University of Tehran

College of Science

School of Mathematics, Statistics, and Computer Science

Topological Data Analysis

By

Farid Hazratian

Supervisor

Dr. Reza Rezavand

Advisor

Dr. Ali Kamalinejad

A thesis submitted to the graduate studies office

in partial fulfillment of the requirements for the degree of Master of Science in Mathematics and Applications.

September 2024

Attention

My original thesis was written in Persian, and I translated it within two days for the purpose of my PhD application for your review. I apologize in advance for any syntactical or grammatical mistakes, as I am currently reviewing and editing the translated text.

Declaration of Originality

I, Farid Hazratian, declare that this thesis is the result of my independent research and investigation. All sources and references used in this research are properly cited. This thesis has not been previously submitted for any degree. If any form of academic dishonesty is discovered, my academic degree will be revoked.

Farid Hazratian

Signature:

Intellectual Property Rights

Contents

List of Figures

Abstract

This thesis examines a key algorithm in topological data analysis. Persistent homology is a fundamental algorithm that provides a robust framework for understanding the shape of data beyond traditional linear methods. In this thesis, we demonstrate that the persistent homology of a refined simplicial complex of dimension d is simply the standard homology of a graded module over a polynomial ring. This research enables us to study a standard algorithm for computing the persistent homology of spaces in arbitrary dimensions and over any field.

Persistent homology analyzes how topological features change across different scales. By refining spaces and tracking the birth and death of features, persistent homology provides a multiscale summary of the topology of the data. The stability theorem, which guarantees that these features are robust against small perturbations in the data, will also be thoroughly examined and studied. Finally, we will explore and explain the mathematical foundations of another algorithm known as zigzag persistence. Zigzag persistence is a generalization of standard persistent homology that allows for more complex sequences of complexes to be studied. While in standard persistent homology, spaces are mapped to each other in a one-directional manner, in zigzag persistence, sequences involving bidirectional (in-and-out) mappings are possible.

Keywords: Simplicial complex, filtration, persistent homology, Betti number, persistence diagram, zigzag persistence.

Chapter 1

Preliminary Concepts

1.1 Introduction and Overview

Topological data analysis is a branch of mathematics that, using tools from algebra, topology, and statistics, analyzes and examines the topological structures of datasets. Simply put, the idea is to assign topological invariants to data. The data itself is represented as a discrete sample called a point cloud, whose topology is not particularly interesting to us. Therefore, we transform the data into a continuous object that, from a topological perspective, resembles the geometric background shape from which the data was sampled. On the other hand, the obtained continuous spaces need to be discretized to be studied using computers, and this discretization process is carried out using simplicial complexesSimplicial Complexes.

At the heart of topological data analysis, we aim to understand data through topological properties and invariants such as connected components, loops, and holes, which remain invariant under continuous transformations. Thus, the ability to extract meaningful insights from complex datasets highlights the importance of this field in the era of big data. Here, complex data refers to high-dimensional data with nonlinear patterns, noise, incompleteness, and even missing values.

One of the fundamental tools in this domain is persistent homologyPersistent Homology, which studies the multi-scale topological properties of spaces. Persistent homology examines the birth and death of topological features at multiple scales and provides a detailed summary of the changes in topological properties visually.

On the other hand, algebraic topology investigates topological spaces using abstract algebra. Results such as Brouwer's fixed-point theorem, the Brouwer–Jordan separation theorem, the Borsuk–Ulam theorem, and even the emergence of category theory are classical examples of achievements in this field. Mathematicians continue to pursue research in this area, with a contemporary example being the Poincaré conjecture, which was proven in 2006. The main algebraic tool we focus on here is known as homology algebra. In this section, we will review the prerequisites of this field by discussing concepts and theorems from algebraic topology. The concepts used in this section are extracted from the references [7], [8], and [5].

1.2 Simplicial Complexes

Consider the points $p_0, p_1, \ldots, p_m \in \mathbb{R}^n$. The notation $[p_0, p_1, \ldots, p_m]$ represents the convex hull of these points, meaning the set of all convex combinations of the points $p_0, p_1, \ldots, p_m$. Furthermore, we say that the points $(p_0, p_1, \ldots, p_m)$ are affinely independent if the set $\{p_1 - p_0, \ldots, p_m - p_0\}$ in $\mathbb{R}^n$ is linearly independent. Equivalently, this means that every element x in the affine set spanned by $\{p_0, p_1, \ldots, p_m\} \in \mathbb{R}^n$ has a unique representation as an affine combination:

$$x = \sum_{i=\theta}^{m} t_i p_i, \quad \sum_{i=\theta}^{m} t_i = 1 \qquad (t_i \in \mathbb{R}).$$

Specifically, this means that affine independence is not affected by the ordering of the points.

The $-m+1$ tuple $(t_0, \ldots, t_m)$ is called the barycentric coordinatesBarycentric Coordinates

of x with respect to $(p_0, \ldots, p_m)$.

Definition 1.2.1. Suppose $(p_0, p_1, \ldots, p_m) \in \mathbb{R}^n$ are affinely independent. The convex hull $s = [p_0, \ldots, p_m]$ is called an m-simplexSimplex with vertices $Vert(s) = \{p_0, \ldots, p_m\}$. Its dimension is m, and the face opposite to p_i is defined as

$$[p_0, \ldots, \hat{p}_i, \ldots, p_m] = \left\{ \sum_{j=0}^{m} t_j p_j | \sum t_j = 1,\ t_j \geq 0, t_i = 0 \right\}$$

for all $i = 0, \ldots, m$.

More generally, a simplex s' is called a face of s if $Vert(s') \subseteq Vert(s)$, which we denote as $s' \leq s$.

If $Vert(s') \nsubseteq Vert(s)$, then s' is called a proper face, denoted by $s' < s$.

Remark 1.2.1. The set

$$\Delta^n = \left\{ (x_1, x_2, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} | x_i \geq 0, \sum_{i=1}^{n+1} x_i = 1 \right\}$$

is called the standard n-simplex.

Definition 1.2.2. A simplicial complex K is a finite collection of simplices in Euclidean space such that for all $s, t \in K$:

1. Every face of s is also in K.

2. The intersection of s and t is either empty or a common face of both.

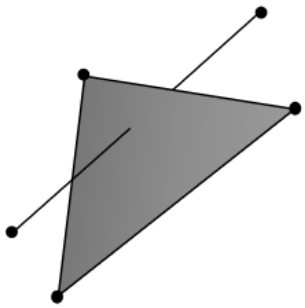
The underlying space $|K|$ is defined as

$$|K| = \bigcup_{s \in K} s.$$

The dimension of K is given by

$$\dim K = \max_{s \in K}(\dim s).$$

Example 1.2.1. A finite undirected simple graph $G = (V, E)$ forms a 1-dimensional simplicial complex with $V = K_0$ and $E = K_1$.



Example 1.2.2. A torus can be triangulated as shown in the figure.

Example 1.2.3. Define the equivalence relation $\sim$ on the Cartesian product $[0, 1]^2$ for each $t \in [0, 1]$ such that $(t, 0)$ is identified with $(t, 1)$ and $(0, t)$ is identified with $(1, t)$. The quotient space

$$\mathbb{T}^2 = [0, 1] \times [0, 1]/\sim$$

is endowed with the topology induced by this construction. Since the function

$$\phi : \mathbb{T}^2 \to \mathbb{S}^1 \times \mathbb{S}^1 : \overline{(s, t)} \to ((\sin(2\pi s), \cos(2\pi s)), (\sin(2\pi t), \cos(2\pi t)))$$

is a well-defined homeomorphism, where $\overline{(s, t)}$ represents the equivalence class generated by (s, t) and $\mathbb{S}^n$ denotes the n-dimensional sphere, it follows that $\mathbb{T}^2$ is a torus.

In Figure 1.1, we see a triangulation of the torus.

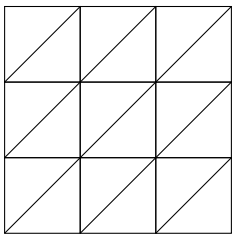


Figure 1.1: Triangulation of the torus

Given the necessity of working with more abstract spaces beyond Euclidean space, we now define abstract simplicial complexesAbstract Simplicial Complex.

Definition 1.2.3. Suppose V is a finite set. An abstract simplicial complex K is a nonempty family of subsets of V, called simplices, such that:

1. If $v \in V$, then $\{v\} \in K$.

2. If $s \in K$ and $t \subseteq s$, then $t \in K$.

V is called the vertex set of K, and we denote it by $Vert(K)$. A subset of K that is itself an abstract simplicial complex is called a subcomplex.

Remark 1.2.2. Clearly, every simplicial complex is also an abstract simplicial complex.

Definition 1.2.4. (Geometric Realization of Simplicial ComplexesGeometric Realization)

Consider the function $\phi : K_0 \to \mathbb{R}^n$ (note that K_0 is the set of vertices of K). The geometric realization of K with respect to ϕ is defined as $|K|_\phi = \bigcup_{\sigma \in K} |\sigma|_\phi$, where for each $\sigma = \{v_0, \ldots, v_k\}$ in K, the set $|\sigma|_\phi \subset \mathbb{R}^n$ is a geometric simplex spanned by the points $\{\phi(v_0), \ldots, \phi(v_k)\}$.

Geometric realization allows us to see beyond the combinatorial aspects of simplicial complexes and explore their geometric and topological structures. This provides a justification for visualizing simplices in dimensions $0, 1, 2, \ldots$ as points, lines, pyramids, triangles, etc.

Lemma 1.2.1. An abstract simplicial complex K of dimension d has a geometric realization in $\mathbb{R}^{2d+1}$.

Proof. Reference [5]. □

Lemma 1.2.2. (Helly's Theorem) Suppose A is a finite family of convex and closed sets in $\mathbb{R}^d$. If the intersection of every $d+1$ sets in the family is nonempty, then the entire family has a nonempty intersection.

Proof. Reference [5]. □

Lemma 1.2.3. (Jung's Theorem) Suppose A is a finite set of points in $\mathbb{R}^d$. Then A is contained in a closed ball of radius $r \leq \delta \cdot \text{diam}(A)$, where $\delta = \sqrt{\frac{d}{2d+1}}$.

Proof. Reference [5]. □

Definition 1.2.5. (Oriented Simplicial Complex)

A simplicial complex K is called oriented if there exists a partial order on $Vert(K)$ such that its restriction to the vertices of each simplex in K is a linear order.

Definition 1.2.6. (n-Chains)

Suppose K is an oriented simplicial complex. For $n \in \mathbb{N}$, $C_n(K)$ is the abelian group generated by $-(n+1)$ tuples $(p_0, \ldots, p_n)$ such that for all $i \in \{0, \ldots, n\}$, $p_i \in V$, where $\{p_0, \ldots, p_n\}$ forms a simplex in K, satisfying the following conditions:

1. $(p_0, \ldots, p_n) = 0$ if $p_i = p_j$ for $i \neq j$.

2. $(p_0, \ldots, p_n) = sign(\sigma)(p_{\sigma(0)}, \ldots, p_{\sigma(n)})$ where σ is a permutation of $\{0, \ldots, n\}$.

Lemma 1.2.4. Let K be an oriented simplicial complex of dimension m.

1. $C_n(K)$ is a free abelian group with basis $< p_0, \ldots, p_n >$, where $\{p_0, \ldots, p_n\}$ forms an n-simplex in K and $p_0 < p_1 < \cdots < p_n$. Furthermore,

$$sign(\sigma) < p_0, \ldots, p_n > = < p_{\sigma(0)}, \ldots, p_{\sigma(n)} > .$$

2. $C_n(K) = 0$ for all $n > m$.

Proof. Reference [8]. □

Remark 1.2.3. Specifically, the above lemma implies that $C_n(K)$ is a free $\mathbb{Z}$-module.

Definition 1.2.7. (Boundary Operator)

Suppose K is an oriented simplicial complex. For each $n \in \mathbb{N}_{\geq 1}$, the nth boundary operator is defined as $\partial_n : C_n(K) \to C_{n-1}(K)$ by:

$$\partial_n(< p_0, \ldots, p_n >) = \sum_{i=0}^{n} (-1)^i < p_0, \ldots, \hat{p}_i, \ldots, p_n > .$$

and is extended linearly. Additionally, ∂_0 is defined as the zero map on $C_0(K)$.

Theorem 1.2.1. Let K be an oriented simplicial complex of dimension n. Then

$$0 \to C_n(K) \xrightarrow{\partial_n} \cdots \xrightarrow{\partial_2} C_1(K) \xrightarrow{\partial_1} C_0(K) \to 0$$

forms a chain complex, denoted by (C_*, ∂). In other words, for all $k \in \mathbb{Z}$, we have $\partial_k \circ \partial_{k+1} = 0$.

Proof. Reference [7]. □

Remark 1.2.4. Note that the above statement is equivalent to $Im\partial_{k+1} \subseteq \ker \partial_k$.

Definition 1.2.8. (Simplicial Homology)

Suppose K is an oriented simplicial complex and $n \in \mathbb{N}$. Then:

$$Z_n(K) = \ker \partial_n$$

is called the group of n-cycles, and

$$B_n(K) = Im\partial_{n+1}$$

is called the group of n-boundaries, and

$$H_n(K) = Z_n(K)/B_n(K)$$

is the nth homology group. Additionally, the nth Betti number is defined as $\beta_n = \operatorname{rank} H_n(K)$.

Remark 1.2.5. The nth Betti number counts the number of $(n+1)$-dimensional holes. For example, β_0 counts the number of connected components of the space.

While we primarily deal with simplicial complexes, it is possible to define homology for any topological space X.

Definition 1.2.9. Consider a topological space X:

1. A continuous map $\sigma : \Delta^n \to X$ is called an n-singular simplex in X.

2. The group of n-singular chains, $C_n(X)$, is the free abelian group generated by n-singular simplices in X, and its elements are called n-singular chains in X.

3. The nth singular boundary operator ∂ is defined as:

$$\partial = \sum_{i=0}^{n} (-1)^i d_i \colon C_n(X) \to C_{n-1}(X).$$

4. The abelian group

$$Z_n(X) = \ker(\partial \colon C_n(X) \to C_{n-1}(X))$$

is called the group of n-singular cycles in X, and the abelian group

$$B_n(X) = Im(\partial \colon C_n(X) \to C_{n-1}(X))$$

is called the group of n-singular boundaries in X.

5. The nth singular homology group is defined as

$$H_n(X) = Z_n(X)/B_n(X).$$

The definition of the homology group $H_n(X)$ is the same as in Definition 1.2.8. It can be shown that for a simplicial complex K,

$$H_n(K) = H_n(|K|)(\forall n \geq 0).$$

For more technical details, refer to [8].

This is why the homology of a polyhedron can be computed by calculating the simplicial homology groups of one of its triangulations. Moreover, this result shows that simplicial homology groups are independent of the partial order of $Vert(K)$.

1.3 Homotopy

Definition 1.3.1. A homotopy between two maps $f, g: X \to Y$ is a map $F: X \times [0,1] \to Y$ such that $F(x,0) = f(x)$ and $F(x,1) = g(x)$ for all $x \in X$. ($X \times [0,1]$ is equipped with the product topology.)

If a homotopy exists between f and g, we say that f and g are homotopic and denote this by $f \simeq g$.

Note: Homotopy defines an equivalence relation.

Definition 1.3.2. Two topological spaces (X, A) and (Y, B) are said to be homotopy equivalent if there exist maps $g: (Y, B) \to (X, A)$ and $f: (X, A) \to (Y, B)$ such that $g \circ f \simeq id_{(X,B)}$ and

$f \circ g \simeq id_{(Y,B)}$.

Theorem 1.3.1. If $f \simeq g\colon |K| \to |L|$, then $f_* = g_*\colon H_n(K) \to H_n(L)$.

Proof. Reference [8]. □

Theorem 1.3.2. If $f\colon |K| \to |L|$ is a homotopy equivalence between polyhedra, then

$$f_* : H_n(K) \to H_n(L)$$

is an isomorphism.

Proof. Reference [8]. □

- A special example of homotopy equivalence is given by the **Nerve Theorem**.

Definition 1.3.3. Suppose F is a finite family of convex and closed sets in Euclidean space. The **nerve** of F is defined as the abstract simplicial complex

$$N(F) := \left\{ G \subseteq F : \bigcap_{F_i \in G} F_i \neq \emptyset \right\}.$$

Example 1.3.1. Consider the set $F = \{F_1, F_2, F_3\}$ shown in Figure 1.2. We have

$$N(F) = \{\{F_1\}, \{F_2\}, \{F_3\}, \{F_1, F_2\}, \{F_2, F_3\}, \{F_1, F_3\}\}.$$

which is isomorphic to the abstract simplicial complex corresponding to the boundary of a 2-simplex. Clearly,

$$F_1 \cup F_2 \cup F_3 \simeq |N(F)|.$$

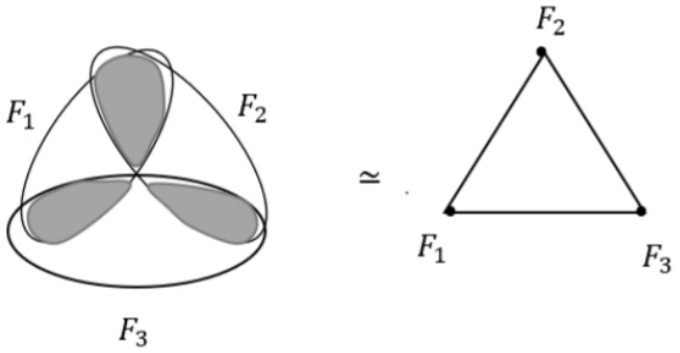


Figure 1.2: The nerve of the set F, which is isomorphic to the boundary of a 2-simplex.

Theorem 1.3.3. (Nerve TheoremNerve Theorem)

1. Let $F = \{F_1, \ldots, F_m\}$ be a family of convex and closed sets in Euclidean space. Then,

$$\bigcup_{F_i \in F} F_i \simeq |N(F)|.$$

2. Consider another family $F' = \{F'_1, \ldots, F'_m\}$ of closed and convex sets such that $F_i \subseteq F'_i$ for each i.

   Let

$$j\colon \bigcup_{F_i \in F} F_i \hookrightarrow \bigcup_{F'_i \in F} F'_i$$

   be the inclusion map and $\sigma\colon N(F) \to N(F')$ the simplicial map sending F_i to F'_i. Then the following diagram commutes:

$$\begin{array}{ccc} H_n\left(\bigcup_{F_i \in F} F_i\right) & \xrightarrow{j_*} & H_n\left(\bigcup_{F'_i \in F} F'_i\right) \\ \cong\downarrow & & \downarrow\cong \\ H_n(N(F)) & \xrightarrow{\sigma_*} & H_n(N(F')). \end{array}$$

Theorem 1.3.4. (Euler-Poincaré Formula)

The Euler characteristicEuler characteristics of a simplicial complex K is the integer $\chi(K)$ defined by:

$$\chi(K) = \sum_{i=0}^{\infty}(-1)^i K_i = \sum_{i=0}^{\infty}(-1)^i \beta_i(K).$$

Proof. Reference [7]. □

Example 1.3.2. The projective planeProjective Plane $\mathbb{R}P^2$ is a closed, non-orientable surface that cannot be embeddedEmbedded in $\mathbb{R}^3$. A triangulation of $\mathbb{R}P^2$ is shown in Figure 1.3. For any field $\mathbb{F} = \mathbb{Z}_2$ and $\mathbb{F} = \mathbb{Z}_3$, we have:

$$\begin{aligned} \beta_0(\mathbb{R}P^2; \mathbb{Z}_3) &= 1 \qquad & \beta_0(\mathbb{R}P^2; \mathbb{Z}_2) &= 1 \\ \beta_1(\mathbb{R}P^2; \mathbb{Z}_3) &= 0 \qquad & \beta_1(\mathbb{R}P^2; \mathbb{Z}_2) &= 1 \\ \beta_2(\mathbb{R}P^2; \mathbb{Z}_3) &= 0 \qquad & \beta_2(\mathbb{R}P^2; \mathbb{Z}_2) &= 1. \end{aligned}$$

Clearly, the Euler characteristic is independent of the choice of the field.

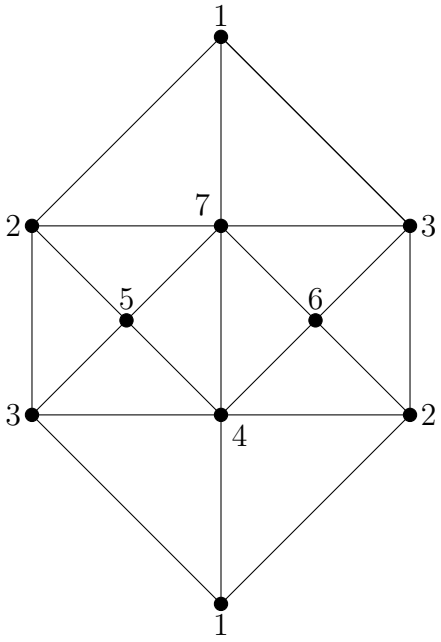


Figure 1.3: Triangulation of the projective plane $\mathbb{R}P^2$

1.4 Special Simplicial Complexes

In real-world scenarios, we often deal with a dataset that is a subset of Euclidean space: $X \subset \mathbb{R}$, where X is finite. Such finite point clouds do not inherently possess an interesting topology for study. For example, the zeroth Betti number represents the number of connected components, which, in the case of a point cloud, is simply the number of data points. All other Betti numbers are zero.

When analyzing the topology of data, we think of the point cloud (the set X) as a sample of the underlying continuous space. Understanding the topology of this continuous space provides valuable intuition about the point cloud.

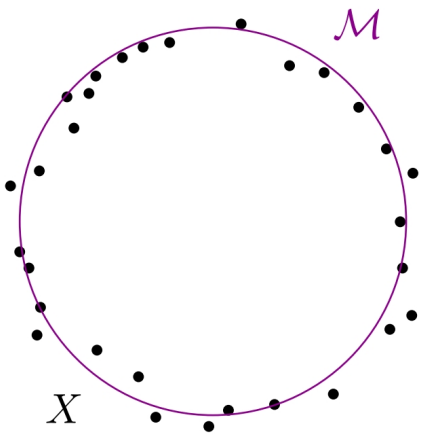


Figure 1.4: An example of a point cloud sampled from a circle

For example, in Figure 1.4, X is a sample from a circle, and M represents the topology of the underlying space. Our goal is to understand this topology, which is the fundamental idea of topological data analysis.

But how can we infer the topology of X when we only have access to X itself?

More formally, consider M as a bounded subset of Euclidean space, and assume we have a finite sample X.

Question: Estimate the homology groups of M from the dataset X.

Idea: We thicken X.

Definition 1.4.1. For each $t \geq 0$, the t-thickening of the set X, denoted by X^t, is defined as the set of points in the ambient spaceambient space that are at most a distance t from X:

$$X^t = \{y \in \mathbb{R}^n \mid \exists x \in X : ||x - y|| \leq t\}$$

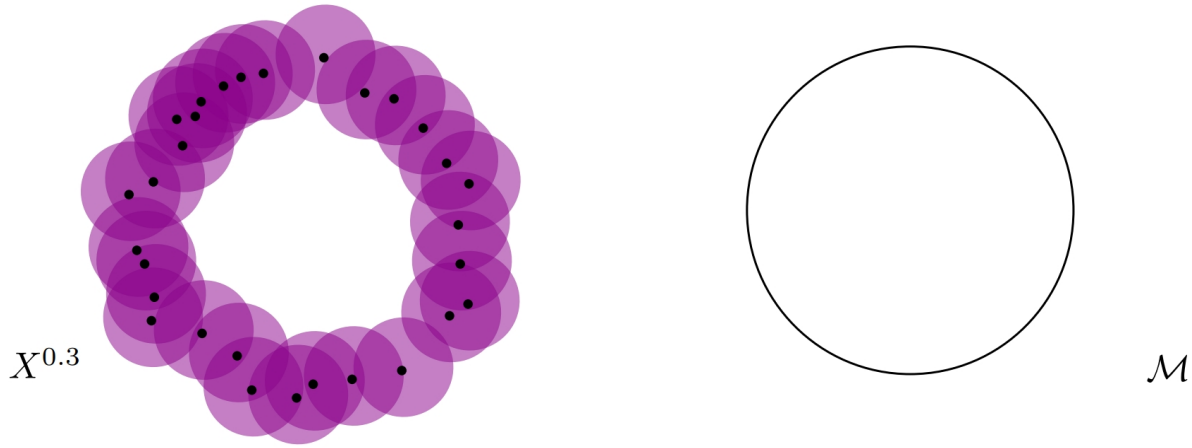


Figure 1.5: Thickening of a point cloud

In Figure 1.5, we observe that at a certain t value, such as $X^{0.3}$, we obtain a subset of Euclidean space that is homotopy equivalent to a circle. Consequently, we can compute the homology of the circle. Now, we face two key questions:

1. How should we choose t so that $X^t \simeq M$?
2. How can we compute the homology groups of X^t?

Definition 1.4.2. The Hausdorff distance between two subsets X and Y of $\mathbb{R}^n$ is defined as:

$$d_H(X,Y) = \inf\{t \geq 0 \mid X \subset Y^t, Y \subset X^t\}$$

For an example of the Hausdorff distance applied to a point sample from a circle, see Figure 1.6.

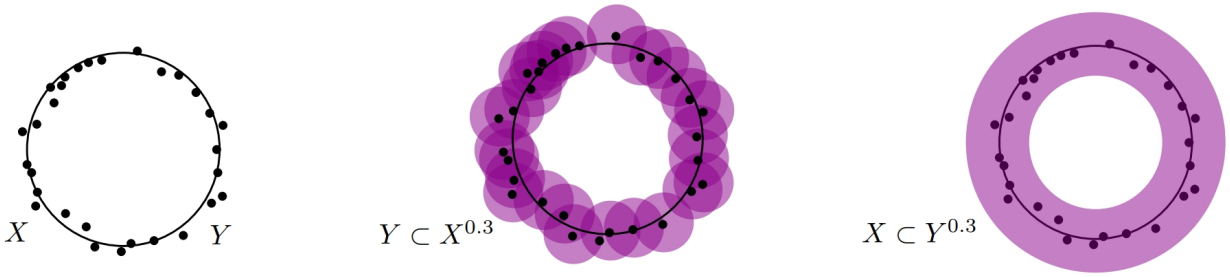


Figure 1.6: Hausdorff distance

Definition 1.4.3. (Voronoi DiagramVoronoi Diagram)

For $p \in S \subset \mathbb{R}^d$, its Voronoi cellVoronoi Cell/Region V_p is defined as:

$$V_p = \{x \in \mathbb{R}^d \mid ||x - p|| \leq ||x - q||, \forall q \in S\}$$

The points satisfying $||x-p|| \leq ||x-q||$ form a closed half-space, and V_p is the intersection of a finite number of such half-spaces. Therefore, it is a convex polygon. Voronoi cells share at most their boundaries and together cover the entire space. The collection of all these Voronoi cells forms the Voronoi diagram of S, as shown in Figure 1.7.

Remark 1.4.1. The Voronoi ball of each p is defined as the intersection of the Voronoi cell with a closed ball of radius r centered at that point.

Simply put, imagine having balls of radius r centered at each point and gradually increasing their radii. At some point, all the balls intersect, forming a structure where the entire space appears as a union of circles. However, in Voronoi diagrams, neighboring circles do not merge. Instead, the circles in the given set begin to touch at a single point (similar to the number 8 in Latin). As they grow further, they seem to squeeze against each other and transform into a line, forming a pattern like Figure 1.7.

Additionally, the boundary of each Voronoi cell is equidistant from two nearest points in the set, and whenever three such lines meet, they form a vertex that is equidistant from three points in the set.

Definition 1.4.4. (Delaunay ComplexDelaunay Complex)

The Delaunay complex of a set $S \subset \mathbb{R}^d$ is homeomorphic to the nerve of the family of Voronoi cells,

$$Delaunay = \left\{ \sigma \subseteq S | \bigcap_{P \in \sigma} V_P \neq \emptyset \right\}$$

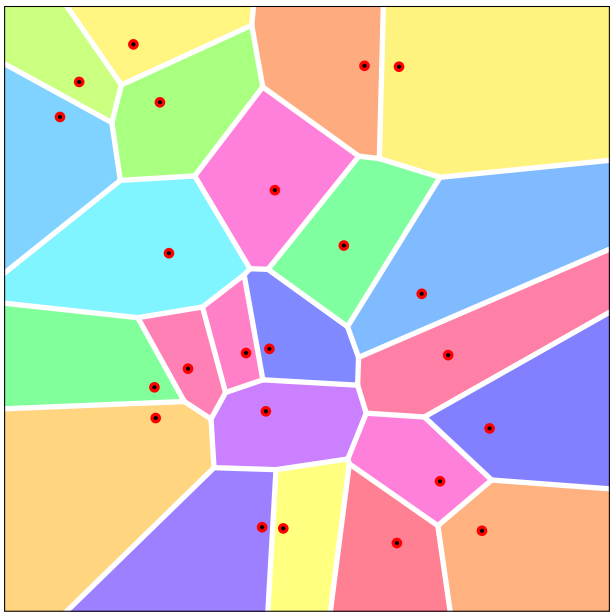


Figure 1.7: Voronoi diagram for a finite set S of points in $\mathbb{R}^2$.

Definition 1.4.5. Suppose $B_r(P), P = \{p_0, \ldots, p_n\} \subset \mathbb{R}^d$ represents the closed ball centered at P with radius r,

1. The Čech complexČech Complex of the set P is defined as:

$$Cech_r(P) = \left\{ \sigma \subseteq P : \bigcap_{P \in \sigma} B_r(P) \neq \emptyset \right\}$$

2. The Vietoris–Rips complexVietoris–Rips complex of the set P is defined as:

$$VR_r(P) = \{\sigma \subseteq P : diam(\sigma) \leq 2r\}$$

Recall that the diameter of a subset $\sigma \subseteq P$ is given by:

$$diam(\sigma) = \max_{P_i, P_j \in P} ||P_i - P_j||$$

Definition 1.4.6. (Filtration) A filtration of a simplicial complex K is a family of sub-

complexes $\{K^i \subseteq K\}_{i=0}^m$ such that

$$K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K.$$

For readability, instead of (S_*, ∂), we use S_* for the given complex. Suppose the simplicial complex K is equipped with a filtration $(K^i)_{i\in\mathbb{N}}$. Then, for each i, we have the corresponding chain complex and homology. Thus, K_n^i, Z_n^i, B_n^i, and H_n^i represent the n-th chain, cycle, boundary, and homology of K^i, respectively. Similarly, we denote the corresponding Betti number by β_n^i.

1.5 Filtration of Čech and Vietoris-Rips Complexes

Čech Filtration

Consider $P \subset \mathbb{R}^d$. By the Nerve Theorem, we have the following diagram of vector spaces and linear maps:

$$\begin{array}{ccccccc}
H_n(P_{r_0}) & \longrightarrow & H_n(P_{r_1}) & \longrightarrow & \cdots & \longrightarrow & H_n(P_{r_m}) \\
\downarrow\cong & & \downarrow\cong & & \downarrow\cong & & \downarrow\cong \\
H_n(\mathrm{Cech}_{r_0}(P)) & \longrightarrow & H_n(\mathrm{Cech}_{r_1}(P)) & \longrightarrow & \cdots & \longrightarrow & H_n(\mathrm{Cech}_{r_m}(P))
\end{array}$$

It is straightforward to show that the persistent Betti numbers of both rows are equal. Therefore, the barcode and persistence diagram of the Čech Filtration have a clear geometric interpretation: they describe the evolution of holes at a given dimension as the balls around points in P grow.

Vietoris-Rips Filtration

For a given radius r, it is not necessarily true that $VR_r(P) \cong P_r$, so interpreting the

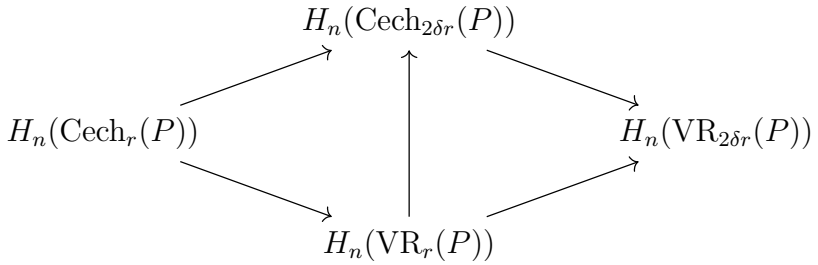
persistent Betti numbers associated with a Filtration of the form:

$$VR_{r_0}(P) \hookrightarrow VR_{r_1}(P) \hookrightarrow VR_{r_2}(P) \hookrightarrow \cdots \hookrightarrow VR_{r_m}(P).$$

is not always straightforward. Thus, from the results of the simplicial complex section, we obtain the inclusion chain:

$$\text{Cech}_r(P) \subseteq \text{VR}_r(P) \subseteq \text{Cech}_{2\delta r}(P) \subseteq \text{VR}_{2\delta r}(P).$$

where $\delta = \sqrt{\frac{d}{2(d+1)}}$. Applying homology yields the following commutative diagram of vector spaces:



The above diagram shows that any class in $H_n(VR_r(P))$ that has a nonzero image in $H_n(VR_{2\delta r}(P))$ must also be nonzero in $H_n(Cech_{2\delta r}(P))$. The converse also holds. In summary, sufficiently long intervals in the barcode of the Čech Filtration lead to intervals in the barcode of the Vietoris-Rips Filtration, and vice versa. These relationships will be made more precise in the next chapter when we discuss embedding.

Why should we consider Vietoris-Rips complexes?

There are two main reasons:

1. Constructing a Čech complex requires the coordinates of points in $\mathbb{R}^d$, while Vietoris-Rips complexes are built solely from pairwise distances. This allows us to construct Filtration for simplicial complexes even when points are not embedded in $\mathbb{R}^d$.

2. Using Vietoris-Rips complexes enables heuristic algorithmsHeuristic algorithm that significantly speed up computations in practice.

1.6 Computing Homology Groups

Suppose K is a simplicial complex. Clearly, $H_n(K)$ is finitely generated for each n and is thus fully classified by its rank and torsion coefficients using the **Structure Theorem**.

Theorem 1.6.1. (Structure Theorem)

Suppose G is a finitely generated abelian group. Then:

1. There exist a free abelian group F of finite rank $r \in \mathbb{N}$ and a finite group T such that:

$$G \cong F \oplus T.$$

We call F the **free part** and T the **torsion part**Torsion of G.

2. There exist unique cyclic groups $C_1, \dots C_k$ such that for $b_i = |C_i|$ with $i = 1, \dots, k$, the divisibility condition $b_1|b_2|...|b_k$ holds, and:

$$T = \bigoplus_{i=1}^{k} C_i.$$

The numbers $b_1, b_2, \dots, b_k$ are called the **torsion coefficients** of G.

3. Two finitely generated abelian groups are isomorphic if and only if they have the same rank and the same torsion coefficients.

Proof. Reference [9]. □

Remark 1.6.1. Let R be a **principal ideal domain (PID)**, and let M be a finitely generated R-module of rank r. Then, the theorem above takes the following form.

There exist elements $b_1, b_2, \ldots, b_k \in R$ such that $b_1|b_2|\ldots|b_k$, where:

$$M \cong R^{r-k} \times R/< b_1 > \times \ldots \times R/< b_k > .$$

Again, $k \in \mathbb{N}$ is uniquely determined, and the torsion coefficients $b_1, \ldots, b_k$ are unique up to multiplication by units in R.

To compute the above decomposition, we use the well-known **Smith normal form** of matrices. Recall that if R is a **Euclidean domain**, then for any matrix A over this domain, there exist invertible matrices P and Q such that $A = PSQ$, where S has the form:

$$S = \left[\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right]$$

with $D = \text{diag}(b_1, \ldots, b_{rank(A)})$ and $b_1|b_2|\ldots|b_{rank(A)}$. The matrix S is called the **Smith normal form** of A.

Theorem 1.6.2. For any oriented simplicial complex K, there exists an algorithm to compute its homology groups.

Proof. Reference [12], Section 3, Page 60. □

Remark 1.6.2. The algorithm works as follows: Each chain group $C_n(K)$ is finitely generated, and thus each boundary operator ∂_n corresponds to a matrix M_n with entries in $\{-1, 0, 1\}$.

Now, let S_n be the Smith normal form of the matrix M_n, let r_n denote the number of nonzero rows in S_n, and let c_n be the number of zero columns in S_n. Then, the elementary divisors of M_n correspond to the torsion coefficients of $H_n(K)$, and we have:

$$\beta_n = \text{rank}\, H_n(K) = c_n - r_{n+1}.$$

The last equality follows from the fact that $C_n(K)$ is a free $\mathbb{Z}$-module. Since $\mathbb{Z}$ is a principal ideal domain, its submodules are also free. Therefore, both Z_n and B_{n+1} are free modules. Applying the **rank-nullity theorem** to the map $Z_n \to Z_n/B_{n+1}$ gives:

$$\operatorname{rank} H_n(K) = \operatorname{rank} Z_n - \operatorname{rank} B_{n+1},$$

which corresponds to the previously stated result.

Chapter 2

Topological Data Analysis

2.1 Introduction to Persistent Homology

The goal of topological data analysis is to understand the geometric structure of a finite set of points. The idea is to use given points as the vertices of a simplicial complex and divide it into a family of ascending subcomplexes. This process allows us to measure the persistence of topological features relative to the subcomplexes, enabling a topological classification of the data.

2.2 Persistent Homology

This chapter begins by constructing the fundamental tool of topological data analysis: persistent homology. The material in this chapter is adapted from sources [12], [5].

Definition 2.2.1. (Subcomplex)

Suppose (S_*, ∂) is a complex. Then, a complex (S'_*, ∂') is a subcomplex of (S_*, ∂) if the following diagram commutes for every $n \in \mathbb{Z}$:

$$\begin{array}{ccc} S'_n & \xrightarrow{\partial'_n} & S'_{n-1} \\ \downarrow{\scriptstyle \eta^n} & & \downarrow{\scriptstyle \eta^{n-1}} \\ S_n & \xrightarrow{\partial_n} & S_{n-1} \end{array}$$

where η^n is an injective map from S'_n into S_n.

Definition 2.2.2. (Persistent Homology Module)

For $j \geq i$, we define the modules:

$$H_n^{i,j} = \frac{Z_n^i}{\left(B_n^j \cap Z_n^i\right)}$$

as the n-th persistent homology module. Additionally, the n-th persistent Betti number is defined as:

$$\beta_n^{i,j} = \text{ rank } H_n^{i,j}.$$

Remark 2.2.1. $H_n^{i,j}$ describes the n-cycles in K^i that are not the boundary of an $(n+1)$-chain in the larger complex K^j. Thus, $H_n^{i,j}$ identifies the $(n+1)$-dimensional holes in K^j that were formed in K^i. Note that these holes persist in all complexes K^l where:

$$i \leq l \leq j.$$

Definition 2.2.3. Consider the complexes (S^*, ∂) and (S'^*, ∂'). A chain of maps:

$$(f^n : \ S'_n \to S_n)_{n \in \mathbb{Z}}$$

is called a chain map if the following diagram commutes for every $n \in \mathbb{Z}$:

$$\begin{array}{ccccccccc} \cdots & \longrightarrow & S'_{n-1} & \xrightarrow{\partial'_{n+1}} & S'_n & \xrightarrow{\partial'_n} & S'_{n-1} & \longrightarrow & \cdots \\ & & \downarrow{\scriptstyle f^{n+1}} & & \downarrow{\scriptstyle f^n} & & \downarrow{\scriptstyle f^{n-1}} & & \\ \cdots & \longrightarrow & S_{n-1} & \xrightarrow{\partial_{n+1}} & S_n & \xrightarrow{\partial_n} & S_{n-1} & \longrightarrow & \cdots \end{array}$$

We briefly denote this as:

$$f = (f_n) : \ (S'^*, \partial') \to (S^*, \partial).$$

In the conditions above, we observe that for the complexes (K^{i+1}_*, ∂) and (K^i_*, ∂), the family of injective functions:

$$Inj^i : \ (K^i_*, \partial) \to \ (K^{i+1}_*, \partial)$$

forms a chain map inducing the maps:

$$\eta^i_n : \ H^i_n \ \to \ H^{i+1}_n$$

on homology groups. Now, we define a persistent complex.

Definition 2.2.4. (Persistent Complex) A sequence of complexes and chain maps (K^i_*, η^i) is called a persistent complex.

Remark 2.2.2.

The following diagram illustrates a part of a persistent complex, where each column represents a chain complex:

$$\begin{array}{ccccccccc}
 & & \vdots & & \vdots & & \vdots & & \\
 & & \downarrow{\scriptstyle \partial^i_{n+1}} & & \downarrow{\scriptstyle \partial^{i+1}_{n+1}} & & \downarrow{\scriptstyle \partial^{i+2}_{n+1}} & & \\
\cdots & \longrightarrow & K^i_n & \xrightarrow{\eta} & K^{i+1}_n & \xrightarrow{\eta} & K^{i+2}_n & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle \partial^i_n} & & \downarrow{\scriptstyle \partial^{i+1}_n} & & \downarrow{\scriptstyle \partial^{i+2}_n} & & \\
\cdots & \longrightarrow & K^i_{n-1} & \xrightarrow{\eta} & K^{i+1}_{n-1} & \xrightarrow{\eta} & K^{i+2}_{n-1} & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle \partial^i_{n-1}} & & \downarrow{\scriptstyle \partial^{i+1}_{n-1}} & & \downarrow{\scriptstyle \partial^{i+2}_{n-1}} & & \\
 & & \vdots & & \vdots & & \vdots & &
\end{array}$$

For $i \leq j$, we define:

$$\eta_n^{i,j} = \eta_n^{j-1,j} \circ \cdots \circ \eta_n^{i+1,i+2} \circ \eta_n^{i,i+1}$$

and thus:

$$H_n^{i,j} = \text{ Im } \eta_n^{i,j}.$$

2.3 Persistence Modules and Graded Rings

Definition 2.3.1. (Persistence Module) A family of homology modules H_n^i and maps η_n^i is called the n-th persistence module ($\mathcal{H}_n$). A persistence module is called of finite type if each module is finitely generated, and there exists an integer m such that for every $i \geq m$, η_n^i is an isomorphism.

Attention. 1) Given the assumption that all simplicial complexes we deal with are finite, all homology modules are of finite type.

2) Consider the function $f : X \to \mathbb{R}$ on an arbitrary topological space X. The n-th homology module $\mathcal{H}_n(f)$ corresponding to it can be defined as

$$\mathcal{H}_n^a = \mathcal{H}_n(f^{-1}((-\infty, a]))$$

for $a \in \mathbb{R}$, with the corresponding morphisms induced by inclusion.

A function f is called **tame**[1] if its corresponding homology module is constant and of finite dimension for all $a \in \mathbb{R}$. Hence, we always assume that f is tame.

2.4 Graded Modules and Rings

Our next goal in this section is to classify homology modules. We briefly recall the concept of graded rings and modules.

Definition 2.4.1. If for each $k \in \mathbb{Z}$, there exists an additive subgroup R_k such that

$$R = \bigoplus_{k \in \mathbb{Z}} R_k$$

and for all $k, \ell \in \mathbb{Z}$,

$$R_k \cdot R_\ell \subseteq R_{k+\ell}$$

then the ring R is called a **$\mathbb{Z}$-graded ring**.

Moreover, an element $a \in R_k$ is called **homogeneous**[2] of degree k, denoted as

$$\deg(a) = k.$$

Remark 2.4.1. From now on, for simplicity, we refer to a $\mathbb{Z}$-graded ring simply as a graded ring.

2.5 Decomposition of Homology Modules

Now, for an arbitrary field $\mathbb{F}$, we define a structure on $\mathcal{H}_n$ so that it becomes a graded module over the polynomial ring $\mathbb{F}[t]$. Here, the idea of chains, cycles, boundaries, and homology

[1] Tame Function
[2] Homogeneous

classes can be similarly defined for any ring or field other than $\mathbb{Z}$.

By definition, we have:

$$\mathcal{H}_n = \bigoplus_{i=0}^{\infty} H_n^i$$

and for an arbitrary variable t, multiplication is defined as:

$$t \cdot \left(\sum_{i=0}^{\infty} \xi^i \right) = \sum_{i=0}^{\infty} \eta_n^i(\xi^i)$$

where $\xi^i \in H_n^i$.

Clearly, $\mathcal{H}_n$ is a graded ring that is also finitely generated over $\mathbb{F}[t]$. Our goal is to classify homology modules using the **Structure Theorem**[3].

Let $\gamma_1, ..., \gamma_r$ be a finite generating system of homogeneous homology module elements with minimal principal number, and for $i = 1, ..., r$, set:

$$d_i = \deg(\gamma_i).$$

We define a grading on $\mathbb{F}[t]^r$ as an $\mathbb{F}[t]$-module:

$$\mathbb{F}[t]^r = \bigoplus_{\ell=0}^{\infty} \left(\bigoplus_{k=1}^{r} \mathbb{F}[t]_{\ell - d_k} \right)$$

Consider the following normal surjective map:

$$\phi : \mathbb{F}[t]^r \to \mathcal{H}_n : (p_1, ..., p_r) \mapsto \sum_{k=1}^{r} p_k \gamma_k$$

Assume $(p_1, ..., p_r)$ is homogeneous, meaning there exists $\ell \in \mathbb{N}$ such that for each k:

$$p_k \in \mathbb{F}[t]_{\ell - d_k}.$$

[3]Structure Theorem

Thus,

$$\phi(p_1, ..., p_r) = \sum_{k=1}^{r} p_k \gamma_k \in H_n^\ell$$

since

$$p_k \gamma_k \in \mathbb{F}[t]_{\ell - d_k} H_n^{d_k} \subseteq H_n^\ell.$$

Therefore, ϕ is a graded homomorphism, and so $U = \ker(\phi)$ is a homogeneous ideal. Since U is a submodule and $\mathbb{F}[t]$ is a principal ideal domain, U is finitely generated by homogeneous elements:

$$U = < t^{c_1}, ..., t^{c_m} >$$

where $c_m, ..., c_1$ are non-negative integers and $m \in \mathbb{N}$.

Using the First Isomorphism Theorem, we obtain:

$$\mathcal{H}_n \cong \mathbb{F}[t]^r / U.$$

Remark 2.5.1. The above decomposition allows us to classify topological features of data sets.

2.6 Computation of Homology Modules

Now, according to Theorem 1-3-2, we discuss an algorithm for computing homology modules and follow the general approach of sources [5], [12], and [10].

To enhance computational capabilities, we make a slight modification in defining n-chains. As usual, we assume K is a finite simplicial complex, and we consider $\mathcal{C}_n(K)$ as the $\mathbb{Z}/2\mathbb{Z}$ vector space generated by the n-dimensional simplices of K instead of the module generated by $\mathbb{Z}$.

Additionally, we define boundary operators by setting $\partial_n(\sigma)$ as the sum of $(n-1)$-

dimensional faces such that σ is an n-dimensional chain, extending this linearly.

Considering a filtration $\sigma_1, ..., \sigma_m$ of K, we define the boundary matrix ∂ where:

$$\partial_{ij} = \begin{cases} 1 & \text{if } \sigma_i \text{ is a face of } \sigma_j \\ 0 & \text{otherwise} \end{cases}$$

The matrix ∂ is upper triangular, and we define the reduced form R using Algorithm 1:

Algorithm 1 Matrix Reduction Algorithm

```
R = ∂
for j = 1 to m do
    while there exists j_0 < j with low (j_0) = low (j) do
        R_{-j} ← R_{-j} + R_{-j_0}
    end while
end for
return R
```

Using the matrix decomposition $R = \partial \cdot V$, we derive the Betti numbers of the complex K using:

$$\text{rank } H_n(K) = \#\ \text{Zero}_n(R) - \#\ \text{Low}_n(R)$$

This decomposition enables efficient computation of persistent homology and topological feature classification.

1) Adding σ_j to K^{j-1} creates a new n-cycle because there is no $(n+1)$-chain in K^j for which σ_j is a face. This new cycle cannot be part of the boundary of an $(n+1)$-chain, thus a new homology class is born.

Moreover, only one new homology class can be created. Every new cycle generated contains σ_j. Therefore, by selecting a new cycle γ and adding it as a new basis element to the previous homology module basis, every new cycle can be written as a linear combination of γ and the older bases. Since σ_j increases the corresponding Betti number β_n^i, we call σ_j positive from now on.

2) Adding σ_j does not create a new cycle. Therefore, the boundary $\partial_n(\sigma_j)$ is a nontrivial

cycle in K^{j-1} and is filled by σ_j, removing exactly one homology class. Since σ_j decreases the Betti number, we call it negative.

By analyzing R, we can determine which of the above cases occurs. First, we observe that adding ∂_{-j_1} to ∂_{-j_2} corresponds to $\sigma_{j_1} + \sigma_{j_2}$.

Since Algorithm 1 only adds columns from left to right, column R_{-j} reaches its final form at the end of the j-th iteration of the `for` loop.

Two cases can occur:

1) If R_{-j} is a zero column, then the n-chain given by the sum of simplices indexed by the nonzero row indices in V_{-j} forms a cycle. Since σ_j is a summand, this cycle must be new, and thus σ_j is positive.

2) If R_{-j} is nonzero, let γ denote the $(n-1)$-chain accumulated in R_{-j}. Then γ is a nontrivial cycle in K^{j-1}; otherwise, R_{-j} could be written as a linear combination of previous columns, making it zero. However, in K^j, it becomes a boundary. Therefore, adding σ_j eliminates the homology class $\bar{\gamma}$. Now, let $k = \text{Low}(j)$. Then, cycle γ was generated by adding σ_k since it is the youngest part of γ, and since we have a filtration, the homology class $\bar{\gamma}$ is born when σ_k is added. Thus, adding σ_j eliminates the homology class γ, which was born in K^k, making σ_j negative. We summarize the algorithmic steps formally.

Assume that K is an n-dimensional simplicial complex and

$$D : c_0(K) \oplus \cdots \oplus c_n(K) \to c_0(K) \oplus \cdots \oplus c_n(K).$$

This defines the boundary operator represented in standard bases by simplices in each dimension.

1) Using Algorithm 1, we compute the matrix decomposition $R = DV$.

2) Columns R_i where $\text{Low}(i) \neq 0$ form the basis:

$$\Sigma_B = \Sigma_{B_0} \cup \cdots \cup \Sigma_{B_n}$$

for $B_0(K) \oplus \cdots \oplus B_n(K)$.

3) Columns V_i such that $R_i = 0$ form the basis:

$$\tilde{\Sigma}_Z = \tilde{\Sigma}_{Z_0} \cup \cdots \cup \tilde{\Sigma}_{Z_n}$$

for $Z_0(K) \oplus \cdots \oplus Z_n(K)$.

4) Define:

$$\Sigma_E = \tilde{\Sigma}_Z \setminus \{V_i \in \tilde{\Sigma}_Z : \exists R_j \mid \text{Low}(R_j) = \text{Low}(V_i) = i\}$$

(a) $\Sigma_E \cup \Sigma_B$ forms a basis for $Z_0(K) \oplus \cdots \oplus Z_n(K)$.

(b) $\{[V_i] : V_i \in V_E\}$ forms a basis for $H_0(K) \oplus \cdots \oplus H_n(K)$.

Remark 2.6.1. The standard algorithm for computing persistent homology is essentially the standard homology algorithm with simplices sorted based on their appearance in the filtration. This is illustrated by computing the barcode of the filtration in Figure 2.1 at the end of this chapter.

Considering the filtration $\{K^i\}_{i=1}^m$ of simplicial complex K, applying simplicial homology to the filtration induces inclusion maps for all $i \leq j$ in homology:

$$f_*^{i,j} : H_n(K_i) \to H_n(K_j)$$

Recalling that the n-th persistent Betti number is given by:

$$\beta_n^{i,j} = \dim \text{Im} f_*^{i,j}$$

Definition 2.6.1. Assume that $0 \neq [c] \in H_n(K_i)$. If $[c] \notin \ mIm f_*^{i-1,i}$, we say that c is born in K^i. If $[c]$ is born in K^i, then it dies in K^j if $f_*^{i,j-1} \notin \ mIm f_*^{i-1,j-1}$, but $f_*^{i,j}([c]) \in \ mIm f_*^{i-1,j}$.

To analyze the evolution of homological features, we introduce the following two numbers:

$$\mu_n^{i,j} = (\beta_n^{i,j-1} - \beta_n^{i,j}) - (\beta_n^{i-1,j-1} - \beta_n^{i-1,j}),$$
$$\mu_n^{i,\infty} = \beta_n^{i,m} - \beta_n^{i-1,m}.$$

The integer $\mu_n^{i,j}$ counts the number of linearly independent homology classes in dimension n that are born at index i and die at index j.

Equivalently,

$$\mu_n^{i,j} = \frac{\text{Im } f_*^{i,j-1} \cap \ker f_*^{j-1,j}}{\text{Im } f_*^{i-1,j-1} \cap \ker f_*^{j-1,j}}.$$

We observe that

$$\dim H_n(K) = \sum_{i=0}^{m} \mu_n^{i,m}.$$

The integers $\mu_n^{i,j}$ are commonly visualized in two ways.

Definition 2.6.2. A **barcode** consists of a set of finite pairs (m, n), where m is a nonnegative integer and n is either a nonnegative integer or $+\infty$. More precisely, a barcode in dimension n is obtained by interpreting each nonnegative $\mu_n^{i,j}$ as an interval $[i, j]$, representing the lifespan of a homology feature, and stacking these intervals on a plane.

Definition 2.6.3. A **persistence diagram** in dimension n is a scatter plot of points (i, j) for each nonnegative $\mu_n^{i,j}$ in the plane.

Example 2.6.1. Consider Figure 2.1, which illustrates a refined simplicial complex. The boundary matrix representation, ordered according to the simplices, results in:

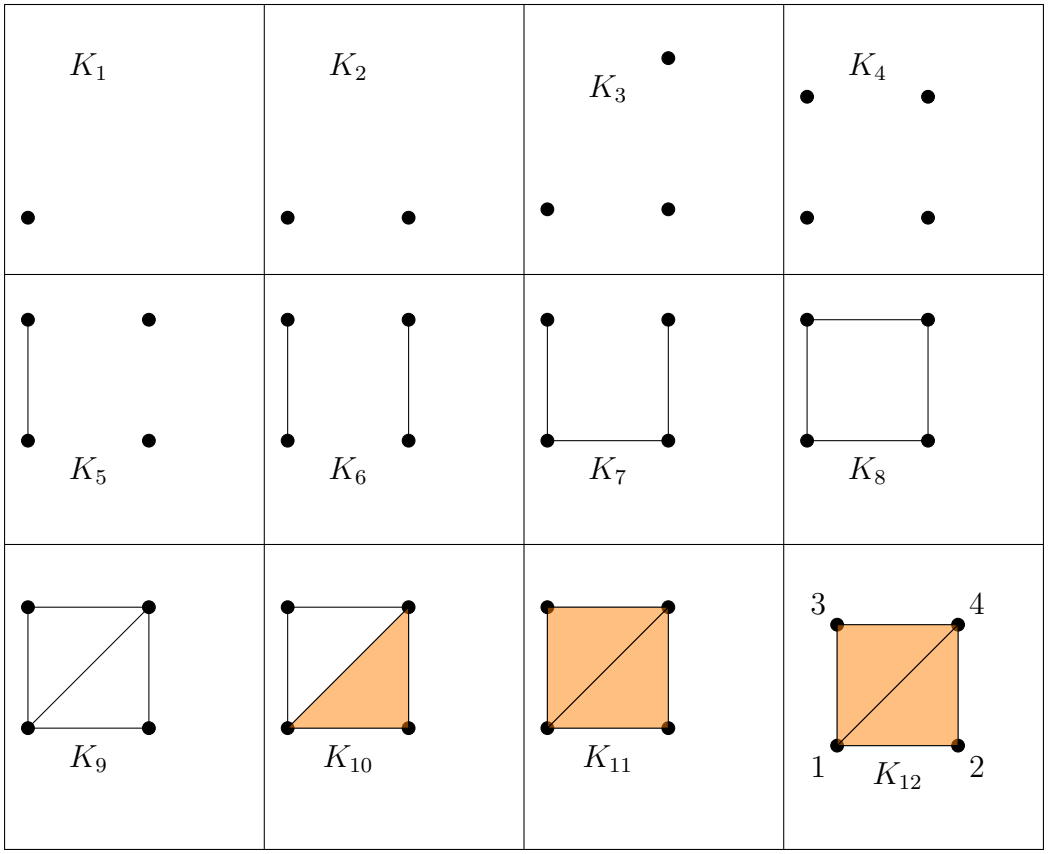


Figure 2.1: A filtered simplicial complex.

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 13 \\ 12 \\ 24 \\ 34 \\ 14 \\ 134 \\ 124 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 13 & 12 & 24 & 34 & 14 & 134 & 124 \end{array} \\ \left[\begin{array}{cccccccccccc} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$$

And the corresponding identity matrix is:

$$V = I_{12\times 12}.$$

The resulting matrices R and V after applying Algorithm 1 to D are:

$$R = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 13 \\ 12 \\ 24 \\ 34 \\ 14 \\ 134 \\ 124 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 13 & 12 & 24 & 34 & 14 & 134 & 124 \end{array} \\ \left[\begin{array}{cccccccccccc} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$$

and

$$V = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 13 \\ 12 \\ 24 \\ 34 \\ 14 \\ 134 \\ 124 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 13 & 12 & 24 & 34 & 14 & 134 & 124 \end{array} \\ \left[\begin{array}{cccccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}$$

We observe that the simplicial homology of K_i — the refined simplicial complex at time i — can be obtained by restricting R and V to the first i columns. Following the described steps, we obtain:

$$\Sigma_{B_0} = \{(10100000000)^T, (11000000000)^T, (01010000000)^T\},$$

$$\tilde{\Sigma}_{Z_0} = \{(10000000000)^T, (01000000000)^T, (00100000000)^T, (00010000000)^T\},$$

$$\Sigma_{B_1} = \{(00001001100)^T, (00001111000)^T\},$$

$$\tilde{\Sigma}_{Z_1} = \{(00001111000)^T, (00000110100)^T\}.$$

The next step is pairing the vectors of $\tilde{\Sigma}_{Z_0}$ with Σ_{B_0} based on their lowest entries:

$$(10000000000)^T \leftrightarrow \emptyset, \quad (01000000000)^T \leftrightarrow (11000000000)^T,$$
$$(00100000000)^T \leftrightarrow (10100000000)^T, \quad (00010000000)^T \leftrightarrow (01010000000)^T.$$

In dimension one:

$$(00001111000)^T \leftrightarrow (00001111000)^T, \quad (00000110100)^T \leftrightarrow (00001001100)^T.$$

Thus, we conclude that $H_0(K_{11}) \cong \mathbb{Z}_2$, which is generated by the vertex v_1, and for $i \geq 1$, $H_i(K_{11}) = 0$.

From the perspective of persistent homology, we focus on specific pairings: A pairing $v \leftrightarrow w$ results in an interval $[t_v, t_w)$ in the barcode, where t_v is the birth time of cycle v in the filtration and t_w is when w becomes a boundary. We determined $t_v = \text{low}(v)$ and $t_w = j$ such that $\text{low}(R_j) = t_v$.

We summarize our observations in the following table:

Dimension	Birth	Death	Representative Cycle	Vertex
0	1	∞	$(10000000000)^T$	1
0	2	6	$(01000000000)^T$	2
0	3	5	$(00100000000)^T$	3
0	4	7	$(00010000000)^T$	4
1	8	11	$(000001111000)^T$	$13 + 12 + 24 + 34$
1	9	10	$(00000110100)^T$	$12 + 24 + 14$

Table 2.1: Summary of persistent homology computations.

Chapter 3

Stability

Bottleneck Distance

In this chapter, we discuss metrics on barcodes/persistence diagrams and show that persistent homology is stable against variations in input data.

This chapter is adapted from the methods in [1] and [2].

Assume that $\mathcal{C}$ and $\mathcal{D}$ are multisets of intervals $< a, b >$ in $\mathbb{R}$, where $< a, b >$ means that the interval can be any well-defined member of $\{[a, b], [a, b), (a, b], (a, b]\}$. A matching between $\mathcal{C}$ and $\mathcal{D}$ is a collection of pairs $X = \{(I, J) \in \mathcal{C} \times \mathcal{D}\}$ such that each I and J appears in at most one pair. Equivalently, a matching is a bijection between subsets of $\mathcal{C}$ and $\mathcal{D}$. If $(I, J) \in X$, then I is matched to J; otherwise, it is unmatched.

Example 3.0.1. Consider the two sets $\mathcal{C} = \{I_1, I_2\}$ and $\mathcal{D} = \{J\}$. If I_2 is unmatched, the pair (I_1, J) defines a matching. Note that $\{(I_1, J), (I_2, J)\}$ does not define a valid matching.

The cost of matching between $I =< a, b >$ and $J =< c, d >$ is defined as:

$$c(I, J) = \max\left(|c - a|, |d - b|\right).$$

The cost of an unmatched interval I is defined as:

$$c(I) = (b - a)/2.$$

The cost of a matching X is given by:

$$c(X) := \max\Big(\sup_{(I,J)\in X} c(I, J), \quad \sup_{\text{unmatched } I\in\mathcal{C}\cup\mathcal{D}} c(I)\Big).$$

If $c(X) \leq \varepsilon_0$, we call X an ε_0-matching.

Remark 3.0.1. A geometric interpretation is as follows:

Consider the intervals in $\mathcal{C}$ and $\mathcal{D}$ as points in the plane $\mathbb{R}^2$. An ε_0-matching finds pairs p and q of intervals from $\mathcal{C}$ and $\mathcal{D}$ such that:

$$\|p - q\|_\infty \leq \varepsilon_0.$$

Thus, each unmatched point is at most ε_0 away from the diagonal in the infinity norm.

Lemma 3.0.1. If there exists an ε_0-matching between $\mathcal{C}$ and $\mathcal{D}$ and an ε_0'-matching between $\mathcal{D}$ and $\mathcal{E}$, then their composition is an $\varepsilon_0 + \varepsilon_0'$-matching between $\mathcal{C}$ and $\mathcal{E}$.

Proof. Let $I =< a_1, b_1 >$, $J =< a_2, b_2 >$, and $K =< a_3, b_3 >$ be such that I is matched to J, and J is matched to K. Then,

$$|a_3 - a_1| \leq |a_3 - a_2| + |a_2 - a_1| \leq \epsilon + \epsilon'.$$

Similarly, the same holds for $|b_3 - b_1|$. If I and J are as above, but J is unmatched in the second matching, then by assumption,

$$b_2 \leq a_2 + 2\epsilon'.$$

Thus,

$$a_1 + 2\epsilon + 2\epsilon' \geq a_2 + \epsilon + 2\epsilon' \geq b_2 + \epsilon \geq b_1,$$

which implies

$$c(I) \leq \epsilon + \epsilon'.$$

A similar argument applies if J is matched to K in the second matching but unmatched in the first. □

Definition 3.0.1. The bottleneck distance between $\mathcal{C}$ and $\mathcal{D}$ is defined as:

$$d_B(\mathcal{C}, \mathcal{D}) = \inf \Big\{ C(X): \quad X \text{ is a matching between } \mathcal{C} \text{ and } \mathcal{D} \Big\}.$$

Example 3.0.2. Suppose $\mathcal{C} = \{[0, 6), [4, 6)\}$ and $\mathcal{D} = \{[2, 8), [10, 12)\}$. The trivial matching that leaves everything unmatched has cost:

$$\frac{|8 - 2|}{2} = 3.$$

Thus,

$$d_B(\mathcal{C}, \mathcal{D}) \leq 3.$$

To improve this bound, we must match $[2, 8)$ with either $[0, 6)$ or $[4, 6)$:

$$c([0, 6), [2, 8)) = c([4, 6), [2, 8)) = 2.$$

If we match $[0, 6)$ with $[2, 8)$, we can either match $[4, 6)$ with $[10, 12)$ or leave it unmatched. The latter has a lower cost:

$$\max \Big\{ c([0, 6), [2, 8)), c([4, 6), [10, 12)) \Big\} = 2.$$

If we match $[0, 6)$ with $[2, 8)$, the best alternative is:

$$\max\Big\{c([4,6),[2,8)), c([0,6),[10,12))\Big\} = 3.$$

Thus, $d_B(\mathcal{C}, \mathcal{D}) = 2$.

As seen in the example, the computations required were quite complex. However, in practice, faster algorithms exist for computing the bottleneck distance. By reformulating the problem as an instance of the bipartite matching problem, modern algorithms in this domain compute $d_B(\mathcal{C}, \mathcal{D})$ in $O(n^{1.5} \log n)$, where $n = |C|+|D|$.

Lemma 3.0.2. Suppose P is finite and V_p has finite dimension for all p. Then,

$$V \cong W^1 \oplus \ldots \oplus W^k$$

such that each W^i is indecomposable.

Proof. Source [9] □

Theorem 3.0.1. Suppose V is an $[n]$-module such that for every $p \in [n]$, $\dim V_p \leq \infty$. Then,

$$V \cong \bigoplus_{[a,b)\in B(V)} I^{[a,b)}$$

where $B(V)$ is a multiset of intervals in $[n]$, called the barcode of V.

Proof. Source [11] □

Theorem 3.0.2. Suppose V is defined as in the previous theorem. Then, $\mu_n^{i,j}$ is the number of occurrences of $[i, j)$ in $B(V)$.

Proof. Source [11] □

As we know from linear algebra, the dimension of a vector space is independent of the choice of basis. The previous result shows that although the decomposition in the theorem is not unique, the multiset $B(V)$ corresponding to it is unique.

3.1 Stability of Persistent Homology

One of the important characteristics of persistent homology is its stability under the bottleneck distance. For example, consider the two graphs shown in Figure 3.1. The zero-dimensional persistence diagram of these graphs is shown on the right.

The persistence diagram of the perturbed graph has two points close to the original persistence diagram's points, along with two points near the diagonal.

Thus, it seems to define a matching where the last two points remain unmatched, and the other two points are matched with bounded cost under changes in the infinity norm C.

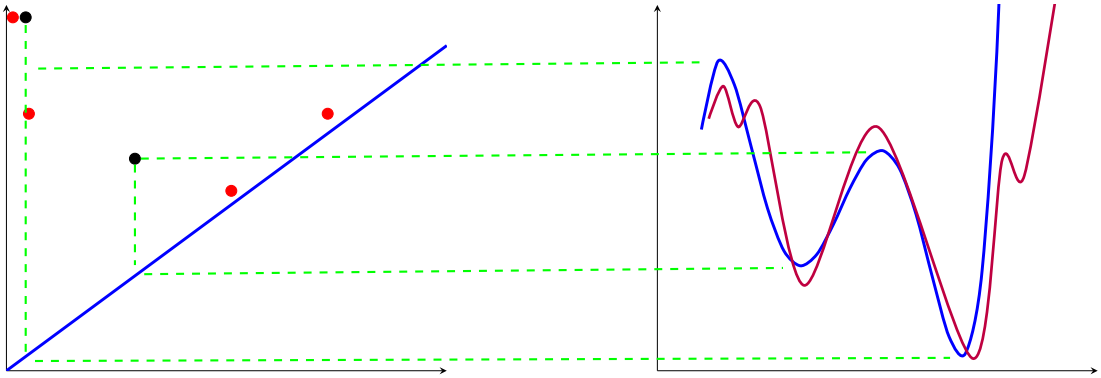


Figure 3.1: A graph and its perturbed version along with the zeroth persistence diagram of the sublevel set of persistent homology.

In general, suppose X is a topological space equipped with two functions $f, g\colon X \to \mathbb{R}$.

Corresponding to these functions, we have the $\mathbb{R}$-modules M^f and M^g as follows:

$$\begin{aligned} M_t^f &= H_i(f^{-1}(-\infty, t]), \\ M_t^g &= H_i(g^{-1}(-\infty, t]). \end{aligned} \tag{3.1}$$

Suppose that for every t, $\dim M_t^f < \infty$ and $\dim M_t^g < \infty$, ensuring well-defined barcodes by Theorem 1.0.3. Now, we have the following theorem:

Theorem 3.1.1.

$$d_B(B(M^f), B(M^g)) \leq \|f - g\|_\infty.$$

Proof. This is proved at the end of the section. □

Recall the Čech and Vietoris-Rips complexes from Chapter 1. Similar stability results can be applied to point sets. For a finite point set P, suppose $H_i(VR(P))$ represents the $\mathbb{R}$-indexed module of persistent homology defined at t via $H_i(VR_t(P))$, and consider the induced map $H_i\left(VR_s(P)\right) \rightarrow H_i\left(VR_t(P)\right)$ via inclusion $VR_s(P) \subseteq VR_t(P)$.

Theorem 3.1.2. Suppose P and Q are finite sets of points in $\mathbb{R}^d$ and that $\sigma\colon P \to Q$ is a bijection such that for all $p \in P$, $\|p - \sigma(p)\| \leq \varepsilon$. Then,

$$d_B(B\left(H_i(VR(P))\right), B\left(H_i(VR(Q))\right)) \leq \varepsilon.$$

A more general result, which we do not discuss, is that using the Hausdorff-Gromov distance, the above theorem can be extended to point sets that do not necessarily have the same cardinality.

The stability theorem for the Čech filtration, in cases where Q and P do not need to have the same cardinality, leads to Theorem 1.2.1. Now, suppose $H_i\left(Cech(P)\right)$ represents the $\mathbb{R}$-module corresponding to the Čech filtration. Consider the function $d_P : \mathbb{R}^d \to \mathbb{R}$. We

have the following filtration from $\mathbb{R}^d$:

$$P_r := d_P^{-1}(-\infty, r] = \bigcup_{p \in P} B_r(p).$$

Theorem 3.1.3. Suppose P and Q are finite sets of points in $\mathbb{R}^d$. If there exists $\varepsilon \geq 0$ such that $P \subseteq Q_\varepsilon$ and $Q \subseteq P_\varepsilon$, then:

$$d_B(B\left(H_i(Cech(P))\right), B\left(H_i(Cech(Q))\right)) \leq \varepsilon.$$

Proof. Define the functions $d_P, d_Q : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$d_P(x) = \min_{p \in P} \|x - p\|, \qquad d_Q(x) = \min_{q \in Q} \|x - q\|$$

Fix $x \in \mathbb{R}^d$ and consider the nearest point $p \in P$ to x. Since $P \subseteq Q_\varepsilon$, there exists $q \in Q$ such that $\|p - q\| \leq \varepsilon$, implying that $d_P(x) \leq d_Q(x) + \varepsilon$. Similarly, $d_Q(x) \leq d_P(x) + \varepsilon$. Consequently,

$$\|d_P - d_Q\|_\infty \leq \varepsilon$$

Now, define the $\mathbb{R}$-modules M^{d_P} and M^{d_Q} as above (*).

From the nerve theorem studied in Chapter 1, we know:

$$M^{d_P} \cong H_i\left(Cech(P)\right),$$

$$M^{d_Q} \cong H_i\left(Cech(Q)\right).$$

Since isomorphism of $\mathbb{R}$-modules implies identical barcodes, we conclude:

$$\begin{aligned} d_B(B\left(H_i(Cech(P))\right), B\left(H_i(Cech(Q))\right)) &= d_B(B\left(M^{d_P}\right), B\left(M^{d_Q}\right)) \\ &\leq \|d_P - d_Q\|_\infty \leq \varepsilon. \end{aligned}$$

□

3.2 Generalizations

The bottleneck distance is sensitive only to the maximum over a matching. However, in practice, other distinguishing distances are often used in data analysis, even though they do not exhibit the same stability as the bottleneck distance.

Definition 3.2.1. Suppose $q \geq 1$, then the q-th Wasserstein distance between persistence diagrams d_1 and d_2 is defined as:

$$W_q(d_1, d_2) = \left(\inf_{\sigma} \sum_{x \in d_1} \|x - \sigma(x)\|_\infty^q \right)^{1/q}$$

where σ is any bijection between d_1 and d_2, and $\|\cdot\|_\infty$ denotes the infinity norm.

Remark 3.2.1.

1. Since we have added the diagonal to the persistence diagram, the set of all bijections from d_1 to d_2 is nonempty.

Another interpretation of the above definition is:

$$\lim_{q \to \infty} W_q(d_1, d_2) = d_B(d_1, d_2)$$

which shows the relationship between the bottleneck distance and it.

2. The Wasserstein distance can also be defined as follows:

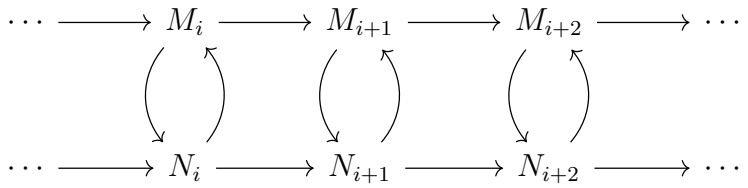
$$W_q(\mathcal{C}, \mathcal{D}) = \inf_{\text{matching } \chi \text{ between } \mathcal{C} \text{ and } \mathcal{D}} \left(\sum_{(I,J)\in X} c(I,J)^q + \sum_{\text{unmatched } I \in \mathcal{C}\cup\mathcal{D}} c(I)^q \right)^{1/q}.$$

3.3 Interleavings

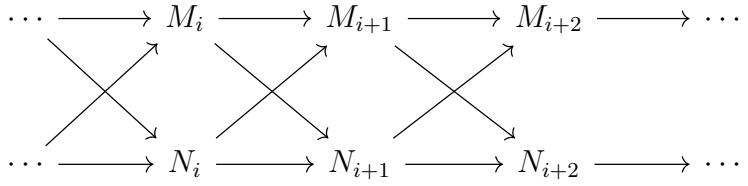
In this section, we introduce the theory of interleavings and an interleaving distance between persistence modules. As we will soon see, the interleaving distance is equal to the bottleneck distance, demonstrating stability in the interleaving framework.

Discrete Case. Suppose we are dealing with persistence modules indexed by integers:

A 0-interleaving is an isomorphism between the families of maps $\{M_i \to N_i\}$ and $\{N_i \to M_i\}$ such that the following diagram commutes:



Similarly, a 1-interleaving is a family of diagonal morphisms such that the following diagram commutes:



Definition 3.3.1. Suppose persistence modules N and M are indexed over real numbers. The ε-shift of M, denoted M^ε, is defined as follows for all $s \le t \in \mathbb{R}$:

$$M^\varepsilon_t = M_{t+\varepsilon}, \quad M^\varepsilon(s \le t) = M(s+\varepsilon \le t+\varepsilon).$$

If $f\colon M \to N$ is a morphism, then its ε-shifted morphism $f^\varepsilon\colon M^\varepsilon \to N^\varepsilon$ is defined by:

$$f_t^\varepsilon = f_{t+\varepsilon}.$$

In general, suppose X is a topological space equipped with two functions $f, g\colon X \to \mathbb{R}$. Corresponding to these functions, we have the $\mathbb{R}$-modules M^f and M^g as follows:

$$\begin{aligned} M_t^f &= H_i(f^{-1}(-\infty, t]), \\ M_t^g &= H_i(g^{-1}(-\infty, t]). \end{aligned} \tag{3.2}$$

Suppose that for every t, $\dim M_t^f < \infty$ and $\dim M_t^g < \infty$, ensuring well-defined barcodes by Theorem 1.0.3. Now, we have the following theorem:

Theorem 3.3.1.

$$d_B(B(M^f), B(M^g)) \leq \|f - g\|_\infty.$$

Proof. This is proved at the end of the section. □

3.3.1 Interleavings

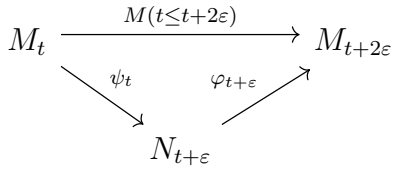
In this section, we introduce the theory of interleavings and an interleaving distance between persistence modules. As we will soon see, the interleaving distance is equal to the bottleneck distance, demonstrating stability in the interleaving framework.

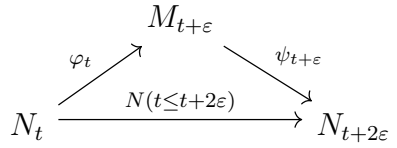
Definition 3.3.2. Suppose that $\eta_M^\varepsilon\colon M \to M^\varepsilon$ is a morphism defined by restriction to M_t as $M(t \leq t + \varepsilon)$. Given $\varepsilon \in [0, \infty)$, an ε-interleaving between N and M consists of a pair

of morphisms $\varphi\colon M \to N^\varepsilon$ and $\psi\colon N \to M^\varepsilon$ such that:

$$\varphi^\varepsilon \circ \psi = \eta_M^{2\varepsilon},$$
$$\psi^\varepsilon \circ \varphi = \eta_N^{2\varepsilon}.$$

Explicitly, the above conditions ensure that the following diagrams commute for each $t \in \mathbb{R}$:





If such a pair exists, we say that N and M are ε-interleaved.

Lemma 3.3.1. If N and M are ε-interleaved, and N and L are ε'-interleaved, then M and L are $(\varepsilon + \varepsilon')$-interleaved.

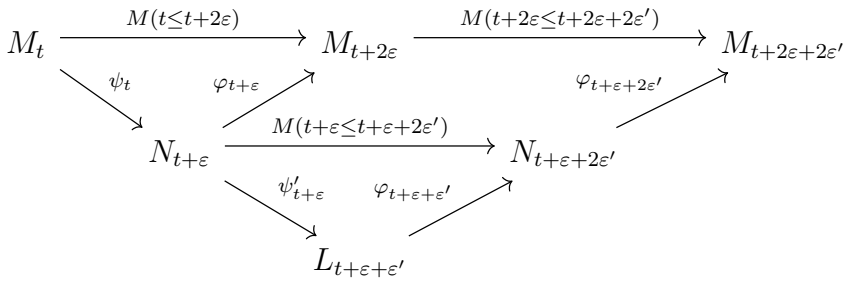
Proof. Suppose $\psi\colon M \to N^\varepsilon$ and $\varphi\colon N \to M^\varepsilon$, as well as $\psi'\colon N \to L^{\varepsilon'}$ and $\varphi'\colon L \to N^{\varepsilon'}$, satisfy the interleaving conditions. Then we define the morphisms:

$$\psi''\colon\ M \to L^{\varepsilon+\varepsilon'}$$
$$\psi''_t = \psi'_{t+\varepsilon} \circ \psi_t$$
$$\varphi''\colon\ L \to M^{\varepsilon+\varepsilon'}$$
$$\varphi''_t = \varphi_{t+\varepsilon'} \circ \varphi'_t$$

The remaining task is to verify:

$$(\varphi''^{\varepsilon+\varepsilon'}) \circ \psi'' = \eta_M^{2(\varepsilon+\varepsilon')}, (\psi''^{\varepsilon+\varepsilon'}) \circ \varphi'' = \eta_L^{2(\varepsilon+\varepsilon')}.$$

The first equation follows from the commutativity of the following diagram, with a similar diagram for the second equation:



□

Definition 3.3.3. The interleaving distance between two $\mathbb{R}$-modules N and M is defined as:

$$d_I(M,N) = \inf\{\varepsilon\colon \text{ an } \varepsilon\text{-interleaving exists between } N \text{ and } M\}.$$

The following non-trivial theorem is a fundamental result in topological data analysis, and we now proceed to its proof.

Theorem 3.3.2. Suppose N and M are $\mathbb{R}$-modules such that M_t and N_t have finite dimension for all $t \in \mathbb{R}$. Then,

$$d_B(B(M), B(N)) = d_I(M,N).$$

Proof. Since $\|f-g\| \leq \varepsilon$, the following commutative diagram of inclusions holds for each

$t \in \mathbb{R}$:

$$\begin{array}{ccccc} f^{-1}(-\infty, t] & \longrightarrow & f^{-1}(-\infty, t+\varepsilon] & \longrightarrow & f^{-1}(-\infty, t+2\varepsilon] \\ & \searrow\nearrow & & \searrow\nearrow & \\ g^{-1}(-\infty, t] & \longrightarrow & g^{-1}(-\infty, t+\varepsilon] & \longrightarrow & g^{-1}(-\infty, t+2\varepsilon] \end{array}$$

Applying H_i results in the following commutative diagram:

$$\begin{array}{ccccc} M_t^f & \longrightarrow & M_{t+\varepsilon}^f & \longrightarrow & M_{t+2\varepsilon}^f \\ & \searrow\nearrow & & \searrow\nearrow & \\ M_t^g & \longrightarrow & M_{t+\varepsilon}^g & \longrightarrow & M_{t+2\varepsilon}^g \end{array}$$

which clearly defines an ε-interleaving. □

Theorem 3.3.3. Suppose $P = \{p_1, \ldots, p_m\}$ and $Q = \{q_1, \ldots, q_m\}$ such that $\|p_i - q_i\| \leq \varepsilon$. If $\sigma = \{p_{i_1}, \ldots, p_{i_m}\}$ is a simplex in $VR_r(P)$, then $\operatorname{diam}(\sigma) \leq 2r$. By the bijection of corresponding simplices, $\tau = \{q_{i_1}, \ldots, q_{i_m}\}$ satisfies $\operatorname{diam}(\tau) \leq 2r + 2\varepsilon$. Thus, inclusion

$$VR_r(P) \subseteq VR_{r+\varepsilon}(Q) \text{ and } VR_r(Q) \subseteq VR_{r+\varepsilon}(P)$$

hold, leading to:

$$VR_r(P) \subseteq VR_{r+\varepsilon}(Q) \subseteq VR_{r+2\varepsilon}(P),$$
$$VR_r(Q) \subseteq VR_{r+\varepsilon}(P) \subseteq VR_{r+2\varepsilon}(Q).$$

Applying H_i defines an ε-interleaving between $H_i(VR(P))$ and $H_i(VR(Q))$, and from Theorem 1.4.1:

$$d_B(B\left(H_i(VR(P))\right), B\left(H_i(VR(Q))\right)) = d_I(H_i\left(VR(P)\right), H_i\left(VR(Q)\right)) \leq \varepsilon.$$

3.4 Zigzag Persistent Homology

So far, in the past two chapters, we have focused on applying homology to filtrations of a topological space. In this section, we see that by considering partially ordered sets whose underlying diagram is not linearly oriented, we obtain richer invariants and study the persistence of topological features across a family of spaces, leading to what is called zigzag persistence.

3.5 Zigzag Persistence Modules

A **zigzag poset** on n vertices is a partial order of the form

$$1 \leftrightarrow 2 \leftrightarrow \ldots \leftrightarrow n-1 \leftrightarrow n$$

where $\leftrightarrow$ denotes that the arrow can be either $\rightarrow$ or $\leftarrow$. A **zigzag persistence module** is a P-module where P is any zigzag poset. Note that if P has n vertices and all the arrows point in the same direction, then a zigzag persistence module is nothing more than an $[n]$-module.

For a zigzag poset P on n vertices, let $[a, b]$ denote the restriction of P to the vertices $i \in [a, b]$. The associated **interval module** $I^{[a,b]}$ is the P-module defined by

$$I_i^{[a,b]} = \begin{cases} k, & \text{if } i \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

together with the identity morphism $\mathrm{id} : I_i^{[a,b]} \rightarrow I_j^{[a,b]}$ whenever $i, j \in [a, b]$ and i and j are comparable.

Example 13.2. If $n = 4$, and the arrows alternate, then $I^{[2,3]}$ is the persistence module:

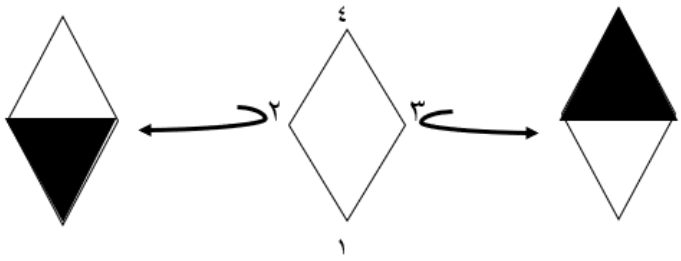


Figure 3.2: A zigzag of simplicial complexes: $K \leftrightarrow K \cap K' \leftrightarrow K'$.

$$0 \to k \leftarrow k \to 0.$$

We state the following result without proof. A proof can be obtained by a variation of the proof of Theorem 6.16.

Theorem 13.3. Let V be a zigzag module on n vertices such that $\dim V_p < \infty$ for all $p \in [n]$. Then

$$V \cong \bigoplus_{[a,b] \in B(V)} I^{[a,b]}$$

where $B(V)$ is a unique multiset of intervals in $[n]$ called the **barcode** of V.

While the situation to a large extent mimics that of standard persistent homology, the barcodes in zigzag persistence can be more refined as the following example illustrates.

Example 13.4. Let K and K' be the simplicial complexes shown in Fig. 47, and let $K \cap K'$ denote the simplex-wise intersection. This defines a zigzag of simplicial complexes:

$$K \leftrightarrow K \cap K' \leftrightarrow K',$$

and a corresponding zigzag persistence module:

$$\mathbb{Z}_2 \cong H_i(K) \leftarrow H_i(K \cap K') \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \rightarrow H_i(K') \cong \mathbb{Z}_2.$$

Note that the 1-cycle $\{\{1,2\}, +\{1,3\} + \{2,3\} + \{3,4\}\}$ is non-trivial in homology in all three complexes, and one may therefore be inclined to think that there is a bar in the barcode spanning all three vertices. That is, however, not the case, as the following choice of a basis for the middle vector space shows:

$$\{\{1,2\} + \{1,3\} + \{2,3\}, \{2,3\} + \{2,4\} + \{3,4\}\}$$

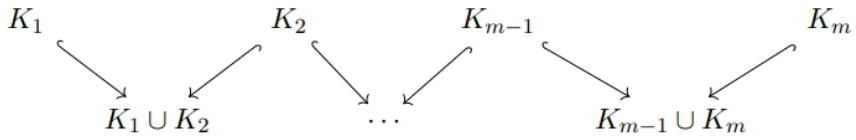
for $H_1(K \cap K')$. Indeed, this gives the decomposition:

$$\mathbb{Z}_2 \xleftarrow{[1,0]} \mathbb{Z}_2 \oplus \mathbb{Z}_2 \xrightarrow{[0,1]} \mathbb{Z}_2 = \left(\mathbb{Z}_2 \xleftarrow{1} \mathbb{Z}_2 \rightarrow 0\right) \oplus \left(0 \xleftarrow{1} \mathbb{Z}_2\right).$$

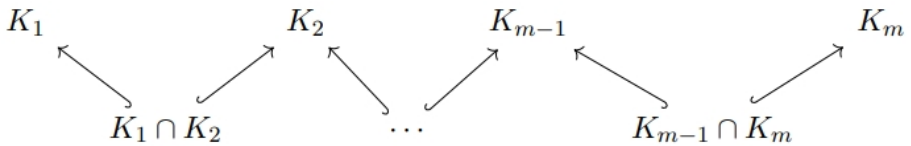
The barcode thus consists of the intervals $\{[1,2],[2,3]\}$.

3.6 The Diamond Principle

Given a sequence of simplicial complexes $\{K_i\}_{i=1}^m$, there are two natural ways of linking K_i's. The union zigzag $K_\cup$:

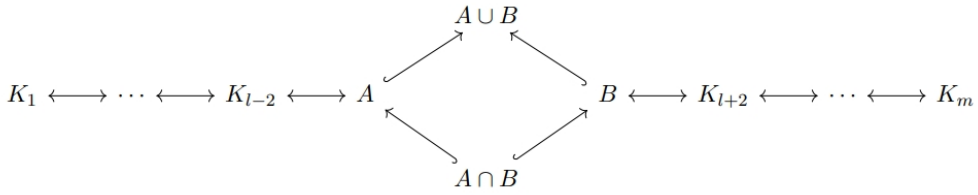


and the intersection zigzag $K_\cap$:



For instance, replacing the middle simplicial complex in Example 13.4 by the union of the

two outer complexes yields a barcode $\{[1,1],[3,3]\}$ in degree 1 homology. The question is thus if there is a method to obtain one barcode from the other. That turns out to be the case if one dispenses with intervals supported only at an intersection or a union. **Theorem 13.5 (The strong diamond principle [?]).** Consider the following diagram of simplicial complexes and simplicial maps where the four middle maps are inclusions.



Let K^+ and K^- denote the zigzags passing through the union and intersection, respectively. Then there is the following correspondence of intervals in the barcodes:

$$[l,l] \in B(H_{p+1}(K^+)) \leftrightarrow [l,l] \in B(H_p(K^-)),$$

In the remaining cases, the matching preserves homological dimension:

$$[b,l] \in B(H_p(K^+)) \leftrightarrow [b,l-1] \in B(H_p(K^-)), \quad \text{if } b < l,$$

$$[l,d] \in B(H_p(K^+)) \leftrightarrow [l+1,d] \in B(H_p(K^-)), \quad \text{if } d > l,$$

$$[b,d] \in B(H_p(K^+)) \leftrightarrow [b,d] \in B(H_p(K^-)), \quad \text{for all other cases.}$$

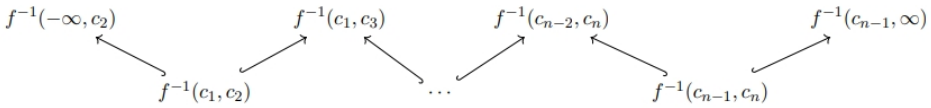
where $*, \uparrow \in \{+,-\}$ and $* \neq \uparrow$.

By iteratively applying the previous theorem we see that $B(H_p(K_\cup))$ is determined by $B(H_p(K_\cap)) \cup B(H_{p-1}(K_\cap))$, and $B(H_p(K_\cap))$ is determined by $B(H_p(K_\cup)) \cup B(H_{p+1}(K_\cup))$. It is left as an exercise to spell out precisely how to transform an interval in one barcode to an interval in the other (possibly shifted by a degree in homology).

3.7 Levelset Zigzag Persistent Homology

The persistent homology of an $\mathbb{R}$-space $f : X \to \mathbb{R}$ studies the evolution of the homology of the sublevel sets. An alternative approach is to study how the homology persists across the level sets $f^{-1}(t)$ as the parameter t sweeps over the real line. In the following, we shall

assume that the pair (X, f) is a constructible $\mathbb{R}$-space. By labeling the critical values $c_1 < c_2 < \cdots < c_n$, we obtained the following zigzag of topological spaces:

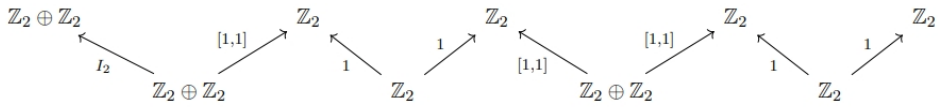


The interest in this diagram comes from the fact that:

- The fibers are constant (up to homeomorphism) between critical values. That is, $f^{-1}(c_i, c_{i+1})$ is homeomorphic to $f^{-1}(s) \times (c_i, c_{i+1})$ with f being the projection onto the second component, and s is any point in $c_i < s < c_{i+1}$.
- $f^{-1}(c_{i-1}, c_{i+1})$ deformation retracts onto $f^{-1}(c_i)$.

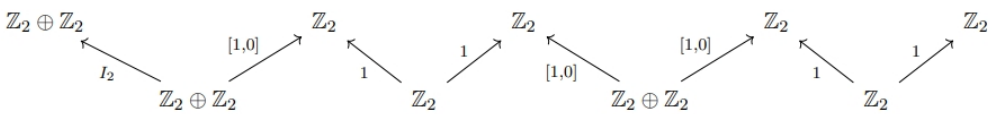
Hence, the topological evolution of the fibers is completely captured up to homotopy by the above zigzag; applying H_p to the above zigzag yields the **levelset zigzag persistence module** in dimension p (associated to f), and the corresponding barcode $ZZ_p(f)$. The intervals in $ZZ_p(f)$ are naturally given in terms of the critical values. To see this, assume that we have a bar born at $H_p(f^{-1}(c_i, c_{i+2})) \cong H_p(f^{-1}(c_i))$ that lives up to (and including) $H_p(f^{-1}(c_j, c_{j+1}))$. This represents a feature that is alive at all fibers $f^{-1}(s)$ for $s \in [c_{i+1}, c_{j+1})$, and the corresponding interval in $ZZ_p(f)$ is thus $[c_i, c_{j+1}) \in ZZ_p(f)$. Likewise, we have intervals (c_i, c_j), $[c_i, c_j]$, and $(c_i, c_j]$.

Example 13.6. Consider the constructible $\mathbb{R}$-space shown in Fig. 48. Proceeding in H_0 with the "obvious basis" we get the following zigzag of vector spaces:



For each copy of $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ above, we replace the basis $\{v_1, v_2\}$ with the basis $\{v_1, v_1 + v_2\}$.

The zigzag then diagonalizes as:



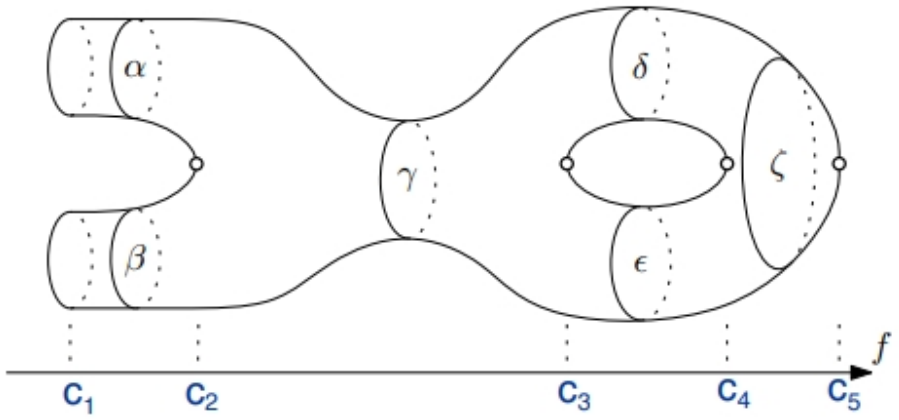
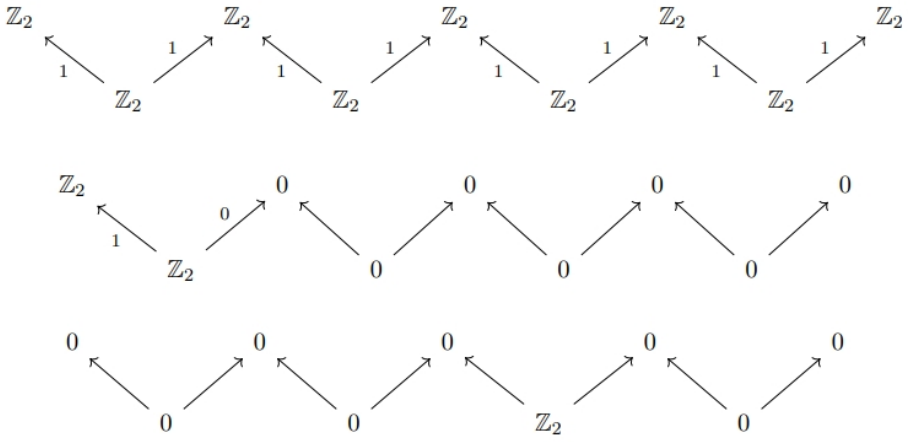


Figure 3.3: A topological space projected onto the horizontal axis. Copied from [2].

which is the direct sum of the following three zigzag modules:



We conclude that

$$ZZ_0(f) = \{[c_1, c_5], [c_1, c_2), (c_3, c_4)\}.$$

And for H_1 (exercise):

$ZZ_1(f) = \{[c_1, c_5], [c_1, c_2], [c_3, c_4]\}$. Let $PH_p(f)$ denote the barcode of the sublevel persistence module M^f in homology dimension p . A simple computation yields:

$$PH_0(f) = \{[c_1, \infty), [c_1, c_2)\} \quad \text{and} \quad PH_1(f) = \{[c_1, \infty), [c_1, c_5], [c_3, \infty), [c_4, \infty)\}.$$

Comparing with the above computations, there seems to be a close connection between the bars in levelset and sublevel persistent homology. We now state a theorem that makes this connection precise.

Theorem 13.7 ([2]). *Assume that* (X, f) *is a constructible* $\mathbb{R}$*-space, and that* $\dim H_p(f^{-1}(t)) < \infty$ *for all* t*. Then the sublevel barcode* $PH_p(f)$ *is given as the following union:*

$$\{[c_i, c_j) : [c_i, c_j) \in ZZ_p(f)\} \cup \{[c_i, \infty) : [c_i, c_j] \in ZZ_p(f)\} \cup \{(c_j, \infty) : (c_i, c_j) \in ZZ_{p-1}(f)\}.$$

Hence, levelset persistent homology is a finer invariant than sublevel persistence. Moreover, and as illustrated by the previous example, the types of endpoints carry concrete information about the topological feature: closed-closed and open-open intervals correspond to global features of X, i.e., features that are present irrespective of the function (but potentially with different birth and death times). The half-open intervals, on the other hand, can be perturbed away.

We end this section by stating an important theorem.

Theorem 13.8. *Assume that* (X, f) *and* (X, g) *are constructible* $\mathbb{R}$*-spaces, such that*

$$\dim H_p(f^{-1}(t)), \dim H_p(g^{-1}(t)) < \infty$$

for all $t \in \mathbb{R}$*. Then,*

$$d_B(ZZ_p(f), ZZ_p(g)) \leq ||f - g||_\infty.$$

In fact, here we can replace the standard bottleneck distance with a more refined version that only allows matchings of intervals with the same type of endpoints, i.e., closed-closed must be matched to closed-closed and so forth. Much more can be said on this topic, but we will not pursue it further

Bibliography

[1] Bjerkevik, H. B. Stability of higher-dimensional interval decomposable persistence modules. *arXiv preprint*, arXiv:1609.02086, 2016.

[2] Carlsson, G., Silva, V. De, and Morozov, D. Zigzag persistent homology and real-valued functions. *Foundations of Computational Mathematics*, pp. 247–256, 2009.

[3] Frosini, P. and Ferri, M. On the use of size functions for shape analysis. *Proceedings IEEE on Qualitative Vision*, 1993.

[4] Frosini, Patrizio. Measuring shapes by size functions. *Proceedings of the SPIE.*, 1607:122–133, 1992.

[5] Herbert, E. and John., H. *Computational topology: an introduction*. American Mathematical Soc, 2010.

[6] Morozov, D. Persistence algorithm takes cubic time in worst case. *BioGeometry News, Dept. Comput. Sci., Duke Univ.*, 20, 2008.

[7] Munkres, J. R. *Elements of algebraic topology*. CRC press, 2018.

[8] Rotman, J. Joseph. *An Introduction to Algebraic Topology*. Springer-Verlag New York Inc., 1988.

[9] Rotman, J. Joseph. *Advanced Modern Algebra*. Prentice Hall, 2nd printing ed. , 2003.

[10] Wang, Y. Computational topology: Theory, algorithms, and applications to data analysis. Lecture notes, Spring 2016, The Ohio State University.

[11] Y.Oudot, Steve. *Persistence Theory: From Quiver Representations to Data Analysis*. American Mathematical Soc., 2015.

[12] Zomorodian, A. and Carlsson, G. Computing persistent homology. *Discrete and Computational Geometry*, pp. 347–356, 2004.