farmer, Edward Ott, and James Yorke. The dimension of chaotic attractors. *Physica D: Nonlinear Phenomena*, 7(1):153–180, 1983.

[35] Gerald Folland. *Real Analysis*. John Wiley & Sons, 1999.

[36] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[37] Peter Grassberger and Itamar Procaccia. Characterization of strange attractors. *Physics Review Letters*, 50(5):346–349, 1983.

[38] Peter Grassberger and Itamar Procaccia. Measuring the Strangeness of Strange Attractors. In *The Theory of Chaotic Attractors*, pages 170–189. Springer, New York, NY, 2004.

[39] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, Cambridge, 2002.

[40] Felix Hausdorff. Dimension und äußeres maß. *Mathematische Annalen*, 79(1-2):157–179, 1918.

[41] Yasuaki Hiraoka, Tomoyuki Shirai, and Khanh Duy Trinh. Limit theorems for persistence diagrams. *The Annals of Applied Probability*, 28(5):2740–2780, 2018.

[42] Patrick Jaillet. On properties of geometric random problems in the plane. *Annals of Operations Research*, 61(1):1–20, 1995.

[43] Matthew Kahle. Random geometric complexes. *Discrete & Computational Geometry*, 45(3):553–573, 2011.

[44] Albrecht M Kellerer. On the number of clumps resulting from the overlap of randomly placed figures in a plane. *Journal of Applied Probability*, 20(1):126–135, 1983.

[45] Harry Kesten and Sungchul Lee. The central limit theorem for weighted minimal spanning trees on random points. *The Annals of Applied Probability*, pages 495–527, 1996.

[46] Gady Kozma, Zvi Lotker, and Gideon Stupp. The minimal spanning tree and the upper box dimension. *Proceedings of the American Mathematical Society*, 134(4):1183–1187, 2006.

[47] Gady Kozma, Zvi Lotker, and Gideon Stupp. On the connectivity threshold for general uniform metric spaces. *Information Processing Letters*, 110(10):356–359, 2010.

[48] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

[49] H Lee, H Kang, M K Chung, B N Kim, and D S Lee. Persistent brain network homology from the perspective of dendrogram. *IEEE Transactions on Medical Imaging*, 31(12):2267–2277, 2012.

[50] Javier Lamar Leon, Andrea Cerri, Edel Garcia Reyes, and Rocio Gonzalez Diaz. Gait-based gender classification using persistent homology. In José Ruiz-Shulcloper and Gabriella Sanniti di Baja, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 366–373, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[51] Robert MacPherson and Benjamin Schweinhart. Measuring shape with topology. *Journal of Mathematical Physics*, 53(7):073516, 2012.

[52] Benoit B Mandelbrot. *The fractal geometry of nature*, volume 1. WH Freeman, New York, 1982.

[53] Pertti Mattila, Manuel Morán, and José-Manuel Rey. Dimension of a measure. *Studia Math*, 142(3):219–233, 2000.

[54] Pat A .P. Moran. Additive functions of intervals and Hausdorff measure. *Proceedings of the Cambridge Philosophical Society*, 42(1):15–23, 1946.

[55] Takashi Owada and Omer Bobrowski. Convergence of persistence diagrams for topological crackle. *arXiv preprint arXiv:1810.01602*, 2018.

[56] Heinz-Otto Peitgen, Hartmut Jürgens, and Dietmar Saupe. *Chaos and fractals: New frontiers of science*. Springer Science & Business Media, 2006.

[57] Mathew Penrose. *Random geometric graphs*, volume 5. Oxford University Press, Oxford, 2003.

[58] Mathew D Penrose. The longest edge of the random minimal spanning tree. *The annals of applied probability*, pages 340–361, 1997.

[59] Mathew D Penrose et al. A strong law for the longest edge of the minimal spanning tree. *The Annals of Probability*, 27(1):246–260, 1999.

[60] Mathew D Penrose and Joseph E Yukich. Central limit theorems for some graphs in computational geometry. *Annals of Applied probability*, pages 1005–1041, 2001.

[61] Yakov B Pesin. *Dimension theory in dynamical systems: contemporary views and applications*. University of Chicago Press, 2008.

[62] Robert Clay Prim. Shortest connection networks and some generalizations. *Bell Labs Technical Journal*, 36(6):1389–1401, 1957.

[63] Alfréd Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Hungarica*, 10(1-2):193–215, 1959.

[64] Alfréd Rényi. *Probability Theory*. North Holland, Amsterdam, 1970.

[65] Vanessa Robins. *Computational topology at multiple resolutions: foundations and applications to fractals and dynamics*. PhD thesis, University of Colorado, 2000.

[66] M.J. Schervish. *Theory of Statistics*. Springer Series in Statistics. Springer New York, 1996.

[67] Benjamin Schweinhart. Persistent homology and the upper box dimension. *arXiv preprint arXiv:1802.00533*, 2018.

[68] Benjamin Schweinhart. The persistent homology of random geometric complexes on fractals. *arXiv preprint arXiv:1808.02196*, 2018.

[69] Benjamin Schweinhart. Weighted persistent homology sums of random Čech complexes. *arXiv preprint arXiv:1807.07054*, 2018.

[70] J Michael Steele. Growth rates of Euclidean minimal spanning trees with power weighted edges. *The Annals of Probability*, pages 1767–1787, 1988.

[71] J Michael Steele. Probability and problems in Euclidean combinatorial optimization. *Statistical Science*, pages 48–56, 1993.

[72] J Michael Steele. Minimal spanning trees for graphs with random edge lengths. In *Mathematics and Computer Science II*, pages 223–245. Springer, 2002.

[73] J Michael Steele, Lawrence A Shepp, and William F Eddy. On the number of leaves of a Euclidean minimal spanning tree. *Journal of Applied Probability*, 24(4):809–826, 1987.

[74] J Michael Steele and Luke Tierney. Boundary domination and the distribution of the largest nearest-neighbor link in higher dimensions. *Journal of Applied Probability*, 23(2):524–528, 1986.

[75] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. Javaplex: A research software package for persistent (co)homology. In *International Congress on Mathematical Software*, pages 129–136, 2014. Software available at http://appliedtopology.github.io/javaplex/.

[76] James Theiler. Estimating fractal dimension. *JOSA A*, 7(6):1055–1073, 1990.

[77] Robert W Vallin. *The elements of Cantor sets: with applications*. John Wiley & Sons, 2013.

[78] Kelin Xia and Guo-Wei Wei. Multidimensional persistence in biomolecular data. *Journal of Computational Chemistry*, 36(20):1502 – 1520, 2015.

[79] Joseph E Yukich. *Probability theory of classical Euclidean optimization problems*. Springer, 2006.
[80] Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI*, pages 1953–1959, 2013.