creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학 박사 학위논문

On the minimal number of sample points and the persistence of the Vietoris-Rips complex

(Vietoris-Rips complex의 최소 표본점과 persistence에 대하여)

2021년 8월

서울대학교 대학원

수리과학부

정효진

On the minimal number of sample points and the persistence of the Vietoris-Rips complex

(Vietoris-Rips complex의 최소 표본점과 persistence에 대하여)

지도교수 Otto van Koert

이 논문을 이학 박사 학위논문으로 제출함

2021년 4월

서울대학교 대학원

수리과학부

정효진

정효진의 이학 박사 학위논문을 인준함

2021년 6월

위 원 장	국 웅
부 위 원 장	Otto van Koert
위 원	임 선 희
위 원	Ernest Ryu
위 원	이 상 율

On the minimal number of sample points and the persistence of the Vietoris-Rips complex

A dissertation
submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
to the faculty of the Graduate School of
Seoul National University

by

Hyojin Jung
Dissertation Director : Professor Otto van Koert

Department of Mathematical Sciences
Seoul National University

August 2021

© 2021 Hyojin Jung

All rights reserved.

Abstract

On the minimal number of sample points and the persistence of the Vietoris-Rips complex

Hyojin Jung

Department of Mathematical Sciences

The Graduate School

Seoul National University

Recently topological data analysis become a popular tool to analyze data. In this paper, we study the behaviour of Vietoris-Rips complex with persistent homology to figure out the shape of data. In particular, we find the minimal number of data points on a sphere such that homology of the Vietoris-Rips complex of those data points is isomorphic to the homology of the sphere.

Key words: Vietoris-Rips complex, persistence module, persistent homology, Nerve theorem, TDA

Student Number: 2012–30869

Contents

Chapter 1

Introduction

Topological data analysis has become a fairly popular tool to analyze data from a qualitative point of view. In particular, the Vietoris-Rips complex is a common tool to analyze the shape of data. The Vietoris-Rips complex generalizes the concept of neighborhood graph, and is very flexible. Roughly speaking, given a finite set $X = \{x_i\}_{i=1}^n$ together with a dissimilarity measure d, we define the Vietoris-Rips complex at scale parameter ϵ, as a simplicial complex whose vertices consist of x_i and whose k-simplices are just $k+1$-tuples $x_{i_0}, \ldots, x_{i_k}$ of points with dissimilarity less than ϵ.

One may wonder what scale parameter to choose, and again there is an idea from TDA dealing with this, namely the idea of persistence. Briefly introduce the concept of persistence: for a parameterized family of spaces, certain topological features which persist over a significant parameter range are to be treated as signal with short-lived features as noise. We can denote these parameter ranges as intervals, say a barcode: for each interval, we consider the left end point as the birth and the right end point as the death of a topological feature. In other words, a barcode is a graphical representation as a collection of horizontal line segments in a plane where the horizontal axis corresponds to the parameter and the vertical axis shows an ordering of homology generators.

We will investigate how reliable these barcodes are, and how many data points are necessary.

Mathematically, we can simply forget about data, and ask the following questions.

- Given the manifold, what is the minimal number of cells that is needed to describe homotopy type correctly? Weaker and easier is the question how many cells are needed to obtain the correct homology groups. This is in some sense a classical invariant that has been studied quite well.

- Another question is what is the minimal size of a simplicial complex that computes the homology correctly? Again, this is a classical question. For example, to compute the homology of S^1 correctly, one can use a simplicial complex with 2 vertices and 2 edges.

- In the setup of the Vietoris-Rips complex we are dealing with a simplicial complex, but there are now restrictions on the complex. For example, the smallest Vietoris-Rips complex that computes the homology of the circle correctly has 4 points and 4 edges.

  In particular, we need at least 4 data points.

This line of reasoning brings us to a lower bound on the number of data points. We will not consider these interesting questions in any detail, but instead address the opposite approach.

Assuming that we have enough data points, we may wonder which barcodes are reliable. The motivating example here is very simple. Consider the unit cube, with vertices $(0,0,0)$, $(1,0,0),\ldots,(1,1,1)$. If we consider the Vietoris-Rips complex of these 8 points with the Euclidean distance, we find that this space has the homology of S^3, which is obviously not desired.

Although one might think at first sight that this is caused by the lack of points, a result by Adamaszek and Adams, [13], shows that there is another

mechanism at work here. Namely, they studied the Vietoris-Rips complex of the circle; they found that as r increases the Vietoris-Rips complex $VR(S^1, r)$ is homotopy equivalent to $S^1, S^3, S^5, S^7, \cdots$ until finally it is contractible.

We shall see that this phenomenon of the wrong barcodes is easily understood using the Nerve theorem. The Nerve theorem tells us that good covers of a space will yield simplicial complexes that are homotopy equivalent to the original space.

Chapter 2

Preliminaries

In this chapter, we review several complexes and its related homology groups with some fixed notation and terminologies. Recall the Nerve theorem which shows certain connectivity of simplicial complex if its subcomplexes have properties of some connectivity conditions.

Definition 2.0.1. A family K of a set X with a collection of finite non-empty subsets of X is called an **abstract simplicial complex** if

(1) $\{v\} \in K$ for all $v \in X$, and

(2) If $\sigma \in K$ and $\emptyset \neq \tau \subseteq \sigma$, then $\tau \in K$ where τ is called a **face of** σ.

The elements of X are called **vertices** and we denote X as $\mathrm{Vert}(K)$. The elements of a simplicial complex K are called **simplices**. A $k-$**simplex** σ consists of $k+1$ vertices and **its dimension** $\dim \sigma = k$. The **dimension of an abstract simplicial** K is the maximum dimension of simplexes of K. A subset $L \subset K$ is called a **subcomplex of** K, denoted $L < K$, if L is a simplicial complex in its own right.

Example 2.0.2. $0-$simplices are vertices.

Remark 2.0.3. Consider a collection of all simplices of K with dimension p or less, i.e.,

$$K^p = \{\ \sigma \in K \mid \dim \sigma \leq p\ \}.$$

This subset $K^p \subset K$ becomes a subcomplex of K and it is called the $p-$**skeleton of** K.

Note that we use the notation $[v_0, \cdots, v_k]$ as a $k-$simplex $\{v_0, \cdots, v_k\}$ to emphasize itself.

Given a simplicial complex K, the p**th chain group** C_p **of** K contains all linear combinations of $p-$simplices in the complex with $\mathbb{Z}_2-$coefficients. Since $\mathbb{Z}_2 = \{0, 1\}$, all elements of C_p are of the form $\sum_j \sigma_j$, for $\sigma_j \in K$. The group operation is addition with $\mathbb{Z}_2$ coefficients. Define the p**th boundary homomorphism** as the homomorphism that assigns each simplex $\sigma = [v_0, \cdots, v_p] \in K$ to **its boundary**

$$\partial_p \sigma = \sum_i [v_0, \cdots, \hat{v_i}, \cdots, v_p]$$

where $\hat{v_i}$ indicates that the simplex excludes the ith vertex. The function $\partial_p : C_p \longrightarrow C_{p-1}$ is a homomorphism between the chain groups. Abbreviate $C = C_p$ and $\partial = \partial_p$.

Lemma 2.0.4. For all p, we have $\partial_{p-1} \circ \partial_p = 0$.

Proof. By expanding the definition with a $p-$simplex $\sigma = [v_0, \cdots, v_p]$, we get

$$\begin{aligned}
\partial_{p-1} \circ \partial_p(\sigma) &= \partial_{p-1}\Big(\sum_i [v_0, \cdots, \hat{v_i}, \cdots, v_p]\Big) \\
&= \sum_{j<i} [v_0, \cdots, \hat{v_j}, \cdots, \hat{v_i}, \cdots, v_p] \\
&\quad + \sum_{i<j} [v_0, \cdots, \hat{v_i}, \cdots, \hat{v_j}, \cdots, v_p] \\
&= 0.
\end{aligned}$$

□

With the above lemma, we can use a useful tool of a sequence and its induced invariants.

Definition 2.0.5. A chain complex $(C,\ \partial)$ is a sequence of abelian groups C_p connected by homomorphisms ∂_p such that the composition of any two consecutive maps is the zero map.

$$\cdots \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

This boundary homomorphism is also called **differentials**.

Definition 2.0.6. Given a chain complex (C, ∂), the image $im(\partial_{p+1})$ of the boundary homomorphism is called the **boundary group** B_p and its elements are called **boundaries**. Similarly, the kernel $ker(\partial_p)$ of the boundary homomorphism is called **cycle groups** Z_p and its elements are called **cycles**. The p**th homomorphism group** H_p is a quotient group defined as

$$\begin{aligned} H_p &= Z_p/B_p \\ &= ker\partial_p/im\partial_{p+1}. \end{aligned}$$

In this paper, the word "map" is used to denote "continuous function."
Let X and Y be two topological spaces.

Definition 2.0.7. Let $f_0, f_1 : X \longrightarrow Y$ be two maps of X into Y. The maps f_0 and f_1 are **homotopic**, denotes as $f_0 \simeq f_1$, if there exist a map $F : X \times [0,1] \longrightarrow Y$ such that $f_0(x) = F(x,0)$ and $f_1(x) = F(x,1)$ for all $x \in X$. We call this map F a **homotopy between** f_0 **and** f_1.

Definition 2.0.8. The spaces X and Y are **homotopy equivalent**, denotes $X \simeq Y$, if there are maps $g : X \longrightarrow Y$ and $h : Y \longrightarrow X$ such that the two

compositions $h \circ g : X \longrightarrow Y \longrightarrow X$ and $g \circ h : Y \longrightarrow X \longrightarrow Y$ are homotopic to the respective identity maps. We call the pair of these maps a **homotopy equivalence**.

Properties 2.0.9 (Homotopy Invariance)**.** For any map $f : X \to Y$, there is an induced homomorphism $H_n(f) : H_n(X) \to H_n(Y)$ by the map $f_\# : (C(X), \partial_X) \to (C(Y), \partial_Y)$ defined as $\sigma \mapsto f \circ \sigma$.

(1) If $f, g : X \to Y$ are homotopic, then $H_n(f) = H_n(g)$.

(2) If f is a homotopy equivalence, then $H_n(f)$ is an isomorphism.

A given set of point $\{p_0, p_1, \cdots, p_k\} \subset \mathbb{R}^k$ is **geometrically independent** (*or* **affinely independent**) if those vectors $p_1 - p_0, p_2 - p_0, \cdots, p_k - p_0$ are linearly independent. A combination $x = \sum_{i=0}^k a_i p_i$ is a **convex combination** if $\sum_{i=0}^k a_i = 1$ and all $a_i \geq 0$. The **convex hull** of a given point set $M = \{q_0, q_1, \cdots, q_n\}$ is the set of all convex combinations, denoted as $Conv(M) = \{\sum_{i=0}^n a_i q_i | \sum_{i=0}^n a_i = 1 \text{ and } a_i \geq 0\}$. If $\{p_0, p_1, \cdots, p_k\}$ is geometrically independent, $\sigma = Conv(\{p_0, p_1, \cdots, p_k\})$ is called a $k-$ **simplex in** $\mathbb{R}^N$ (*or* $k-$**simplex spanned by** $\{p_0, p_1, \cdots, p_k\}$), denoted $\langle p_0, p_1, \cdots, p_k \rangle$, and a simplex τ spanned by a subset of $\{p_0, p_1, \cdots, p_k\}$ is a **face of** σ, denoted $\tau < \sigma$.

Definition 2.0.10. A **simplicial complex** K **in** $\mathbb{R}^N$ is a collection of simplices in $\mathbb{R}^N$ such that

(1) if $\tau < \sigma \in K$, then $\tau \in K$, and

(2) if $\sigma, \tau \in K$, then $\sigma \cap \tau < \sigma$ and $\sigma \cap \tau < \tau$.

Remark 2.0.11. For a simplicial complex K, consider $|K| = \cup_{\sigma \in K} \sigma \subset \mathbb{R}^N$ for some N. Define a topology on $|K|$ with the following properties.

(1) Each of σ has the usual induced subspace topology in $\mathbb{R}^N$.

(2) $A \subset |K|$ is closed if $A \cap \sigma$ is closed in σ for all $\sigma \in K$.

(3) $A \subset |K|$ is open if $A \cap \sigma$ is open in σ for all $\sigma \in K$.

This topology on K is called a **weak topology**. We call $|K|$ with a weak topology the **underlying space** (*or* a **polytope**) **of** K.

Given an abstract simplicial complex, we can also define *an underlying space* like the underlying space of a simplicial complex in $\mathbb{R}^N$.

Let K be an abstract simplicial complex and $\sigma = \{v_0, \cdots, v_n\} \in K$. Define $|\sigma| = \{\sum_{i=0}^n t_i v_i | \sum_{i=0}^n t_i = 1,\ t_i \geq 0\}$ and $|K| = \cup_{\sigma \in K} |\sigma|$. In formal,

$$|K| = \{x : \mathrm{Vert}(K) \to [0,1] \mid \sum_{\{v \in \mathrm{Vert}(K) | x(v) \neq 0\} \in K} x(v) = 1\}$$

and $|\sigma| = \{x \in |K| \mid x(v) = 0 \text{ if } v \notin \sigma\ \}$. The topology of $|\sigma|$ is induced by the following distance metric d: for $x, y \in \sigma,\ x = \sum t_i v_i,\ y = \sum s_i v_i$,

$$d(x,y) = \sqrt{\sum (t_i - s_i)^2}.$$

Then $|\sigma|$ and $\langle e_1, \cdots, e_{n+1} \rangle \subset \mathbb{R}^{n+1}$ are isometric, where e_i denotes the vector with a 1 in the ith coordinate and 0's elsewhere. Also $|\sigma|$ is homeomorphic to any affine simplex $\langle a_0, \cdots, a_n \rangle \subset \mathbb{R}^N$ with the subspace topology and this affine simplex $\langle a_0, \cdots, a_n \rangle$ is called a **geometric realization of** $|\sigma|$. Now we define a topology of $|K|$ as a weak topology generated by $\{|\sigma| \mid \sigma \in K\}$. Note that $|\sigma| \cap |\tau|$ is clearly closed in $|\sigma|$ and in $|\tau|$.

We now talk about the Čech complex which is commonly used in practice. To reach the Čech complex, it is a good start from the nerve of an open covering. We refer to the paper by Aleksandroff [1].

Definition 2.0.12. Let X be a topological space, a **cover $\mathcal{U}$ of** X is a collection of subsets of X

$$\mathcal{U} = \{U_v \subset X \mid v \in \mathcal{V}\}$$

such that $\bigcup_{v\in\mathcal{V}} U_v = X$. Moreover if each subset U_v is an open set, then the cover $\mathcal{U}$ is called an **open cover of** X.

Definition 2.0.13. Given an open cover $\mathcal{U} = \{U_v\}_{v\in\mathcal{V}}$ of a topological space X, the **nerve $\mathcal{N}(X,\mathcal{U})$ of $\mathcal{U}$** is an abstract simplicial complex defined as

$$\mathcal{N}(X,\ \mathcal{U}) = \{\sigma \in \mathcal{V} \mid \bigcap_{v\in\sigma} U_v \neq \emptyset\}.$$

We sometimes abbreviate $\mathcal{N}(X,\ \mathcal{U})$ as $\mathcal{N}(\mathcal{U})$.

Remark 2.0.14. The nerve $\mathcal{N}(\mathcal{U})$ is well-defined, i.e., if $\tau \subset \sigma, \sigma \in \mathcal{N}(\mathcal{U})$, then $\tau \in \mathcal{N}(\mathcal{U})$.

Definition 2.0.15. Let X be a topological space, a **refinement of a cover $\mathcal{U}$ of** X is another cover $\mathcal{R}$ of X such that for each $R \in \mathcal{R}$ there exists some $U \in \mathcal{U}$ satisfying $R \subset U$.

Remark 2.0.16. If two covers of a space refine each other, then their nerves have the same homotopy type.

We need several definitions to make a useful notation $K(\mathcal{U})$ for some expression in Nerve Lemma.

Definition 2.0.17. Given a set A, a **(non-strict) partial order** is a homogeneous binary relation $\leq$ over a set A satisfying axioms below: for all $x, y, z \in A$,

(1) (reflexivity) $x \leq x$,

(2) (antisymmetry) if $x \leq y,\ y \leq x$, then $x = y$, and

(3) (transitivity) if $x \leq y,\ y \leq z$, then $x \leq z$.

A set with a partial order is called a **partially ordered set** (*also called* a **poset**).

Definition 2.0.18. Let $\leq$ be a partial order on a set A. This partial order is called a **total order** if for all $x, y \in A$, either $x \leq y$ or $y \leq x$. A set with a total order is called a **totally ordered set**.

A partially ordered set may have subsets that are totally ordered and such subsets are called **chains**. The length l of a finite chain C is one less than the number of elements in the chain:

$$l(C) = |C| - 1.$$

We shall use the following notation: given a open cover $\mathcal{U}$ of a topological space X, let $K(\mathcal{U})$ denote the complex whose vertices are the members of $\mathcal{U}$ and whose simplexes are the finite totally ordered subcollections of $\mathcal{U}$, where $\mathcal{U}$ is partially ordered by inclusion.

Now we have some definitions to describe conditions of Nerve Lemma.

Definition 2.0.19. Given a space X, let a collection of sets $\mathcal{U} = \{U_v\}_{v \in \mathcal{V}}$ be an open cover of a space X. The open cover $\mathcal{U}$ is called **locally finite** if for each point $x \in X$, there exists some neighborhood W_x of x such that the set $\{v \in \mathcal{V} \mid U_v \cap W_x \neq \emptyset\}$ is finite.

Definition 2.0.20. Let X be a space and let a collection of sets $\mathcal{U} = \{U_v\}_{v \in \mathcal{V}}$ be an open cover of a space X. An open cover $\mathcal{U}$ is called to be **basis-like** if for any $v_1, v_2 \in \mathcal{V}$, there exist a subset $\mathcal{W} \in \mathcal{V}$ such that

$$U_{v_1} \cap U_{v_2} \subset \bigcup_{w \in \mathcal{W}} U_w.$$

Definition 2.0.21. A map $f : X \longrightarrow Y$ is called a **weak homotopy equivalence** if

1) f induces an isomorphism of connected components, i.e,

$$\pi_0(f)\colon \pi_0(X) \xrightarrow{\simeq} \pi_0(Y),$$

2) for all $x \in X$ and for all $n \geq 1$, f induces an isomorphism on homotopy groups, i.e.,

$$\pi_n(f, x)\colon \pi_n(X, x) \xrightarrow{\simeq} \pi_n(Y, f(x)).$$

Remark 2.0.22. Homotopy equivalence is an equivalence relation on spaces which is well-defined since homotopy is an equivalence relation on the set of maps. A **homotopy type** is an equivalence class of homotopy equivalent spaces.

Definition 2.0.23. A space X is **homotopically trivial** if $\pi_i(X, x) = 0$ for all $i \geq 0$.

Remark 2.0.24. A contractible space has the homotopy type of a point space; all the homotopy groups of a contractible space are trivial. In other words, any contractible space is homotopically trivial.

The following is obtained from [2], but for ease of reading this connection is discussed in appendix 7.

Theorem 2.0.25. *Let X be a space and let $\mathcal{U} = \{U_v\}_{v \in \mathcal{V}}$ be a locally finite, basis-like, open cover of X with contractible sets U_v. Then there exist a weak homotopy equivalence $f : |K(\mathcal{U})| \longrightarrow X$.*

Note that the result in the theorem 2.0.25 is about not 'homotopy equivalence' but 'weak homotopy equivalence'.

Definition 2.0.26. Let X be a space. A subspace A of X is called a **deformation retract of** X if there is a homotopy $F : X \times [0,1] \longrightarrow X$ such that for all $x \in X$ and $a \in A$,

(1) $F(x,0) = x$,

(2) $F(x,1) \in A$, and

(3) $F(a,1) = a$.

We call this map a **deformation retraction of** X onto A.

Lemma 2.0.27. Let $\mathcal{U}$ be a cover of a space such that the intersection of any finite subcollection of $\mathcal{U}$ is either empty or a member of $\mathcal{U}$. Then $|K(\mathcal{U})|$ is a deformation retract of $|\mathcal{N}(\mathcal{U})|$.

The proof is given in the section 7.

Corollary 2.0.28. (Nerve Lemma) Let X be a space and let $\mathcal{R}$ be a locally finite open cover of X such that the intersection of any (finite) subcollection of $\mathcal{R}$ is contractible. Then there exists a weak homotopy equivalence $|\mathcal{N}(\mathcal{R})| \longrightarrow X$.

The idea of the proof follows the paper [2].

Proof. Let $\mathcal{U}$ be the collection of all nonempty intersection of finite subcollections of $\mathcal{R}$. Observe that the collection $\mathcal{U}$ is a locally finite, basis-like open cover of X. Hence there exists a weak homotopy equivalence $|K(\mathcal{U})| \longrightarrow X$. Since $\mathcal{R} \subset \mathcal{U}$, $\mathcal{U}$ is a refinement of $\mathcal{R}$. By the remark 2.0.16, $|\mathcal{N}(\mathcal{R})|$ is a deformation retract of $|\mathcal{N}(\mathcal{U})|$. Since we have a deformation retraction of $|K(\mathcal{U})|$ onto $|\mathcal{N}(\mathcal{U})|$ by Lemma 2.0.27, the proof is completed. □

Remark 2.0.29. If X has the homotopy type of a CW-complex such as a topological manifold, then we may conclude from a theorem of Whitehead [7] that f is an actual "homotopy equivalence".

Definition 2.0.30. A topological space X is said to be **paracompact** if every open cover has a locally finite open refinement.

The following is a version of Nerve theorem with a paracompact space.

Theorem 2.0.31 (Corollary 4G.3, [18])**.** *If $\mathcal{U}$ is an open cover of a paracompact space X such that every nonempty intersection of finitely many sets in $\mathcal{U}$ is contractible, then X is homotopy equivalent to the nerve $\mathcal{N}(\mathcal{U})$.*

We call such a cover a **good cover**, i.e., an open cover in which all sets and all intersections of finitely many sets are contractible.

Remark 2.0.32. Nerve theorem gives conditions under which the nerve of a cover is equivalent to the underlying space. We can see several examples related to this theorem: one for closed covers by K. Borsuk [8] and another one for open covers by A.Weil[9], both for homotopy equivalences. After their attribution, several generalizations were made. First generalizations by W. Holsztynski [10] and J.N.Haimov [11] relaxed conditions on the cover. Moreover, weak homotopy equivalence were studied in the context by M.McCord[2] and weak $n-$homotopy equivalence by A.Bjorner [12].

Definition 2.0.33. Let M be a metric space. Given a point set X in M and a number $\epsilon > 0$, the **Čech complex** $\check{C}(X, \epsilon)$ is the simplicial complex whose simplices are formed as follows: for each subset $S \subset X$ of points, form a $(\epsilon/2)-$ball around each point in S, and include S as a simplex of dimension $|S|$ if there is a common point contained in all of the balls in S.

Remark 2.0.34. The Čech complex is well-defined as a simplicial complex since any subsets $\sigma \subset S$ of a simplex S is a simplex.

Remark 2.0.35. Notice that the Čech complex is the nerve of the set of $(\epsilon/2)-$balls centered at points of X. By the Nerve theorem, the Čech complex is homotopy equivalent to the union of the balls.

Since it is not easy to figure out the intersections of balls, it is better to use a variant version of the Čech complex, a **Vietoris-Rips complex**.

Definition 2.0.36. Let (X, d) be a metric space, the **Vietoris-Rips complex** $VR(X, \epsilon)$ for X with the parameter ϵ, is the simplicial complex whose vertex set is X and a finite subset $\{x_0, \cdots, x_k\}$ spans a $k-$simplex $[x_0, \cdots, x_k]$ if $d(x_i, x_j) \leq \epsilon$ for all $0 \leq i, j \leq k$.

Note that the notation $VR_\epsilon(X)$ is also used instead of $VR(X, \epsilon)$. We some times use notations $VR(X, d, \epsilon)$ and $VR_\epsilon(X, d)$ to denote the distance metric d on the space X.

By the triangle inequality, we have the following lemma.

Lemma 2.0.37. $\check{C}(X, \epsilon) \subset VR(X, 2\epsilon) \subset \check{C}(X, 2\epsilon)$.

We can also represent a Vietoris-Rips complex as Categorical object. Recall some definitions in Category.

Definition 2.0.38. A **colimit in a category** $\mathcal{C}$ is the same as a limit in the opposite category $\mathcal{C}^{op}$.

Definition 2.0.39. Let $\mathcal{C}$ be a category with weak equivalences and let D be a (small) diagram category. The **homotopy colimit of a functor** $F : \mathcal{D} \to \mathcal{C}$ is, if it exists, the image of F under the left derived functor of the colimit functor $\mathrm{colim}\mathcal{D} : [\mathcal{D}, \mathcal{C}] \to \mathcal{C}$ with respect to the given weak equivalences on $\mathcal{C}$ and the objectwise weak equivalences on $[\mathcal{D}, \mathcal{C}]$:

$$\mathrm{hocolim}_{\mathcal{D}} F = (\mathbb{L}\mathrm{colim}_{\mathcal{D}})F.$$

Remark 2.0.40. Let (X, d) be metric space. Denote $F(X)$ as the poset of all finite subset of X ordered by inclusion. We can consider the following: for any $r > 0$, $VR(-, r) : F(-) \longrightarrow \mathcal{T}op$ is a functor. Hence

$$VR(X, r) = \mathrm{colim}_{Y \in F(X)} VR(Y, r) \simeq \mathrm{hocolim}_{Y \in F(X)} VR(Y, r).$$

Note this homotopy equivalence is induced from inclusions of closed subcomplex $VR(Y,r) \hookrightarrow VR(Y',r)$ for any $Y \subseteq Y'$, i.e., cofibrations.

2.1 Nerve theorem with Vietoris-Rips complex

In this section we observe some results of combining Nerve theorem and Vietoris-Rip complexes in metric space.

Hausmann proved the following, which appears in [20].

Theorem 2.1.1. *For a closed Riemannian manifold X and ϵ sufficiently small the geometric realization $|VR_\epsilon(X)|$ of this complex is homotopy equivalent to X.*

Given a metric space X, the set of closed sets of X supports a metric, *the Hausdorff metric*. For a set $A \in X$ and a number $r > 0$, define **the $r-$neighborhood**(*or* **the $r-$thickening**) of A as the set

$$A^{(r)} = \bigcup_{x \in A} B_x(r)$$

where $B_x(r)$ is the open ball of radius r centred at x.

Definition 2.1.2. Let $A, B \subset X$ be closed sets. The **Hausdorff distance** $d_H(A,B)$ is defined by

$$d_H(A,B) = \inf\{r > 0 \mid B \subset A^{(r)},\ A \subset B^{(r)}\}.$$

Recall *the Gromov-Hausdorff distance* is the extension of the Hausdorff distance.

Definition 2.1.3. Given two closed metric spaces A and B, the **Gromov-Hausdorff distance** is defined as

$$d_{GH}(A, B) = \inf_{f,g} d_H(f(A),\ g(B))$$

where maps $f : A \to X$ and $g : B \to X$ are isometric embeddings and the infimum is taken over all possible such embeddings.

The following is a well known property about Gromov-Hausdorff distance.

Proposition 2.1.4. Let (X, d_X) and (Y, d_Y) be metric spaces that admit compact exhaustions. If $d_{GH}(X, Y) = 0$, then (X, d_X) and (Y, d_Y) are isometric.

The following is obtained from a theorem of Latschev [21].

Theorem 2.1.5. *Let X be a closed Riemannian manifold. Then there exists ϵ_0 such that for every $0 \leq \epsilon \leq \epsilon_0$ there exists a $\delta > 0$ such that the geometric realization $|VR_\epsilon(Y)|$ of the $\epsilon-$complex of any metric space Y which has Gromvov -Hausdorff distance less than δ to X is homotopy equivalent to X.*

Here is the unsurprising formulation: if we have enough sample points, and look at small scale, then the Vietoris-Rips complex will reproduce the correct homology. It make us move on to the problem of finding what scale is small enough.

Chapter 3

Persistence

Nerve Theorem has an beneficial effect on the area of topological data analysis (TDA). The purpose of TDA is to obtain information about the topology of a space which is given as a discrete sample of the space, commonly called *point cloud data*. A powerful tool in TDA is *persistent homology*, which computes not the homology of a single space but the homology of a filtration. Using the homology functor, we get a persistence module. Computing homology with field coefficients, we can obtain a complete topological invariant called *persistence barcode* or *persistence diagram*. In this chapter, we handle the persistent homology with Vietoris-Rips complex.

3.1 Persistent homology

Persistent homology is a way to record the homology of a so-called filtered space at different filtration levels. It plays a major role in topological data analysis, and can be applied to many different settings.

Let us start by recalling some definitions.

Definition 3.1.1. Given a subset $T \subset \mathbb{R}$ (or more generally some ordered set), a **filtration** F over T is a family of objects (topological spaces, rings, etc) X_i parametrized by T such that $X_i \subset X_j$ whenever $i \leq j$.

This can be phrased with categorical language. A filtered space will be a filtration of topological spaces. Simply put

$$X_0 \subset X_1 \subset X_2 \subset \dots,$$

where $X_0 = \emptyset$. We will often assume that the filtration stabilizes. Basic examples are

- sublevel sets of a (continuous) function $f : X \to \mathbb{R}$, and
- the Vietoris-Rips complex associated with a metric space (X, d), namely $VR_\epsilon(X, d)$. We have inclusions $VR_{\epsilon'}(X, d) \to VR_\epsilon(X, d)$ for $\epsilon \geq \epsilon'$.

Given a fixed filtration level, say X_i of a filtered space $\{X_i\}_i$, we can of course compute the homology, but this homology does not carry so much information. For example, if f_{ij} is the inclusion $X_i \to X_j$, we also want to the effect of this map of homology to see which cycles survive. This leads to the following definition.

Definition 3.1.2. A **persistence module** or **persistence vector space** consists of

1. a family of vector spaces $\{V^r\}_{r \in \mathbb{R}_{\geq 0}}$,
2. for $r' \geq r$, linear maps $f^{r,r'} : V^r \to V^{r'}$ satisfying

$$f^{r_2,r_3} \circ f^{r_1,r_2} = f^{r_1,r_3}.$$

We will call these maps *continuation maps*.

We call a persistence module **q-tame** (quadrant-tame) if the ranks of the continuation maps are finite.

In order to use naturality, we need to define a category of persistence vector spaces. While we can define morphisms between persistence vector spaces directly, it is more economical to first recast the above definition in more categorical language.

Let $\mathcal{P}$ denote the category $Poset(\mathbb{R}_{>0}, <)$ or $Poset(\mathbb{N}_0, <)$.

Definition 3.1.3. A **persistence object** P in a category $\mathcal{C}$ is just a functor $P : \mathcal{P} \to \mathcal{C}$.

With this in mind, a persistence vector space is just a persistence object in the category of vector spaces.

The **category of persistence vector spaces** is then the functor category of persistence objects in the category of vector spaces. This means that morphisms between a persistence vector spaces $\mathbb{V} = (V_i, f_{ij})$ are just natural transformations. Written out, this means that a morphism between persistence vector spaces $\{V^r, I^{r,r'}\}_{r,r'}$ and $\{W^r, J^{r,r'}\}_{r,r'}$ is a collection of linear maps $f^r : V^r \to W^r$ such all the following diagrams commute,

$$\begin{array}{ccc} V^r & \xrightarrow{I^{r,r'}} & V^{r'} \\ \downarrow{\scriptstyle f^r} & & \downarrow{\scriptstyle f^{r'}} \\ W^r & \xrightarrow{J^{r,r'}} & W^{r'}. \end{array}$$

3.1.1 Tameness and barcodes

Finite dimensional vector spaces over a field k are isomorphic to k^r for some $r \in \mathbb{N}_0$, so we have a practical model which we can always work with. We want to have a similar description for persistence vector spaces. The basic building

block will be the so-called **interval module** $I(a, b)$ for an interval $[a, b)$,

$$\{\mathcal{I}(a,b)\}^r = \begin{cases} k & r \in [a,b), \\ 0 & \text{otherwise.} \end{cases}$$

with maps f^{r_1,r_2} satisfying

$$f^{r_1,r_2} = \begin{cases} id & r_1, r_2 \in [a,b), \\ 0 & \text{otherwise.} \end{cases}$$

The concept of direct sum can be defined on persistence vector spaces. If $\mathbb{V} = (V_i, v_{ij})$ and $\mathbb{W} = (W_i, w_{ij})$ are persistence vector spaces, then we define the direct sum $\mathbb{V} \oplus \mathbb{W}$ by

$$(V_i \oplus W_i, v_{ij} \oplus w_{ij}).$$

Definition 3.1.4. We call a persistence module $\mathbb{W}$ **indecomposable** if the only decomposition $\mathbb{W} = \mathbb{U} \oplus \mathbb{V}$ is the trivial decomposition, i.e., $\mathbb{U}$ or $\mathbb{V}$ is trivial.

Definition 3.1.5. Finitely generated persistence vector spaces are isomorphic to a direct sum of of $\mathcal{I}(a, b)$ or $\mathcal{I}(c, \infty)$.

Theorem 3.1.6 (Gabriel)**.** *Suppose that $\mathbb{V}$ is a persistence module over a set $T \subset \mathbb{R}$. Assume that one of the following is true:*

1. *the set T is finite, or*
2. *V_t is finite-dimensional for all $t \in T$.*

Then $\mathbb{V}$ is isomorphic to a direct sum of interval modules.

We call such a persistence module $\mathbb{V}$, which is the direct sum of interval modules, **interval-decomposable**.

The above assumptions can be restrictive, and the class of q-tame filtrations can be better when dealing with stability.

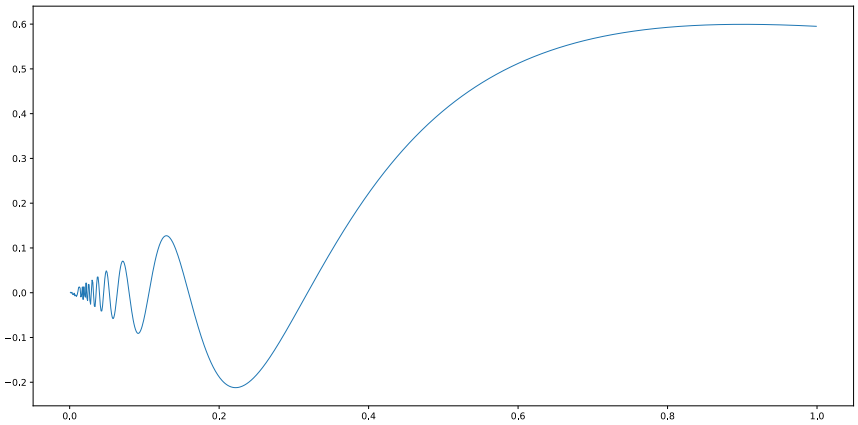


Figure 3.1: A dampened topologist's sine curve

For example, consider the example in Figure 3.1: let X be the graph of the function $x \mapsto \frac{x}{\sqrt{x^2+1}} \sin(1/x)$, and filter this space by

$$X_t = \{(x, y) \mid y = \frac{x}{\sqrt{x^2+1}} \sin(1/x) < t\}.$$

The sublevel set X_0 has infinitely many components, so we can not directly apply the above theorem. The concept of q-tameness is better for this as we have.

Theorem 3.1.7. *Let X be a compact simplicial complex, i.e., topological space, and assume that $f : X \to \mathbb{R}$ is a continuous function. Then the persistent homology of the sublevel set filtration of f, namely $H(X^{f \leq \cdot})$, is q-tame.*

Remark 3.1.8. Of course with data, we only have finitely many points, so then Theorem 3.1.6 applies. However, for stability one needs Theorem 3.1.7.

We will describe persistence in detail in this section. Most of the material is extracted from [16].

Definition 3.1.9. Let $\mathcal{T}$ denote any totally ordered set and R be a ring. **A $R-$persistence module parameterized by $\mathcal{T}$** is a family of $R-$modules $\{M_t\}_{t\in\mathcal{T}}$ together with homomorphisms of $R-$modules $\varphi_{t,t'} : M_t \to M_{t'}$ for all $t \leq t'$ such that the homomorphisms are compatible, i.e.,

$$\varphi_{t,t'} \circ \varphi_{t',t''} = \varphi_{t,t''} \quad \text{whenever} \quad t \leq t' \leq t''.$$

Persistence modules bring with the information contained in the homomorphisms, and also are computable in the time required for computing a single homology group.

Example 3.1.10. Assume a family of spaces X_ϵ is parameterized by the real valued parameter ϵ satisfying $X_\epsilon \subseteq X'_\epsilon$ if $\epsilon \leq \epsilon'$. Then the family of Abelian groups $H_n(X_\epsilon, R)$, where R is a ring, becomes a $R-$persistence module parameterized by $\mathbb{R}$. Moreover, if R is a field, then this family is a $R-$persistence vector space parameterized by $\mathbb{R}$ over the field R.

Definition 3.1.11. Given $\mathcal{T} = \mathbb{R}$, a persistence module $\{V_t\}_t$ over a field F is **of finite type** if there are a finite number of unique finite-dimensional vector spaces in the persistence module.

Let I be an interval. Define *a persistence vector space $Q(I)$ over a field F* as

$$Q(I_t) = \begin{cases} F, & \text{if } t \in I, \\ 0, & \text{otherwise,} \end{cases} \tag{3.1.1}$$

where the homomorphism is the identity within each interval.

We recall the following property due to Carlsson.

Theorem 3.1.12 (Proposition 5.2., [16])**.** *A persistence module parameterized by* $\mathbb{R}$ *over a field* F *of finite type is isomorphic to one of the form*

$$\bigoplus_{t=1}^{n} Q(I_t),$$

where each interval I_t *is bounded from below, and the description is unique up to the order of the intervals.*

Now it is natural to name the intervals in the above theorem.

Definition 3.1.13. A barcode is a finite multiset of intervals that are bounded below.

Remark 3.1.14. From a barcode, we have comprehension of the space in the sense of treating the intervals as the life times of non-trivial cycles in a growing complex. In detail, we consider the left endpoint of an interval as the birth of a new topological attribute, and the right endpoint as its death. Intuitively, the longer the interval, the more important the topological attribute, since it persists in being a feature of the complex.

Persistent homology can be used to measure the scale or resolution of a topological feature. Let's introduce persistent homology in detail.

Definition 3.1.15. Given a simplicial complex K, a function $f : K \to \mathbb{R}$ is called to be **monotonic** if it is non-decreasing along increasing chains of faces, i.e., $f(\sigma) \geq f(\tau)$ whenever σ is a face of τ.

For a simplicial complex K, if we have a monotonic function $f : K \to \mathbb{R}$, then this monotonicity implies that the sublevel set, $K(a) = f^{-1}(-\infty, a]$, is a subcomplex of K for every $a \in \mathbb{R}$. Indeed if $a_1 < a_2 < \cdots < a_n$ are the

function values of the simplices in K with $a_0 = -\infty$, define $K_i = K(a_i)$ for each i. Hence we get an increasing sequence

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K.$$

This sequence of complexes is the **filtration** of f. More than in the sequence of complexes, we are interested in the topological evolution, as expressed by the corresponding sequence of homology groups. For every $i \leq j$ we have an inclusion map from the underlying space of K_i to that of K_j and therefore an induced homomorphism,

$$f_p^{i,j} : H_p(K_i) \to H_p(K_j), \quad \text{for each dimension } \ p.$$

The filtration thus corresponds to a sequence of homology groups connected by homomorphisms,

$$0 = H_p(K_0) \to H_p(K_1) \to \cdots \to H_p(K_n) = H_p(K),$$

for each dimension p. As we go from K_i to K_{i+1}, we gain new homology classes and we lose some when they become trivial or merge with each other. We collect the classes that are born at or before a given threshold and die after another threshold in groups.

Definition 3.1.16. The $p-$**th persistent homology groups** are the images of the homomorphisms induced by inclusion, $PH_p^{i,j} = \text{im} f_p^{i,j}$, for $0 \leq i \leq j \leq n$. The corresponding $p-$th **persistent Betti numbers** are the ranks of these groups, $\beta_p^{i,j} = \text{rank} PH_p^{i,j}$.

Persistence diagrams. We visualize the collection of persistent Betti numbers by drawing points in two dimensions. Some of these points may have infinite coordinates and some might be the same, so we want to consider a

multiset of points in the extended real plane $\overline{\mathbb{R}}^2$. Denote $\mu_p^{i,j}$ as the number of $p-$dimensional classes born at K_i and dying entering K_j, we have

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j}),$$

for all $i < j$ and all p. In detail, $\beta_p^{i,j-1} - \beta_p^{i,j}$ counts the classes that are born at or before K_i and die entering K_j, and $\beta_p^{i-1,j-1} - \beta_p^{i-1,j}$ counts the classes that are born at or before K_{i-1} and die entering K_j. Drawing each point (a_i, a_j) with multiplicity $\mu_p^{i,j}$, we get the $p-$**th persistence diagram** of the filtration, denoted as $\mathrm{Dgm}_p(f)$. It represents a class by a point whose vertical distance to the diagonal is the diagonal is the persistence. Since the multiplicities are defined only for $i < j$, all points lie above the diagonal. In other words, none of classes can be born before dying. Moreover we can take $\beta_p^{i,j}$ as the number of points in the upper, left quadrant with corner point (a_k, a_l). Note that a class which is born at K_i and dies entering K_j is counted if and only if $a_i \leq a_k$ and $a_j > a_l$.

Theorem 3.1.17. *(Fundamental Lemma of Persistent Homology). Let* $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$ *be a filtration. For every pair of indices* $0 \leq k \leq l \leq n$ *and every dimension* p*, the* $p-$*th persistent Betti number is* $\sum_{i \leq k} \sum_{j > l} \mu_p^{i,j}$.

Interpretation of this theorem is the diagram encodes the entire information about persistent homology groups.

3.2 The Isometry theorem

In this section we discuss the metric relationship between persistence modules and their persistence diagrams.

Let $\mathbb{U}$ and $\mathbb{V}$ be persistence modules over $\mathbb{R}$, and let δ be any real number.

A **homomorphism of degree** δ is a collection Φ of linear maps

$$\phi_t : U_t \to V_{t+\delta}$$

for all $t \in \mathbb{R}$, such that the diagram

$$\begin{array}{ccc} U_s & \xrightarrow{u_t^s} & U_t \\ {\scriptstyle \phi_s}\downarrow & & \downarrow{\scriptstyle \phi_t} \\ V_{s+\delta} & \xrightarrow{v_{t+\delta}^{s+\delta}} & V_{t+\delta} \end{array}$$

commutes whenever $s \leq t$. Denote

$$\mathrm{Hom}^\delta(\mathbb{U}, \mathbb{V}) = \{ \text{ homomorphisms } \mathbb{U} \to \mathbb{V} \text{ of degree } \delta \ \}.$$

Let $\delta \geq 0$. Two persistence modules $\mathbb{U}$ and $\mathbb{V}$ are said to be $\delta-$**interleaved** if there are maps

$$\Phi \in \mathrm{Hom}^\delta(\mathbb{U}, \mathbb{V}),\ \Psi \in \mathrm{Hom}^\delta(\mathbb{U}, \mathbb{V})$$

such that

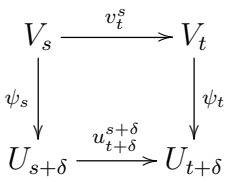
$$\Psi\Phi = 1_{\mathbb{U}}^{2\delta},\ \Phi\Psi = 1_{\mathbb{V}}^{2\delta}.$$

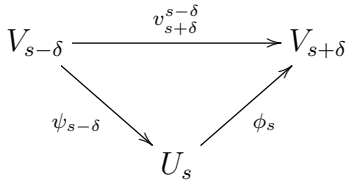
More expansively, there are maps

$$\phi_t : U_t \to V_{t+\delta} \quad \text{and} \quad \psi_t : V_t \to U_{t+\delta}$$

defined for all t, such that the following diagrams

$$\begin{array}{ccc} U_s & \xrightarrow{u_t^s} & U_t \\ {\scriptstyle \phi_s}\downarrow & & \downarrow{\scriptstyle \phi_t} \\ V_{s+\delta} & \xrightarrow{v_{t+\delta}^{s+\delta}} & V_{t+\delta} \end{array} \qquad \begin{array}{ccc} U_{s-\delta} & & \xrightarrow{u_{s+\delta}^{s-\delta}} & & U_{s+\delta} \\ & {\scriptstyle \phi_{s-\delta}}\searrow & & \nearrow{\scriptstyle \psi_s} & \\ & & V_s & & \end{array}$$





commute for all $s \leq t$.

Lemma 3.2.1. (Interpolation Lemma) Suppose $\mathbb{U}$ and $\mathbb{V}$ are a $\delta-$interleaved pair of persistence modules. Then there exists a $1-$parameter family of persistence modules $(\mathbb{U}_x \mid s \in [0, \delta])$ such that $\mathbb{U}_0, \mathbb{U}_\delta$ are equal to $\mathbb{U}, \mathbb{V}$ respectively, and $\mathbb{U}_x, \mathbb{U}_y$ are $|y - x|-$interleaved for all $x, y \in [0, \delta]$. Moreover, if $\mathbb{U}$ and $\mathbb{V}$ are $q-$tame, then the $(\mathbb{U}_x)$ may be assumed $q-$tame also.

We say that two persistence modules $\mathbb{U}$ and $\mathbb{V}$ are δ^+-**interleaved** if there are $(\delta + \epsilon)-$interleaved for all $\epsilon > 0$.

Definition 3.2.2. The **interleaving distance** between two persistence modules is defined as

$$\begin{aligned} \mathrm{d}_i(\mathbb{U}, \mathbb{V}) &= \inf\{\ \delta \mid \mathbb{U} \text{ and } \mathbb{V} \text{ are } \delta - \text{interleaved }\} \\ &= \min\{\ \delta \mid \mathbb{U} \text{ and } \mathbb{V} \text{ are } \delta^+ - \text{interleaved }\}. \end{aligned}$$

If there is no $\delta-$interleaving between $\mathbb{U}$ and $\mathbb{V}$ for any value of δ, then $\mathrm{d}_i(\mathbb{U}, \mathbb{V}) = \infty$.

A **partial matching** between A and B is a collection of pairs

$$M \subset A \times B$$

such that

1. for every $\alpha \in A$ there is at most one $\beta \in B$ such that $(\alpha, \beta) \in M$, and

2. for every $\beta \in B$ there is at most one $\alpha \in A$ such that $(\alpha, \beta) \in M$.

We Say that a partial matching M is a $\delta-$**matching** if all of the following are true:

1. if$(\alpha, \beta) \in M$ then $d^{\infty}(\alpha, \beta) \leq \delta$;
2. if $\alpha \in A$ is unmatched then $d^{\infty}(\alpha, \Delta) \leq \delta$;
3. if $\beta \in B$ is unmatched then $d^{\infty}(\beta, \Delta) \leq \delta$.

Definition 3.2.3. The **bottleneck distance** between two multisets A, B in the extended half-plane is

$$\mathrm{d}_b(A, B) = \inf(\ \delta \mid \text{there exists a } \delta - \text{matching between } \ A \text{ and } \ B).$$

Theorem 3.2.4. *Let $\mathbb{U}, \mathbb{V}$ be q-tame persistence modules. Then*

$$d_i(\mathbb{U}, \mathbb{V}) = d_b(dgm(\mathbb{U}), dgm(\mathbb{V})).$$

Theorem 3.2.5 (The Isometry Theorem/ Stability theorem)**.** *Let X be a topological space homeomorphic to a finite simplicial complex, and let $f, g : X \to \mathbb{R}$ be continuous functions. Then*

$$d_b(dgm(f), dgm(g)) \leq ||f - g||_{\infty}.$$

This result can be interpreted by saying that barcodes are resistant to noise.

Chapter 4

2n-problem

We consider a topological space X, say a CW-complex so that the Eilenberg-Steenrod axioms suffice to compute its homology. We have the following natural questions:

- what is the minimal number of cells or points of a CW-complex with the same (isomorphic) homology as that of X?
- what is the minimal number of points of a simplicial complex with the same homology as that of X?

In this chapter, we will observe how to find a proper number of data point for the persistent homology of the standard space.

We define

$$N(X) := \min\{\#(Y, d_Y) \mid \text{there is } \epsilon > 0 \ st.\ H_*(VR_\epsilon(Y, d_Y); F) \cong H_*(X; F)\}$$

and

$$N(X,d) := \min\{\#Y \subset X \mid \text{there is } \epsilon > 0 \ st.\ H_*(VR_\epsilon(Y,d);F) \cong H_*(X;F)\}.$$

Indeed, $N(X)$ means the minimal number of cardinality of vertex sets such that the persistent homology of the related Vietoris-Rips complexes has the same homology of the space X for some parameter. Similarly, we can consider $N(X,d)$ with a given metric.

Remark 4.0.1. If (X,d) is a metric CW complex, then we only have

$$N(X,d) \geq N(X),$$

and many cases this inequality will be strict.

Example 4.0.2. By the definition of Vietoris-Rips complex, three vertices have a cycle which is represented as the boundary of a $2-$simplex. We can consider the set $\{A=(0,0), B=(1,0), C=(1,1), D=(0,1)\}$ over $\mathbb{R}^2$: for $1 \leq \epsilon < \sqrt{2}$, there is no $2-$simplex and one cycle $\{A,B\}+\{B,C\}+\{C,D\}+\{D,A\}$. Thus we have $N(S^1)=4$.

4.1 Examples

Terminology & Notation

1. The **cubic shaped sphere** $\mathcal{C}$ is the boundary of the unit cube $I^3 = [0,1]^3$, i.e., the union of 6 faces of the unit cube.

2. $N_{\mathcal{C}}(S^2)$ is the notation of the least number of points on the cubic shaped sphere $\mathcal{C}$ having a 2-dimensional barcode of the persistent homology of S^2.

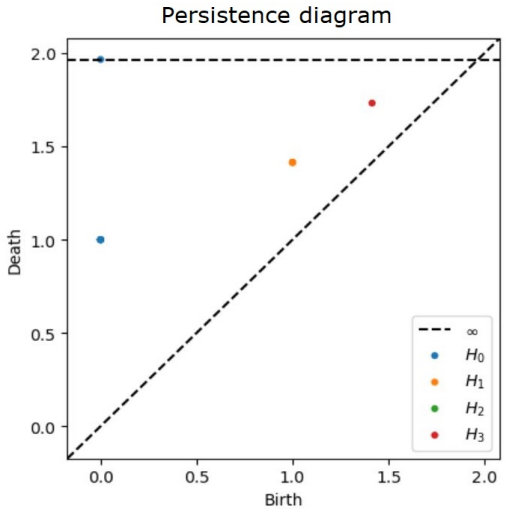
Question On the cubic shaped sphere $\mathcal{C}$, at least how many points do we need to get a 2-dimensional barcode of the persistent homology like 2-dimensional generator of the homology of the unit sphere S^2?

We use the python module *ripser.py* to compute barcodes for samples of *data8* and *data6* in this section. Instead of the module ripser.py, you may use the python module *gudhi.py* or the c++ library *gudhi* which offer a powerful tool for computing various complexes and their homology.

Example 4.1.1. data8 = [[0,0,0], [0,0,1], [0,1,0], [0,1,1], [1,0,1], [1,1,1], [1,1,0], [1,0,0]] is the collection of 8 corner points in $\mathcal{C}$. The result of the persistent homology of data8 with the Euclidean distance is the following:

value range: $[1, \sqrt{3}]$

distance matrix with 8 points
persistence intervals in dim 0:
[0,1)
[0,1)
[0,1)
[0,1)
[0,1)
[0,1)
[0,1)
$[0, \infty)$

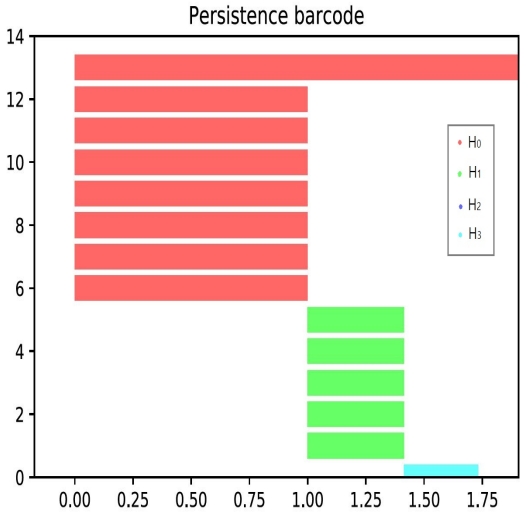


persistence intervals in dim 1:
$[1, \sqrt{2})$
$[1, \sqrt{2})$
$[1, \sqrt{2})$
$[1, \sqrt{2})$
$[1, \sqrt{2})$

persistence intervals in dim 2:

persistence intervals in dim 3:
$[\sqrt{2}, \sqrt{3})$

persistence intervals in dim 4:



Remark 4.1.2. We can observe that there is no 2-dimensional generator in the sample of data8.

Example 4.1.3. data6 = [[0,1/2,1/2], [1,1/2,1/2], [1/2,0,1/2], [1/2,1,1/2], [1/2,1/2,0], [1/2,1/2,1]] is the collection of 6 central points on each face of the unit cube. The result of the persistent homology of data6 with the Euclidean distance is the following:

value range: $[\sqrt{2}/2,1]$

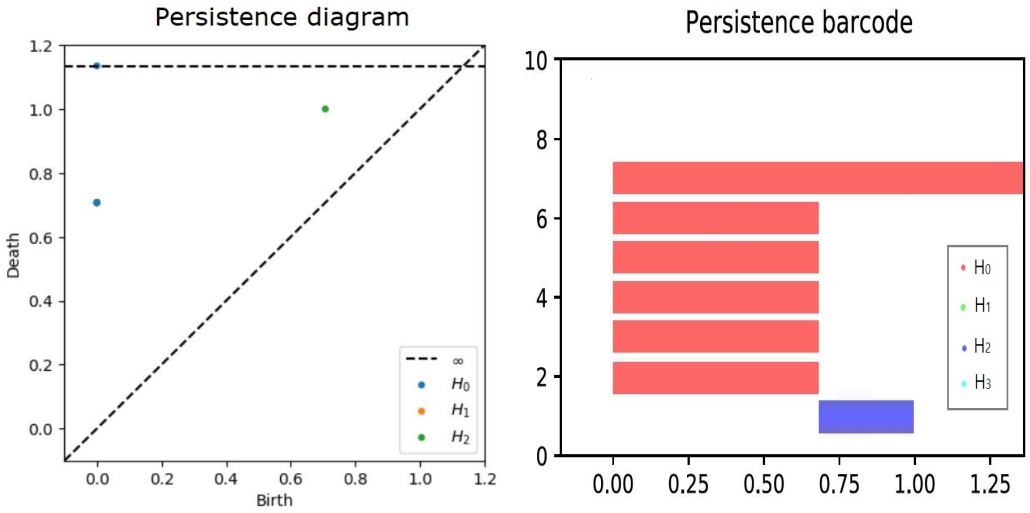
distance matrix with 6 points
persistence intervals in dim 0:
$[0, \sqrt{2}/2)$
$[0, \sqrt{2}/2)$
$[0, \sqrt{2}/2)$
$[0, \sqrt{2}/2)$
$[0, \sqrt{2}/2)$
$[0, \infty)$

persistence intervals in dim 1:

persistence intervals in dim 2:
$[\sqrt{2}/2, 1)$

persistence intervals in dim 3:

persistence intervals in dim 4:



Remark 4.1.4. Since data6 has a 2-dimensional barcode, we get $N_{\mathcal{C}}(S^2) \leq 6$.

4.2 Minimal Construction for S^2

In this section, we will find the answer of the question for finding a proper number of points such that the Vietoris-Rips complex of those points gets the same homology of S^2.

Notation Let $\mathcal{C}_r$ be a Vietoris-Rips complex with the parameter r.

Lemma 4.2.1. There exist at least four 2−simplices for 2−dimensional generator of the persistent homology.

STRATEGY Check case by case according to the number of intersected 1-simplices among 2-simplices and don't care intersected 0-simplices since we are looking for 2-dimensional cycles.

Proof.
Suppose there exists only one 2-simplex $[A, B, C]$, see Figure 1. Its boundary would be the union of three distinct 1-simplices and so such a $2-$simplex is not even a cycle.

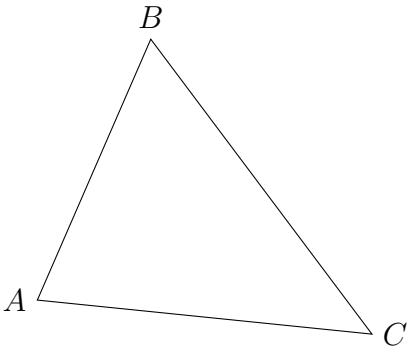


Figure 1

Suppose there exist two 2-simplices. There are two cases of two 2-simplices. One is two 2-simplices with one intersected 1- simplex , see Figure 2, and the other are two 2-simplices without a intersected 1-simplex, see Figure 3 & 4. The boundary of the former case is the union of four distinct 1-simplice and the latter case is the union of six distinct 1-simplices. None of them are a 2-dimensional cycle.

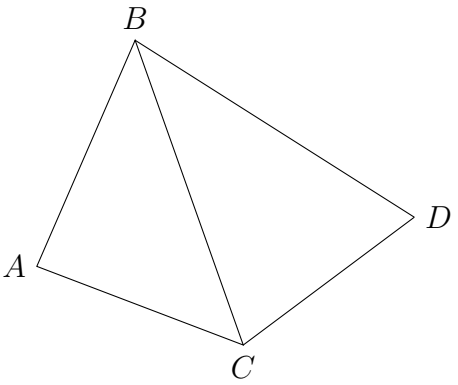


Figure 2

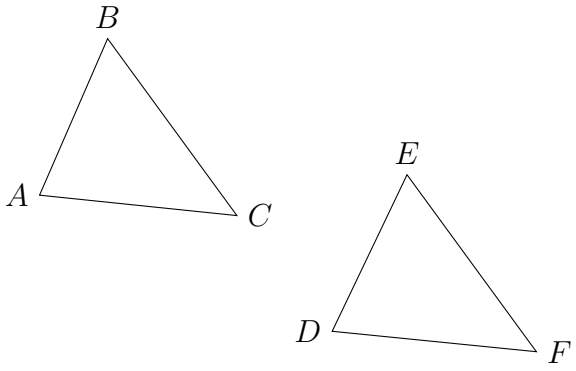


Figure 3

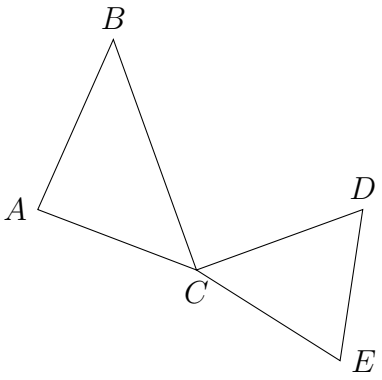


Figure 4

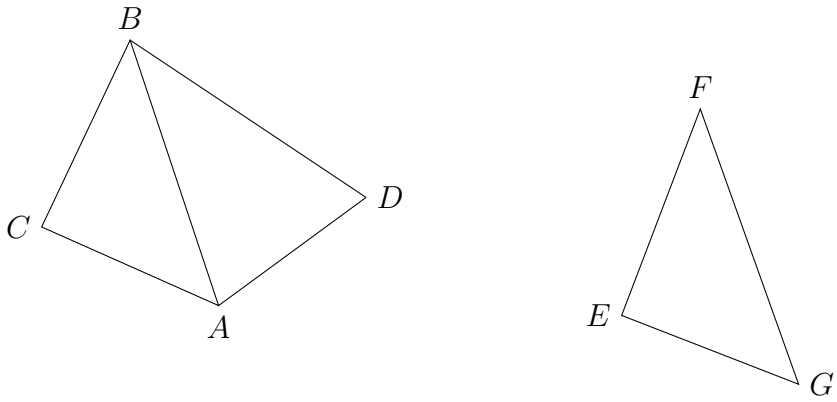


Figure 5

Suppose there exist three 2-simplices. Automatically if there is no intersected 1−simplex among them, then it is not a cycle by its boundaries. Assume there exists one intersected 1-simplex, say {A,B} so that there are two cases of three 2-simplices. The first case is only two 2-simplices with one intersected 1-simplex, see Figure 5,6, &7 and the second case is three 2-simplices having the same 1-simplex, see Figure 8. The boundary of both cases is the union of seven 1-simplices and it is not a cycle.

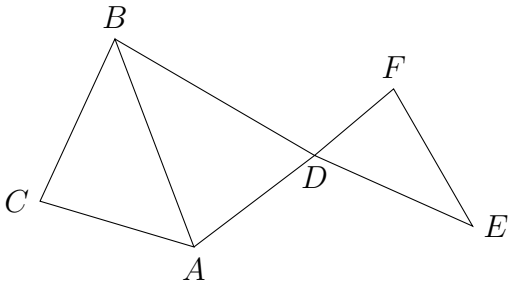


Figure 6

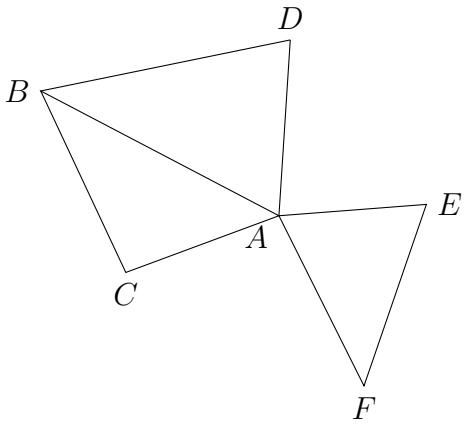


Figure 7

And if there exists two intersected 1-simplices, say {A,B} and {A′, B′}, then we have two pairs of 2-simplices associated with intersected 1-simplices. Since there are only three 2-simplices, two pairs have a common 2-simplex, i.e., {A=A′, B, B′}, see Figure 9. Hence its boundary is the union of five 1-simplices and so it is not a cycle. Therefore, we need more than three 2-simplices to have $H_2(\mathcal{C}_r) \neq 0$. □

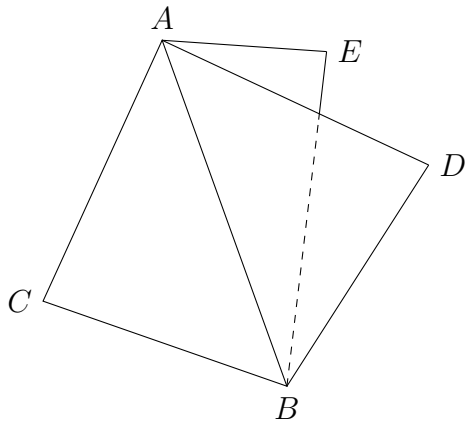


Figure 8

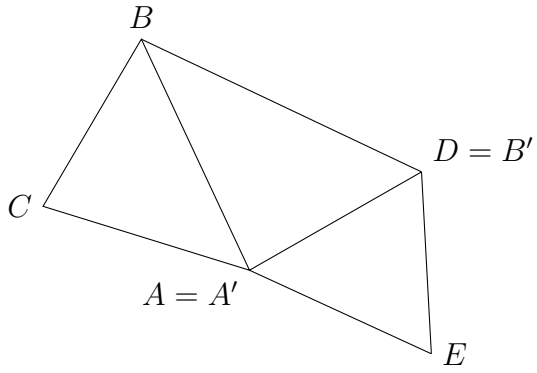


Figure 9

Remark 4.2.2. A tetrahedron has no 2-dimensional generator of persistent homology. Since a tetrahedron is a complete graph, all edges are measured among 4 points, say v_i for $i = 0, 1, 2, 3$. Let r be the maximum length of edges of this tetrahedron. Consider balls of radius r with center v_i for $i = 0, 1, 2, 3$, then we have four 2-simplices which become a cycle of 2-dimensional persistent homology. However this 2-dimensional cycle is the boundary of the 3-simplex $\{v_0, v_1, v_2, v_3\}$. Hence there is no 2-dimensional generator of a tetrahedron.

Theorem 4.2.3. *Then there are at least 6 points such that the induced Vietoris-Rips complex has a* $2-$*dimensional generator in its persistent homology.*

Proof. By the Lemma 4.2.1, we have at least four $2-$simplices. The minimum number of points having four $2-$simplices is 4 vertices since ${}_4\mathbf{C}_3 = 4$. Thus we get a least 4 points and there is only one $2-$dimensional cycle, *the boundary of a tetrahedron* induced from 4 points, which is not a $2-$generator. Thus we have at least 5 points to get a $2-$dimensional generator in its persistent homology. Now it is enough to check that 5 vertices cannot have 2-dimensional generator of persistent homology. Our strategy starts from the most popular 1-simplex and the most popular 0-simplex among complices.

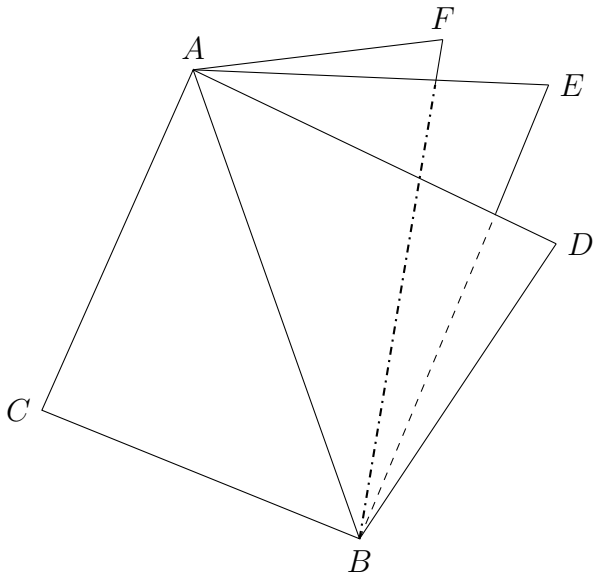


Figure 10

Let σ be a k-simplex. Define $\mathcal{M}_\sigma = \{\tau | \sigma \subset \tau, \tau : \text{a } (\mathrm{k}+1) - \text{simplex}\}$ as a collection of 1 dimension higher complices containing σ and $\mu_\sigma = |\mathcal{M}_\sigma|$ as its number of elements. Consider $M_k = \max_{\sigma \,:\, \text{a k}-\text{simplex}} \mu_\sigma$.

If $M_1 \leq 1$, then at least four 2-simplices have the boundary of twelve 1-simplices. This number is more than the total number of 1-simplices with 5 vertices, i.e., ${}_5\mathcal{C}_2 = 10$ (ten) 1-simplices. On the other hand, if $M_1 \geq 4$, then there should be more than 5 vertices, see Figure 10. Hence we observe that $M_1 = 2$ or 3.

Assume $M_1 = 3$. Consider it with Figure 8. Since we have at least four 2-simplices, there should be at least one 1-simplex among {C,D}, {D,E} and {C,E}. Without loss of generality, if there exists only one of these 1-simplices, then it turns to a tetrahedron with a connected 2-simplex, see Figure 11. If there are only two of them, then it turns to an object of two tetrahedrons having a common face, see Figure 12.

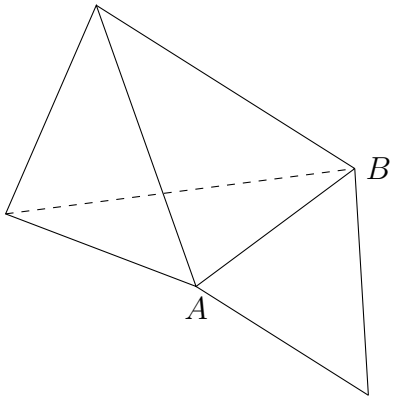


Figure 11

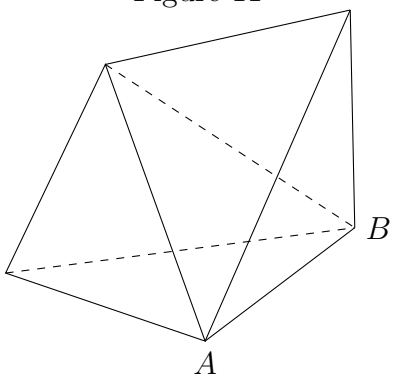


Figure 12

Finally, if there are all of them, then it turns to a complete graph, the maximal complex with 5 vertices, see Figure13.

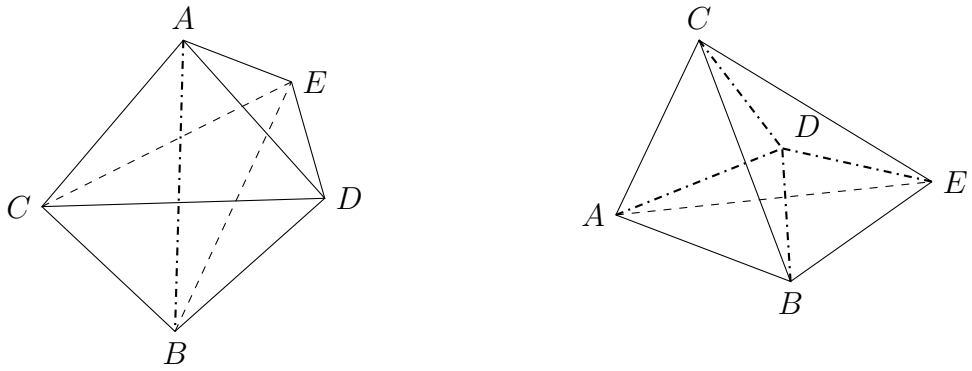


Figure 13

Suppose $M_1 = 2$. Let σ be a k-simplex. Define $\mathcal{N}_\sigma = \{\tau | \sigma \subset \tau, \tau :$ a $(\mathrm{k}+2) - \mathrm{simplex}\}$ as a collection of 2 dimension higher complices containing σ and $\nu_\sigma = |\mathcal{N}_\sigma|$ as its number of elements.

Consider $N_k = \max_{\sigma \,:\, \text{a k}-\text{simplex}} \nu_\sigma$.

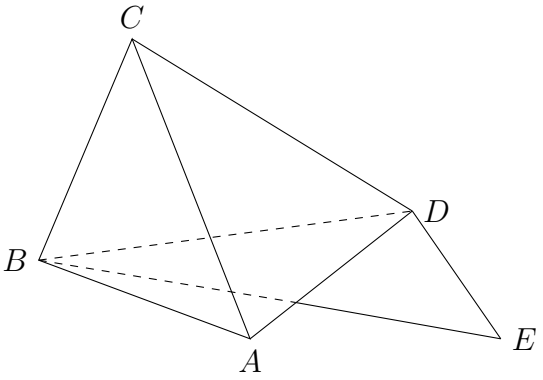


Figure 14

We focus on the following cases:

case1: $N_0 = 1$)
All 2-simplices are disjoint and so at least 12 vertices exist. This is a contradiction to given 5 vertices.

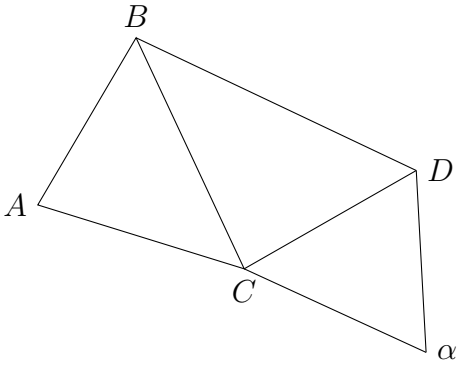


Figure 15

case2: $N_0 = 2$)
Let A be a vertex such that $\mathcal{N}_A = \{\{A, B, C\}, \{A, C, D\}\}$ and $\nu_A = 2$ and E be the last vertex. Then the 3rd 2-simplex should be $\{B, D, E\}$. Consequently, $N_A \geq 3$ with $\{A, B, D\}$. This is a contradiction to $N_0 = 2$.

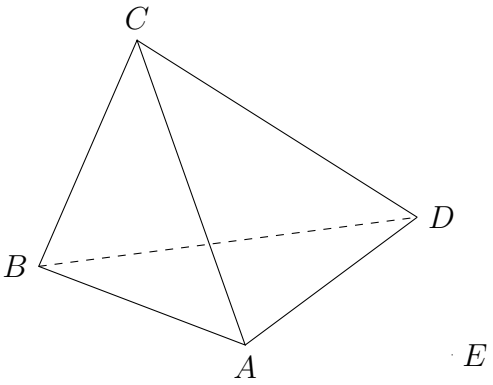


Figure 16

case3: $N_0 = 3$)
Let A be a vertex such that $\mathcal{N}_A = \{\{A, B, C\}, \{A, C, D\}, \{A, D, \alpha\}\}$ and $\nu_A = 3$. Then α can be one of vertices B and E.
If $\alpha = B$, then we automatically get $\{B, C, D\}$ and there are no 2-simplex containing the vertex E.
If $\alpha = E$, then there should be one of 1-simplices $\{B, D\}, \{B, E\}$ and $\{C, E\}$ for the 4th 2-simplex. Consequently, $N_A \geq 4$ and it is contradic to $N_0 = 3$.

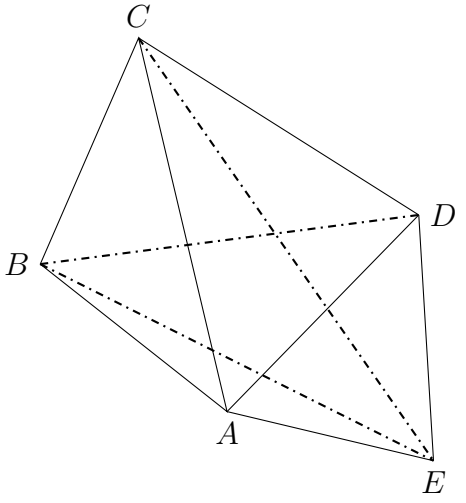


Figure 17

case4: $N_0 = 4$)
Let A be a vertex such that $\mathcal{N}_A = \{\{A, B, C\}, \{A, C, D\}, \{A, D, E\}, \{A, E, \alpha\}\}$ and $\nu_A = 4$. Then α can be one of vertices B and C. If $\alpha = C$, then $M_{\{A,C\}} \geq 3$ which is a contradiction to $M_1 = 2$. Hence $\alpha = B$, see Figure 19.

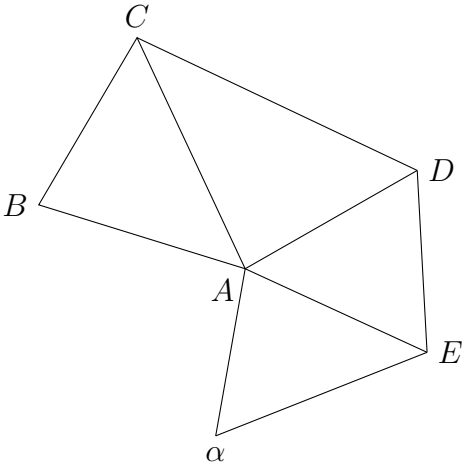


Figure 18

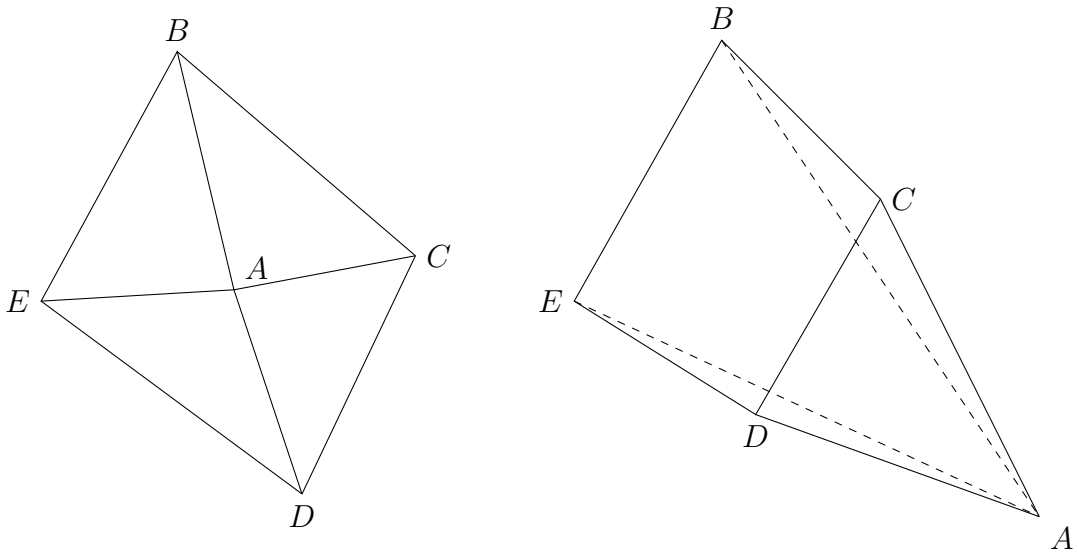


Figure 19

case5: $N_0 \geq 5$)
Let A be a vertex such that
$\mathcal{N}_A = \{\{A, B, C\}, \{A, C, D\},$
$\{A, D, E\}, \{A, E, \alpha\},$
$\{A, \alpha, \beta\}, \ldots \ldots \}$
and $\nu_A \geq 5$. Then α can be one of vertices B and C. If $\alpha = B$ or C, then $M_{\{A,B\}} \geq 3$ or $M_{\{A,C\}} \geq 4$ which is a contradiction to $M_1 = 2$. Thus, there is no complex which satisfies $N_0 \geq 5$ with $M_1 = 2$.

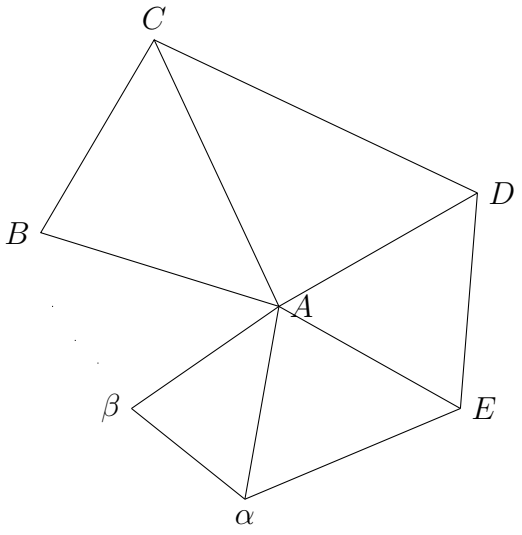


Figure 20

In summary there are only 5 kinds of complices with 5 vertices: Figure

11, Figure 12, Figure 13, Figure 16 and Figure 19. None of them have 2-dimensional generators. In other words, 5 points are not enough to contribute a 2−dimensional element of persistent homology. Therefore, there exist at least 6 point such that the induced Vietoris-Rips complex has a 2−dimensional generator in its persistent homology. □

Remark 4.2.4. About the answer of the main question, we conclude 6 is the minimal number of points for the cubic shaped sphere $\mathcal{C}$ having a 2−dimensional barcode of its persistent homology as the homology of S^2, i.e., $N_{\mathcal{C}}(S^2) = 6$.

4.3 Another proof of Minimal Construction for S^2

Our goal is to show there is no 2-dimensional generator of persistent homology with 5 vertices.

Strategy
Find all possible 2-cycles and show they are boundaries of some 3-simplices.

Observe the following:
F1 There are at most three 2-simplices associated to a 1-simplex with 5 vertices since the more 2-simplices, the more vertices.
F2 A boundary of a 2-simplex is the sum of three 1-simplices.
F3 Let σ be a 2-cycle. Each 1-simplex of σ should be in a pair of 2-simplices by F1 since our coefficient ring of persistent homology is $\mathbb{Z}_2$. In other words, there exist only two 2-simplices containing the same 1-simplex.
F4 There are ten 2-simplices combining 5 vertices (i.e., ${}_5\mathbf{C}_2 = 10$).
F5 There are ten 1-simplices combining 5 vertices (i.e., ${}_5\mathbf{C}_3 = 10$).

Claim

All 2-cycles with 5 vertices are a surface of a tetrahedron or a contractible union of two tetrahedrons.

Proof. Without loss of generality, we don't consider a subcomplex having two same simplices because of $\mathbb{Z}_2$ coefficient ring of persistent homology.

A 2-cycle σ consists of an even number l of 2-simplices by F2. Hence $l =$ 2,4,6,8 or 10 by F4. If $l = 2$, σ is just a copies of the same 2-simplex. If $l = 10$, a complete graph, then every 1-simplex of σ is on three 2-simplices. Thus it is a contradiction to 2-cycle. If $l = 8$, then there are twenty four$(= 8 \times 3)$ of 1-simplices as a boundary of σ. Since there are at most ten 1-simplices by F5, there exist at least four 1-simplices such that each 1-simplex is on some three 2-simplices. Thus it is a contradiction to 2-cycle.

Assume $l = 4$. It is easy to imagine the surface of a tetrahedron with four 2-simplices. By the way, we can show that a tetrahedron is the only 2-cycle with four 2-simplices. Suppose that σ is a 2-cycle of four 2-simplices associated with 5 vertices. There are twelve$(= 4 \times 3)$ 1-simplices of $\partial\sigma$ and six$(= 12/2)$ distinct 1-simplices. Because of 5 vertices, two 2-simplices of σ have at least a common vertex. If it has only one common vertex, then σ is a union of only 2-simplices having one common vertex with six 1-simplices which is a contradiction to 2-cycle. Otherwise if it has two common vertices, then this two 2-simplices with one common 1-simplex with 4 vertices and five 1-simplices should be connected to the 5th vertex with one 1-simplex which is the last 1-simplex and it is not a 2-cycle. Hence a 2-cycle with a surface of a tetrahedron which is a boundary of a 3-simplex.

Assume $l = 6$. Denote 5 vertices as A, B, C, D, E. There are nine(=

$6 \times 3/2$) distinct 1-simplices in σ which are all of ten in F5 except one 1-simplex, say $\{A, B\}$. Then we can have four 2-simplices in F4 which are not in σ as $\{A, B, C\}, \{A, B, D\}, \{A, B, E\}$, and $\{C, D, E\}$. In other words, σ is the boundary of a union of two 3-simplices $\{A, C, D, E\}$ and $\{B, C, D, E\}$. □

Lemma 4.3.1. A tetrahedron has no 2-dimensional generator of persistent homology.

Proof. Since a tetrahedron is a complete graph, all edges are measured among 4 points, say v_i for $i = 0, 1, 2, 3$. Let r be the maximum length of edges of this tetrahedron. Consider balls of radius r with center v_i for $i = 0, 1, 2, 3$, and then we have four 2-simplices which become a cycle of 2-dimensional persistent homology. However this 2-dimensional cycle is the boundary of the 3-simplex $\{v_0, v_1, v_2, v_3\}$. Hence there is no 2-dimensional generator of a tetrahedron. □

In conclusion, there is no 2-dimensional generator of persistent homology with 5 vertices by above Lemma and Claim.

4.4 6 points probability for S^2

In the previous subsections, we show 6 points in Vietoris-rips complex is the minimum number of points that can have 2-dimensional generator of persistent homology. In this subsection, we want to figure out 6 points probability, i.e., the probability of 6 points in Vietoris-Rips complex holding 2-dimnesional generator of persistent homology by *ripser* module which is an algorithm to compute persistent homology.

4.4.1 Script for 6 points probability

On the cubical sphere, we can measure the possible region of 6 points can make form 2-dimensional persistent homology such that we get the 6 points

probability. The following pseudocode shows how to compute it using Boost library and Ripser library.

6 points probability

```
N ← the number of times for splitting intervals
LowerBound=0 ← the lower bound of 6 points possiblity
UpperBound=1 ← the upper bound of 6 points possiblity
INTERVAL ← boost::numeric::interval<double>
CUBE ← std::vector<INTERVAL>
CUBELIST ← collection of CUBE
Check2dim(CUBE) ← using ripser library,
                  check the number of 2-dimensional interval;
                  if all values(numbers) on CUBE > 0
                        return 1
                  else if all values(numbers) on CUBE = 0
                        return -1
                  else
                        return 0
SplitCUBE(CUBE) ← collection of subcubes of CUBE
                  by dividing all intervals into half size
volume=0 ← the possible region volume
complementvolume=0 ← the impossible region volume
initialCUBE ← initialize the cubical sphere
CUBELIST ← initialCUBE
for (i in 1:N)
    preCUBELIST ← another collection of CUBE
        for j in CUBELIST
            if Check2dim(j)=1
                volume += size(j)
```

```
            else if Check2dim(j)=-1
                complementvolume += size(j)
            else
                preCUBELIST ← SplitCUBE(j)
    CUBELIST ← updated by preCUBELIST
LowerBound ← volume/size(initialCUBE)
UpperBound ← 1− complementvolume/size(initialCUBE)
```

Remark 4.4.1. When N increases, we get the accurate value of 6 point probability. However, we need a large memory in computer and it takes a lot of time to get proper a range of probability.

Another method for 6 point probability is sampling which is much faster than optimization. The following pseudocode shows how to compute the probability using Ripser library.

6 points sampling probability

```
SN ← sampling number
S=0 ← count the number of successful case
for (i in 1:SN)
    p=GenerateRandompt(i) ← randomly pick 6 points on the cubic sphere
    Check2dim(p) ← using ripser library,
                   check the number of 2-dimensional interval;
                   if the numbers > 0
                       s++;
Result ← S/SN
```

Remark 4.4.2. In practical, we get the following result of 6 points sampling probability.

6point sampling probability results				
SN	1,000	10,000	100,000	1,000,000
Trial 1	0.007	0.0053	0.00651	0.006904
Trial 2	0.003	0.0068	0.00697	0.006945
Trial 3	0.006	0.0064	0.00648	0.00684
Trial 4	0.009	0.007	0.00678	0.00703
Trial 5	0.012	0.0078	0.00663	0.006929
Trial 6	0.008	0.0061	0.00664	0.007006
Trial 7	0.008	0.0068	0.00692	0.006899
Trial 8	0.006	0.0055	0.00668	0.006836
Trial 9	0.005	0.0087	0.00725	0.007036
Trial 10	0.007	0.0061	0.00691	0.006923

By the law of large numbers, when SN $\geq$ 100,000, we may get the expectation of 6 points probability as the number around 0.007. However, how much can we trust this expectation? Now we want to state *confidence intervals* in terms of proportions or percentages. In next subsection, we will find confidence intervals with the Bootstrap method.

4.4.2 Bootstrap Confidence Intervals

In this subsection, we will apply the same technique to find out the bootstrap confidence intervals and we use the same notation as in the book [22] of Larry Wasserman, see 7.2 in Appendix.

Suppose we draw an independent and identically distributed (IID) sample $X_1, ..., X_B$ from a distribution F. By the law of large numbers,

$$\overline{X}_B \xrightarrow{P} \int x d\mathrm{F(x)} = \mathbb{E}(X)$$

as $B \to \infty$. So if we draw a large sample from F, we can use the sample mean $\overline{X}_B$ to approximate $\mathbb{E}(X)$. In a simulation, we can make B as large as we like, in which case, the difference between $\overline{X}_B$ and $\mathbb{E}(X)$ is negligible. More generally, if h is any function with finite mean then

$$\frac{1}{B}\sum_{j=1}^{B} h(X_j) \xrightarrow{P} \int h(x)d\mathrm{F(x)} = \mathbb{E}(h(x))$$

as $B \to \infty$. In particular,

$$\begin{aligned}\frac{1}{B}\sum_{j=1}^{B}(X_j - \overline{X}_B)^2 &= \frac{1}{B}\sum_{j=1}^{B} X_j^2 - \Big(\frac{1}{B}\sum_{j=1}^{B} X_j\Big)^2 \\ &\xrightarrow{P} \int x^2 d\mathrm{F(X)} - \Big(\int x d\mathrm{F(x)}\Big)^2 = \mathbb{V}(X)\end{aligned}$$

as $B \to \infty$. Hence, we can use the sample variance of the simulated values to approximate $\mathbb{V}(X)$.

In most practical research, the standard deviation is not known. In this case, the standard deviation is replaced by the estimated standard deviation, also known as the standard error *se*.

Bootstrap Variance Estimation

1.Draw $X_1^*, \cdots, X_n^* \sim F_n^*$.
2.Compute $T_n^* = g(X_1^*, \cdots, X_n^*)$.
3.Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \cdots, T_{n,B}^*$.
4.Let

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}\Big(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\Big)^2.$$

The following pseudocode shows how to use the bootstrap to estimate the standard error of the median.

Bootstrap for The Expectation

Given data $X = [X[1], \cdots, X[n]]$:
$T \leftarrow \mathbb{E}(X)$
$T_{boot} \leftarrow$ vector of length B
for (i in a:B):
    $X_{star} \leftarrow$ sample of size n from X (with replacement)
    $T_{boot}[i] \leftarrow \mathbb{E}(X_{star})$
se $\leftarrow$ sqrt(variance(T_{boot}))

Here we discuss the simplest method, *the Normal Interval*

$$T_n \pm z_{\alpha/2}\ se_{boot}$$

where $se_{boot} = \sqrt{v_{boot}}$ is the bootstrap estimate of the standard error, $T_n = g(X_1, \cdots, X_n)$. Note that this interval is not accurate unless the distribution of T_n is close to Normal.

Remark 4.4.3. In experimental with $SN = 1,000,000$, $B = 1,000$ and $\alpha = 0.005$, $z_{\alpha/2} = 1.96 \approx 2$, we get the 95% confidence interval

$$0.007081 \pm 2 * 8.254727372845205e{-}5$$

$$= (0.006915905452543096,\quad 0.007246094547456903)$$

4.5 Vietoris-Rips complex for S^n

In this section, we observe the phenomenon of higher dimensional sphere and get some generalization results related to previous section.

Example 4.5.1. Consider

$$\text{data2}(N+1) := \{\ c_{i,j} \mid 1 \leq i \leq N+1,\ \text{and } j = 0, 1\}$$

where $c_{i,j}$ denotes the vector with a δ_j in the ith coordinate and $\frac{1}{2}$'s elsewhere and $\delta_j = 0$ or 1. The set data2$(N+1)$ is the collection of $2(N+1)$ central points on each face of the $(n+1)-$dimensional unit cube. The result of the persistent homology of this example with the Euclidean distance is the following:
value range: $[\frac{\sqrt{2}}{2}, 1]$
distance matrix with $2(n+1)$ points
persistence intervals in dim 0:
$2n+1$ copies of the interval $[0, \frac{\sqrt{2}}{2})$ and one interval of $[0, \infty)$
persistence intervals in dim n:
$[\frac{\sqrt{2}}{2}, 1)$
There is no persistence intervals other dimension. From above observation, we get

$$N(S^n) \leq 2n + 2.$$

c.f. When the parameter ϵ is in $[\frac{\sqrt{2}}{2}, 1]$, there are 2^{n+1} number of $n-$simplices which represents the combination number of choosing $N+1$ times one elements in each two elementary set.

Remark 4.5.2. Consider

$$\text{data}2^{N+1} := \{\delta_i\ e_i \mid \delta_i = 0 \text{ or } 1,\ \text{and } 1 \leq i \leq N+1\}$$

where e_i denotes the vector with a 1 in the ith coordinate and 0's elsewhere. We get its persistent homology with the Euclidean distance through value range: $[1, \sqrt{n+1}]$. Clearly, we have persistence intervals in dim 0: $2^{n+1} - 1$ copies of the interval $[0, 1)$ and one interval of $[0, \infty)$. However, it's already hard to compute other dimensional persistence intervals by hands after 4 vertices;

whenever parameter ϵ grows through $\sqrt{k}$ for $1 \leq k \leq n+1$, we should consider all combination of vertices whose distance less than $\sqrt{k}$.

Property 4.5.3. No $n+2$ vertices of Vietoris-Rips complex can form a $n-$dimensional generator in its persistent homology.

Proof. Without loss of generality, we have a $n+1$ vertices $\{v_0, \cdots, v_n\}$ which form a $n-$simplex $[v_0, \cdots, v_n]$. Consider a distinct vertex v_{n+1} which is not on this simplex. Then there is only one $n-$dimension cycle using all v_i, $0 \leq i \leq n+1$, i.e., the boundary of $(n+1)$-simplex $[v_0, \cdots, v_{n+1}]$. Since this cycle is exact, there is no $n-$dimensional generator in its persistent homology with $n+2$ vertices. □

Remark 4.5.4. Construction for S^3 of Vietoris-Rips complex is hard to do in higher dimension because additional axes give spaces not like lower dimension.

Even though the direct construction is not possible but still we get an remarkable result.

Notation For a n- simplex σ and a collection of n-simplices $\Sigma\sigma_i$,

$$< \sigma, \Sigma\sigma_i >= k \bmod 2$$

where the number k is a cardinality of the set $\{\ i \mid \sigma_i = \sigma \text{ for some } i\}$.

Theorem 4.5.5. *Given numbers n, r and a finite point set X with a distance d, assume a Vietoris-Rips complex $VR_r(X)$ has the same homology of S^n with $\mathbb{Z}_2$-coefficients. Suppose $\sigma + \Sigma\sigma_i$ is a collection of n-simplices of $VR_r(X)$ such that $[\sigma + \Sigma\sigma_i]$ is a generator of $H_n(VR_r(X))$. Say, $\sigma = [0, 1, \cdots, n]$.*

If the n-simplex σ satisfies $< \sigma, \partial\tau >= 0$ for any $(n+1)$ -simplex τ in $VR_r(X)$, then there exist $n+1$ additional points in X.

Proof. Since we have the $\mathbb{Z}_2$ - coefficients, we may assume $\sigma \neq \sigma_i$ for any i. Consider the boundary of σ, i.e., $\partial\sigma = \Sigma_{j=0}^{n}[0, 1, \cdots, \hat{j}, \cdots, n]$. Because $\sigma + \Sigma\sigma_i$ is closed, there should be at least one n-simplex σ_l for some l such that $< [0, 1, \cdots, \hat{j}, \cdots, n],\ \partial\sigma_l >= 1$, i.e., $\sigma_l = [0, 1, \cdots, p_j, \cdots, n]$ for a point $p_j \neq j$. If the distance between j and p_j is less than r, then we get a $(n+1)$-simplex $[0, 1, \cdots, n, p]$. This is a contradiction to the assumption $< \sigma, \partial\tau >= 0$ for any $(n+1)$ -simplex τ in $VR_r(X)$. Hence the distance $d(j, p_j) \geq r$ and $d(k, p_j) < r$ for $k \in \{0, 1, \cdots, \hat{j}, \cdots, n\}$. In other words, we have some points $\{p_j\}_{j=0}^{n}$ such that $p_j \neq p_k$ for $j \neq k$. Note $p_j \notin \{0, 1, \cdots, n\}$ for all j. Therefore X has $n+1$ additional points. □

Corollary 4.5.6. Given numbers n, r and a finite point set X with a distance d, assume a Vietoris-Rips complex $VR_r(X)$ has the same homology of S^n with $\mathbb{Z}_2$-coefficients. Suppose $\Sigma\sigma_i$ is a collection of n-simplices of $VR_r(X)$ such that $[\Sigma\sigma_i]$ is a generator of $H_n(VR_r(X))$. If all $(n+1)-$simplices τ satisfy $< \sigma_i, \partial\tau >\leq 1$, then there exist at least $2n+2$ points in X.

Conjecture Given numbers n, r and a finite point set X with a distance d, assume a Vietoris-Rips complex $VR_r(X)$ has the same homology of S^n with $\mathbb{Z}_2$-coefficients. Then there exist at least $2n+2$ points in X.

Chapter 5

The Vietoris-Rips complex on a circle S^1

In this chapter, we observe homotopy types from finite points on a circle S^1 as Vietoris-Rips complex. We will follow the technique and notation from the Adamaszek [13].

Definition 5.0.1. A subset X of a metric space M is an $\epsilon-$ **covering** if every point of M is within distance less than ϵ from some point in X.

Remark 5.0.2.

1. A finite subset $X \subseteq S^1$ is an $\epsilon-$ covering of S^1 if and only if every two cyclically consecutive points in X are less than 2ϵ apart.

2. If $0 < r < \frac{1}{3}$ and $X \subseteq S^1$ is a finite subset, then $VR(X, r) \simeq S^1$ if and only if an $(\frac{1}{2})-$ covering of S^1.

Now we follow notations from Adamaszek[14].

Definition 5.0.3. For $n \geq 1, i, j \in \mathbf{Z}$ with $i \geq j$, *the discrete circular arc* $[i, j]_n$ is the image of the set $\{i, i+1, \cdots, j\}$ under the quotient map $\mathbf{Z} \longrightarrow \mathbf{Z}/n$ sending $z \mapsto z \bmod n$.

For $n \geq 1$ and $k \geq 0$, denote $\mathcal{N}(n,k)$ as a simplicial complex consisting of the vertex set $\{0, 1, \cdots, n-1\}$ and its set of maximal simplices $\{[i, i+k]_n | i = 0, \cdots, n-1\}$.

Observe that for $k \leq n-2$ $\mathcal{N}(n,k)$ has n maximal simplices given by the n rotation of $[0,k]_n$. The simplicial complex $\mathcal{N}(n,k)$ is the $(n-1)-$ simplex if $k \geq n-1$.

Suppose S^1 has the circumference 1. For $0 \leq k < n$,

$$\mathcal{U}_{n,k} = \{[\frac{i}{n}, \frac{i+k}{n}]_{S^1} | i = 0, \cdots, n-1\},$$

i.e., a set of n evenly-spaced arcs of length $\frac{k}{n}$ in S^1.

Remark 5.0.4. If $X_n \subseteq S^1$ is a set of n evenly-spaced points, then

$$\check{C}(X_n, \frac{k}{2n}) \simeq \mathcal{N}(\mathcal{U}_{n,k}) \simeq \mathcal{N}(n,k).$$

Properties 5.0.5.

1. $\mathcal{N}(n,0) = \vee^{n-1} S^0$ is the disjoint union of n points.
2. For $1 \leq k < \frac{n}{2}$, $\mathcal{N}(n,k) \simeq S^1$.
3. $\mathcal{N}(n, n-2)$ is the boundary of a $(n-1)-$simplex.
4. For $k \geq n-1$, $\mathcal{N}(n,k)$ is a $(n-1)-$ simplex.

Definition 5.0.6. Let K and L be simplicial complexes with disjoint vertex sets.

1. The **join** $K * L$ is the simplicial complex whose faces are all the union $\sigma \cup \tau$ for $\sigma \in K, \tau \in L$.
2. The **unreduced suspension** $\sum K$ is $S^0 * K$.

Lemma 5.0.7. Let M be a simplicial complex such that $M = M_1 \cup M_2$ with contractible subcomplexes M_i for $i = 1, 2$. Then

$$M \simeq \sum(M_1 \cap M_2).$$

Proposition 5.0.8. For $\frac{n}{2} \leq k < n$, $\mathcal{N}(n,k)$ and $\sum^2 \mathcal{N}(k, 2k-n)$ are homotopy equivalent.

Proof. Consider the maximal simplices of $\mathcal{N}(n,k)$, denoted as $\sigma_i = [i, i+k]_n$ for $i = 0, 1, \cdots, n-1$. Then we have the following expression:

$$\mathcal{N}(n,k) = (\cup_{i=0}^{n-k-2}\sigma_i) \cup (\cup_{j=n-k-1}^{n-1}\sigma_j).$$

Say, $A = (\cup_{i=0}^{n-k-2}\sigma_i)$ and $B = (\cup_{j=n-k-1}^{n-1}\sigma_j)$. Note that σ_j contains $(n-1)$ figure point and σ_i contains k figure point since $n-k-2 \leq k$. Moreover, A and B are cones and σ_i does not contain $n-1$ figure point. By Lemma 5.0.7, we get

$$\mathcal{N}(n,k) \simeq \sum K$$

where K is a simplicial complex with the vertex set $\{0, 1, \cdots, n-2\}$ and its simplex $\{\sigma_i \cap \sigma_j | i = 0, \cdots, n-2, j = n-k-1, \cdots, n-1\}$.

Observe the shape of simplices of K:

Type1) If $0 \leq i \leq j+k-n \leq i+k < j \leq n-1$, then $\sigma_i \cap \sigma_j = \{i, \cdots, j+k-n\}$. Hence

$$\sigma_i \cap \sigma_j \subseteq \sigma_0 \cap \sigma_{n-1} = \{0, \cdots, k-1\}.$$

Type2) If $0 \leq j+k-n < i \leq j \leq i+k \leq n-2$, then $\sigma_i \cap \sigma_j = \{j, \cdots, j+k\}$. Thus

$$\sigma_i \cap \sigma_j \subseteq \sigma_{n-k-2} \cap \sigma_{n-k-1} = \{n-k-1, \cdots, n-2\}.$$

Type3) If $i \leq j+k-n$ and $j \leq i+k$, then

$$\begin{aligned}\sigma_i \cap \sigma_j &= \{i, \cdots, j+k-n\} \cup \{j, \cdots, i+k\} \\ &\nsubseteq \sigma_{i'} \cap \sigma_{j'} \quad \text{for any } i', j'.\end{aligned}$$

Now we have the following maximal simplices of K:

$$\begin{aligned}\tau &= \{0, \cdots, k-1\}, \\ \tau' &= \{n-k-1, \cdots, n-2\}, \\ \tau_{i,j} &= \{i, \cdots, j+k-n\} \cup \{j, \cdots, i+k\} \\ &\qquad \text{for } 0 \leq i \leq j+k-n \text{ and } j \leq i+k \leq n-2.\end{aligned}$$

Observe the subcomplex $T = \tau' \cup (\cup_{i,j} \tau_{i,j})$ is contractible. By the Lemma 5.0.7 and $K = \tau \cup T$,

$$K \simeq \sum(\tau \cap T).$$

Since $\tau \cup T = \mathcal{N}(k, 2k-n)$, we have

$$\begin{aligned}\mathcal{N}(n,k) &\simeq \sum(A \cap B) \\ &\simeq \sum(\tau \cup T) \\ &\simeq \sum\sum(\tau \cap T) \\ &\simeq \sum{}^2 \mathcal{N}(k, 2k-n).\end{aligned}$$

□

Note that $\mathcal{N}(n,0) \simeq \vee^{n-1} S^0$ and $\mathcal{N}(n,k) \simeq S^1$ for $1 \leq k < \frac{n}{2}$. Applying Proposition 5.0.8 repeatly, we get the following theorem.

Theorem 5.0.9. *Let* $0 \leq k \leq n-2$. *Then*

$$\mathcal{N}(n,k) \simeq \begin{cases} S^{2l+1}, & \text{if } \frac{l}{l+1} < \frac{k}{n} < \frac{l+1}{l+2} \text{ for } l = 0, 1, \cdots, \\ \vee^{\mathfrak{c}} S^{2l}, & \text{if } \frac{k}{n} = \frac{l}{l+1}. \end{cases}$$

Definition 5.0.10. Let G be a simple, loopless, undirected graph. The **clique complex** $\mathrm{Cl}(G)$ is to be a simplicial complex whose vertices are the vertices of G and the simplices are the cliques (complete subgraphs) of G.

Definition 5.0.11. For $n \geq 1, k \geq 0$, the **clique complex** $\bar{\mathcal{N}}(n,k)$ is the maximal simplicial complex with 1-skeleton $\mathcal{N}(n,k)$, i.e., $\mathrm{Cl}(\mathcal{N}(n,k)^{(1)})$.

Remark 5.0.12. If $X_n \subseteq S^1$ is a set of n evenly-spaced points, then

$$VR(X_n, \frac{k}{n}) \simeq \bar{\mathcal{N}}(\mathcal{U}_{n,k}) \simeq \bar{\mathcal{N}}(n,k).$$

Lemma 5.0.13. Let $n \geq 1$ and $k \geq 0$. The map $f : \{0, \cdots, n+k-1\} \longleftarrow \{0, \cdots, n-1\}$ assigning $i \mapsto i \bmod n$ induces a simplicial, surjective homotopy equivalence

$$f : \bar{\mathcal{N}}(n,k) \xrightarrow{\simeq} \mathcal{N}(n,k).$$

By the above Lemma 5.0.13, we have the following theorem.

Theorem 5.0.14. *Let* $0 \leq k < \frac{n}{2}$. *Then*

$$\bar{\mathcal{N}}(n,k) \simeq \begin{cases} S^{2l+1}, & \text{if } \frac{l}{2l+1} < \frac{k}{n} < \frac{l+1}{2l+3} \text{ for } l \geq 0, \\ \vee^{n-2k-1} S^{2l}, & \text{if } \frac{k}{n} = \frac{l}{2l+1} \text{ for some } l \end{cases}$$

Remark 5.0.15. It is also the same result in Adamaszek [15], Corollary 6.7.

From above theorem we have the following theorem, see Adamaszek [13](Theorem 7.6).

Theorem 5.0.16. *Let X be a subset which is dense in S^1 and $0 \leq r < \frac{1}{2}$. Then*

$$VR(X,r) \simeq \begin{cases} S^{2l+1}, & \text{if } \frac{l}{2l+1} < r < \frac{l+1}{2l+3}, l = 0, 1, \cdots, \\ \vee^{\mathfrak{c}} S^{2l}, & \text{if } r = \frac{l}{2l+1}, \end{cases}$$

where $\mathfrak{c}$ is the cardinality of the continuum.

In short, for large r, $VR(X,r)$ does not have the homotopy of X.

Chapter 6

Reliable barcodes

In the previous chapter, we observe that even a dense set of sample points cannot suggest the underlying space. Then the issue is hence the choice of radius in the Vietoris-Rips complex. In short, for small radii we get a good cover; by Nerve theorem we have the correct homology of the underlying space. The problem is now to determine how small the radius should be.

6.1 On the length of barcodes

We consider a topological space X. The Nerve theorem tells us that the nerve of a good cover of X is homotopy equivalent to X, so this motivates us to find good covers. Since we have some form of metric or dissimilarity measure when working with data, we can look at ϵ-balls. We can use this to either look at the Čech complex or the Vietoris-Rips complex. We consider the latter for a metric space (X, d).

Recall that the $VR(X, d, \epsilon)$ can be defined as the nerve of a cover, namely

$$VR(X, d, \epsilon) = \mathcal{N}(X, \mathcal{U} = \{B_\epsilon(x)\}_{x \in X}).$$

We may hence wonder for what values of ϵ the cover $\mathcal{U}$ is a good cover. From the results on the Vietoris-Rips complex for S^1 we know that $\mathcal{U}$ will not be a good cover for large ϵ, since the homotopy type changes. We also note that for general topological spaces there is no good cover. For example, the quasi-circle has no good cover, since that space is connected, but not path-connected. In particular, it is not homotopy equivalent to a simplicial complex.

We will restrict ourselves to smooth manifolds, because many data sets may be regarded as samples from some underlying manifold. From a mathematical point of view, there are many methods from Riemannian geometry available, so this is a relatively good setting to obtain results.

Remark 6.1.1. We will work with the intrinsic distance on Riemannian manifolds. This is from a mathematical point of view the most natural. However, when working with data, we usually don't have direct access to this information, and instead we have only direct access to the coordinates of some embedding. Although it is in principle possible to reconstruct the intrinsic distance by using sufficiently many data points and good optimization, this is not a very practical approach.

The reader should keep this in mind when reading our arguments.

6.1.1 Mission impossible

Before we give some criteria to find good covers, we point out that these criteria involve assumptions on the data which impossible to check in practice.

- the topology of the space: we will assume often that the space is a manifold of some form. Although one can argue that this is natural (regular values of a smooth function being dense), there is no way to verify this based on only finitely many points

- bounds on the curvature or reach: it is possible approximate curvature numerically, but given a finite sample sets of points $p_i \in M \subset \mathbb{R}^n$, we

can never be sure that some small region where the absolute value of the curvature is large has been missed.

6.1.2 Basic assumptions

We may assume that the sample points lie on or close to a smooth manifold M; this is a reasonable assumption, since sample points are commonly assumed to be chosen from some probability distribution. If the associated probability density function p is smooth, then sets where a large part of the support lies are generically smooth manifolds. Sample points can have numerical and categorical features, but after indexing the categorical values, we can always assume that the sample points lie in $\mathbb{R}^n$, so $M \subset \mathbb{R}^n$. By the way, this choice is not natural. Moreover, we have a distance/dissimilarity measure d for the sample points. This is an extrinsic distance on M, meaning that it is already defined on $\mathbb{R}^n$.

6.1.3 Further assumptions

To get clear results, we want work with an intrinsic distance, which we first assume to come from a Riemannian metric. Here are the key points for this choice: First, it is mathematically natural. Second, at small scales, the extrinsic and intrinsic distances are approximately equal. Third, intrinsic distances are "observer" independent. To give a concrete example, indexing categorical variables is not natural, leading to arbitrary choices, which can be dealt with (to some extent) using this intrinsic distance Lastly, a downside is that the intrinsic distance is not readily available; it can however be deduced from the second point. Therefore, in this Riemannian setting (including some generalizations), we will see that we can obtain upper bounds on the the admissible radii for the Vietoris-Rips complex.

6.1.4 Convex balls and curvature

Suppose that (M, g) is a Riemannian manifold.

Definition 6.1.2. The **injectivity radius** of (M, g) at $x \in M$ is

$$inj(M, x) := \sup\{r \in \mathbb{R} \cup \{\infty\} \mid \exp_x : B_r(0) \to M \text{ is embedding } \}.$$

The injectivity radius of (M, g) is then $\inf_p inj(p)$.

Clearly, compact Riemannian manifolds without boundary have positive injectivity radius.

Definition 6.1.3. A subset B in M is called **strongly convex** if for every pair $p, q \in B$, there is a unique minimal geodesic segment $\gamma : I \to M$ that is entirely contained in M. The **convexity radius of** M **at** p is defined as

$$r(p) := \max\{R > 0 \mid B_s(p) \text{ is strongly convex for } s \in (0, R)\}.$$

The **convexity radius of** (M, g) is then $\inf_p conv(p)$.

Remark 6.1.4. We note that a cover consisting of strongly convex balls $\mathcal{U} = \{B_i\}_{i \in I}$ will give a good cover. Indeed, consider any finite non-empty intersection $C = B_{i_1} \cap \ldots \cap B_{i_k}$. Take $c \in C$, which will serve as a center. By convexity, we find for all $x \in C$ a unique geodesic segment $\gamma_x : I \to M$ connecting x to c that is completely contained in C. Now define the contraction

$$H : C \times I \longrightarrow C, \quad (x, t) \longmapsto \gamma_x(t).$$

This map is continuous by continuous dependence on initial conditions of geodesics, see Milnor [23].

From the above we see that a cover by balls whose radius is less than the convexity radius will be a good cover. We apply this observation together with the Nerve theorem to obtain the following, simple, yet useful proposition.

Proposition 6.1.5. Suppose that (M, g) is a Riemannian manifold with the property that there is R_0 such that the convexity radius $conv(p) > R_0$ for all $p \in M$. Then $VR(M, d, R_0)$ is homotopy equivalent to M.

If we understand a Riemannian manifold sufficiently well, then we can compute its convexity radius. For example, the convexity radius of the round sphere $S^n \subset \mathbb{R}^{n+1}$ is π. On the other hand, if we consider the surface of revolution

$$S_f = \{(x, y, z) \in \mathbb{R}^3 \mid f(z) = \sqrt{x^2 + y^2}\}$$

for a a function f as sketched in Figure 6.1, we find the much smaller convexity radius r_0.

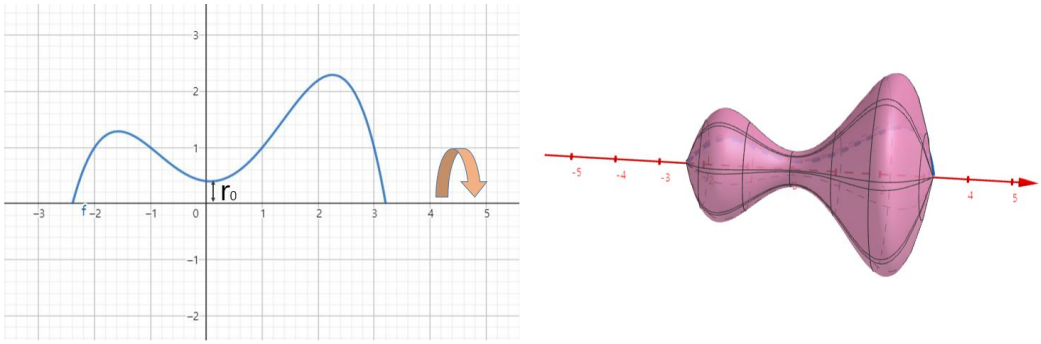


Figure 6.1: $f(z) = \sqrt{x^2 + y^2}$

In general, the best we can try to do, is to bound the convexity radius from below. From the literature we know that there is no universal bound without further assumptions. Indeed, we have the following theorem due to Dibble, [4].

Theorem 6.1.6. *Fix $n \geq 2$. Then for every $\epsilon > 0$ there exists a compact n-dimensional Riemannian manifold with $conv(M)/inj(M) < \epsilon$.*

To deal with this, we can impose curvature bounds, but we will need more.

For example, the lens space $L(p,1) = S^3/\mathbb{Z}_p$ has a much smaller convexity radius despite having the same curvature as its covering space S^3.

From Chavel's book on Riemannian geometry, Theorem IX.6.1, [17], we find the following.

Theorem 6.1.7. *Assume that (M,g) is a Riemannian manifold whose sectional curvature is bounded from above by K_M. Put $r_1 = \min\left(inj(M)/2, \frac{\pi}{2\sqrt{K_M}}\right)$. Then $B_{r_1}(x)$ is a strongly convex ball.*

By a result of Klingenberg, [3, Corollary 5.7], we know that

Theorem 6.1.8 (Klingenberg)**.** *If the sectional curvature of (M,g) is bounded from above by K_M, then*

$$inj(M) \geq \min\left(\pi/\sqrt{K_M}, \ell_g(\gamma)/2 \text{ where } \gamma \text{ is the shortest periodic geodesic}\right).$$

This gives the following bound,

$$conv(M) \geq \min\left(\ell_g(\gamma)/2 \text{ where } \gamma \text{ is the shortest periodic geodesic}, \frac{\pi}{2\sqrt{K_M}}\right).$$

In some cases, we can remove the dependence on the shortest geodesic, which is a global condition, rather than a local one.

Theorem 6.1.9. *Suppose that M is an even-dimensional, orientable manifold whose sectional curvature is positive and bounded from above by K_M. Then*

$$conv(M) \geq \frac{\pi}{2\sqrt{K_M}}.$$

Note that K_M can be determined or approximated from local measurements.

Proof. By a theorem of Synge, see [3, Theorem 5.9, part 2], we see that M is simply-connected. A theorem of Klingenberg, [3, Theorem 5.10], then tells use

that $inj(M) \geq \pi/\sqrt{K_M}$. Now suppose that γ is the shortest, periodic geodesic. Clearly, we have

$$\ell(\gamma)/2 \geq inj(M) \geq \pi/\sqrt{K_M}.$$

Hence the two terms in Theorem 6.1.7 have equal bounds, so we obtain the claimed conclusion. □

6.1.5 Background from metric geometry

In the following we describe several concepts which originate from Riemannian geometry to describe these bounds. Some of these concepts can be generalized to the metric setting, which will be convenient for later use.

Definition 6.1.10. Suppose that (X, d) is a metric space, and $\gamma : [a, b] \to X$ a curve. Given a partition $Y = \{y_0 = a, y_1, \ldots, y_N = b\}$ of $[a, b]$, define

$$\Sigma(Y, \gamma) := \sum_{i=1}^{N} d(\gamma(y_{i-1}, \gamma(y_i)\,).$$

The **length** of γ is then defined as

$$L_d(\gamma) = \sup_Y \Sigma(Y, \gamma).$$

Call the curve **rectifiable** if $L_d(\gamma)$ is finite.

In a Riemannian manifold (M, g), the logic goes the other way around. Given a smooth curve γ in a Riemannian manifold (M, g), we define the **length of the curve** as

$$\ell_g(\gamma) := \int_a^b \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t)\,)}dt.$$

After that, the **distance function on** (M, g) is defined as

$$d_g(x, y) := \inf_{\{\gamma \in C^1([0,1],M)\ \mid\ \gamma(0)=x, \gamma(1)=y\}} \ell_g(\gamma).$$

One can show that the Riemannian and metric concepts agree.

In a Riemannian manifold, we have locally a unique shortest path, i.e., given $x \in M$, there is a neighborhood U of x such that for all $y \in U$, there is a unique curve $\gamma_{x,y}$, up to reparametrization, such that

- $\gamma_{x,y}(0) = x$, $\gamma_{x,y}(1) = y$, and
- $\ell_g(\gamma_{x,y}) = d_g(x, y)$.

The concept of shortest path can obviously be defined in metric spaces as well.

Definition 6.1.11. Suppose that (X, d) is a metric space. We will call a subset B in X **strongly convex** if for every pair $p, q \in B$, there is a unique shortest path $\gamma : I \to X$ that is entirely contained in B. The **convexity radius** of X at p is defined as

$$r(p) := \sup\{R > 0 \mid B_s(p) \text{ is strongly convex for } s \in (0, R)\}.$$

Definition 6.1.12. Suppose that (X, d) is a metric space. The **cut locus** of $x \in X$ is

$$\begin{aligned} CL(x) := \{y \in X \mid &\text{there exist distinct paths } \gamma, \gamma' : I \to X \\ &\text{with } \gamma(0) = x = \gamma'(0),\ \gamma(1) = y = \gamma'(1), L(\gamma) = L(\gamma') = d(x, y)\}. \end{aligned}$$

We will denote by $\mathfrak{C}(p)$ the infimum of the distance of a point p to its cut locus $CL(p)$, and let $\mathfrak{C} = \inf_p \mathfrak{C}(p)$ By definition, we can use $\mathfrak{C}$ to bound the convexity radius, but the computation of $\mathfrak{C}$ is next to impossible in general.

Hence we will consider a simplicial complex X, which we will equip with the following distance function:

- prescribe distances between the vertices (this is natural if the vertices are sample points),

- embed each simplex affinely in $\mathbb{R}^N$ such that the Euclidean distance is the prescribe distance,
- give each simplex the flat Riemannian structure inducing the prescribed distances, and
- define the distance function

$$d(x, y) = \inf\{\ell(\gamma) \mid \gamma \text{ piecewise linear curve from } x \text{ to } y\}.$$

This definition is of course inspired by the Vietoris-Rips complex for a finite metric space. The convexity radius can then in principle be determined by investigating the combinatorics of the line segments.

6.1.6 Persistent homology of the Vietoris-Rips complex

Suppose that (X, d) is a metric space with convexity radius r_c, and consider a cover $\mathcal{U} = \{B_\epsilon(x)\}_x$. If $\epsilon < r_c$, then this is a good cover.

Now let us consider the Vietoris-Rips complex $VR_\epsilon(X, d)$. Consider the persistent homology of the resulting filtered space $\{VR_\epsilon(X, d)\}_\epsilon$. By Lemma 2.0.37, we have $VR(X, r) \subset \check{C}(X, \epsilon)$.

By convexity we know that this cover will be a good cover for $\epsilon < r_c$, so the Nerve theorem will tell us that $PH_*(VR_\epsilon(X, d)) \cong H_*(X)$ for $\epsilon < r_c$. On the other hand, for a larger radius, we know from the cube example that for larger ϵ, we have, in general, $PH_*(VR_\epsilon(X, d)) \ncong H_*(X)$. Inspired by this, we call a barcode **reliable** if its endpoint is shorter than the convexity radius.

6.2 Application to data

Let us first consider a situation where we sample data from a manifold, and we have the necessary information.

6.2.1 Revisiting the cube

Let us consider the hyper unit cube C spanned $V = [0,1]^{n+1} \subset \mathbb{R}^{n+1}$ as in the introduction. Instead of the extrinsic metric coming from $\mathbb{R}^{n+1}$, we consider the intrinsic metric defined following piecewise smooth curves on the hypersurface of C, so for example $d((1/2,0,1),(1/2,1,0)) = 1+1 = 2$. In this case, the curvature is actually infinite at the vertices, so we use the metric space description instead. The convexity radius is 1. The barcodes of the persistent homology of $VR_\epsilon(V,d)$ are plotted in

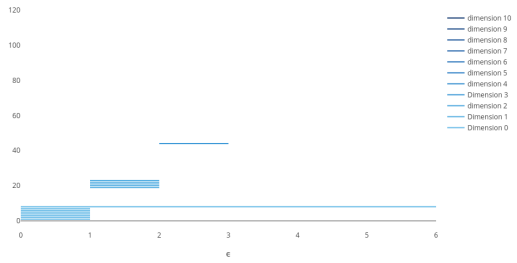


Figure 6.2: Persistent homology of the Vietoris-Rips complex of the corners of the cube, data8.

We see that the incorrect 3-homology barcode is not reliable according to this criterion.

Now let us take the 6 face center points, data6, again with the intrinsic distance. As before, the convexity radius is 1, so the (correct) 2-dimensional homology is unreliable.

In the natural setting of the theorem, i.e., with many data points, we do, of course, get reliable barcodes for the 2-dimensional homology, and we can rule out the incorrect 3-dimensional with this criterion.

6.2.2 Discretized curvature

So far, we apply the above criterion to a situation where we know the underlying manifold. Let us now consider the situation where we have sample points coming from a manifold, but we do not have more information (other than the dimension of the manifold). As we indicated earlier, it is now impossible to give any guarantees, but we can try to estimate the curvature from the data. We suggest a couple of approaches:

- A direct approach via the discretized derivative: this approach suffers from numerical instability.
- Define the Gauss curvature via the angle defect. In other words, apply the local Gauss-Bonnet theorem to define the curvature.
- Use volume/area defect to define the scalar curvature. This is based on the following classical result.

  Proposition 6.2.1. Let (M, g) be an n-dimensional Riemannian manifold with scalar curvature S. Then

  $$\frac{Vol(B_\epsilon(p) \subset (M, g))}{Vol(B_\epsilon)(0) \subset (\mathbb{R}^n, g_{Euclidean})} = 1 - \frac{S(p)}{6(n+2)}\epsilon^2 + o(\epsilon^2).$$

Let us make the above a little more concrete. We describe the volume approach, and assume that we have sample points $\{x_0, \ldots, x_n\}$ from which we have constructed a simplicial complex (for example the Vietoris-Rips complex). For convenience, assume that this is a 2-dimensional manifold.

Take $p = x_i$. We want to define S near p. The proposition suggests that we should take the limit $\epsilon \to 0$, but since that will yield infinite scalar curvature in curvature, we choose ϵ to be half the edge length to closed neighbor of p. Compute the area of $B_\epsilon(p)$ in the simplicial complex, and solve $S(p)$ from the

approximate equation

$$\frac{Area}{\pi\epsilon^2} = 1 - \frac{S}{24}\epsilon^2.$$

If we assume that the curvature is approximately constant near p, we obtain an approximate value for S on $B_\epsilon(p)$.

For higher-dimensional data, the same approach can be pursued, but we need to fix a 2-dimensional "geodesic" subcomplex first. We can then apply the above approach to this subcomplex to obtain the scalar/Gauss curvature for this subcomplex. The sectional curvature can then be bounded be repeating this procedure for all 2-dimensional subcomplexes.

Here is a possible approach to generate these 2-dimensional "geodesic" subcomplexes Δ:

- Fix a simplex containing p, and consider a 2-simplex σ contained in this simplex.
- We follow all geodesics in σ until they enter another 2-simplex τ. Include τ in Δ if τ contains p. More practically, consider the adjacent 2-simplices containing p
- Repeat this procedure for the new τ. Stop when we get back a a simplex sufficiently close to σ.

Example 6.2.2. Take a set of 60 random points on a cube.

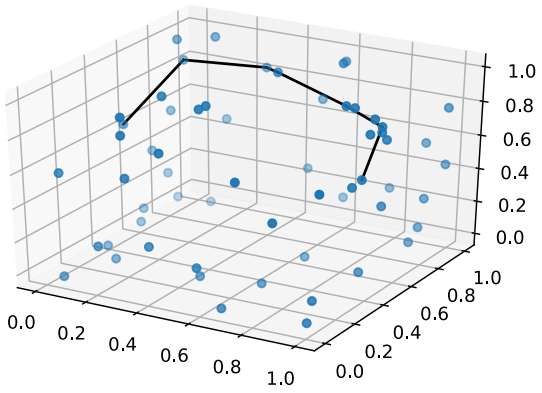


Figure 6.3: Random points on a cube and an approximate geodesic

We use this and other approximate geodesics to determine the convexity radius: we find 1.0123.

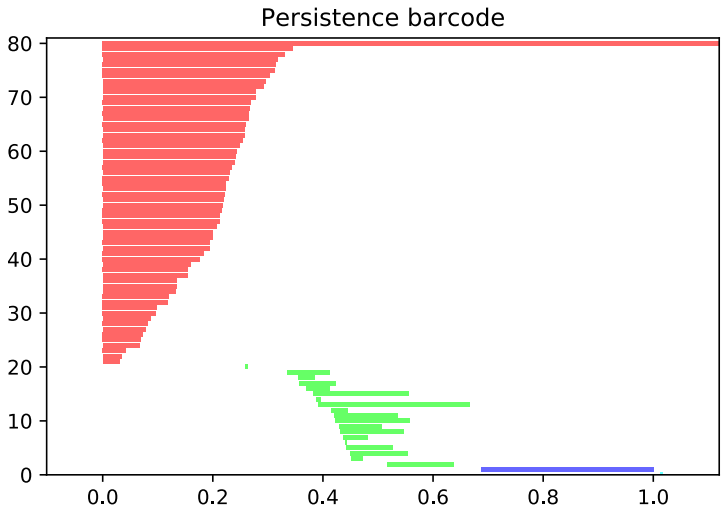


Figure 6.4: Barcodes of the persistent homology of the Vietoris-Rips complex 60 random points on a cube

With the estimate on the convexity radius, we can dismiss the barcode in degree 3,

$$(3, \text{with the interval } (1.0127, 1.0180))$$

The same approach works also for higher dimensional data, where visualization is not possible. The upshot is the following:

- In TDA, small barcodes are commonly treated as corresponding to noise, and long barcodes as the true signal.
- However, barcodes can be too long, and result in nonsense homology groups. In higher dimensions, visualization is not possible without loss of information, and we cannot easily recognize these meaningless barcodes.
- Our above techniques allow us to eliminate some of these meaningless barcodes.

Chapter 7

Appendix

7.1 Notation and Conventions

The $n-$**dimensional disk**(*or* the $n-$**disk**) is

$$D^n = \{x \in \mathbb{R}^n \mid |x| \leq 1\},$$

where $|\cdot| : \mathbb{R}^n \to [0, \infty)$ is the standard norm on $\mathbb{R}^n$. Then the $n-$disk is closed subset in $\mathbb{R}^n$. The **open** $n-$**disk**(*or* the **interior of** D^n **in** $\mathbb{R}^n$) is

$$\text{int}(D^n) = \{x \in \mathbb{R}^n \mid |x| < 1\}.$$

The **boundary of** D^n **in** $\mathbb{R}^n$ is the $(n-1)$-sphere

$$S^{n-1} = \{x \in \mathbb{R}^n \mid |x| = 1\}.$$

Note that the $0-$disk $D^n = \mathbb{R}^n = \{0\}$ and its interior $\text{int}(D^0) = \{0\} = D^0$.

Definition 7.1.1. An $n-$**cell** is a space which is homeomorphic to the open $n-$disk $\text{int}(D^n)$. A **cell** is a space which is an $n-$cell for some $n \geq 0$.

Since $\text{int}(D^m)$ and $\text{int}(D^n)$ are homeomorphic if and only if $m = n$, we can define the dimension of a cell, i.e., an $n-$cell has dimension n.

Definition 7.1.2. A **cell-decomposition of a space** X is a collection $\mathcal{C} = \{e_\alpha | \alpha \in J\}$ of subspaces of X such that each e_α is a cell and X is a disjoint union of such cells

$$X = \coprod_{\alpha \in J} e_\alpha.$$

The $n-$**skeleton of** X is the subspace

$$X^n = \coprod_{\{\beta \in J |\ \dim(e_\beta) \leq n\}} e_\beta.$$

Remark 7.1.3. A cell-decomposition of a space can have many different dimensional cells. Since there are no restrictions on the number of cells in a cell-decomposition, a cell-decomposition of a space is not unique.

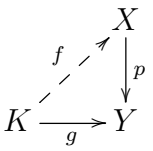
Definition 7.1.4. Let X be a Hausdorff space. A **CW-complex** is a pair $(X, \mathcal{C})$ consisting of a space X and a cell-decomposition $\mathcal{C}$ of X such that the following axioms are satisfied:

(A1) For each $n-$cell $e \in \mathcal{C}$, there is a map $f_e : D^n \to X$ such that the restriction $f_e|_{\text{int}(D^n)} \to e$ is a homeomorphism and $f_e(S^{n-1}) \subset X^{n-1}$. We call such maps f_e **characteristic maps**.

(A2) For any cell $e \in \mathcal{C}$, the closure $\bar{e}$ of e intersects only a finite number of other cells in $\mathcal{C}$.

(A3) A subset $A \subseteq X$ is closed if and only if $A \cap \bar{e}$ is closed in X for each $e \in \mathcal{C}$.

Lemma 7.1.5. Let K be a CW-complex and X, Y be topological spaces. Assume a map $p : X \to Y$ is a weak homotopy equivalence in the following diagram.



Then there exist a map f such that $p \circ f$ and g are homotopic. Moreover, if g is a weak homotopy equivalence, then f is also a weak homotopy equivalence.

The proof of this lemma can be found in the paper [5].

We recall the following lemma due to McCord [6].

Lemma 7.1.6. Let X, Y be topological spaces. Suppose p is a map of X into Y and there exist a basis-like open cover $\mathcal{W}$ of Y such that for each $W \in \mathcal{W}$, the restriction $p|_{p^{-1}(W)} : p^{-1}(W) \to W$ is a weak homotopy equivalence. Then p is a weak homotopy equivalence.

Now we prove Theorem 2.0.25 as in [2].

Proof. (Theorem 2.0.25)
Suppose X is a topological space and $\mathcal{U}$ is a locally finite, basis-like open cover of X. Since $\mathcal{U}$ is partially ordered by inclusion, we make $\mathcal{U}$ into a topological space. Indeed, for $U \in \mathcal{U}$, let $\lfloor U \rfloor = \{V \in \mathcal{U} \mid V \subset U\}$. Then the collection $\{\lfloor U \rfloor \mid U \in \mathcal{U}\}$ is a basis for the required topology of $\mathcal{U}$. First, define a map $g : |K(\mathcal{U})| \to \mathcal{U}$ as follows : for $x \in |K(\mathcal{U})|$, we have the unique open simplex $[U_0, \cdots, U_n]$ for $|K(\mathcal{U})|$ satisfying and $x \in U_0$ and $U_0 \subset \cdots \subset U_n$, i.e., $g(x) = U_0$. Note that g is continuous. By the Lemma 7.1.6, g is a weak homotopy equivalence. Second, define a map $p : X \to \mathcal{U}$ as follows : for each $x \in X$, let $p(x)$ be the smallest element of $\mathcal{U}$ containing x. This map is well-defined since $\mathcal{U}$ is locally finite and basis-like. Now observe the following:

$$\text{For each } U \in \mathcal{U},\ p^{-1}(\lfloor U \rfloor) = U. \tag{a}$$

Note that p is continuous since the collection $\{\lfloor U \rfloor \mid U \in \mathcal{U}\}$ forms a basis for the space $\mathcal{U}$. Now let's show that p is a weak homotopy equivalence. For the basis-like open cover of $\mathcal{U}$, consider the basis $\mathcal{W} = \{\lfloor U \rfloor \mid U \in \mathcal{U}\}$. Note that all $\lfloor U \rfloor$ are a contractible subset of $\mathcal{U}$. By the eq. (a), $p^{-1}(\lfloor U \rfloor) = U$. Since U is homotopically trivial, $p|_U : U \to \lfloor U \rfloor$ is a weak homotopy equivalence. By the lemma 7.1.6, p is a weak homotopy equivalence. Therefore, we obtain a weak homotopy equivalence $f : |K(\mathcal{U})| \to X$ by the lemma 7.1.5.

□

Definition 7.1.7. Let K be a simplicial complex and $\sigma = [v_0, \cdots, v_n]$ be a $n-$simplex of K.

(1) The **barycenter** $b(\sigma)$ of σ is the point

$$b(\sigma) = \frac{1}{n+1} \sum_{0 \le i \le n} v_i \in |K|.$$

(2) The **barycentric subdivision** of σ consists of $n-$dimensional simplices $\sigma_i = [p_0^i, \cdots, p_n^i]$, $1 \le i \le (n+1)!$ defined as follows: for each permutation τ_i of the set $\{0, 1, \cdots, n\}$ and the ordered set of vertices $\{v_{\tau_i(0)}, \cdots, v_{\tau_i(n)}\}$, let

$$p_j^i = b([v_{\tau_i(0)}, \cdots, v_{\tau_j(n)}]).$$

Definition 7.1.8. Let K, L be simplicial complexes. A map $\varphi : K \longrightarrow L$ is called to be a **simplicial map** if it has the property that the images of the vertices of simplex always span a simplex.

By definition, a simplicial map assigns vertices to vertices. In other words, simplicial maps are determined by their effects on vertices.

Definition 7.1.9. A **carrier** is a function Carr : $\mathcal{F} \longrightarrow \mathcal{G}$ from a cover $\mathcal{F}$ of a space X into a collection $\mathcal{G}$ of subsets of a topological space such that for each $U_i \in \mathcal{F}$ if $\bigcap_i U_i \neq \emptyset$, then $\bigcap_{U_i} \text{Carr } (U_i) \neq \emptyset$.

We are ready to prove Lemma 2.0.27.

Proof. (Lemma 2.0.27)
Let N be the first barycentric subdivision of $\mathcal{N}(\mathcal{U})$ and its nerve $\mathcal{N}(\mathcal{U})$. Define a simplicial map $\varphi : N \to K(\mathcal{U})$ as follows:

$$\varphi(b(\sigma)) = \text{Carr } (\sigma)$$

where $b(\sigma)$ as the barycenters of the simplexes σ of $\mathcal{N}(\mathcal{U})$. Note that this is well-defined because of Carr (σ), i.e., the intersection of the vertices of σ is not a empty set but a member of $\mathcal{U}$, a vertex of $K(\mathcal{U})$. Observe that every simplex of N is spanned by vertices $b(\sigma_0), \cdots, b(\sigma_n)$ with $\sigma_0 \subset \cdots \subset \sigma_n$. Then Carr $(\sigma_0) \supset \cdots \supset$ Carr (σ_n) and these vertices span a simplex of K. Thus φ is simplicial.

$\varphi : |\mathcal{N}(\mathcal{U})| = |N| \to |K(\mathcal{U})|$ is clearly a retraction, i.e., $\varphi|_{|K(\mathcal{U})|} = \mathbf{id}_{|K(\mathcal{U})|}$. If we show that for every simplex τ of N, both τ and $\varphi(\tau)$ are subsets of some simplex σ of $\mathcal{N}(\mathcal{U})$, then φ is a deformation retraction. Let the vertices of τ be $b(\sigma_0), \cdots, b(\sigma_n)$, where $\sigma_0 \subset \cdots \subset \sigma_n$, and the vertices of σ_n be $U_0, \cdots, U_r$. Let σ be the simplex of $\mathcal{N}(\mathcal{U})$ spanned by the vertices

$$\{U_0, \cdots, U_r, \text{Carr } (\sigma_0), \cdots, \text{Carr } (\sigma_n)\}.$$

Then this simplex σ is desired one. Therefore, the proof is completed. □

7.2 Background from Probability

First we recall some statistics definitions and theorems from the book [22] of Larry Wasserman.

Denote Ω as the set of possible outcomes of an experiment, called the **sample space**.

Definition 7.2.1. A function $\mathbb{P}$ that assigns a real number $\mathbb{P}(A)$ to each event $A \in \omega$ is called to be a **probability distribution** or a **probability measure** if it satisfies the following three axioms:

(A1) $\mathbb{P}(A) \geq 0$ for every A,

(A2) $\mathbb{P}(\Omega) = 1$, and

(A3) If $A_1, A_2, \cdots$, are disjoint, then

$$\mathbb{P}\Big(\cup_{i=1}^{\infty} A_i \Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

A real number $\mathbb{P}(A)$ to every event A is called the **probability** of A.

Definition 7.2.2. A **random variable** is a mapping

$$X : \Omega \to \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω.

Definition 7.2.3. The **cumulative distribution function**, or **CDF**, is the fuction $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

Definition 7.2.4. A random variable X is **continuous** if there exists a function f_X such that

(1) $f_X(x) \geq 0$ for all x,

(2) $\int_{-\infty}^{\infty} f_X(x)dx = 1$, and

(3) for every $a \leq b$, $\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$.

The function f_X is called the **probability density function** or **PDF**. We have that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

and $f_X(x) = F_X'(x)$ at all points x, i.e., F_X is differentiable at x.

Sometimes we write $\int f(x)dx$ to mean $\int_{\infty}^{\infty} f(x)dx$.

Example 7.2.5. A continuous random variable X has a **Normal** (or **Gaussian**) **distribution** with parameters μ and σ, denoted by $X \sim N(\mu, \sigma^2)$, if its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \Big\},\ x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. We say X has a **standard Normal distribution** if $\mu = 0$ and $\sigma = 1$.

Given continuous random variables $(X_1, \cdots, X_n)$, we call a function $f(x_1, \cdots, x_n)$ a PDF if it satisfies the following:

(1) $f(x_1, \cdots, x_n) \geq 0$ for all $(x_1, \cdots, x_n)$,

(2) $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \cdots, x_n)dx_1 \cdots dx_n = 1$, and

(3) for any set $A \subset \mathbb{R} \times \cdots \times \mathbb{R}$,

$$\mathbb{P}((x_1, \cdots, x_n) \in A) = \int \cdots \int_A f(x_1, \cdots, x_n) dx_1 \cdots dx_n.$$

The **marginal densities** are

$$f_{X_i}(x_i) = \int \cdots \int f(x_1, \cdots, x_n) dx_1 \cdots \hat{dx_i} \cdots dx_n, \quad \text{for } 1 \leq i \leq n,$$

where $\hat{dx_i}$ indicates that dx_i excludes and there are $n-1$ copies of $\int$. The **corresponding marginal distribution functions** are denoted by $F_{X_1}, \cdots$ and F_{X_n}.

We say that $X_1, \cdots, X_n$ are **independent** if for every $A_1, \cdots, A_n$,

$$\mathbb{P}(X_1 \in A_1, \cdots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Note that if $f(x_1, \cdots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$, then $X_1, \cdots, X_n$ are independent.

Definition 7.2.6. If $X_1, \cdots, X_n$ are independent and each has the same marginal distribution with CDF F, we say that $X_1, \cdots, X_n$ are **independent and identically distributed** or **IID** and we write

$$X_1, \cdots, X_n \sim F.$$

If F has density f, we also write $X_1, \cdots, X_n \sim f$. We also call $X_1, \cdots, X_n$ a **random sample of size** n **from** F.

Definition 7.2.7. Suppose a random variable X is continuous with its probability density function $f(x)$ and $\int xf(x)dx$ is well-defined. The **expected**

value, or **mean**, or **expectation** of X is defined to be

$$\mathbb{E}(X) = \int x d\mathrm{F(x)} = \int x f(x) dx.$$

Definition 7.2.8. Let X be a random variable with mean μ. The **variance** of X is defined by

$$\mathbb{V}(X) = \mathbb{E}(X - \mu)^2 = \int (x - \mu)^2 d\mathrm{F(x)},$$

assuming this expectation exists. The **standard deviation** is $\mathrm{sd}(X) = \sqrt{\mathbb{V}(X)}$.

Property 7.2.9. Let X be a random variable with mean μ. If the variance is well-defined, then

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2.$$

Definition 7.2.10. Let $X_1, X_2, \cdots$, be a sequence of random variables and let X be another random variable. We say X_n **converges to** X **in probability**, written $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

Definition 7.2.11. Let $X_1, X_2, \cdots$, be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X. Then X_n **converges to** X **in distribution**, written $X_n \rightsquigarrow X$, if

$$\lim_{n \to \infty} F_n(t) = F(t) \quad \text{at all} \quad t,$$

i.e., F is continuous at t.

Theorem 7.2.12. *(The Weak Law of Large Numbers) Let $X_1, X_2, \cdots, X_n$ be an IID sample, let $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathbb{V}(X_1)$. Then*

$$\overline{X}_n \xrightarrow{P} \mu \quad \textit{as} \quad n \to \infty$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *is the sample mean.*

Proof. Assume that $\sigma < \infty$. By Chebyshev's inequality,

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\mathbb{V}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which approach to 0 as n goes to infinity. □

The Weak Law of Large Numbers means the distribution of $\overline{X}_n$ becomes more concentrated around μ as n increases.

Theorem 7.2.13. *(The Central Limit Theorem) Let Let* $X_1, X_2, \cdots, X_n$ *be an IID sample with mean* μ *and variance* σ^2*. Let* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *and* $Z_n = \frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}(\overline{X}_n)}}$*. Then* $Z_n \rightsquigarrow Z$ *where* $Z \sim N(0,1)$*.*

Definition 7.2.14. A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a,b)$ where $a = a(X_1, \cdots, X_n)$ and $b = b(X_1, \cdots, X_n)$ are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all} \quad \theta \in \Theta.$$

In words, (a,b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the **coverage** of the confidence interval.

7.3 Scripts to compute persistent homology

Here is a simple `julia`-script to compute persistent homology.

```
using eirene

function add_bits(k)
    result = 0
```

```
    for i in (1:32)
        if k & 1 == 1
            result += 1
        end
        k = k >> 1
    end
    result
end

dim = 3
N = 2^dim
D = Array{Float64,2}(undef, N, N)
for k1 in (0:N-1)
    for k2 in (0:N-1)
        k = k1 ⩒ k2
        n = add_bits(k)
        D[k1+1,k2+1] = sqrt(n)
        D[k2+1,k1+1] = D[k1+1,k2+1]
    end
end
cpx = eirene(D, maxdim = 10)
plotbarcode_pjs(cpx)
```

Bibliography

[1] Alexandroff, P., *Über den allgemeinen Dimensionsbegriff und seine Beziehungen zur elementaren geometrischen Anschauung* (1928), Mathematische Annalen, Vol. 98, pp. 617-635.

[2] McCord, M.C., *Homotopy type comparison of a space with complexes associated with its open covers* (1967), the American Mathematical Society, Vol. 18, No. 4, pp. 705-708.

[3] Cheeger, J., and Ebin, David G., *Comparison theorems in Riemannian geometry*. Revised reprint of the 1975 original. AMS Chelsea Publishing, Providence, RI, 2008. x+168 pp. ISBN: 978-0-8218-4417-5 53C20

[4] Dibble, J., *The convexity radius of a Riemannian manifold*, Asian J. Math. 21 (2017), no. 1, 169–174.

[5] Dold, A., and Thom, R., *Quasifaserungen und unendliche symmetrische Produkte* (1958), Annals of Mathematics, pp.239-281.

[6] McCord, M.C., *Singular homology groups and homotopy groups of finite topological spaces* (1966), Duke Mathematical Journal, 33(3), pp.465-474.

[7] Whitehead, J.H., *Combinatorial homotopy. I. Bulletin of the American Mathematical Society* (1949), 55(3), pp.213-245.

[8] Borsuk, K., *On the imbedding of systems of compacta in simplicial complexes* (1948), Fundamenta Mathematicae, 35(1), pp.217-234.

[9] Weyl, A., *Sur les théorèmes de de Rham* (1952), Comm. Mathem. Helvetici, 26, pp.119-145.

[10] Holsztynski, W., *On spaces with regular decomposition* (1964), BULLETIN DE L ACADEMIE POLONAISE DES SCIENCES-SERIE DES SCIENCES MATHEMATIQUES ASTRONOMIQUES ET PHYSIQUES, 12(10), pp.607-611.

[11] Haimov, J. N., *The homotopy type of a space having a brick decomposition* (1979), In Dokl. Akad. Nauk Tadzhik. SSR (Vol. 22, No. 1, pp.25-29).

[12] Björner, A., *Nerves, fibers and homotopy groups* (2003), Journal of Combinatorial Theory, Series A, 102(1), pp.88-93.

[13] Adamaszek, M., and Adams, H., *The Vietoris–Rips complexes of a circle* (2017), Pacific Journal of Mathematics, 290(1), pp.1-40.

[14] Adamaszek, M., Adams, H., Frick, F., Peterson, C., and Previte-Johnson, C., *Nerve complexes of circular arcs* (2016), Discrete & Computational Geometry, 56(2), pp.251-273.

[15] Adamaszek, M., *Clique complexes and graph powers*(2013), Israel Journal of Mathematics, 196(1), pp.295-319.

[16] Carlsson, G., Zomorodian, A., Collins, A., and Guibas, L.J., *Persistence barcodes for shapes* (2005), International Journal of Shape Modeling, 11(02), pp.149-187.

BIBLIOGRAPHY

[17] Chavel, I., *Riemannian geometry. A modern introduction*, Second edition. Cambridge Studies in Advanced Mathematics, 98. Cambridge University Press, Cambridge, 2006. xvi+471 pp. ISBN: 978-0-521-61954-7; 0-521-61954-8

[18] Hatcher, A., *Algebraic Topology*(2002), Cambridge University Press, Cambridge. Also available online at http://www. math. cornell. edu/∼ hatcher/AT/ATpage. html.

[19] Govc, D., and Skraba, P., *An approximate nerve theorem*(2018), Foundations of Computational Mathematics, 18(5), pp.1245-1297.

[20] Hausmann, J.C., *On the Vietoris-Rips complexes and a cohomology theory for metric spaces* (1995), Annals of Mathematics Studies, 138, pp.175-188

[21] Latschev, J., *Vietoris-Rips complexes of metric spaces near a closed Riemannian manifold*(2001), Archiv der Mathematik, 77(6), pp.522-528.

[22] Wasserman, L., *All of statistics: a concise course in statistical inference*(2013), Springer Science & Business Media.

[23] Milnor, J., *Morse Theory (AM-51)*(2016), Volume 51. Princeton university press.

국문초록

최근 위상적 자료분석 방법은 데이터 분석에 각광받고 있다. 이 논문에서는 데이터의 모형을 분석하기 위하여 Vietoris-Rips complex와 persistent 호몰로지를 연구하였다. 특별히, Vietoris-Rips complex구조가 주어진 데이터가 n차원 구와 같은 호몰로지를 갖을 수 있다고 할 때, 필요한 최소한의 데이터 양을 산출해냈다.

주요어휘: Vietoris-Rips complex, persistence 모듈, persistent 호몰로지, Nerve 이론, 위상적 자료 분석
학번: 2012–30869