

THE PERSISTENCE OF DATA: A ROAD MAP

by

Shrunal Pothagoni

A Honors Thesis

Submitted to the

Faculty

of

George Mason University

In Partial fulfillment of

The Requirements for the Degree

of

Bachelor of Science with Honors in

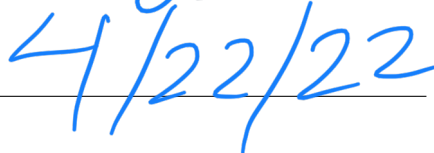
Mathematics

Committee:



Dr. Sean D Lawton, Honors Thesis Director

Date:



Spring Semester 2022
George Mason University
Fairfax, VA

Copyright © 2022 by Shrinal Pothagoni
All Rights Reserved

Dedication

I dedicate this dissertation to my family, friends, and loved ones. Thank you for putting up with my nonsense for the past four years.

Acknowledgments

I would first and foremost like to thank Dr. Sean Lawton for his continuous support, dedication, and commitment for the past year. Thank you for making the past year of research incredibly fun and rewarding. I would also like to thank Arun and Ajay Krishnavajjala for their help with the coding aspect of this thesis and creating the visuals used in this thesis. I would also like to thank Julian Benali for our insightful conversations on Algebraic Topology and Category Theory. Lastly, I would like to thank all of my mentors and colleagues that I have met through the Mason Experimental Geometry Lab. Thank you everyone.

Table of Contents

| | Page |
|--|------|
| List of Figures | vi |
| Abstract | vii |
| 1 An Introduction | 1 |
| 2 Preliminary Background and Results | 3 |
| 2.1 Basic Group Theory | 3 |
| 2.2 Vector Spaces | 4 |
| 2.3 Topology | 6 |
| 2.4 Posets and Basic Category Theory | 8 |
| 3 Simplicial Complexes | 10 |
| 3.1 Simplicial Complexes | 10 |
| 3.1.1 Čech and Rips Complex | 11 |
| 3.2 Simplicial Homology | 17 |
| 3.2.1 Finitely Generated Abelian Groups | 17 |
| 3.2.2 Chain Complexes | 22 |
| 4 Persistence of Data | 28 |
| 4.1 Computing Homology | 28 |
| 4.2 Smith Normal Form and The Standard Algorithm | 31 |
| 4.3 Filtration to Barcodes | 36 |
| 5 Algorithmic Implimentation | 43 |
| 5.1 Deriving Complexes From Point Clouds | 44 |
| 6 Multi-Parameter Filtrations and Persistent Modules | 49 |
| 6.1 Bifiltration and Bipersistence Modules | 50 |
| 6.2 No Good Barcodes | 54 |
| Bibliography | 55 |

List of Figures

| Figure | Page |
|--|------|
| 3.1 Visualization of Simplicies | 11 |
| 3.2 Examples of Non-Simplicial Complexes | 11 |
| 3.3 A 3-Dimensional Simplicial Complex | 12 |
| 3.4 Construction of the Čech Complex given a point cloud | 13 |
| 3.5 Here we can see the relationship between the Čech complex (left) and the Rips complex (right). Image taken from [Ghr08] | 16 |
| 3.6 Visual Intuition of the Boundary Map | 23 |
| 3.7 Visual Intuition of a Simplicial Chain Complex | 25 |
| 4.1 A 3-Dimensional Simplicial Complex | 29 |
| 4.2 This image illustrates the filtration of a given point cloud \mathcal{P} | 36 |
| 4.3 This barcode illustrates of how the homology classes are changing with respect to the filtered simplicial complex. Image taken from [Ghr08] | 40 |
| 4.4 Persistent Diagram. Image taken from [FC16] | 41 |
| 4.5 The distance between two persistent diagrams using the bottleneck distance. Image taken from [FC16] | 42 |
| 5.1 A noisy set of points around a circle of radius 3. | 44 |
| 5.2 The resulting 1-skeleton. | 45 |
| 5.3 This visual diagram illustrates how the computer is being programmed to complete the simplex or ignore it. | 47 |
| 6.1 This visual illustrates how barcodes are not stable with respect to noise. Notice that a slight introduction of noise from the left most image to the middle derails the persistence in the barcodes produced. Image credits to [LW15] | 50 |
| 6.2 Visualization of a tendril data set. Image taken from [KW18] | 51 |
| 6.3 An example of a bifiltration of simplicial complexes | 52 |

Abstract

THE PERSISTENCE OF DATA: A ROAD MAP

Shrunal Pothagoni, B.S.

George Mason University, 2022

Honors Thesis Director: Dr. Sean D Lawton

The purpose of data mining is to use advanced mathematical and statistical techniques to extract quantitative information from large data sets. These tools are incredibly powerful and in conjunction with machine learning algorithms allow for extremely accurate pattern prediction. However, there are various datasets that have qualitative properties that cannot be discerned using classic data mining techniques. Topological Data Analysis (TDA) is a field developed within the last two decades that uses methods in topology to extract such qualitative features. In this paper we will study how to use abstract simplicial complexes on point cloud data sets to find their most ‘optimal’ topology using computational homology.

Chapter 1: Introduction

Data mining techniques are commonly used to find patterns or extract meaning within data. With the abundance of large and rich datasets, the ability to extract meaning from data is more prevalent than ever. This information is processed using various statistical and mathematical methods. However, an unfortunate consequence of using such methods is that most data mining techniques provide no explanation as to what attributes or features in a data set are responsible for a particular anomaly.

For instance, one might design a machine learning algorithm to find a particular object within a set of images and given a large enough sample of data one can execute this algorithm to output predictions with a high level of accuracy. However, this machine learning algorithm cannot provide any explanation on what attributes are present in the image that allow it help classify this set of data. This isn't an issue when a data mining technique is used for pattern prediction. But in the context of testing scientific hypotheses, these statistical algorithms have a disadvantage.

To address this issue we need to examine what qualitative attributes our data set possess rather than its quantitative attributes. The purpose of this thesis is to study an emerging branch of mathematical data science known as Topological Data Analysis (TDA). TDA uses methods in topology to extract information from sets that are often high dimensional, noisy, and incomplete. A particular method in TDA is Persistent Homology (PH), a computational tool that allows users to extract qualitative features that persist from a dataset across multiple scales. There have been many applications of PH, such as signal analysis, material science, and shape recognition to name a few. Our goal is to begin by first building the theoretical background of persistence homology. Once this pipeline has been established we will then move onwards and layout a proper road map so that if one were to decide to algorithmically implement these tools, they could.

Chapter 2 includes all the preliminary knowledge required to move forward throughout this thesis. The first section will include sufficient background on basic concepts from linear algebra and group theory and should be review to anyone who has completed (or nearly completed) a traditional undergraduate mathematics degree. The last two sections will cover some basic point set-topology as well as a basic introduction into category theory. It is advised that one read through this chapter thoroughly as these concepts will be used throughout this thesis.

Once the preliminaries have been established, we will move forward into the realm of simplicial complexes in Chapter 3. Simplicial complexes are pivotal in the field of PH. The first section will focus on defining a simplicial complex and how various complexes, such as the Čech complex and Vietoris-Rips complex, are used to give datasets a topological structure. From here we will then explore the most important section of this thesis: simplicial homology. Homological algebra is a necessary tool used to articulate the qualitative features of a topological structure. This topological structure endowed upon the data set will allow us to describe it's global behavior.

At this point we have all of the relevant information to begin computing homology to find persistent features within our dataset. Chapters 4 provides an outline for how the theory discussed in the last few chapters can be synthesized to summarize this information into a compact pictographic known as a barcode. Chapter 5 will be a basic guide on how to implement persistence theory. This includes discussions on computational complexity, outlines for algorithms, and a few toy examples.

Lastly, we will finish this thesis off with a small introduction into what is the current state of research in persistent theory. These topics will include a brief introduction into multi-parameter persistent homology, noise filtration, and topological statistics.

Chapter 2: Preliminary Background and Results

The techniques and methods developed to study persistent homology require the understanding of algebraic topology. Although it is often the case that many universities don't offer this course at the undergraduate level, those that have taken courses in linear algebra, group theory, and topology have the necessary prerequisite knowledge to skip ahead to Section 2.4 if desired. Otherwise, I would recommend reading the following definitions and background as it will provide the foundation for the ideas that will be explored in this thesis.

2.1 Basic Group Theory

Definition 2.1. A nonempty set G endowed with a closed operation $*$ (i.e. if $a, b \in G$, then $a * b \in G$), is a *group*, denoted by $(G, *)$, if it satisfies the following axioms:

- 1.) For all $a, b, c \in G$, $a * (b * c) = (a * b) * c$ (Associativity).
- 2.) G has an *identity* element e_G that is for all $g \in G$, $e_G * g = g * e_G = g$.
- 3.) For every element $g \in G$ there exists an element $g^{-1} \in G$ such that $g * g^{-1} = g^{-1} * g = e_G$.

Remark 2.1. The operation $*$ is often omitted and we instead write the operation of two elements $g * h$ as just gh .

Example 1. The most known example of a group is the set of integers \mathbb{Z} under addition (one can verify that if we change the operation to multiplication (\mathbb{Z}, \cdot) is not a group). However, there are a large variety of groups that range in complexity and interest. An

example of a nontrivial group that is often introduced in most introductory algebra courses is the dihedral group D_n , the set of symmetries on a regular polygon with n -sides

Definition 2.2. A group G is said to be *abelian* if for all $g, h \in G$, $gh = hg$

Definition 2.3. A *group homomorphism* φ is a map between two groups $(G, *)$ and (H, \cdot) , denoted $\varphi: G \rightarrow H$, such that $\varphi(g * g') = \varphi(g) \cdot \varphi(g')$ for all g and g' in G . If φ is also a bijective map between G and H then φ is known as an *isomorphism* of groups.

Definition 2.4. G is said to be a *cyclic group* if there exists an element $a \in G$ such that every other element in G can be generated by repeated operations on a . Likewise, a is called the *generator* of G .

Remark 2.2. Of course, not every group is cyclic in nature. However, it may be the case that G has multiple elements that generate it. This is known as a *generating set* of G . Consequently, if this set happens to consist of finitely many elements we say that G is *finitely generated*

2.2 Vector Spaces

Definition 2.5. A *Vector Space* V over a field k is a nonempty set equipped with addition and scalar multiplication satisfying the following:

- V is an abelian group under addition (that is, it is closed under addition, addition is associative, there is an additive identity, each element has an additive inverse, and addition is commutative);
- V is closed under scalar multiplication, scalar multiplication is associative, the element $1 \in k$ is the identity scalar, and scalar multiplication distributes over scalar addition and vector addition.

Definition 2.6. Let V be a vector space over a field k . A set of nonzero vectors $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is said to be *linearly independent* if

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \dots + \alpha_n \mathbf{v}_n = \mathbf{0}$$

for some $\alpha_1, \dots, \alpha_n$ in k if and only if $\alpha_1 = \dots = \alpha_n = \mathbf{0}$

Definition 2.7. Let V be a vector space over a field k . A set of vectors $\mathcal{B} \subset V$ is a spanning set for V if, for every vector $\mathbf{v} \in V$, there exists a set of vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ in \mathcal{B} and $\alpha_1, \dots, \alpha_n$ in k such that $\mathbf{v} = \alpha_1 \mathbf{v}_1 + \dots + \alpha_n \mathbf{v}_n$.

Definition 2.8. A basis \mathcal{B} of a given vector space V is a nonzero set of vectors that are linearly independent and are a spanning set for V .

A nontrivial result that arises as a consequence of assuming the axiom of choice is that every nonzero vector space has a basis (a linearly independent spanning set) [Axl97]. One of the fundamental results of linear algebra is that each basis for a vector space V has the same size, and this size is referred to as the dimension of V . By convention, we say that the dimension of the zero vector space is zero.

One of the standard results from linear algebra is that given a finite dimensional vector space V and a nonzero k , V is isomorphic to k^n if and only if $n = \dim V$.

Definition 2.9. A *linear transformation* of a k vector spaces is a function

$$f : V \rightarrow W$$

that preserves vector addition and scalar multiplication. That is, $f(u + v) = f(u) + f(v)$ and $f(au) = af(u)$ for all $u, v \in V$ and $a \in k$. If a linear transformation is also bijective (i.e. injective and surjective), it is called an *isomorphism* of vector spaces.

Although there are various examples of interesting vector spaces, the contents of this thesis will primarily focus on studying \mathbb{R}^n . However, these definitions will play a crucial role in Section 3.2 in which we extend the notion of bases with respect to groups.

2.3 Topology

Definition 2.10. A *topology* on a set X is a collection \mathcal{T} of subsets of X have the following properties:

- 1.) \emptyset and X are in \mathcal{T}
- 2.) The union of any arbitrary subcollection of \mathcal{T} is in \mathcal{T} .
- 3.) The intersection of any finite subcollection of \mathcal{T} is also in \mathcal{T} .

A set X in which a topology \mathcal{T} is defined is called a *topological space*. Often we denote this as an order pair (X, \mathcal{T}) where \mathcal{T} is the topology on the set X . However this notation is often dropped within context if there is no confusion on what topology is being used. Furthermore, all sets $U \in \mathcal{T}$ are called *open sets*.

Remark 2.3. For the rest of this section we will assume X to be a topological space.

Definition 2.11. A set C is said to be *closed* if its compliment is open, that is, $X - C \in \mathcal{T}$

Definition 2.12. A function between two topological spaces X, Y , denoted $f: X \rightarrow Y$, is *continuous* if $f^{-1}(U)$ is open in X for all open $U \subset Y$. Furthermore, a continuous function f is a *homeomorphism* if it has a continuous inverse.

Proposition 1. Suppose f is a function between two topological spaces X, Y . f is continuous if and only if for every closed set $C \subseteq Y$, $f^{-1}(C)$ is closed in X .

Proof. Suppose that f is continuous and $C \subseteq Y$ is closed. It suffices to show that $f^{-1}(C)$ is closed in X . Given that $X - C$ is open in Y we have that $f^{-1}(X - C) = X - f^{-1}(C)$. However, we know that f is continuous. So, $X - f^{-1}(C)$ is open. Thus, $f^{-1}(C)$ is closed. Conversely, assume that the inverse image of closed is closed under f . Given an open set $U \subseteq Y$, $Y - U$ is a closed. Then $f^{-1}(Y - U) = X - f^{-1}(U)$. But this implies that $f^{-1}(U)$ is open and f is continuous. □

Definition 2.13. A *metric space* is an ordered pair (X, d_X) where X is a set endowed with a function $d_X: X \times X \rightarrow \mathbb{R}$ such that for any $x, y, z \in X$:

- 1.) $d(x, y) = 0$ iff $x = y$ and $d(x, y) \geq 0$ (positive definite)
- 2.) $d(x, y) = d(y, x)$ (symmetry)
- 3.) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

Remark 2.4. Suppose (X, d_X) is a metric space. We define open set to be any set that is the union of ϵ -balls where an ϵ -ball, is any set of the form:

$$\mathbb{B}_\epsilon(x) = \{y \in X : d(x, y) < \epsilon\}$$

Consequently, this defines a topology on X known as the metric topology [Mun00]. In particular, once a metric has been defined for a given set X , its topology is determined by the set of open balls and closed sets are defined by closed balls.

Example 2. The consider the taxi cab metric on \mathbb{R}^2 defined by $d(x, y) = |x_1 - y_1| + |x_2 - y_2|$. This defines a metric topology on \mathbb{R}^2 whose open balls look like a square rotated by 45° .

Definition 2.14. A *cover* of a set E is a collection of sets, $\{U_\alpha : \alpha \in \Lambda\}$ for some index set Λ , such that $E \subseteq \bigcup_{\alpha \in \Lambda} U_\alpha$.

Remark 2.5. Within the context of this thesis we will be referring to covers with respect to a topological space X . Furthermore, if the covering of a topological space X consists of a collection of open sets it is called an *open cover* of X .

Remark 2.6. If $\{U_\alpha : \alpha \in \Lambda\}$ is a cover of X and there exists a subcollection of sets from the cover that also covers X , it is called a *subcover*.

Definition 2.15. Let $E \subset X$. E is said to be *compact* if for every open cover of E there exists a *finite subcover*.

2.4 Posets and Basic Category Theory

Definition 2.16 ([Rie17]). We define \mathcal{C} to be a *category* if

- there is a class of object denoted $\text{Ob } \mathcal{C}$
- for all $x, y \in \text{Ob } \mathcal{C}$ there is a class of morphisms denoted as $\text{Hom}(X, Y)$

such that:

- Each morphism has a specified domain and codomain between objects (i.e. $f: X \rightarrow Y$ specifies a morphism between an object X to Y).
- Each object has a unique identity morphism $I_X: X \rightarrow X$.
- For any pair of morphism f and g , if the codomain of f is the domain of g , then there exists a composite morphism $g \circ f$ whose domain is equal to the domain of f and codomain is that of g , i.e.,:

$$f: X \rightarrow Y, \quad g: Y \rightarrow Z, \quad g \circ f: X \rightarrow Z$$

These axioms are subject to the condition that the composition of morphisms is associative and that unital with respect identity morphism given by the to the two-sided identity.

Remark 2.7. Classic examples of categories include:

- The category **Top**, whose objects are topological spaces and morphisms are continuous functions
- The category **Grp**, whose objects are groups and morphisms are group homomorphisms
- Lastly, the category **Vect**, whose objects are vector spaces and morphisms are linear transformations.

There are many other examples of categories. However, we will limit our focus to these specific categories as they will appear again in the last chapter of this thesis.

Definition 2.17. Let (P, \leq) be a preorder set (i.e. \leq is reflective and transitive). A *partially order set* (often called *poset*) is a category such that

- the objects of this category is P itself;
- for all x and y in P there is single unique morphism in $\text{Hom}(x,y)$ if and only if $x \leq y$.
Otherwise $\text{Hom}(x,y)$ is empty.

Example 3. Suppose $P = \mathbb{N}$. Then \mathbb{N} can be represented by following "Hasse diagram"

$$1 \rightarrow 2 \rightarrow 3 \rightarrow \dots$$

and \mathbb{N}^2 can be represented as

$$\begin{array}{ccccccc}
 & \vdots & & \vdots & & \vdots & \\
 & \uparrow & & \uparrow & & \uparrow & \\
 (3,1) & \longrightarrow & (3,2) & \longrightarrow & (3,3) & \longrightarrow & \dots \\
 & \uparrow & & \uparrow & & \uparrow & \\
 (2,1) & \longrightarrow & (2,2) & \longrightarrow & (2,3) & \longrightarrow & \dots \\
 & \uparrow & & \uparrow & & \uparrow & \\
 (1,1) & \longrightarrow & (1,2) & \longrightarrow & (1,3) & \longrightarrow & \dots
 \end{array}$$

Definition 2.18. Given two categories \mathcal{C} and \mathcal{D} , a *functor* , $F: \mathcal{C} \rightarrow \mathcal{D}$ satisfies:

- An object $F(x) \in \text{Ob } \mathcal{D}$ for every $x \in \text{Ob } \mathcal{C}$
- A morphism $F(\gamma) \in \text{Hom}(F(x), F(y))$ for each $\gamma \in \text{Hom}(x, y)$

such that F respects the composition operation in \mathcal{C} and \mathcal{D} , i.e. $F(f \circ g) = F(f) \circ F(g)$ and that $F(\text{Id}_x) = \text{Id}_{F(x)}$ for all $x \in \text{Ob } \mathcal{C}$

Chapter 3: Simplicial Complexes

3.1 Simplicial Complexes

Notation. Throughout the rest of this thesis we will use I to denote the unit interval $[0, 1]$.

Definition 3.1. Let V be a \mathbb{R} -vector space and C be a subset of V . C is said to be a *convex* if every pair of points $c_1, c_2 \in C$ the line segment between c_1 and c_2 is in C , that is,

$$c_1 + (c_2 - c_1)t \in C$$

for all $t \in I$ [ST15].

We can think of a convex set as a shape where, given any two points in the set, there is a linear path from one point to the other point within the shape itself.

Definition 3.2. Let $\{v_0, \dots, v_k\}$ be a set of vectors in V . This set is said to be *convex independent* or *c-independent* if $\dim(\text{Span}\{v_0 - v_i, \dots, v_k - v_i\}) = k$ for any $0 \leq i \leq k$.

Definition 3.3. Let V be a vector space over \mathbb{R} . A convex set generated by c -independent vectors $\{v_0, v_1, \dots, v_k\}$ is called a (closed) k -*simplex* denoted by $[v_0, v_1, \dots, v_k]$.

Definition 3.4. A *simplicial complex* K (Euclidean) is a finite set of open simplices in some \mathbb{R}^n such that:

- (1) if $(s) \in K$ then all open faces of $[s] \in K$;
- (2) if $(s_1) \cap (s_2) \neq \emptyset$ then $(s_1) = (s_2)$ [ST15].

Remark 3.1. All zero dimensional simplices are referred to as the *vertices* of a simplicial complex. Furthermore, given a k -simplex defined by vertices $\{v_{i_0}, \dots, v_{i_k}\}$ then any subset of these vertices form a *face* which is also a simplex in the simplicial complex.

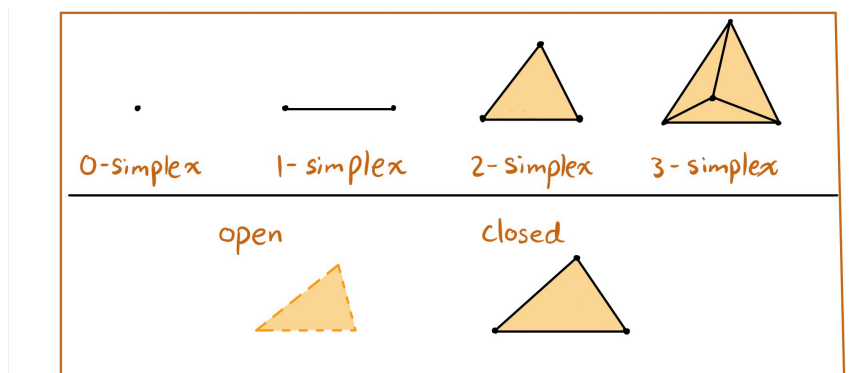


Figure 3.1: Visualization of Simplicies

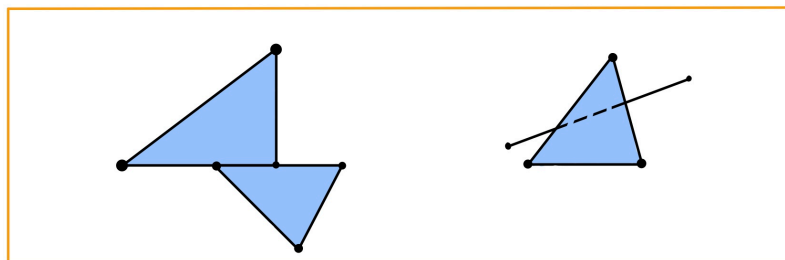


Figure 3.2: Examples of Non-Simplicial Complexes

3.1.1 Čech and Rips Complex

Given a point cloud dataset there are multiple ways in which one can derive a simplicial complex. However, the choice of which simplicial complex to construct is highly dependent on the computational constraints of the problem. In this thesis we will primarily focus on the Čech Complex and the Rips Complex.

Definition 3.5. Given a set of points $\mathcal{K} = \{k_1, \dots, k_n\} \subset \mathbb{R}^d$ and a real value $\epsilon > 0$, a n -simplex $\sigma = [k_{i_0}, \dots, k_{i_n}]$ is in the Čech complex $\check{C}ech_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ if and only if

$$\bigcap_{0 \leq j \leq n} \mathbb{B}(k_{i_j}, \epsilon) \neq \emptyset.$$

In particular, the Čech complex is determined by the parameter ϵ . 3.4 shows how a particular ϵ results in a particular simplicial complex. It will become apparent in Section 3

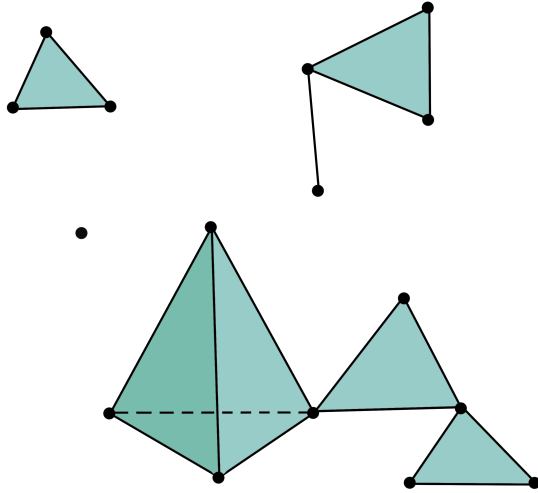


Figure 3.3: A 3-Dimensional Simplicial Complex

that variations to ϵ will effect the "birth" and "death" of certain simplices and the complex as a whole, thus, changing their topological characteristics.

Definition 3.6. Recall that a topological space X is said to *second countable* or *completely separable* if there exists a countable basis.

Definition 3.7. A topological space X is said to be *locally compact* if at every point $x \in X$ there exists a compact neighborhood around it. In particular $x \in U \subset K$ where U is open and K is compact.

Definition 3.8. A cover $\{U_i\}_{i \in \mathcal{I}}$ of X is said to be *locally finite* if for all points $x \in X$, there exists a neighborhood U_x that contains it such that it intersects only finitely many elements of the cover.

Definition 3.9. Suppose that $\{U_i\}_{i \in \mathcal{I}}$ is an open cover of X . A *refinement* of this open cover is a set of open subset $\{V_k\}_{k \in \mathcal{J}}$ which is still an open cover of X such that for each $j \in \mathcal{J}$ there exists an $i \in \mathcal{I}$ such that $V_j \subset U_i$

Definition 3.10. A topological space X is said to be *paracompact* if every cover has a finite local refinement.

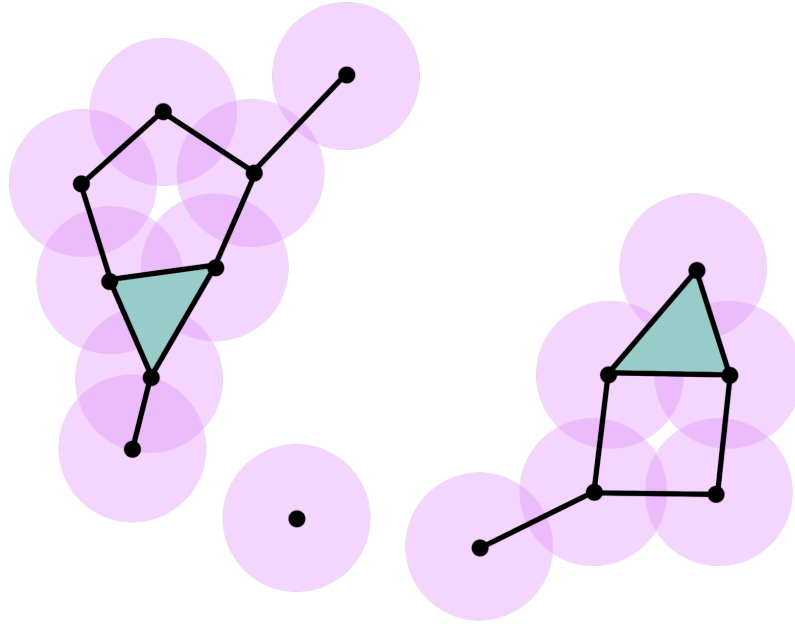


Figure 3.4: Construction of the Čech Complex given a point cloud

Definition 3.11. Consider two topological spaces X and Y with continuous maps $f, g: X \rightarrow Y$. A *homotopy* from f to g is a function $F: X \times I \rightarrow Y$ such that $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$ for all $x \in X$. Furthermore, X and Y are considered to be *homotopy equivalent* if there exists two continuous functions $f: X \rightarrow Y$ and $g: Y \rightarrow X$, such that $f \circ g$ is homotopic to Id_Y and $g \circ f$ is homotopic to Id_X .

One should think of homotopy between X and Y as a continuous deformation from one function to the other function. A simple illustration of this idea is that any simple closed loop in \mathbb{R}^2 can be continuously deformed into a circle in \mathbb{R}^2 and vice versa.

Definition 3.12. A topological space X is said to be *contractible* if X is homotopy equivalent to a point.

Definition 3.13. A *topological vector space* is a vector space V over a field k is endowed with a topology such that vector addition and scalar multiplication are continuous functions, that is,

$$+: V \times V \rightarrow V \quad \text{and} \quad \cdot: k \times V \rightarrow V$$

are continuous functions.

Proposition 2. If V be a Topological Vector Space over \mathbb{R} and $C \subset V$ be a convex subset, then C is contractible.

Proof. Let $c \in C$. Consider the continuous function $F: C \times [0, 1] \rightarrow C$ defined by $F(x, t) = ct - (1 - t)x$. By construction, this yields us a homotopy between Id_C at $t = 0$ and the constant map at c for $t = 1$. By convexity of C all inputs into F will be contained in C . Then for all $x \in C$, $F(x, 0) \cong \text{Id}_C$ and $F(x, 1) \cong c$ (where c is the constant function). Thus, $\text{Id}_C \simeq c$. □

Definition 3.14. Given an open cover $\mathcal{U} = \{U_\alpha\}$ of a given topological space X there is an associated simplicial complex defined as the *nerve* of \mathcal{U} denoted as $\mathcal{N}(\mathcal{U})$. This simplicial complex is constructed so that each vertex v_α corresponds to each open set U_α and every k -simplex is defined by $k + 1$ nonempty intersections of the corresponding U_α .

Remark 3.2. It follows by definition that the Čech Complex is equivalent to the nerve of a given set of vertices \mathcal{K} . However, it should be noted that since $\mathcal{N}(K)$ is determined by the intersection of the open sets, $\mathcal{N}(K)$ changes with respect to the chosen ϵ for the Čech Complex.

Theorem 3.1. 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 $\mathcal{N}(\mathcal{U})$.

Unfortunately, the proof for this theorem is beyond the scope of this thesis. Anyone interested in reading the details of this proof and its construction can find it in Hatcher's Algebraic Topology [Hat02]. Instead, we will be using this theorem to prove a rather strong condition that follows in the use of the Čech Complex.

Corollary 1. The Čech Complex $\check{C}ech_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ is homotopy equivalent to the union of balls $U_{\mathbb{R}^d}(\mathcal{K}, \epsilon) := \bigcup_{k_i \in \mathcal{K}} \mathbb{B}(k_i, \epsilon)$

Proof. By definition, $\check{C}ech_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ is equivalent to the $\mathcal{N}(\mathcal{K})$. It suffices to prove that $U_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ is a paracompact space and that every finite nonempty intersection of sets from it are contractible. First, \mathbb{R}^d is a locally compact space. It follows that since subspaces of locally compact spaces are locally compact, $U_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ must be locally compact. Second, $U_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ is an open cover on the set of points \mathcal{K} . Furthermore, $U_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ is a finite open cover on \mathcal{K} , implying second countability. Therefore, the union of balls on a finite dataset is a paracompact space. Furthermore, every open set in $U_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ is defined as an open ball, which is convex. But every finite intersection of convex sets are also convex. Thus, a finite intersection of sets from $U_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ are contractible. \square

Definition 3.15. Given a set of points $\mathcal{K} = \{k_1, \dots, k_n\} \subset \mathbb{R}^d$ and a real value $\epsilon > 0$, a n -simplex $\sigma = [k_{i_0}, \dots, k_{i_n}]$ is in the *Vietoris-Rips complex* (or often called Rips complex) $Rips_{\mathbb{R}^d}(\mathcal{K}, \epsilon)$ if and only if

$$\mathbb{B}(k_{i_j}, \epsilon) \cap \mathbb{B}(k_{i_{j'}}, \epsilon) \neq \emptyset$$

for any $j, j' \in \{0, 1, \dots, n-1, n\}$.

Definition 3.16. Given a graph \mathcal{G} , a *flag complex* or *clique complex* of \mathcal{G} is the maximal simplicial complex that has the graph as its 1-skeleton.

In essence, the intuition is that given a set of vertices and edges if it appears that they form simplex (the cliques) in the 1-skeleton, then the simplex is contained within the complex. The Rips complex is an example of a flag complex. From a computational stand point this serves as a strong advantage for the Rips complex. In particular, the relevant information for the simplicial complex is implicitly encoded within the 1-skeleton. Therefore it isn't necessary to store all of the simplices of the simplicial complex.

The main difference between the Čech complex and the Rips complex is that one

must find the common intersection for all $\mathbb{B}(p_{i_j}, \epsilon)$ to compute a simplex in the Čech Complex. This inevitable will lead to the Čech Complex being computationally more expensive than the Rips complex. In particular, the time complexity for computing the Čech complex is $O(n^{d+1})$ where n is the number of points used and d is the d -skeleton whereas the time complexity for the 1-skeleton of the Rips complex is $O(n^2)$ [DI12].

This computational inefficiency is the primary reason to chose the Rips complex over the Čech complex. This is also further amplified by the following relationship from Ghrist's paper [Ghr08] :

$$\check{C}ech_{\mathbb{R}^d}(\mathcal{K}, \epsilon) \subset Rips_{\mathbb{R}^d}(\mathcal{K}, \epsilon) \subset \check{C}ech_{\mathbb{R}^d}\left(\mathcal{K}, \sqrt{\frac{2n}{n+1}} \epsilon\right)$$

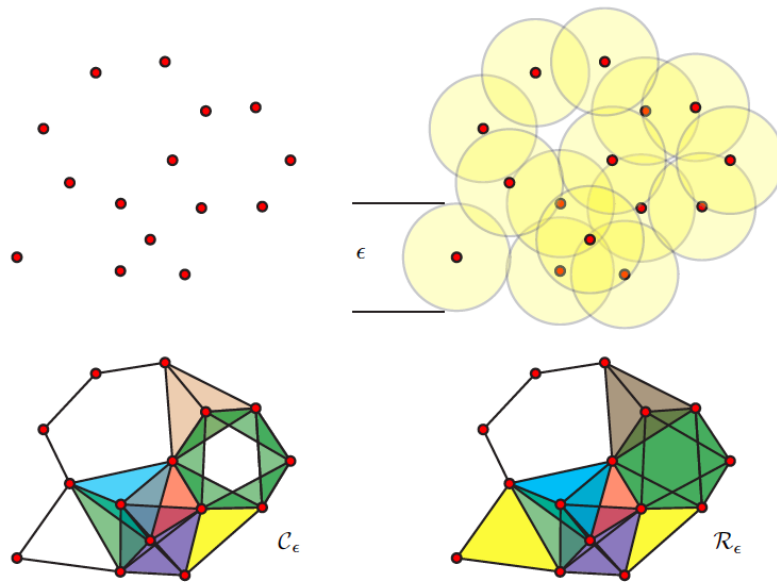


Figure 3.5: Here we can see the relationship between the Čech complex (left) and the Rips complex (right). Image taken from [Ghr08]

3.2 Simplicial Homology

Up till this point we have discussed what is a simplicial complex and how one is constructed. However, our ability to define and further talk about the geometric characteristics of a simplicial complex is contingent on applying group theoretic properties to the complex. To begin, we will give a gentle introduction into finitely generated abelian groups. We will then conclude with studying the homology of a given simplicial complex.

3.2.1 Finitely Generated Abelian Groups

Definition 3.17. Let G be an abelian group. G is said to be a *free abelian group* if G is isomorphic to $\bigoplus_{\alpha \in \Lambda} \mathbb{Z}$ for some index set Λ .

Remark 3.3. Abelian groups can have bases as well. The definition of a basis for group is similar to that of basis of a vector space. It must still generate all of G and be linearly independent, however, rather than using coefficients from a field, its coefficients are derived from \mathbb{Z} .

Proposition 3. An abelian group has a basis if and only if G is a free abelian group.

Proof. Let \mathcal{B} be a basis for G . Consider the map

$$f: \bigoplus_{x \in \mathcal{B}} \mathbb{Z} \rightarrow G, \quad f((\alpha_x)_{x \in \mathcal{B}}) = \sum_{x \in \mathcal{B}} \alpha_x x.$$

This function is a homomorphism since

$$\begin{aligned} f((\alpha_x)_{x \in \mathcal{B}} + (\beta_x)_{x \in \mathcal{B}}) &= \sum_{x \in \mathcal{B}} (\alpha_x + \beta_x) x \\ &= \sum_{x \in \mathcal{B}} \alpha_x x + \sum_{x \in \mathcal{B}} \beta_x x \\ &= f((\alpha_x)_{x \in \mathcal{B}}) + f((\beta_x)_{x \in \mathcal{B}}). \end{aligned}$$

Furthermore, the map

$$f^{-1}: G \rightarrow \bigoplus_{x \in \mathcal{B}} \mathbb{Z}, \quad f^{-1} \left(\sum_{x \in \mathcal{B}} \alpha_x x \right) = (\alpha_x)_{x \in \mathcal{B}}.$$

defines an inverse function for f . Thus, f is an isomorphism.

Conversely, suppose that

$$f: \bigoplus_{i \in \mathcal{I}} \mathbb{Z} \rightarrow G$$

is an isomorphism. Consider the set $\{\delta_j\}_{j \in \mathcal{I}}$ where $\delta_j = (n_i)_{i \in \mathcal{I}}$ such that $n_i = 1$ if $i = j$ otherwise $n_i = 0$. The set $\mathcal{B} = \{f(\delta_j)\}_{j \in \mathcal{I}}$ generates all of G as it is the image of a generating set for $\bigoplus_{i \in \mathcal{I}} \mathbb{Z}$. Also,

$$\alpha_1 f(\delta_1) + \dots + \alpha_n f(\delta_n) = 0$$

if and only if $\alpha_1 = \dots = \alpha_n = 0$ by construction. Thus, \mathcal{B} is a basis of G .

□

Proposition 4. If G is an abelian group generated by n elements and F is a free abelian group of rank n , then

$$G \cong F/H$$

where H is a subgroup of F

Proof. Let $\langle g_1, \dots, g_n \rangle$ be the generating set for G . Suppose that $\{x_1, \dots, x_n\}$ is the basis for F . Define

$$f: F \rightarrow G, \quad f(x_i) = g_i.$$

By construction f is surjective. Take $H = \ker(f)$ and by the first isomorphism theorem

$$G \cong F/H. \quad \square$$

Definition 3.18 (Smith Normal Form). Suppose that $A \in M_{m \times n}(\mathbb{Z})$. Then there exists a pair of invertible matrices S and T such that SAT is equivalent to a matrix of the form,

$$\begin{pmatrix} \alpha_1 & 0 & 0 & \cdots & 0 \\ 0 & \alpha_2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & 0 \\ \vdots & & & \alpha_r & \vdots \\ & & & & 0 \\ & & & & \ddots \\ 0 & & \cdots & & 0 \end{pmatrix}$$

where each α_i is on the diagonal of the matrix and $\alpha_i | \alpha_{i+1}$ for all i where $1 \leq i \leq r$

Remark 3.4. A matrix that has a SNF is not necessarily diagonalizable. For example,

$$\text{SNF} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

since

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

However, $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is not diagonalizable.

Remark 3.5. Every matrix with integer entries has a Smith Normal Form and can be obtained via the following elementary row operations

- *Row recombine.* Replace the i th row by itself plus k times another j th row where $k \in \mathbb{Z}$. Symbolically denoted as $r_i \leftarrow r_i + kr_j$
- *Row Scaling.* Every i th row can be scaled by -1 . Symbolically denoted as $r_i \leftarrow -r_i$
- *Row Transposition.* Exchanging or swapping the i th and j th row. Symbolically denoted as $r_i \leftrightarrow r_j$

This is similarly true for operations done with the columns.

Theorem 3.2. Let F be a free abelian group of rank n and let H be a subgroup of F . Given that $\{x_1, \dots, x_n\}$ is a basis for F there exists $d_1, \dots, d_r > 0$ such that,

- $d_i | d_{i+1}$ for $1 \leq i \leq r$ and
- $\{d_1 x_1, \dots, d_r x_r\}$ is a basis for H .

Proof. Let $\{f_1, \dots, f_n\}$ be a basis for F and $\{h_1, \dots, h_m\} \subseteq H$ is a generating set of H . Of course,

$$h_i = a_{i1}f_1 + \dots + a_{in}f_n$$

for some $a_{ij} \in \mathbb{Z}$. Consider the matrix transformation from $A: F \rightarrow H$ by

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{r1} & \dots & a_{rn} \end{pmatrix}.$$

since the columns of A correspond to the basis of F and the rows of A are the generators of H .

Given a set of elementary row operations, we are able to obtain A 's Smith Normal Form

$$B = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}.$$

where D is the diagonal matrix as defined before. Consequently the columns of B also define a basis for F and the rows of B , $\{d_1x_1, \dots, d_rx_r, 0, \dots, 0\}$, defines a generating set for H . Thus, $\{d_1x_1, \dots, d_rx_r\}$ is a basis of H . \square

Theorem 3.3 (Fundamental Theorem of Finite Abelian Groups). If G is a finitely generated free abelian group then

$$G \cong \mathbb{Z}^k \oplus \bigoplus_{i=1}^n \mathbb{Z}/d_i\mathbb{Z}$$

where $k \geq 0$ and $d_i|d_{i+1}$

Corollary 2. Let A, B be two matrices of size $m \times n$ and $l \times n$ respectively whose entries are composed of integers such $BA = 0$. Then

$$\ker(B)/\text{im}(A) \cong \mathbb{Z}^{m-r-s} \oplus \bigoplus_{i=1}^r \mathbb{Z}/\alpha_i\mathbb{Z}$$

where the $\text{Rank}(B) = s$, the $\text{Rank}(A) = r$, and each α_i are generated from the diagonal entries of A 's smith normal form.

Proof. $\ker(B)$ is a subgroup of \mathbb{Z}^m as well as a finitely generated abelian group. Then $\ker(B) \cong \mathbb{Z}^{m-s}$ by Rank-Nullity. Furthermore, given that $\text{im}(A) \leq \ker(B)$ it suffices to prove that $\text{im}(A)$ has a compatible basis with respect to $\ker(B)$. The image of A has a basis generated by the entries of its Smith Normal Form. In particular, $\text{im}(A) \cong \bigoplus_{i=1}^r \alpha_i\mathbb{Z}$. Thus,

$$\ker(B)/\text{im}(A) = \mathbb{Z}^{m-s} / \bigoplus_{i=1}^r \alpha_i \mathbb{Z} \cong \mathbb{Z}^{m-r-s} \oplus \bigoplus_{i=1}^r \mathbb{Z} / \alpha_i \mathbb{Z}$$

□

3.2.2 Chain Complexes

Definition 3.19. Let $[s] = [v_0, v_1, \dots, v_\ell]$ be an ℓ -simplex. We will say that $[v_{i_1}, v_{i_2}, \dots, v_{i_\ell}] \sim [v_{j_1}, v_{j_2}, \dots, v_{j_\ell}]$ if there exists an *even* permutation that maps $(i_1, i_2, \dots, i_\ell) \rightarrow (j_1, j_2, \dots, j_\ell)$. Since all permutations are either even or *odd*, this implies that \sim is an equivalence relation. An *oriented* simplex denoted as $\langle s \rangle = \langle v_0, v_1, \dots, v_\ell \rangle$ is a simplex endowed with a choice of one of these two equivalence classes.

Definition 3.20. Let K be a simplicial complex, \mathcal{G} be an abelian group. We define the group of ℓ -chains to be

$$C_\ell(K, \mathcal{G}) = \left\{ \sum_{\sigma_i \in K} m_i \sigma_i : m_i \in \mathcal{G} \right\}$$

where σ_i is an ℓ -simplex in K with coefficients in \mathcal{G} .

Remark 3.6. Every ℓ -chain is a free abelian group with each basis corresponding uniquely to the number of ℓ -simplexes in K . Consequently, if \mathcal{G} is a field, then every ℓ -chain is a vector space. The importance of this feature will become apparent later in this chapter.

Definition 3.21. Consider the oriented ℓ -simplex $\langle s \rangle = \langle v_0, v_1, \dots, v_\ell \rangle$. The *boundary map*, ∂ , is defined as

$$\partial(\langle s \rangle) = \sum_{i=0}^{\ell} (-1)^i \langle v_0, v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_\ell \rangle.$$

Remark 3.7. In essence, the boundary map is quite literally taking a k -simplex and removing its interior to expose the shell of the k -simplex. This is of course composed of the

$k - 1$ -simplices. The following image provides a visual intuition of this map.

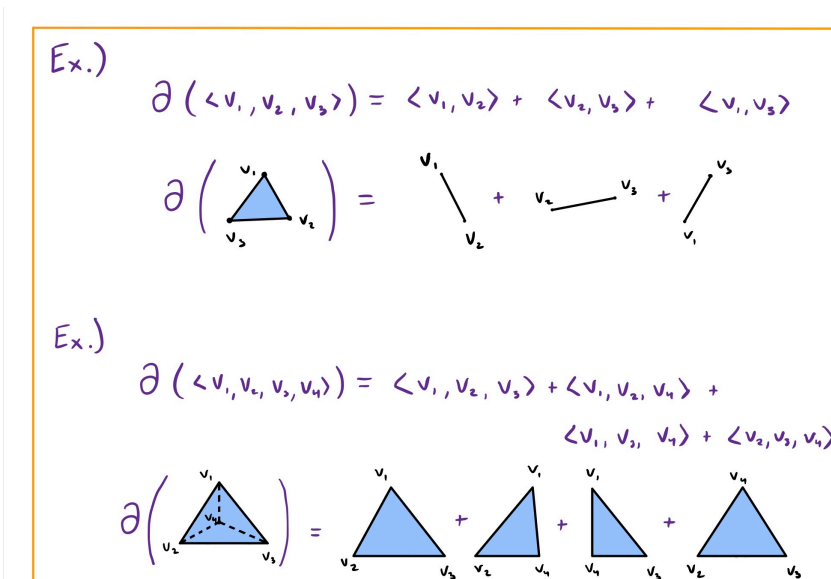


Figure 3.6: Visual Intuition of the Boundary Map

Proposition 5. Given an oriented ℓ -simplex $\langle s \rangle$, $\partial^2(\langle s \rangle) = \partial(\partial(\langle s \rangle)) = 0$

Proof. Consider the simplicial chain complex

$$0 \xrightarrow{\partial} \cdots \xrightarrow{\partial} C_{\ell+1}(K, \mathcal{G}) \xrightarrow{\partial} C_{\ell}(K, \mathcal{G}) \xrightarrow{\partial} C_{\ell-1}(K, \mathcal{G}) \xrightarrow{\partial} \cdots \xrightarrow{\partial} 0.$$

∂^2

Let $\langle s \rangle$ be an $\ell + 1$ -simplex. Notice that

$$\begin{aligned} \partial^2(\langle s \rangle) &= \partial \left(\sum_{k=0}^{\ell+1} (-1)^k \langle v_0, v_1, \dots, \hat{v}_k, \dots, v_{\ell}, v_{\ell+1} \rangle \right) \\ &= \sum_{k=0}^{\ell+1} (-1)^k \partial(\langle v_0, v_1, \dots, \hat{v}_k, \dots, v_{\ell}, v_{\ell+1} \rangle) \end{aligned}$$

However, this divides our current sum into the following two sums with the relation

$$\begin{aligned}
\partial^2(\langle s \rangle) &= \sum_{k=0}^{l+1} (-1)^k \left(\sum_{j=0}^{k-1} \langle v_0, v_1, \dots, \hat{v}_j, \dots, \hat{v}_k, \dots, v_\ell, v_{\ell+1} \rangle \right. \\
&\quad \left. + \sum_{j=k+1}^{\ell+1} \langle v_0, v_1, \dots, \hat{v}_k, \dots, \hat{v}_j, \dots, v_\ell, v_{\ell+1} \rangle \right) \\
&= \sum_{j < k} (-1)^{j+k} \langle v_0, v_1, \dots, \hat{v}_j, \dots, \hat{v}_k, \dots, v_\ell, v_{\ell+1} \rangle \\
&\quad + \sum_{k < j} (-1)^{j+k-1} \langle v_0, v_1, \dots, \hat{v}_k, \dots, \hat{v}_j, \dots, v_\ell, v_{\ell+1} \rangle
\end{aligned}$$

implying that

$$\begin{aligned}
\partial^2(\langle s \rangle) &= \sum_{j < k} \left((-1)^{j+k} + (-1)^{j+k-1} \right) \langle v_0, v_1, \dots, \hat{v}_j, \dots, \hat{v}_k, \dots, v_\ell, v_{\ell+1} \rangle \\
&= 0 \quad \square
\end{aligned}$$

and that the $\text{im } \partial_i \subseteq \ker \partial_{i+1}$.

Definition 3.22. A *chain complex* is a sequence of abelian groups C_i along with the group homomorphisms $\partial_i : C_i \rightarrow C_{i+1}$ satisfying $\text{im } \partial_i \subseteq \ker \partial_{i+1}$ (this is true since $\partial^2(\langle s \rangle) = 0$). In general a chain complex may be infinite in one direction, or infinite in both directions, however, since our chain complex is constructed from a simplicial complex, it will be finite. Within the context of this thesis we will refer to these chain complexes as *simplicial chain complexes*. We refer to the entire complex with the notation C_\bullet and write

$$C_\bullet : \dots \xrightarrow{\partial_i} C_i \xrightarrow{\partial_{i+1}} C_{i+1} \xrightarrow{\partial_{i+2}} C_{i+2} \xrightarrow{\partial_{i+3}} \dots$$

Remark 3.8. In the creation of simplicial chain complexes, since they are all finite free abelian groups under addition they are isomorphic to $\bigoplus_{i=1}^n \mathbb{Z}$ where n represents the number

of k -simplexes of K . For example, consider the simplicial complex represented by Figure 3.1. The resulting simplicial chain complex is

$$0 \rightarrow \mathbb{Z} \rightarrow \mathbb{Z}^8 \rightarrow \mathbb{Z}^{23} \rightarrow \mathbb{Z}^{18} \rightarrow 0$$

given that there are exactly 18 vertices, 23 edges, 8 faces, and one tetrahedron.

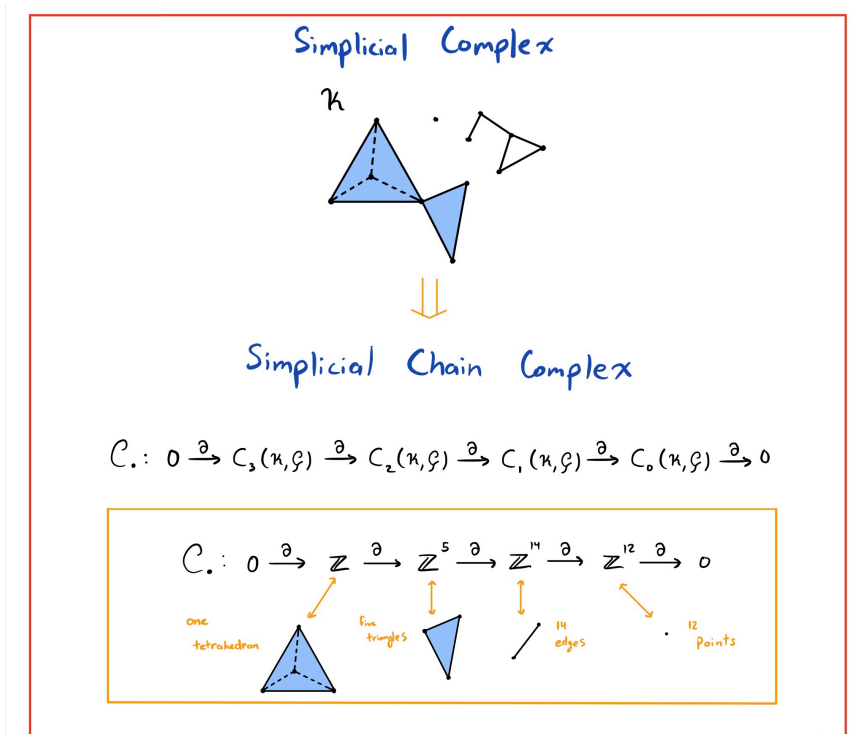


Figure 3.7: Visual Intuition of a Simplicial Chain Complex

Definition 3.23. Given $i \in \mathbb{Z}^+$, the i -th homology of C_\bullet is the quotient group $H_i(C_\bullet) = \ker(\partial_i) / \text{im}(\partial_{i+1})$. If each homology group is 0, the complex is said to be *exact*.

Definition 3.24. For a nonnegative integer i the i -th Betti number, β_i , is defined as the rank of the i -th homology of C_\bullet .

Betti numbers are relevant in determining if two topological spaces are potentially different from each other. Their utility will become apparent in the following chapter when we

compare how the Rips Complex as ϵ changes. Also note that although the homology groups might be abelian, it is not the case that they are free abelian groups. However since all finitely generated abelian group are isomorphic to the product of cyclic group, the following theorem holds.

Corollary 3. Let $H_i(C_\bullet)$ be the homology group of a chain complex C_\bullet generated by a simplicial complex. Then

$$H_i(C_\bullet) = \ker(\partial_i)/\text{im}(\partial_{i+1}) = \mathbb{Z}^{\beta_i} \oplus \bigoplus_k (\mathbb{Z}/m_k\mathbb{Z})$$

for integers $m_1 \leq m_2 \leq \dots \leq m_n$ generated from the Smith Normal Form of $\text{im}(\partial_{i+1})$ where each integer m_k is a divisor of its latter m_{k+1} . This secondary component $\bigoplus_i \mathbb{Z}/m_i\mathbb{Z}$ is known as the torsion subgroup of $H_i(C_\bullet)$.

Remark 3.9. Over \mathbb{Z} , the following provides an intuition for how homology allows us identify characteristics of topological space:

- $H_0(C_\bullet)$ can be interpreted as the number of connected component;
- $H_1(C_\bullet)$ gives the abelianization of π_1 ;
- $H_{\dim(\mathcal{M})}(C_\bullet)$ gives the orientability of a manifold \mathcal{M} without it's boundaries.

Although it is common practice to calculate the homology over \mathbb{Z} , since we are given the flexibility to use any free abelian group, we will be using \mathbb{Z}_2 (field of order 2). In the absence of torsion, the Betti numbers under \mathbb{Z}_2 are the same as those under \mathbb{Z} , according to the Universal Coefficient Theorem [Hat02]. Thus, we have a clear benefit with respect to computational time complexity if we chose to use \mathbb{Z}_2 . In fact, when computing the Betti numbers over \mathbb{Z}_2 our calculations simplify to

$$\beta_i = \text{Rank}(H_i(\mathcal{K}_\epsilon, \mathbb{Z}_2)) = \text{Rank}(\ker(\partial_i)) - \text{Rank}(\text{im}(\partial_{i+1}))$$

However, this does pose a problem in the presence of torsion. In particular, although we may still use \mathbb{Z}_2 coefficients, these answers may differ from those computed using \mathbb{Z} coefficients. Luckily this issue can be circumvented entirely by using various other finite fields as presented by [ZC05].

Chapter 4: Persistence of Data

The goal of this chapter is to establish the framework to find the persistence of a given dataset. The first step is to begin with creating a simplicial complex from a given set of data; the details of which were already mentioned in 3.1.1. Thus, we will instead begin by computing the homology via the boundary maps of the derived simplicial complex. Consequently, it is important that we then encode the homological information in a way that we can observe its changes with respect to the increasing complexity of the simplicial structure. This is often done with pictographic diagrams known as barcodes or persistence diagrams which are then used to analyze the data and interpret the results.

4.1 Computing Homology

In the previous chapter we laid out the relevant background and theory of calculating the homology of a simplicial complex. Moving forward in this section we will explicitly compute the homology of a given complex. This computation will allow us to generalize how we will be able to algorithmically implement these calculations to any complex.

Example 4. To illustrate how we will be computing the homology over the constructed simplex of a given point cloud let us calculate the homology of the nontrivial simplicial 3-complex shown below.

Notice that there is exactly one 3-simplex, five 2-simplex, 13 1-simplex, and 11 vertices. This yields us the simplicial chain complex

$$C_{\bullet} : 0 \xrightarrow{\partial_4} \mathbb{Z}_2 \xrightarrow{\partial_3} \mathbb{Z}_2^5 \xrightarrow{\partial_2} \mathbb{Z}_2^{13} \xrightarrow{\partial_1} \mathbb{Z}_2^{11} \xrightarrow{\partial_0} 0.$$

Since our choice of \mathcal{G} is \mathbb{Z}_2 , a finite field, every ℓ -simplex in its corresponding ℓ -chain

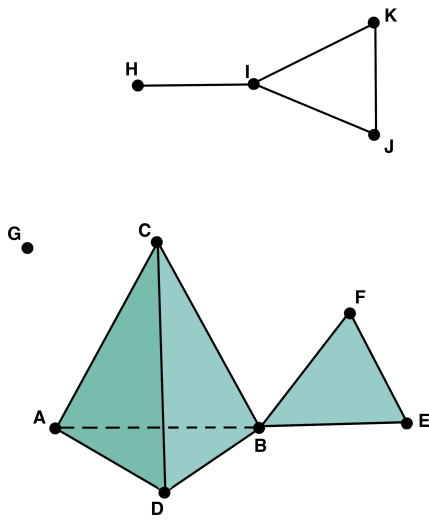


Figure 4.1: A 3-Dimensional Simplicial Complex

can be uniquely classified (up to permutation) to a standard basis element in \mathbb{Z}_2^n , where n is the number of ℓ -simplexes. This allows us to represent the boundary maps, ∂_i , as a matrix transformation from one finite basis to another. In particular, the boundary maps of our chain complex C_\bullet are given by the following:

$$\begin{array}{r}
\partial_3: \\
\begin{array}{l}
ABC \\
BCD \\
ACD \\
ABD \\
BEF
\end{array}
\end{array}
\begin{array}{c}
ABCD \\
\left[\begin{array}{c}
1 \\
1 \\
1 \\
1 \\
0
\end{array} \right]
\end{array}
\begin{array}{r}
\partial_2: \\
\begin{array}{l}
AB \\
AC \\
AD \\
BC \\
BD \\
BE \\
BF \\
CD \\
EF \\
HI \\
IJ \\
IK \\
JK
\end{array}
\end{array}
\begin{array}{c}
ABC \quad ACD \quad ABD \quad BCD \quad BEF \\
\left[\begin{array}{ccccc}
1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array} \right]
\end{array}$$

$$\partial_1: \begin{array}{c} A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \\ J \\ K \end{array} \begin{array}{cccccccccccc} AB & AC & AD & BC & BD & BE & BF & CD & EF & HI & IJ & IK & JK \\ \left[\begin{array}{cccccccccccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array} \right] .$$

As one might notice, the construction of the boundary matrices given by ∂_i corresponds to the decomposition of each ℓ -simplex into the corresponding $\ell-1$ -simplexes. To be specific, each row of the boundary map will represent a particular $\ell-1$ -simplex and each column will represent ℓ -simplex. So if $\langle s \rangle$ is a ℓ -simplex in the j -th column, $\langle k \rangle$ is an $\ell-1$ -simplex in the i -th row, and $\langle k \rangle$ is in $\partial(\langle s \rangle)$, then $\partial_{i,j} = 1$, zero otherwise.

For example, consider the matrix for ∂_2 . Notice that ABC is a 2-simplex in our simplicial complex. Since $\partial(ABC) = AB + AC + BC$, column one and rows one, two and four all have ones whereas all other rows are zero.

4.2 Smith Normal Form and The Standard Algorithm

Once the boundary maps have been constructed one might assume it suffices to reduce them into echelon form and calculate the dimension of the Rank and Null Space to find the

the Betti Numbers. However, one must be careful with calculating the homology. This is especially true since we are generally calculating the homology with coefficients in a finite field or principle integral domain (PID). Here I will outline how to reduce a matrix into this Smith Normal Form as discussed in Chapter 3.2.1, the results of which give rise to the matrix reduction algorithm known as the *Standard Algorithm* when our free abelian group is \mathbb{Z}_2 .

Recall that every relation matrix can be reduced into Smith Normal Form using the list of elementary operations as described in Remark 3.4. The only distinct difference that is apparent between simplifying this matrix via the elementary operations described in Remark 3.4 versus row reduction taught in an introductory linear algebra course is that we cannot scale a row by an arbitrary coefficient. In the case of Principle Ideal Domains (PIDs) such as \mathbb{Z} we are only allowed to scale by -1 (or trivially by 1). This is, of course, because scaling the entries of the rows or columns by other PID coefficients will change the greatest common divisors of the entries of the matrix. Thus, changing the homological calculation.

To begin, suppose that a group G describes a set of n nontrivial linear relationship on its set of m generators as described by

$$\sum_{i=1}^n \sum_{j=1}^m a_{ij} x_m.$$

With respect to homology, this relationship describes the linear decomposition of the $\ker(\partial)$ and $\text{im}(\partial)$. Of course, this linear relation can be represented as a matrix

$$A = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix}$$

Begin operating on the matrix by using the elementary column and row operations until a_{11} becomes the smallest possible integer entry. Once this has been established, reduce all

other entries in its subsequent row and column to zero to yield the matrix

$$\left[\begin{array}{c|ccc} d_1 & 0 & \dots & 0 \\ \hline 0 & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} & \dots & a_{nm} \end{array} \right].$$

If this is done correctly, $1 \leq d_1$ should divide any a_{ij} from A for $2 \leq i \leq m$ and $2 \leq j \leq n$. From there, all that is left is to simplify every diagonal entry of A until reduced into the form

$$\left[\begin{array}{cccc|ccc} d_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & d_r & 0 & \dots & 0 \\ \hline 0 & \dots & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 0 & \dots & 0 \end{array} \right].$$

Now that we have defined how to compute the Smith Normal Form of any matrix with entries from a PID, we will now define a *reduction* algorithm for computing persistent homology. This algorithm was first introduced in [ZC05]. We will not focus too much on the details of how to explicitly write the code, but rather give the general outline of the algorithm from [Les19]. To begin, let R be an $n \times m$ matrix. For $j \in \{1, \dots, n\}$, we define the *pivot* of $R_{*,j}$ (i.e. the j -th column of R) by

$$\rho_j^R := \begin{cases} \mathbf{null} & \text{if } R_{*,j} = \mathbf{0} \\ \max\{i: R(i,j) \neq 0\} & \text{otherwise} \end{cases}$$

Algorithm 1 Standard Reduction Algorithm

Require: B is an $n \times m$ **Ensure:** A reduced matrix R given by left-to-right column addition on B $R \leftarrow B$ **for** $j \in \{1, \dots, n\}$ **do** **while** $\exists k < j$ such that $\mathbf{null} \neq \rho_j^R = \rho_k^R$ **do** add $-\frac{R(\rho_j^R, j)}{R(\rho_j^R, k)} R_{*,k}$ to $R_{*,j}$ **end while****end for**

In general, one must be careful of course with the scaling coefficient $-\frac{R(\rho_j^R, j)}{R(\rho_j^R, k)}$ with regard to the PID in question. The input matrices that are relevant to us will only have entries of either 0 or 1 and so the coefficient is neglectible. With this, our algorithm is further simplified to just include left-to-right column addition according to the position of the pivots. This is often referred to as the *Standard Algorithm*.

Example 5. Consider the following matrix over \mathbb{Z}_2

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

By applying the Reduction Algorithm the following steps will occur:

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \xrightarrow[\text{col. 3}]{\text{add col. 1 to col. 3}} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \xrightarrow[\text{col. 4}]{\text{add col. 3 to col. 4}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Next we will take our reduced matrix R and normalize it. This is outlined from [ZC05].

The algorithmic implementation is as stated below:

Algorithm 2 Normalize Reduced Matrix Algorithm

Require: Reduced matrix R

Ensure: A normalized matrix N obtained via upwards row addition

$N \leftarrow R$

for $i \in \{m, \dots, 1\}$ in descending order **do**
 if \exists column j of N such that it's pivot is i **then**
 for $k \in \{i - 1, \dots, 1\}$ in descending order **do**
 if $N_{k,j} \neq 0$ **then**
 add $-\frac{N_{k,j}}{N_{i,j}}R_{i,*}$ to $R_{k,*}$
 end if
 end for
 end if
end for

Example 6. Using the reduced matrix from Example 4 we can normalize it using Algorithm 2 to yield the following:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \xrightarrow[\text{row 2}]{\text{add row 4 to row 2}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \xrightarrow[\text{row 1}]{\text{add row 4 to row 1}} \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \xrightarrow[\text{row 1}]{\text{add row 3 to row 1}} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \xrightarrow[\text{row 1}]{\text{add row 2 to row 1}} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

4.3 Filtration to Barcodes

Assume we are given an arbitrary point cloud in a finite metric space X . The goal of this section is introduce methods to define how we will find the qualitative measurements of a given point cloud regardless of small variances in the data (more broadly known as *Topological Inference*) [OPT⁺17]. In particular, suppose we are given an experimental dataset $\mathcal{K} = \{k_i\}_{i=0}^n$ which can be represented as a point cloud. This point cloud has a natural simplicial complex defined by $\check{C}ech_X(\mathcal{K}, \epsilon)$ but can also be ascribed via $Rips_X(\mathcal{K}, \epsilon)$. However we arrive at two fundamental questions: what is an appropriate choice of ϵ and why? [Ghr08] To begin, we must first examine how our simplicial complex changes as a function of ϵ .

Definition 4.1. Let \mathcal{S} be a finite simplicial complex and let $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \dots \subset \mathcal{S}_k = \mathcal{S}$ is sequence of subcomplexes called the *filtered simplicial complex*.

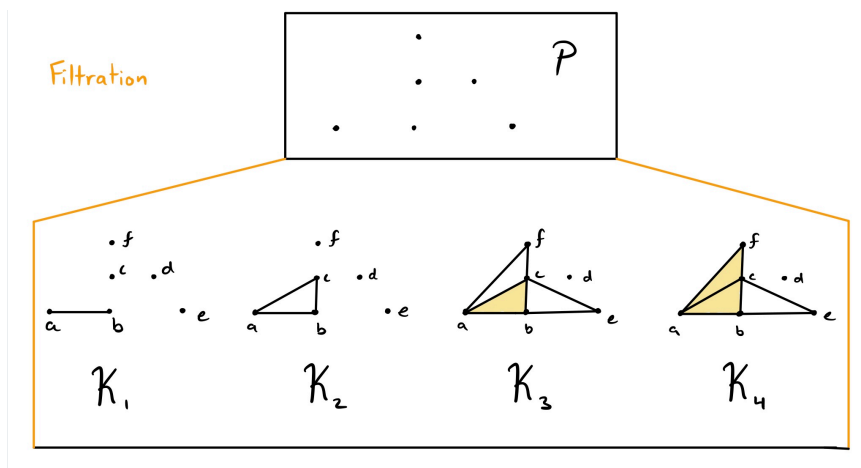


Figure 4.2: This image illustrates the filtration of a given point cloud \mathcal{P}

Remark 4.1. With respect to the Vietoris-Rips Complex, since $Rips_X(\mathcal{K}, \epsilon_1) \subseteq Rips_X(\mathcal{K}, \epsilon_2)$ given that $\epsilon_1 \leq \epsilon_2$, we can describe its filtration as

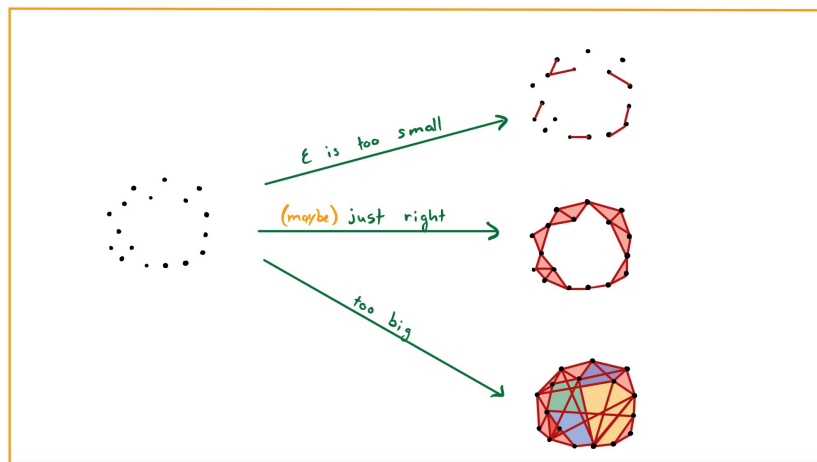
$$Rips_X(\mathcal{K}) := \{Rips_X(\mathcal{K}, \epsilon_i)\}_{i \in \mathbb{N}}.$$

The same reasoning can be applied for the Čech complex as well.

Remark 4.2. In particular, given that $\mathcal{S}_i = \check{C}ech_X(\mathcal{K}, \epsilon_i)$ varies for a given ϵ_i , there arises a natural filtered simplicial complex. Consequentially, we can apply homology to all of the subsequent subcomplexes. For all n , there exists an inclusion map $\iota: \mathcal{S}_i \rightarrow \mathcal{S}_j$ and an induced \mathbb{Z}_2 -linear map $f_{i,j}: H_n(\mathcal{S}_i) \rightarrow H_n(\mathcal{S}_j)$. It follows by functoriality that we have the following definition.

Definition 4.2. Let $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \dots \subset \mathcal{S}_k = \mathcal{S}$ be a filtered simplicial complex. The n th persistent homology of \mathcal{S} is given by the pair $(\{H_n(\mathcal{S}_i)\}, f_{i,j})$ where $i, j \in \{1, 2, \dots, k\}$ with $i \leq j$, and $f_{i,j}: H_n(\mathcal{S}_i) \rightarrow H_n(\mathcal{S}_j)$ induced by the inclusion map $\iota: \mathcal{S}_i \rightarrow \mathcal{S}_j$. In particular, it is a functor $F: \mathbf{Simp} \rightarrow \mathbf{AbGrp}$ where **Simp** is the category of filtered simplicial complexes and **AbGrp** is the category of abelian groups.

By creating a filtration on a point cloud we are able to now calculate their homology groups to study the evolution of the simplex. In essence, as ϵ changes our sequence of homology classes give us an immediate representation of how the homology of the complex is evolving. If ϵ is too small of a parameter, then the sequence of homology classes indicates to us that the structure of the data is discrete and separated whereas if ϵ is far too large of a parameter, the homology classes will indicate that everything is interconnected and trivial as illustrated below.



Of course, neither of these extremes are of use for classification. Rather, like most things, the solution is somewhere in the middle. Our goal is to illustrate the *birth* of new topological feature within the data and measure how long they will persist before they eventually *die*. This can be parameterized (by ϵ) to create a long horizontal bar for each homology class. The 'birth' of a new topological feature will begin the horizontal bar and persist for its corresponding homology class until the topological feature 'dies', in which the bar will end.

Definition 4.3. The Boundary Matrix of a simplicial complex K , denoted as $[\partial]$, is the matrix of all the boundary maps $\partial: C_*(K, \mathcal{G}) \rightarrow C_*(K, \mathcal{G})$ with the basis given by all of the simplices $\{\sigma_1, \dots, \sigma_n\}$ of K .

Remark 4.3. One should not confuse the boundary matrix to be the same as the boundary maps. Although the boundary matrices encode all the same information as the boundary maps, the boundary matrix is a square matrix that describes the decomposition of every simplex.

Example 7. Consider the triangulation of S^1 described by $\{[A], [B], [C], [A B], [A C], [B C]\}$ over the free abelian group \mathbb{Z}_2 . Then

$$C_\bullet: 0 \rightarrow C_1(S^1, \mathbb{Z}_2) \xrightarrow{\partial_1} C_2(S^1, \mathbb{Z}_2) \xrightarrow{\partial_0} 0$$

where $C_1(S^1, \mathbb{Z}_2) \cong \mathbb{Z}_2^3$ with respect to the basis $\{[A],[B],[C]\}$ and $C_2(S^1, \mathbb{Z}_2) \cong \mathbb{Z}_2^3$ with respect to the basis $\{[AB],[BC],[AC]\}$. Of course, ∂_0 is the zero map and ∂_1 can be described with respect to these ordered bases as,

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

The resulting boundary matrix is as follows:

$$\begin{array}{c}
 \\
 \\
[\partial]: \\
 \\
 \\

\end{array}
\begin{array}{c}
A \\
B \\
C \\
AB \\
AC \\
BC
\end{array}
\begin{array}{cccccc}
A & B & C & AB & AC & BC \\
\left[\begin{array}{cccccc}
0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{array} \right]
\end{array}$$

Definition 4.4. A *barcode* of \mathcal{B} is a multiset of intervals I_i with given multiplicities $m_i \in \mathbb{N}$ often denoted as $\{(I_i, m_i)\}$. All intervals of finite form are represented as $(a, b]$ whereas the infinite forms are denoted as $[a, \infty)$ [PRSZ20].

Theorem 4.1 ([ZC05]). Let R be the reduced matrix obtained from $[\partial]$. Then

$$\begin{aligned}
\mathcal{B}_i([\partial]) &= \{[j, k) : \rho_k^R = j, \dim(\sigma_j) = i\} \\
&\cup \{[j, \infty) : R_{*,j} = 0, \dim(\sigma_j) = i, \nexists k \text{ such that } \rho_k^R = j\}
\end{aligned}$$

These pictographic images are the barcodes of a filtered simplicial complex as illustrated by Figure 4.3.

Remark 4.4. Given the boundary matrix $[\partial]$ from Example 7, it follows by Theorem 4.1 that

$$\mathcal{B}_0(S^1) = \{[0, \infty), [2, 4), [3, 5)\}$$

$$\mathcal{B}_1(S^1) = \{[6, 7)\}.$$

Unfortunately, given a large enough sample of data one will notice that visualizing the persistent homology using barcodes will become quite cluttered. To avoid this we create

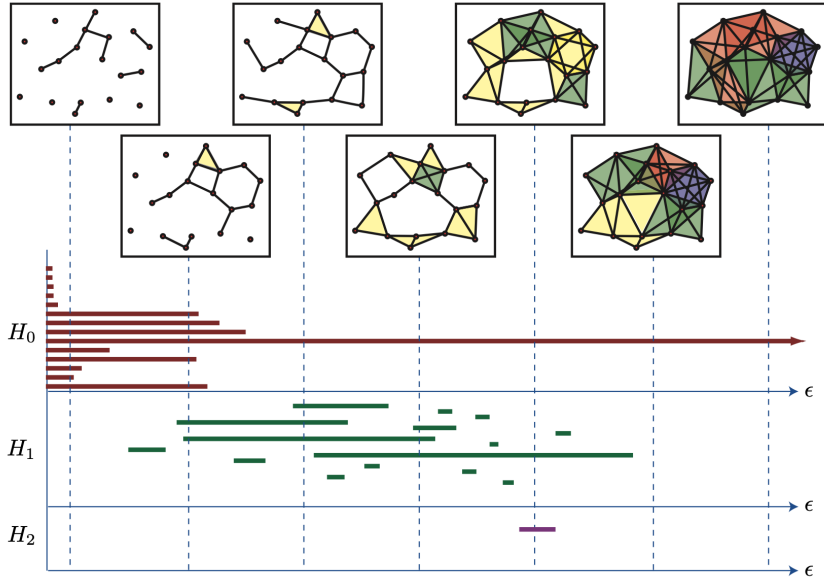


Figure 4.3: This barcode illustrates of how the homology classes are changing with respect to the filtered simplicial complex. Image taken from [Ghr08]

what are known as *Persistent Diagrams*.

Notation 1. We denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$

Definition 4.5. A *Persistent Diagram* is a multiset that is the union of a finite multiset of points in $\overline{\mathbb{R}}^2$ with the multiset of points on the diagonal $\Delta = \{(x, y) \in \mathbb{R} : x = y\}$, where each point on the diagonal has infinite multiplicity. [OPT⁺17]

Both barcodes and persistent diagrams encode the same topological information of a given filtered simplicial complex. To be specific, once a bar, $[a, b)$, has been retrieved and stored in \mathcal{B}_i , it is as simple as using a as the x -coordinate and b as the y -coordinate.

In addition, it is also a lot clearer to read and interpret the results of a persistent diagram as opposed to barcodes. For example, it is likely that in the beginning of the filtration many short lived topological features will begin and die. This is nothing more than noise in our dataset relative to the more prominent global features we are trying to identify. With a persistent diagram these short lived features we identify as topological noise can be ignored by measuring their distance to Δ .

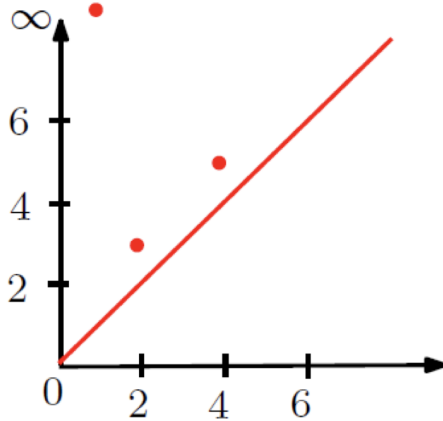


Figure 4.4: Persistent Diagram. Image taken from [FC16]

Now that we have defined how to find a persistent diagram $X_{\mathcal{K}}$ of a given dataset \mathcal{K} , it naturally follows that we would like to measure the distance between another persistent diagram $Y_{\mathcal{L}}$. However, this necessitates that the cardinality of $X_{\mathcal{K}}$ and $Y_{\mathcal{L}}$ must be the same in order to bijectively study the relationship of each point to another.

Definition 4.6. The p th Wasserstein distance between two persistent diagrams $X_{\mathcal{K}}$ and $Y_{\mathcal{L}}$ is defined as

$$W_p[d](X_{\mathcal{K}}, Y_{\mathcal{L}}) := \inf_{\psi \in \Gamma} \left[\sum_{x \in X_{\mathcal{K}}} d[x, \psi(x)]^p \right]^{1/p}$$

for $p \in [1, \infty)$ and

$$W_{\infty}[d](X_{\mathcal{K}}, Y_{\mathcal{L}}) := \inf_{\psi \in \Gamma} \left[\sup_{x \in X_{\mathcal{K}}} d[x, \psi(x)] \right]$$

for $p = \infty$, where d is a metric on \mathbb{R}^2 (usually the L_p norm) and Γ is the set of all bijective functions between $X_{\mathcal{K}}$ and $Y_{\mathcal{L}}$.

Remark 4.5. The most commonly implemented distance function used between persistent

diagrams is called the *bottleneck distance* $W_\infty[L_\infty]$. [OPT⁺17]

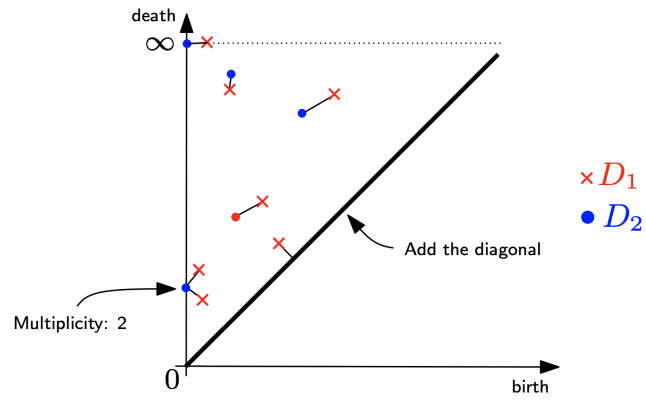
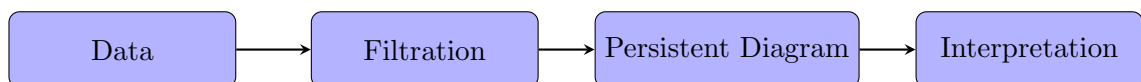


Figure 4.5: The distance between two persistent diagrams using the bottleneck distance. Image taken from [FC16]

Chapter 5: Algorithmic Implimentation

In our previous chapter we have outlined how we use homological algebra to find the persistence of global topological features in datasets. The primary focus of this chapter will be to outline the algorithm implementation of such tools so that anyone with interest in studying persistent homology can do so. Our currently pipeline for computing PH is given by the following flow chart.



During the filtration process one will need to decide on what type of simplicial structure they will want to build from a given dataset. In this thesis we have outlined two such examples; the Čech complex and Vietoris-Rips complex. However, rather than focusing on both the Čech and Rips complex we will primarily focus on developing the persistent barcodes and diagrams for the Rips complex. There are two reasons for this decision. First, there is a clear inclusion between the Rips complex and the Čech complex given a particular size ϵ given by 3.1.1. Second, the Čech Complex will require us to check the common intersection of all of $\mathbb{B}(p_i, \epsilon)$. This subtle difference between the pair-wise intersection from the Rips complex makes a large computational difference when computing their skeletons as stated in Chapter 3.

I do not claim that any of these algorithms or implementations are unique or even computationally efficient. Anyone interested in understanding how to create fast and computational efficient algorithms to compute the Vietoris-Rips Complex and its respective homology I implore you to read [Zom10] and [DI12]. Think of this section as a guide on how to create a simplicial filtration and the corresponding persistent diagram.

5.1 Deriving Complexes From Point Clouds

To begin, let us first illustrate how to create the 1-skeleton of the Rips-Complex. This is relatively straightforward since all that we need to check is whether or not $\mathbb{B}(p_i, \epsilon) \cap \mathbb{B}(p_j, \epsilon) = \emptyset$. If it is not empty then we can append an edge to the two vertices. The explicit algorithm is as shown below.

Algorithm 3 1-Skeleton Algorithm

Require: A point cloud $\mathcal{K} = \{k_i\}_{i=0}^n$ and a choice of ϵ

Ensure: $sk_1(\mathcal{K}, \epsilon)$

```

edges = []
for  $k_i \in \mathcal{K}$  do
  for  $k_j \in \mathcal{K} \setminus k_i$  do
    if  $d(k_i, k_j) < 2\epsilon$  then
      continue
    else
      edges.append( $\langle i, j \rangle$ )
    end if
  end for
end for
return edges

```

Example 8. Let's consider the following dataset $\mathcal{K} = \{k_i\}_{i=1}^{50}$ where each point $k_i \in \mathcal{K}$ is a randomly dispersed around a circle of radius $2.5 < r < 3$ centered at the origin

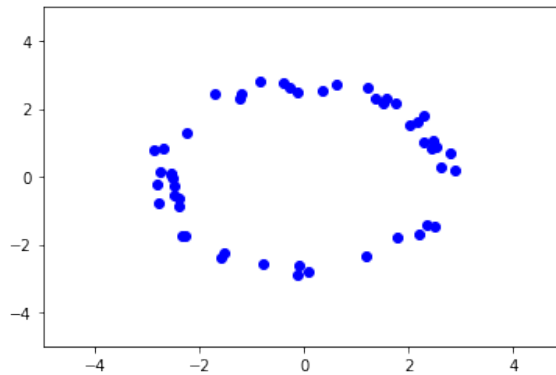


Figure 5.1: A noisy set of points around a circle of radius 3.

Then the following filtration of simplicial complexes is for $\epsilon \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$

Notation. We will denote the list Δ_k to be the set of k -simplicies. Furthermore, the list

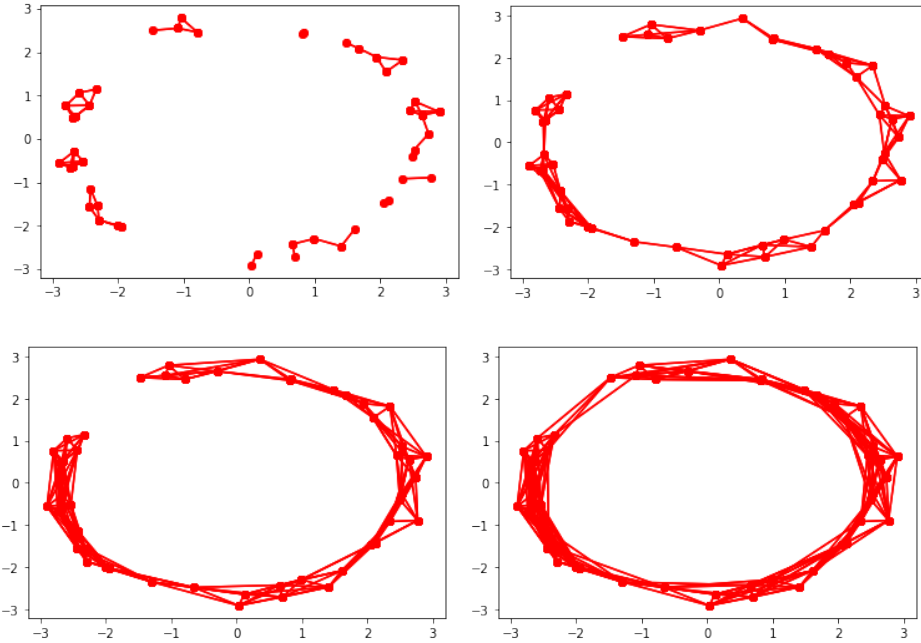


Figure 5.2: The resulting 1-skeleton.

of all simplices from our current simplicial complex will be denoted as $\Delta = \{\Delta_1, \dots, \Delta_n\}$. This list is ordered by the size of the simplex (i.e. all of the 0-simplices will be listed first, then the 1-simplices, etc.).

Now that we have created an algorithm to create the 1-skeleton the subsequent step would be to find the higher dimensional simplices to form the entirety of the Rips complex. To do this we must first break our approach down into two separate algorithms—an algorithm to compute the boundary map and an algorithm to compute the cliques. In essence, we want to be able to look at a list of simplices determined by our data and accurately assess if there are enough present that we can append this new simplex into our complex. To begin let us define the explicit algorithm for computing the boundary map of a given simplex.

Given the input, the algorithm is taking the list of simplices that are stored and iteratively removing a vertex (index value if you are writing this code in python) for each simplex. This reduces the size of the original list (the starting simplex) by one and we append this new smaller list (our boundaries) to a new list denoted Δ_{n-1} .

The purpose for including an algorithm to compute the boundary operator of a simplex is

Algorithm 4 Boundary Operator Algorithm

Require: A set of oriented n -simplex $\Delta_n = \{\sigma_n^1, \sigma_n^2, \dots, \sigma_n^m\}$

Ensure: $\Delta_{n-1} = \{\partial(\sigma_n^i)\}_{i=1}^m$

$\Delta_{n-1} = []$

for $\sigma_n^i \in \Delta_n$ **do**

for $j \in \{0, 1, \dots, n\}$ **do**

 Remove v_j from σ_n^i such $\sigma_n^i \rightarrow \sigma_{n-1}^j$

 Append σ_{n-1}^j to Δ_{n-1}

end for

end for

return Δ_{n-1}

to double check our predictions on how to build higher dimensional simplicies. For instance, suppose we find the simplicies $\langle v_1, v_2, v_3 \rangle$ and $\langle v_1, v_3, v_4 \rangle$ within the set of all 3-simplicies. These simplicies are the boundary of the 4-simplex $\langle v_1, v_2, v_3, v_4 \rangle$. However, we cannot assume this simplex exists in our complex unless all 4 of the faces are found in the set of 3-simplicies. Thus, the following algorithm is required for an accurate construction of the Rips complex.

Algorithm 5 Simplicial Construction Algorithm

Require: A set of oriented n -simplex $\Delta_n = \{\sigma_n^1, \sigma_n^2, \dots, \sigma_n^m\}$

Ensure: Δ_{n+1} constructed by finding the corresponding cliques

$\Delta_{n+1} = []$

for $\sigma_n^i \in \Delta_n$ **do**

for $\sigma_n^j \in \Delta_n$ **do**

if $\langle v_0^i, \dots, v_k^i, \dots, v_n^i \rangle = \langle v_0^j, \dots, v_s^j, \dots, v_n^j \rangle$ for $k \leq s$ **then**

 Create Simplex $\sigma_{n+1} = \langle v_0^i, \dots, v_k^i, \dots, v_s^j, \dots, v_n^i \rangle$

if $\partial(\sigma_{n+1}) \in \Delta_n$ **then**

 Append σ_{n+1} to Δ_{n+1}

end if

end if

end for

end for

It suffices to use the previous algorithms in conjunction to build the Vietoris-Rips Complex. Once a 1-skeleton has been created from a point cloud we want to recursively build higher dimensional simplicies on top of it. Thus, we outline the Rips Complex algorithm below.

Algorithm 6 k -Skeleton of the Vietoris-Rips Complex

Require: $sk_1(\mathcal{K}, \epsilon)$ and desired largest n -skeleton where $n \geq 1$

Ensure: $sk_n(\mathcal{K}, \epsilon) \subseteq Rips_{\mathbb{R}^n}(\mathcal{K}, \epsilon)$

Complex = []

if $n=1$ **then**

 Append $sk_1(\mathcal{K}, \epsilon)$ to Complex

else

 Append $sk_1(\mathcal{K}, \epsilon)$ to Complex

 Rips-Complex($n - 1$, s -builder($sk_1(\mathcal{K}, \epsilon)$))

end if

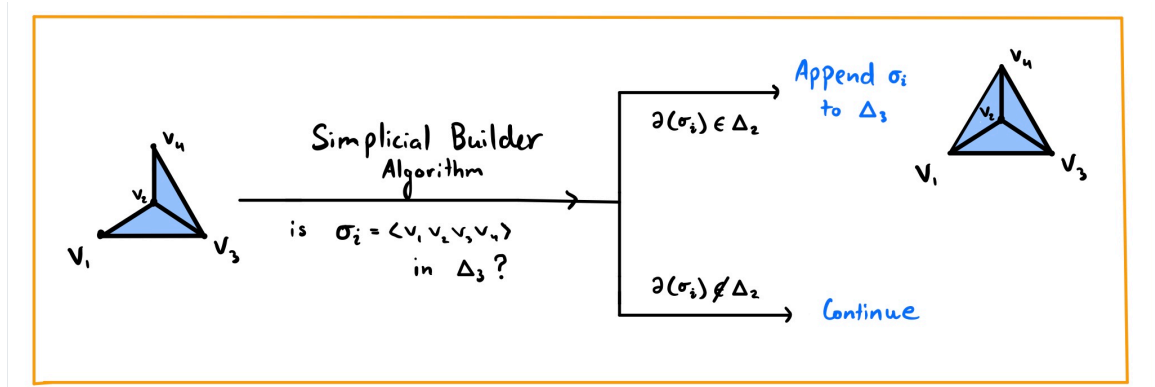


Figure 5.3: This visual diagram illustrates how the computer is being programmed to complete the simplex or ignore it.

In practice, it is often the case that we only desire the lower dimensional skeletons of the of the Rips Complex. This and the additional fact that computing higher dimensional simplicies is computationally expensive is the reason we include the ability to terminate the recursive algorithm at a desired size. If one does however want to build the entire complex it can be attained by setting $n = |\mathcal{K}|$.

The final step left in computing the homology is creating the boundary matrix. This can be achieved by initializing the zero matrix whose size is the same as the total number of simplicies generated by the complex and then altering it using the boundary operator algorithm.

Now that the boundary matrix has been achieved, apply the standard reduction algorithm and we have our homology explicitly expressed. We will not include an algorithm to

Algorithm 7 Boundary Matrix Algorithm

Require: The list of all simplices $\Delta = \{\Delta_1, \dots, \Delta_n\}$

Ensure: The Boundary Matrix $[\partial]$.

Initialize $[\partial]$ to be the zero matrix

for $k \in \{1, \dots, n\}$ **do**

for $\sigma_i \in \Delta_k$ **do**

for $\sigma_j \in \Delta_{k+1}$ **do**

if $\sigma_i \in \partial(\sigma_j)$ **then**

$[\partial]_{i,j} = 1$

else

continue

end if

end for

end for

end for

compute the barcode as it is outlined by Carlsson and Zomorodian's Theorem in Chapter

4.

Chapter 6: Multi-Parameter Filtrations and Persistent Modules

Until this point we have defined all of the ideas necessary for studying the persistence of what is known as a *1-Parameter* Persistent Homology. Unfortunately the applications of 1-parameter persistent homology are quite narrow and not robust with respect to noisy data. These various limitations point in the direction of studying what is known as *multi-parameter* persistent homology. Here are a few notable reason for the need to use multi-parameter persistent homology.

1-Parameter Homology is sufficiently useful when the point cloud in hand is clean (i.e. the variance of the data is limited). However, anyone with experience in data science is aware that datasets can be very noisy. Rather than not using the dataset, it is far more suitable to find methods to potentially de-noise the data so that it can be properly integrated into a data mining algorithm. Unfortunately it is known that 1-parameter persistent homology is not robust with noisy data.

Of course, if there was an additional parameter that could be used to filter out noise, we could control this parameter and optimize it to find an ‘optimal’ topological structure.

Another application in which multi-parameter persistent homology is useful is in the case where the dataset has instances of ‘spikes’ or ‘tendrils’ given by Figure 6.2. This was first introduced in [ZC05], one could remove a portion of data to make the tendril or spike ‘disjoint’ from the main body. The ability to do this requires that there is a second parameter to control the amount of data that should be removed from the main body.

There are various other examples that may require the need for a second parameter and possibly many more that require multiple parameters. Thus, we are left with the task of providing a more general definition of filtrations and persistent modules.

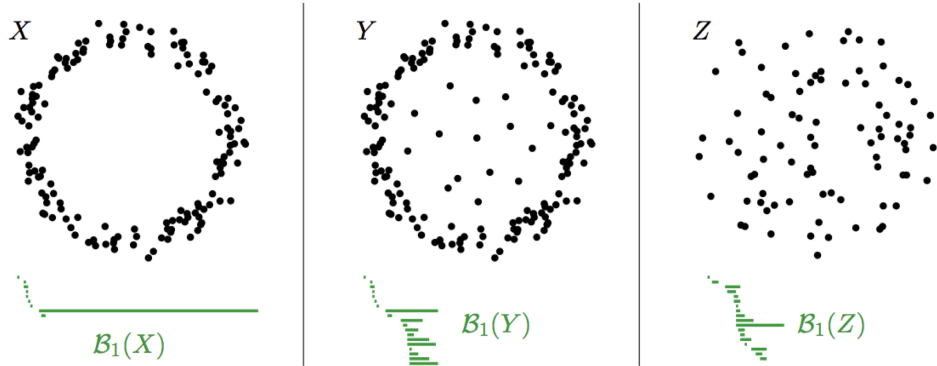


Figure 6.1: This visual illustrates how barcodes are not stable with respect to noise. Notice that a slight introduction of noise from the left most image to the middle derails the persistence in the barcodes produced. Image credits to [LW15]

6.1 Bifiltration and Bipersistence Modules

Definition 6.1 ([BL22]). Let \mathcal{J} be one of the poset of categories \mathbb{N} , \mathbb{Z} , or \mathbb{R} . Then a \mathcal{J} -indexed *filtration* is a functor $F: \mathcal{J} \rightarrow \mathbf{Top}$ such that $F_s \subset F_t$ for all $s \leq t$ in \mathcal{J} . Similarly, a \mathcal{J} -indexed *persistent module* is a functor such that $F: \mathcal{J} \rightarrow \mathbf{Vect}$.

Remark 6.1. A d -parameter filtration is a functor $F: \mathbb{N}^d \rightarrow \mathbf{Top}$ such that $F_s \subset F_t$ for all $s \leq t$ in \mathbb{N}^d (with respect to the product partial ordering on \mathbb{N}^d). In the case we are dealing with a 2-parameter filtration, often referred to as a *bifiltration*, we yield the following commutative diagram below.

$$\begin{array}{ccccccc}
 \vdots & & \vdots & & \vdots & & \\
 \uparrow & & \uparrow & & \uparrow & & \\
 \mathcal{K}_{3,1} & \hookrightarrow & \mathcal{K}_{3,2} & \hookrightarrow & \mathcal{K}_{3,3} & \hookrightarrow & \cdots \\
 \uparrow & & \uparrow & & \uparrow & & \\
 \mathcal{K}_{2,1} & \hookrightarrow & \mathcal{K}_{2,2} & \hookrightarrow & \mathcal{K}_{2,3} & \hookrightarrow & \cdots \\
 \uparrow & & \uparrow & & \uparrow & & \\
 \mathcal{K}_{1,1} & \hookrightarrow & \mathcal{K}_{1,2} & \hookrightarrow & \mathcal{K}_{1,3} & \hookrightarrow & \cdots
 \end{array}$$

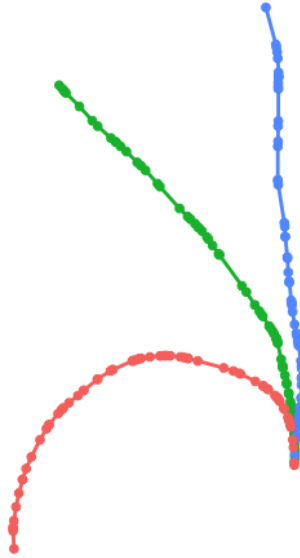


Figure 6.2: Visualization of a tendril data set. Image taken from [KW18]

Remark 6.2. Quite similarly, a *bipersistence module* is a functor $F: \mathbb{N}^2 \rightarrow \mathbf{Vect}$ with the commutative diagram

$$\begin{array}{ccccccc}
 & \vdots & & \vdots & & \vdots & \\
 & \uparrow & & \uparrow & & \uparrow & \\
 \mathcal{M}_{3,1} & \longrightarrow & \mathcal{M}_{3,2} & \longrightarrow & \mathcal{M}_{3,3} & \longrightarrow & \cdots \\
 & \uparrow & & \uparrow & & \uparrow & \\
 \mathcal{M}_{2,1} & \longrightarrow & \mathcal{M}_{2,2} & \longrightarrow & \mathcal{M}_{2,3} & \longrightarrow & \cdots \\
 & \uparrow & & \uparrow & & \uparrow & \\
 \mathcal{M}_{1,1} & \longrightarrow & \mathcal{M}_{1,2} & \longrightarrow & \mathcal{M}_{1,3} & \longrightarrow & \cdots
 \end{array}$$

Depending on the contextual setting of the problem there arises a multitude of functions that can be used to filter through the dataset. Regardless of filtration method used, all of these methods begin with the use of what are known as sublevel and superlevel filtrations.

Definition 6.2 (Sublevel/Superlevel Filtration [BL22]). Suppose \mathcal{T} is a topological space and $\gamma: \mathcal{T} \rightarrow \mathbb{R}$ (not necessarily continuous). The *Sublevel Filtration* $\mathcal{S}^\uparrow(\gamma)$ is the \mathbb{R} -indexed filtration given by

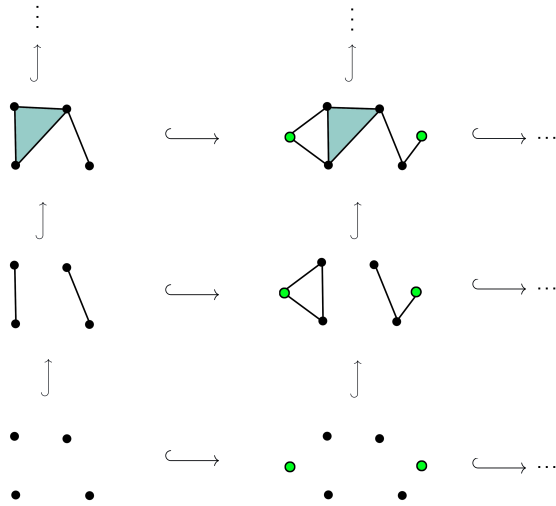


Figure 6.3: An example of a bifiltration of simplicial complexes

$$\mathcal{S}^\uparrow(\gamma) = \{p \in \mathcal{T} : \gamma(p) \leq r\}$$

and the *Superlevel Filtration* $\mathcal{S}^\downarrow(\gamma)$ is the \mathbb{R}^{op} -indexed filtration given by

$$\mathcal{S}^\downarrow(\gamma) = \{p \in \mathcal{T} : \gamma(p) \geq r\}$$

In particular, we use these definitions to define how to control another parameter within our bifiltration. Suppose that P is a finite metric space such that $\gamma: P \rightarrow \mathbb{R}$. The sublevel-Rips bifiltration $\mathcal{S}^\uparrow(\gamma): \mathbb{R}^2 \rightarrow \mathbf{Simp}$ is defined by $\mathcal{S}^\uparrow(\gamma)_{a,r} = \gamma^{-1}(-\infty, a]$ and the superlevel-Rips bifiltration, $\mathcal{S}^\downarrow(\gamma): \mathbb{R}^{\text{op}} \times \mathbb{R} \rightarrow \mathbf{Simp}$ is defined by $\mathcal{S}^\downarrow(\gamma)_{a,r} = \gamma^{-1}[a, \infty)$.

Now we are left with the choice of how to define our function γ . Both Matthew Wright and Michael Lesnick have stated in their documentation for RIVET (a multiparameter persistent homology software) that this lead us to three natural choices of γ [The20].

Example 9. The *Ball Density function* is one of the first prominently known filtration examples. It is defined by $\gamma: P \rightarrow \mathbb{R}$ such that

$$\gamma_r(x) = |\{y \in P: d(x, y) \leq r\}|.$$

This choice of γ surveys regions around the metric space P . In doing so, based on the parameter r (often called the bandwidth parameter) we get a reading on how points cluster around each other. If the returned value is high for a particular value of x we assume that the density is high and statistically significant. In the case that the returned value is low, we assume that the data is sparse and likely represents noise that might be removed.

Example 10. Another suitable choice is the *Gaussian density function*,

$$\gamma_\sigma(x) = \alpha \sum_{y \in P} \exp \left[-\frac{d(x, y)^2}{2\sigma} \right]$$

where α is some normalizing constant.

Example 11. Another suitable choice is the *Gaussian Density function* given by

$$\gamma_\sigma(x) = \alpha \sum_{y \in P} \exp \left[-\frac{d(x, y)^2}{2\sigma} \right]$$

where α is some normalizing constant. With respect to image processing, this density function provides a robust measurement for how topological features present in an image blur. Over time if the image loses sharpness it will inevitably also lose its topological features.

Example 12. Lastly, consider the *Eccentricity function* given by

$$\gamma(x) = \frac{1}{|P|} \sum_{y \in P} d(x, y).$$

This function determines the average distance between a point x and the rest of the dataset P . In the case of a dataset that has similar features to that of a tendril or spike, by

isolating the points whose average distance is the largest we are able to filter out the central core. As a byproduct our dataset has now changed so that the branched components of the central object is now separated into multiple components.

6.2 No Good Barcodes

One might naturally assume that given a bifiltration and a bipersistent module there must also be an equivalent 2 dimensional barcode that could be represented via bounded rectangular regions—maybe even generalized further for multiparameter filtrations and persistent modules. However this is not necessarily true. Let us begin by formalizing the notion of a ”good” barcode.

Definition 6.3. A *good* barcode for an \mathbb{N}^2 indexed persistent module \mathcal{M} is a collection $\mathcal{B}_{\mathcal{M}}$ of subsets in \mathbb{R}^2 such that for each $a \leq b$ in \mathbb{R}^2

$$\text{Rank}\mathcal{M}_{a,b} = |\{S \in \mathcal{B}_{\mathcal{M}}: a, b \in S\}|$$

Unfortunately, there are many examples of multiparameter persistent modules that have no good barcodes. We will not go further into discussing why this is the case as it requires a deeper look into quiver representation theory [Oud17]. But it is worth noting that as a result of not having a well defined measurement of barcodes in generalized persistent theory requires the development and research of new qualitative summaries are required.

One might even further ask that such summaries can be embedded into a Banach or a Hilbert space so that they can be used in conjunction with standard data mining techniques [B⁺15]. This is the current area of research that is most notably being studied. These methods include the use of Hilbert functions, fiber barcodes, and graded betti numbers [LW15].

Bibliography

Bibliography

- [Axl97] Sheldon Axler, *Linear algebra done right*, Springer Science & Business Media, 1997.
- [B⁺15] Peter Bubenik et al., *Statistical topological data analysis using persistence landscapes.*, J. Mach. Learn. Res. **16** (2015), no. 1, 77–102.
- [BL22] Magnus Bakke Botnan and Michael Lesnick, *An introduction to multiparameter persistence*, arXiv preprint arXiv:2203.14289 (2022).
- [DI12] Stefan Dantchev and Ioannis Ivrissimtzis, *Efficient construction of the čech complex*, Computers & Graphics **36** (2012), no. 6, 708–713.
- [FC16] Bertrand Michel Frederic Chazal, *Persistent homology in tda*, June 2016.
- [Ghr08] Robert Ghrist, *Barcodes: the persistent topology of data*, Bulletin of the American Mathematical Society **45** (2008), no. 1, 61–75.
- [Hat02] Allen Hatcher, *Algebraic topology*, Cambridge University Press, 2002.
- [KW18] Martin Karpfors and James Weatherall, *The tendril plot—a novel visual summary of the incidence, significance and temporal aspects of adverse events in clinical trials*, Journal of the American Medical Informatics Association **25** (2018), no. 8, 1069–1073.
- [Les19] Michael Lesnick, *Lecture notes for math 840: Multiparameter persistence*, Springer, 2019.
- [LW15] Michael Lesnick and Matthew Wright, *Interactive visualization of 2-d persistence modules*, arXiv preprint arXiv:1512.00180 (2015).
- [Mun00] James R Munkres, *Topology*, vol. 2, Prentice hall Upper Saddle River, 2000.
- [OPT⁺17] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington, *A roadmap for the computation of persistent homology*, EPJ Data Science **6** (2017), 1–38.
- [Oud17] Steve Y Oudot, *Persistence theory: from quiver representations to data analysis*, vol. 209, American Mathematical Soc., 2017.
- [PRSZ20] Leonid Polterovich, Daniel Rosen, Karina Samvelyan, and Jun Zhang, *Topological persistence in geometry and analysis*, vol. 74, American Mathematical Soc., 2020.

- [Rie17] Emily Riehl, *Category theory in context*, Courier Dover Publications, 2017.
- [ST15] Isadore Manuel Singer and John A Thorpe, *Lecture notes on elementary topology and geometry*, Springer, 2015.
- [The20] The RIVET Developers, *Rivet*, 2020.
- [ZC05] Afra Zomorodian and Gunnar Carlsson, *Computing persistent homology*, Discrete & Computational Geometry **33** (2005), no. 2, 249–274.
- [Zom10] Afra Zomorodian, *Fast construction of the vietoris-rips complex*, Computers & Graphics **34** (2010), no. 3, 263–271.

Curriculum Vitae

Shrunal Pothagoni is a senior currently attending George Mason University. He will be graduating in Spring of 2022 where he will receive his Bachelor of Science in Mathematics. He has worked on numerous research projects in commutative algebra, graph theory, and mathematical data science. Shrunal will continue studying topics in mathematical data this summer through his research internship at the Center for Mathematics and Artificial Intelligence. His current plan is to continue his education in mathematics as a PhD student this coming Fall 2022 at George Mason University.