Persistent Homology in Topological Data Analysis

Altan T. Haan

June 4, 2018

1 Introduction

In this paper, we build up the theory and machinery required to understand *persistent homology*, and give an introductory overview to its tools and applications. At the highest level, persistent homology is an extension of homology, which is itself an extremely broad algebraic and topological field of study. Briefly, homology theory aims to concretely classify and rigorously define the notion of holes, boundaries, and volumes. These concepts are intuitively topological; for example, a sphere has a missing volume whilst a ball does not, and a torus has another hole in the middle that a sphere lacks. It turns out that these topological differences can be attacked algebraically, by essentially breaking the spaces down into closed 'cycles' of various dimensions, which may or may not be the boundaries of regions.

The novelty of persistent homology is in applying homology theory to 'uncertain' spaces. The precise meaning is this: with homology, we are given a space, and then tasked with finding its holes. Persistent homology (and more generally the area of *topological data analysis*) aims to recover a topological space (characterized by its holes), given some set of incomplete information. The most usual case, which persistent homology deals with, would be sets of point cloud data, which we can consider to be finite subsets of arbitrary metric spaces. Obviously, we can give the data the discrete metric, but this might as well be topologically trivial. Instead, we want to image the data at various spatial resolutions; more precisely, we want to fatten up the point cloud space by looking at the covering space generated by taking larger and larger balls around each point. One can imagine, for example, a set of points sampled from a torus. As we increase the size of the balls around the points, the covering space gets fatter and fatter until we have something resembling a 'bumpy' torus. This is topologically a torus nonetheless, and we have recovered (in some loose sense, with some uncertainty) the original topological structure of the data.

Unlike traditional methods of data analysis, which can also perform these kinds of classification tasks, persistent homology as constructed is dimensionless. We need not constrain ourselves to lower dimensional datasets, and can instead construct or

directly deal with high-dimensional datasets. The resultant persistent topological spaces will likely also be of high-dimension, but this is not a problem, as we welcome any persistent structure.

A natural question of representation then arises. How do we characterize and encode this persistent homological information? The answer lies in tracking the holes over time, and looking at the homological differences between the images of the space at different times (where time represents the fatness parameter). This information can then be visually represented in a *persistence diagram*, which pairs the births and deaths of the homological features over time. We give a brief definition of this at the end of the paper.

2 Fundamentals

The theory of homology sits at the intersection of algebra and topology, so we first give some preliminary definitions.

2.1 Group Theory

Definition 2.1. A *group* is a tuple $(G, \cdot)$ where G is a set and $\cdot : G \times G \to G$ is a function so that

(i) For all $a, b \in G$, $a \cdot b \in G$.

(ii) For all $a, b, c \in G$, $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.

(iii) There is an element $e \in G$ so that $a \cdot e = e \cdot a = a$ for all $a \in G$.

(iv) For all $a \in G$, there is an element $a^{-1} \in G$ so that $a \cdot a^{-1} = a^{-1} \cdot a = e$.

A group whose operation is commutative (i.e., $a \cdot b = b \cdot a$ for all $a, b \in G$) is called an *abelian* group, and its operation is conventionally denoted $+$. We often refer to the underlying set as the group, and specify the operation.

Definition 2.2. Let $(G, \cdot)$ and $(H, \circ)$ be abelian groups. Then the *direct sum* of G and H is the abelian group formed from $G \times H$ with operation $+$ defined by $(a, b) + (c, d) = (a \cdot c, b \circ d)$ for all $(a, b), (c, d) \in G \times H$. We write $G \oplus H$ to denote the direct sum.

Definition 2.3. Let $(G, \cdot)$ be a group. A subset $H \subseteq G$ that is also a group under $\cdot$ is a *subgroup* of G.

Groups are generalizations of common algebraic structures, e.g. $\mathbb{Z}$ over addition and $\mathbb{R} \setminus \{0\}$ over multiplication (in fact these are abelian).

Definition 2.4. Let $(G, \cdot)$ be a group, and H be a subgroup. For each $a \in G$, the *left coset* of H in G with respect to a is the set $aH = \{a \cdot h : h \in H\}$. Similarly, the *right coset* is defined to be $Ha = \{h \cdot a : h \in H\}$. H is said to be *normal* if $aH = Ha$ for each $a \in G$.

Definition 2.5. Let $(G, \cdot)$ be a group, and H be a normal subgroup of G. Then the group G/H is the set $\{aH : a \in G\}$ with operation $(aH) \times (bH) = (a \cdot b)H$. We call this group the *quotient group* of G by H.

Remark. We require H to be normal in order to satisfy associativity over $\times$.

Definition 2.6. Let G and H be groups with operations $\cdot$ and $\times$. A function $f : G \to H$ is a *group homomorphism* if $f(a \cdot b) = f(a) \times f(b)$ for all $a, b \in G$. If f is a bijection, then we call f an *isomorphism*, and say G is *isomorphic* to H. We denote this relation by $G \simeq H$.

We can think of group homomorphisms as functions that preserve the group structure.

2.2 Topology

Definition 2.7. Let X be a set. A *topology* τ on X is a collection of subsets of X such that

(i) $X \in \tau$ and $\varnothing \in \tau$,

(ii) $\bigcup_{U \in \kappa} U \in \tau$ for any $\kappa \subseteq \tau$,

(iii) $\bigcap_{i=1}^{n} U_i \in \tau$ for any finite collection of $U_i \in \tau$.

The elements of τ are called *open sets*. We call the ordered pair (X, τ), or simply X if the choice of τ is clear or arbitrary, a *topological space*.

Definition 2.8. Let X and Y be topological spaces. A function $f : X \to Y$ is said to be *continuous* if $f^{-1}(U)$ is open in X for each open $U \in Y$, where $f^{-1}(U) = \{x \in X : f(x) \in U\}$.

These definitions allow us to reason more abstractly about the (usually) geometric notions of 'nearness' (but not distance) and continuity; indeed, we don't need a metric space for these properties to make sense, although a metric induces a topology (cf. the ϵ, δ definition of continuity in Euclidean space, which ultimately reduces to reasoning about open 'balls').

Definition 2.9. Let M be a set, and let $d : M \times M \to \mathbb{R}$ satisfy for all $x, y, z \in M$:

(i) $d(x, y) \geq 0$,

(ii) $d(x, y) = 0$ if and only if $x = y$,

(iii) $d(x, y) = d(y, x)$,

(iv) $d(x, z) \leq d(x, y) + d(y, z)$.

We say that d is a *metric* on M, and that the pair (M, d) is a *metric space*. Often we simply refer to the underlying set M as the metric space, when the choice of metric is clear.

A metric tells us the 'distance' between elements of a set, and satisfies intuitive geometric properties. We are familiar with the Euclidean metric on $\mathbb{R}^n$, which allows us to define the standard topology more directly using open balls. This applies generally, where we can define an open ϵ-ball about x on (M, d) as the set $B_\epsilon(x) = \{y \in M : d(x, y) < \epsilon\}$. We can then define the open sets to be those whose points have open balls contained fully within the set.

Definition 2.10. Let X and Y be topological spaces. A continuous, bijective function $f : X \to Y$ whose inverse $f^{-1} : Y \to X$ is also continuous is called a *homeomorphism*. We say that the spaces X and Y are *homeomorphic* if there exists a homeomorphism between them.

Homeomorphisms preserve topological properties of spaces, as they directly link their open sets. By themselves, however, homeomorphisms do not immediately tell us much about the structure of a topological space (just that two spaces behave the same way topologically, which can be vague). If want to understand and characterize the structure of a particular space more concretely, we need to develop some machinery.

3 Homology Theory

We now give an exposition of homology theory and its goals, via the straightforward construction of simplicial homology.

3.1 Motivation

Consider the classic example of a ball and a torus. One should hopefully immediately notice that these spaces are different in some concrete way — the ball is 'complete' in some sense, whilst the torus has a 'hole' in the middle. A more familiar 2-dimensional analogue would be a disk and an annulus in $\mathbb{R}^2$. At the level of elementary complex analysis, the differences between these two spaces already become imperative in the statement of various theorems. For example, many theorems require domains to be *simply connected*, which simply stated is the property of being able to 'continuously shrink' every closed loop in the domain to a point in the domain. The annulus (and torus) fails in this regard, as no loop enclosing the inner circle can be shrunk to a point

in the annulus. Homology theory aims to concretely classify these topological holes, by passing topological spaces to algebraic structures that reveal more information.

3.2 The Chain Complex Structure

Homology theory is fundamentally the study of certain algebraic structures called *chain complexes.* We first give some formal definitions.

Definition 3.1. Let $(G, +)$ be an abelian group. The *rank* of G is the cardinality of the largest linearly independent subset $S \subseteq G$. That is, the largest set $S = \{a_k\}$ so that every finite integral linear combination $\sum_k n_k a_k = 0$ only if each n_k is 0. We denote this by $\text{rank}(G)$.

Definition 3.2. Let $S = \{b_1, \cdots, b_n\}$ be a finite set. The *free abelian group* generated by S is the abelian group $(G, +)$ where G is the set of formal sums of S:

$$G = \{a_1 b_1 + \cdots + a_n b_n : a_j \in \mathbb{Z} \text{ for } 1 \leq j \leq n\}.$$

Here $a_i b_i$ is defined as $b_i + b_i + \cdots + b_i$ or $-b_i - b_i \cdots - b_i$ (i.e, repeated application of the group operation a_i times on b_i or $b_i^{-1} = -b_i$, depending on sign).

It may be tempting to assign significance to these formal sums, but one should mainly take it as a way of quickly generating an abelian group, unless additional information is attached to these sums.

Lemma 3.1. *Let G be the free abelian group generated by $\{b_1, \cdots, b_n\}$. Then $G \simeq \mathbb{Z}^n$, where $\mathbb{Z}^n = \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}$ is the direct sum of n copies of $\mathbb{Z}$ under addition.*

Proof. We define the isomorphism $f : \mathbb{Z}^n \to G$ by

$$f[(x_1, \cdots, x_n) + (y_1, \cdots, y_n)] = (x_1 b_1 + \cdots + x_n b_n) + (y_1 b_1 + \cdots + y_n b_n),$$

$$f^{-1}[(x_1 b_1 + \cdots + x_n b_n) + (y_1 b_1 + \cdots + y_n b_n)] = (x_1, \cdots, x_n) + (y_1, \cdots, y_n).$$

□

This isomorphism will help us assign some basic numeric meaning to the groups that arise in homology theory. We now come to the central definition.

Definition 3.3. A *chain complex* is a tuple (A_*, ∂_*), where $A_* = (\cdots, A_0, A_1, A_2, \cdots)$ is a sequence of abelian groups, and $\partial_* = (\cdots, \partial_0, \partial_1, \partial_2, \cdots)$ is a sequence of group homomorphisms (here called the *boundary operators*) with each $\partial_n : A_n \to A_{n-1}$ satisfying $\partial_n \circ \partial_{n+1} = 0_n$. Here $0_n : A_n \to 0$ denotes the homomorphism to the trivial group 0 with only the identity element. We sometimes refer to the underlying group sequence A_* as the chain complex, and supply the boundary operator.

We can visualize this complex as a diagram:

$$\cdots \xrightarrow{\partial_3} A_2 \xrightarrow{\partial_2} A_1 \xrightarrow{\partial_1} A_0 \xrightarrow{\partial_0} \cdots$$

In isolation, chain complexes seem removed from topological constructions. However, we can assign topological meaning to each group and homomorphism.

We give an informal example and discussion. Consider the graph-like topological space X (perhaps as a subspace of $\mathbb{R}^2$) depicted in Figure 1. Suppose we want to

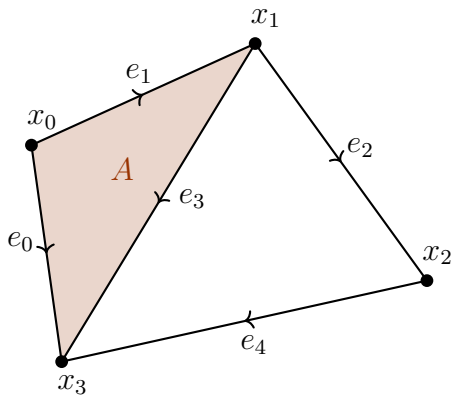

Figure 1: The topological space X, consisting of points $\{x_0, \cdots, x_3\}$, (directed) edges $\{e_0, \cdots, e_4\}$, and area A. In this example, the orientations of the edges are induced by the vertex index order.

classify how many holes X has. Looking at the structure, it seems like X has a hole enclosed by the edges e_2, e_3, e_4 (compare this to area A enclosed by e_0, e_1, e_3, which is contained in X). We can form the chain complex

$$\cdots \xrightarrow{\partial_4} 0 \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0 \xrightarrow{\partial_{-1}} \cdots$$

where C_2 is the free abelian group generated by $\{A\}$, C_1 the one generated by $\{e_0, \cdots, e_4\}$, and C_0 the one generated by $\{x_0, \cdots, x_3\}$. As the name suggests, let's make the boundary operator ∂_2 take A to its edge boundary, given by the formal sum $e_1 + e_3 - e_0$. We subtract e_0 here to mean a reversal of orientation, making the sum denote a cycle from $x_0 \to x_1 \to x_3$. This agrees with our geometric intuition of a boundary being a closed curve. Then for each edge e, we make the operator ∂_1 take e to $v_1 - v_0$, where v_0 is the starting point and v_1 the ending point. The signs attach a direction to each point (i.e. origin vs. destination), which don't have orientations themselves.

We can check that the boundary operators satisfy the required identities:

$$\partial_1\partial_2(A) = \partial_1(e_1 + e_3 - e_0) = \partial_1(e_1) + \partial_1(e_3) - \partial_1(e_0) = x_1 - x_0 + x_3 - x_1 - x_3 + x_0 = 0,$$

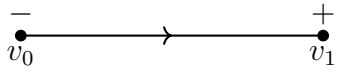

Figure 2: Point orientation on an edge.

$$\partial_0 \partial_1(e) = \partial_0(v_1) - \partial_0(v_0) = 0.$$

If we consider the areas of X as "2-cells", line segments as "1-cells", and points as "0-cells", we can consider C_2 as the group of "2-chains" (i.e., integral linear combinations of 2-cells), and likewise for C_1, C_0. The boundary operators tell us that

(i) a cycle of n-cells has no boundary, and

(ii) the boundary of an n-cell is a cycle of $(n-1)$-cells.

Then, to find n-dimensional holes in a space (by which we mean holes enclosed by n-chains), we just need to find n-cycles that are not the boundaries of $(n+1)$-cells.

More precisely, the group of n-dimensional holes enclosed by n-chains is given by the quotient group

$$H_n(X) = \ker \partial_n / \operatorname{im} \partial_{n+1} = Z_n / B_n,$$

where $Z_n \subseteq C_n$ is the subgroup of n-cycles in C_n, and $B_n \subseteq C_n$ is the subgroup of n-cycles in C_n that are also the boundary of $(n+1)$-cells in C_{n+1}. For spaces composed of finitely many cells, each $H_n(X) \simeq \mathbb{Z}^k$ for some k. In either case, we have a general measure for the n-dimensional holes of X. For X, a computation shows that

$$H_1(X) \simeq \mathbb{Z}^2/\mathbb{Z} \simeq \mathbb{Z},$$

which can be interpreted as there being a single 1-dimensional hole. This agrees with our intuition.

General topological spaces require a general theory of homology in order to produce such chain complexes; this turns out to be somewhat involved, and is covered by the theory of *singular homology.* Singular homology groups are usually much harder to deal with computationally — the n-chains of a singular chain complex are not even finitely generated. For our purposes, we consider a much simpler homology theory that formalizes the computations we did above.

3.3 Simplicial Homology

Unlike singular homology, which is applicable to general topological spaces, we restrict our discussion to spaces called *simplicial complexes.* Intuitively, one can think of a simplicial complex as a space composed of points, lines, triangles, tetrahedrons, and their higher dimensional counterparts. We first define these shapes concretely in Euclidean space, but remark that they carry over to arbitrary metric spaces.

Definition 3.4. Let $S = \{x_0, \cdots, x_m\}$ be a finite set of points in $\mathbb{R}^n$. The *convex hull* of S in $\mathbb{R}^n$ is the set

$$\mathrm{Conv}(S) = \{\alpha_0 x_0 + \cdots + \alpha_m x_m : \alpha_0 + \cdots + \alpha_m = 1, \alpha_i \geq 0 \text{ for each } i\}.$$

One can think of $\mathrm{Conv}(S)$ as the smallest convex set so that $S \subseteq \mathrm{Conv}(S)$.

Definition 3.5. Let $S = \{x_0, \cdots, x_n\} \subseteq \mathbb{R}^k$ be a set of points such that the set $\{x_0 - x_1, \cdots, x_0 - x_n\}$ is linearly independent. Then we call $\mathrm{Conv}(S)$ an *n-simplex*, and we call the elements of S the *vertices* of the simplex.

Definition 3.6. Let $S = \{x_0, \cdots, x_n\} \subseteq \mathbb{R}^k$ be the vertices of an n-simplex Δ, and let $F \subseteq S$. Then the convex hull $\mathrm{Conv}(F)$ is a *face* of Δ. More specifically, if $|F| = m$, we say that $\mathrm{Conv}(F)$ is an m-face of Δ. Notice that $\mathrm{Conv}(F)$ is itself an m-simplex. If dimension is omitted, we often say that the ith face of Δ is the $(n-1)$-simplex $\Delta[i]$ formed from $S \setminus \{x_i\}$.

Simplices generalize the notion of triangles and tetrahedrons to arbitrary dimensions.

Definition 3.7. Given an n-simplex S with vertices $\{x_0, \cdots, x_n\}$, an *orientation* of S is a permutation of the vertices $\tau : \{0, \cdots, n\} \to \{0, \cdots, n\}$. Two orientations τ_1, τ_2 are the same if $\mathrm{sgn}(\tau_1) = \mathrm{sgn}(\tau_2)$ (i.e., they have the same parity), so simplices have two orientations. We denote an oriented simplex by $[x_{\tau(0)}, \cdots, x_{\tau(n)}]$, but often just $[x_0, \cdots, x_n]$ when the vertex order suffices.

Lemma 3.2. *An orientation on an n-simplex $\Delta = [x_0, \cdots, x_n]$ induces an ordering on its faces.*

Proof. Simply orient the ith face by $[x_0, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n]$. Proceeding recursively induces an orientation on each m-face. □

Remark. When the faces of an n-simplex $\Delta = [x_0, \cdots, x_n]$ are given the induced orientation, we write $\Delta[i] = [x_0, \cdots, \hat{x}_i, \cdots, x_n]$, where the circumflex denotes deletion.

We can now define a well-behaved space constructed from simplices.

Definition 3.8. Let $\mathcal{K}$ be a finite set of oriented simplices in $\mathbb{R}^n$. We say that $\mathcal{K}$ is a *simplicial complex* if:

(i) for all $\sigma \in \mathcal{K}$, each face of σ is a simplex in $\mathcal{K}$,

(ii) for all $\sigma_1, \sigma_2 \in \mathcal{K}$, the intersection $\sigma_1 \cap \sigma_2$ is a face of both σ_1 and σ_2 (or the empty set).

If k is the cardinality of the largest vertex set of all simplices in $\mathcal{K}$, then we say $\mathcal{K}$ is a simplicial k-complex. The *underlying space* of $\mathcal{K}$ is the union of all its simplices.

Simplicial complexes admit a chain complex structure of finitely generated abelian groups, along with natural boundary operators that agree with geometric intuition. This makes the computation of their homology groups fairly straightforward.

Lemma 3.3. *Let $\mathcal{K}$ be a simplicial complex, and let C_n be the free abelian group generated by the n-simplices of $\mathcal{K}$, with $-\sigma$ indicating the opposite orientation of a simplex σ. For $n < 0$, let $C_n = 0$. Additionally, define the boundary operator $\partial_n : C_n \to C_{n-1}$ for $n \geq 0$ by*

$$\partial_n(\sigma) = \sum_{i=0}^{n} (-1)^i \sigma[i],$$

where $\sigma \in C_n$ is a formal sum with only one term (i.e., an n-simplex). Then (C_*, ∂_*) with $C_* = (\cdots, C_n, C_{n-1}, \cdots)$ and $\partial_* = (\cdots, \partial_n, \partial_{n-1}, \cdots)$ is a chain complex.

Proof. Since each C_n is abelian by construction, we simply check the boundary operators. Let $\sigma = [x_0, \cdots, x_n] \in C_n$ be as above; then

$$\begin{aligned}
\partial_{n-1} \circ \partial_n(\sigma) &= \partial_{n-1} \left(\sum_{i=0}^{n} (-1)^i \sigma[i] \right) \\
&= \sum_{i=0}^{n} (-1)^i \left(\sum_{j=0}^{n-1} (-1)^j \sigma[i][j] \right). \\
&= \sum_{i=0}^{n} \sum_{j=0}^{n-1} (-1)^{i+j} \sigma[i][j].
\end{aligned}$$

When $i \leq j$, the deletion of the ith vertex shifts the j index up by 1 in relation to the vertices of σ. Thus, $\sigma[i][j] = \sigma[j+1][i]$ when $i \leq j$. Splitting the sum on $i \leq j$ and $i > j$, reindexing gives

$$\begin{aligned}
\sum_{i=0}^{n} \sum_{j=0}^{n-1} (-1)^{i+j} \sigma[i][j] &= \sum_{i=0}^{n} \sum_{j=i}^{n-1} (-1)^{i+j} \sigma[i][j] + \sum_{j=0}^{n-1} \sum_{i=0}^{j} (-1)^{i+j+1} \sigma[j+1][i] \\
&= \sum_{i=0}^{n} \sum_{j=i}^{n-1} (-1)^{i+j} \sigma[i][j] - \sum_{j=0}^{n-1} \sum_{i=0}^{j} (-1)^{i+j} \sigma[j+1][i] \\
&= 0.
\end{aligned}$$

Furthermore, if C_n is generated by $\{\sigma_1, \cdots, \sigma_n\}$, then an arbitrary formal sum $a_1\sigma_1 + \cdots + a_n\sigma_n$ satisfies

$$\begin{aligned}
\partial_{n-1} \circ \partial_n(a_1\sigma_1 + \cdots + a_n\sigma_n) &= \partial_{n-1}(a_1\partial_n(\sigma_1) + \cdots + a_n\partial_n(\sigma_n)) \\
&= a_1\partial_{n-1} \circ \partial_n(\sigma_1) + \cdots + a_n\partial_{n-1} \circ \partial_n(\sigma_n) \\
&= 0.
\end{aligned}$$

□

The boundary operators on this chain complex quite literally take simplices to their topological boundaries with orientation. We now give the following definitions that formalize our intuitive discussion of homology in the previous section.

Definition 3.9. The elements of C_n are called *simplicial n-chains*, and each $c \in C_n$ satisfying $\partial_n(c) = 0$ is called an *n-cycle*. This forms a subgroup Z_n of C_n.

Definition 3.10. The elements of C_{n+1} under ∂_{n+1} are called *n-boundaries*, and they form a subgroup B_n of C_n.

Definition 3.11. The nth *homology group* of a simplicial complex $\mathcal{K}$ is the quotient group $H_n(\mathcal{K}) = Z_n/B_n$.

Definition 3.12. The nth *Betti number* β_n of $\mathcal{K}$ is the rank of the nth homology group. That is, $\beta_n = \operatorname{rank}(H_n(\mathcal{K}))$.

The Betti numbers tell us precisely the number of n-dimensional holes in $\mathcal{K}$.[1] Additionally, since each C_n is finitely generated, we have a guarantee that each $H_n(\mathcal{K})$ is isomorphic to some direct sum of $\mathbb{Z}$. In particular, each homology generator (i.e., each distinct 'hole') corresponds to a copy of $\mathbb{Z}$. One can easily check that $\operatorname{rank}(\mathbb{Z}^k) = k$, so the Betti numbers for simplicial complexes are finite.

Although simplicial complexes are rigid in their construction, it stands to reason that homeomorphic spaces have the same homology groups. For example, the 2-sphere S^2 is homeomorphic to a simplicial complex composed of the faces of a 3-simplex. Computing the associated homology groups shows that $H_1(S^2) \simeq 0$, and $H_2(S^2) \simeq \mathbb{Z}$, which agrees with the sphere 'missing' its 3-dimensional volume but not having any holes in its surface.

At this point, we remark that the above construction of simplicial homology did not use any explicit simplicial-geometric properties, apart from ensuring that simplices did not intersect each other in sets that were not themselves simplices. If we forget about this requirement for now, the only required property was that each face of a simplex (which is also a simplex) be contained in the complex. This allows us to make the following construction.

Definition 3.13. An *abstract simplicial complex* is a finite collection of (finite) ordered sets $\mathcal{A}$, so that for all $\sigma \in \mathcal{A}$, $\tau \subseteq \sigma$ implies $\tau \in \mathcal{A}$. The dimension of $\mathcal{A}$ is $\max\{\dim \sigma : \sigma \in \mathcal{A}\}$, where $\dim \sigma = |\sigma| - 1$.

Any simplicial complex can be taken to an abstract simplicial complex by considering the vertex sets of each simplex, as opposed to the convex hull spanned by them. Abstract simplicial complexes are useful theoretically, as they characterize the simplicial structure without needing a geometric realization. More practically however, this abstract construction allows us to pass a finite set of point data (along with some metric information) to a geometric simplicial complex, which suggests we can directly deal with the homology of abstract complexes.

[1] An argument shows that β_0 gives us the number of path-connected components of $\mathcal{K}$.

Lemma 3.4 (Geometric Realization)**.** *Let $\mathcal{A}$ be an abstract simplicial complex of dimension k. Then $\mathcal{A}$ admits a geometric realization as a simplicial complex in $\mathbb{R}^{2k+1}$.*

Proof. The proof of this lemma is simple, so we give a brief sketch. The idea is to send the vertex set $V(\mathcal{A}) = \bigcup_{\sigma\in\mathcal{A}} \sigma$ injectively to a set of points $f(V(\mathcal{A}))$ in $\mathbb{R}^{2k+1}$ that are in *general position*, i.e. such that any $2k+2$ points (or less) form a valid geometric simplex. Then for $\sigma, \tau \in \mathcal{A}$, we know that $|\sigma\cup\tau| = |\sigma|+|\tau|-|\sigma\cap\tau| \leq 2k+2-|\sigma\cap\tau| \leq 2k+2$, so $\operatorname{Conv} f(\sigma \cup \tau)$ forms a simplex in $\mathbb{R}^{2k+1}$. Thus each $x \in \operatorname{Conv} f(\sigma \cup \tau)$ is uniquely determined by some linear combination of the vertices, so $x \in \operatorname{Conv} f(\sigma\cap\tau)$ if and only if $x \in \operatorname{Conv} f(\sigma)$ and $x \in \operatorname{Conv} f(\tau)$. It follows that $\operatorname{Conv} f(\sigma) \cap \operatorname{Conv} f(\tau)$ is either empty or $\operatorname{Conv} f(\sigma \cap \tau)$, so $\mathcal{K} = \{\operatorname{Conv} f(\sigma) \subseteq \mathbb{R}^{2k+1} : \sigma \in \mathcal{A}\}$ forms a simplicial complex. □

4 Persistent Homology

The central tool of topological data analysis is that of *persistent homology.* The idea is to consider topological features of a sampled point cloud space that emerge and decay at different spacial resolutions, and to classify the features that persist the longest. More concretely, given a set of point cloud data, we construct progressively 'fatter' simplicial complexes by increasing the connectedness of the points. The resulting sequence of homology groups lets us see how the structure of the space changes over 'time,' and persistent homology gives us tools for comparing these progressions. For a more general overview of persistent homology, we refer the reader to [3].

4.1 Filtrations and the Čech, Vietoris-Rips Complexes

In order to precisely define what it means to have a progression of simplicial complexes, we define the notion of a filtration.

Definition 4.1. Let $\mathcal{K}$ be a simplicial complex. Then, a *filtration* of $\mathcal{K}$ is a sequence of simplicial complexes $(\mathcal{K}_0, \cdots, \mathcal{K}_n)$ such that

$$\varnothing \subseteq \mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \cdots \subseteq \mathcal{K}_n \subseteq \mathcal{K}.$$

We can think of each complex in the filtration as being built on top of its predecessor, by adding simplices whilst preserving the simplicial complex structure. The logical next step is to develop methods for constructing filtrations from point cloud data (by which we mean a finite set of points in a metric space). For this, we introduce the following complexes:

Definition 4.2. Let (M, d) be a metric space, and let $\{x_1, \cdots, x_m\} \in M$. The *Čech complex* $\mathcal{C}_\epsilon$ (with parameter ϵ) is the abstract simplicial complex whose k-simplices

are formed from suitably ordered[2] $(k+1)$-subsets $[x_1, \cdots, x_{k+1}]$, satisfying

$$\bigcap_{i=1}^{k+1} B_\epsilon(x_i) \neq \varnothing, \quad B_\epsilon(x_i) = \{x \in M : d(x, x_i) \leq \epsilon/2\}.$$

The Čech complex $\mathcal{C}_\epsilon$, by way of the *nerve lemma*, is a topologically representative simplicial complex of the covering space formed from the $\epsilon/2$ balls around each point. A more detailed exposition can be found in the proof of Corollary 4G.3 in [6]. Essentially, the simplicial homology of $\mathcal{C}_\epsilon$ captures the holes of the fattened point cloud space. A downside of the Čech complex is its computational complexity, as its construction effectively requires examining the distances between all points in conjunction for each simplex. This difficulty is alleviated by the following 'coarser' but more easily computed complex.

Definition 4.3. Let (M, d) be a metric space, and let $\{x_1, \cdots, x_m\} \in M$. The *Vietoris-Rips complex* $\mathcal{R}_\epsilon$ (with parameter ϵ) is the abstract simplicial complex whose k-simplices are formed from suitably ordered $(k+1)$-subsets $[x_1, \cdots, x_{k+1}]$, such that the x_i's are pairwise within distance ϵ.

We will usually refer to the above construction as the Rips complex for brevity. This complex is coarser in the sense that $\mathcal{C}_\epsilon \subseteq \mathcal{R}_\epsilon$; more specifically, the Rips complex $\mathcal{R}_\epsilon$ is entirely determined by its 1-simplices (which coincide with those of $\mathcal{C}_\epsilon$). This means that the Rips complex associated with a set of point cloud data can be easily (re)constructed from an adjacency matrix representing connected points in a graph.[3] In this sense, it is the 'largest' simplicial complex that can be formed from the given 1-simplicial skeleton, whilst the Čech complex is more refined. This can be seen in Figure 3, where the Čech complex preserves the hole in the upper right portion of the covering but the Rips complex does not.

Despite this, these complexes are closely connected, and the Rips complex turns out to be a decent approximation of the Čech complex. The following theorem precisely relates the accuracies of the Čech and Rips complexes.

Theorem 4.1. *Let X be a point cloud in $\mathbb{R}^n$, and let $\mathcal{C}_\epsilon$ and $\mathcal{R}_\epsilon$ be its associated Čech and Rips complexes. Then we have*

$$\mathcal{R}_{\epsilon'} \subseteq \mathcal{C}_\epsilon \subseteq \mathcal{R}_\epsilon \quad \textit{when} \quad \frac{\epsilon}{\epsilon'} \geq \sqrt{\frac{2n}{n+1}}.$$

The proof of this theorem is given in [1]. Essentially, we now have a guarantee that the Rips complexes associated to a set of point cloud data are sufficiently accurate with enough ϵ-samples, as we can always squeeze Čech complexes between Rips

[2]Faces should have the induced ordering.

[3]A side effect of this is that the Rips complex is homologically invariant under embeddings of the point cloud into different metric spaces, whereas the Čech complex depends on the specific embedding (as we consider neighborhoods which depend on the specific space).

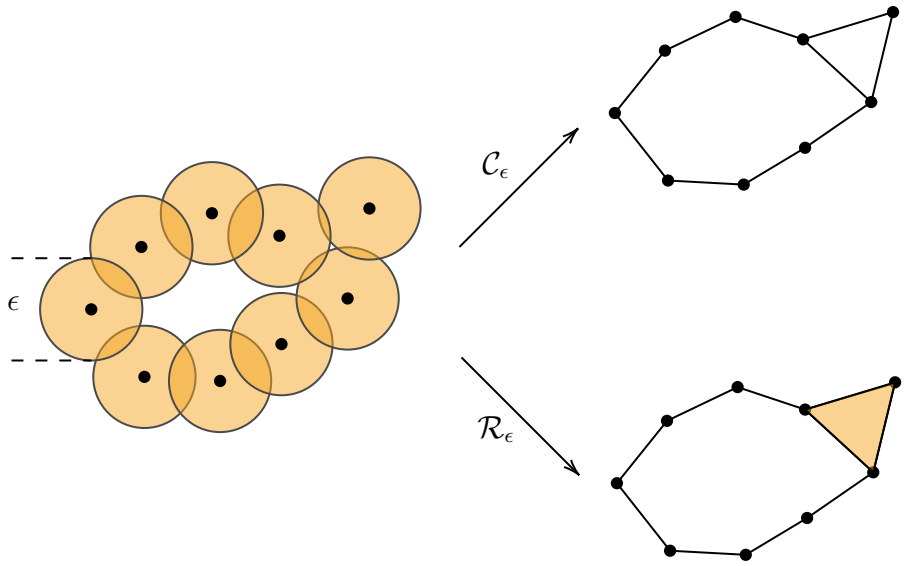

Figure 3: Example Čech and Rips complexes of a set of points in $\mathbb{R}^2$.

complexes (in $\mathbb{R}^n$, so one may wish to first obtain a decent embedding of the data into Euclidean space).

4.2 Persistence

Working with Rips complexes for simplicity, we can associate a natural filtration to each complex in order of increasing ϵ, e.g.:

$$\varnothing \subseteq \mathcal{R}_{\epsilon_0} \subseteq \mathcal{R}_{\epsilon_1} \subseteq \cdots \subseteq \mathcal{R}_{\epsilon_{N-1}} \subseteq \mathcal{R}_{\epsilon_N}, \quad \epsilon_i \leq \epsilon_j \text{ for } i < j.$$

These filtrations come with natural *inclusion maps* $\mathcal{R}_{\epsilon_{k-1}} \hookrightarrow \mathcal{R}_{\epsilon_k}$, which take the simplices of $\mathcal{R}_{\epsilon_{k-1}}$ to their embeddings in the larger complex $\mathcal{R}_{\epsilon_k}$. These inclusion maps further induce maps $H_n(\mathcal{R}_{\epsilon_i}) \to H_n(\mathcal{R}_{\epsilon_j})$ on the homology groups, which take the generators of $H_n(\mathcal{R}_{\epsilon_i})$ to their representations in $H_n(\mathcal{R}_{\epsilon_j})$ if applicable. This map is not necessarily injective; generators can 'die', in which case they are sent to 0. We can thus concretely classify 'persistence', by dropping the dimension and considering the full homology of each complex, which we encode in a *persistence* complex.

Definition 4.4. Let $\mathsf{C} = \{C_*^i\}_1^N$ be a sequence of chain complexes with inclusion maps $C_k^n \hookrightarrow C_k^{n+1}$, and induced group homomorphisms $\phi_*^{i,j} : H_k(C_*^i) \to H_k(C_*^j)$. The (i,j)-*persistent homology* of C is $H_*^{i,j}(\mathsf{C}) = \operatorname{im} \phi_*^{i,j}$. Equivalently, $H_*^{i,j}(\mathsf{C}) = Z_*^i / (B_*^j \cap Z_*^i)$.

A consequence of this definition is a natural extension of the standard Betti numbers.

Definition 4.5. The nth (i,j)-*persistent Betti number* is $\beta_n^{i,j} = \operatorname{rank} H_n^{i,j}$.

These parametrized Betti numbers record the number of holes that persist from the ith complex to the jth complex. However, if we wish to track entire lifetimes of homological features, we need to construct a new representation in the form of a *persistence diagram.* We can begin by counting the number of n-generators that are born at C^i_* which die at C^j_*, by

$$\mu^{i,j}_n = (\beta^{i,j-1}_n - \beta^{i,j}_n) - (\beta^{i-1,j-1}_n - \beta^{i-1,j}_n).$$

Then, considering i as the 'birth time' and j as the 'death time', we form the pair (i, j), which we insert into a multiset D_n with multiplicity $\mu^{i,j}_n$.[4] Taking $\mathbb{R}^2_\Delta = \{(x, y) \in \mathbb{R}^2 : x = y\}$ and $\mathcal{D}_\Delta = \{\mathbf{x} \in \mathbb{R}^2_\Delta : \text{mult}(\mathbf{x}) = \infty\}$, we can form the multiset $\mathcal{D}_n = D_n \cup \mathcal{D}_\Delta$. A multiset of this form is a persistence diagram, and ranging over all suitable $i < j$ encodes the persistence data of $H^{0,N}_n(\mathsf{C})$.

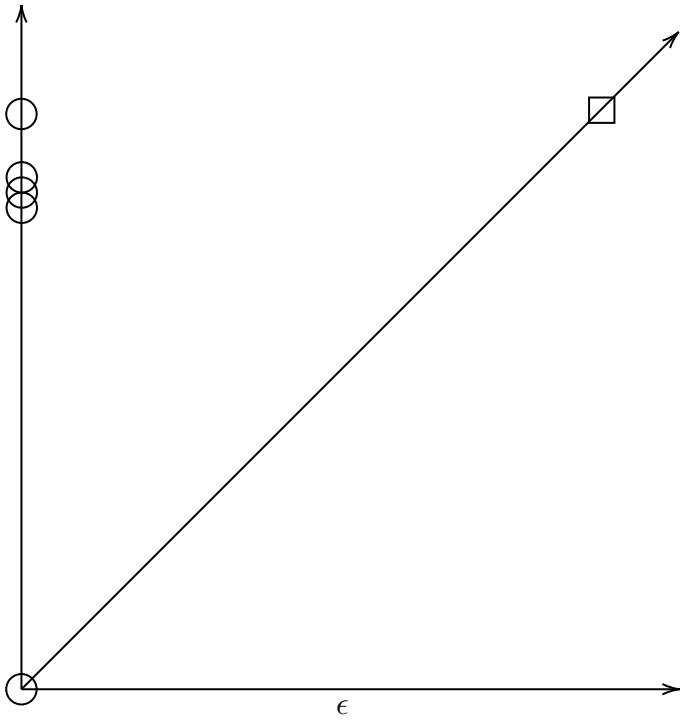

Figure 4: A persistence diagram (where the circles represent pairs in D_0, and squares in D_1), perhaps corresponding to 5 points sampled from a circle. Note the unpaired circle, representing a persistent connected component, and the unpaired square, representing a possibly persistent missing area.

[4] A multiset is a regular set S, along with a function mult : $S \to \mathbb{N} \cup \{\infty\}$ that counts the number of occurrences mult(x) for each $x \in S$.

Persistence diagrams encapsulate the persistent homologies of simplicial filtrations, and allow for a qualitative topological analysis of point cloud datasets. They form a central object of study in persistent homology, and are stable under perturbations of the input dataset. We hope that this paper has provided the reader with sufficient background informations and constructions to begin understanding persistent homology at a deeper level.

References

[1] V. De Silva, R. Ghrist, et al. "Coverage in sensor networks via persistent homology". In: *Algebraic & Geometric Topology* 7.1 (2007), pp. 339–358.

[2] D. S. Dummit and R. M. Foote. *Abstract Algebra*. Third. John Wiley & Sons, Inc., 2004.

[3] H. Edelsbrunner and J. Harer. "Persistent Homology — a Survey". In: *Contemporary Mathematics* 453 (2008), pp. 257–282.

[4] Edelsbrunner, Letscher, and Zomorodian. "Topological Persistence and Simplification". In: *Discrete & Computational Geometry* 28.4 (Nov. 2002), pp. 511–533. ISSN: 1432-0444. DOI: 10.1007/s00454-002-2885-2. URL: https://doi.org/10.1007/s00454-002-2885-2.

[5] R. Ghrist. "Barcodes: the persistent topology of data". In: *Bulletin of the American Mathematical Society* 45.1 (2008), pp. 61–75.

[6] A. Hatcher. *Algebraic Topology*. 2001. URL: http://www.math.cornell.edu/~hatcher/AT/AT.pdf.

[7] C. Hofer et al. "Deep Learning with Topological Signatures". In: *Advances in Neural Information Processing Systems*. 2017, pp. 1633–1643.

[8] J. J. Rotman. *An Introduction to Algebraic Topology*. Graduate Texts in Mathematics. Springer New York, 1988.