UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Homology and Persistent Homology

Autor: Fritz Pere Nobbe Fisas

Director: Dr. Francisco Belchí

Realitzat a: Department of Mathematics and Computer Science

Barcelona, 19 de gener de 2020

Contents

Abstract

Extracting information from data sets that are high-dimensional, incomplete and noisy is generally challenging. The aim of this work is to explain a homology theory for data sets, called *Persistent Homology*, and the topology and algebra behind it. Moreover, we will show different ways to represent it and finally computing some examples with the help of the GUDHI software for Python.

2010 Mathematics Subject Classification. 55NXX, 55N35, 55-04, 62-07

Chapter 1

Introduction

Motivation

The motivation of this work comes from the many real world applications that persistent homology has given, merging, what I thought it was the most abstract field of mathematics, topology, with data analysis.

Persistent Homology has been successfully applied to many fields, such as medicine [7, 10], chemistry [8, 9] or others like, detecting adversaries in artificial neural networks [4].

Using algebraic topology and other mathematical tools, persistent homology allows us to study the "shape" of data in a mathematical rigorous way and extract more information than never before since it is a relatively new field.

Background

There are two fundamental results that gave to persistent homology the robustness needed to be a rigorous mathematical theory with real world applications.

The first one is the structural theorem [12], that explains how the homology of a sequence of topological spaces can be interpreted as a barcode or a persistence diagram.

The second one is the stability theorem [5], which shows that small perturbations in the data only produce at most small perturbations in the corresponding barcode or diagram.

This two theorems provide robustness against noise and justify the use of barcodes and persistent diagrams. It also gives a formal way to estimate the homology groups of the underlying subspace of a discrete set of points giving a large enough sample [5].

Work Development

The aim of this work, is to understand how and why persistent homology works, giving first, the intuitive idea from the algebraic and topological tools needed and then giving a formal definition of them.

First of all there was a good amount of documentation, reading different applications of persistent homology with the help of my tutor, which would finally take me to choose persistent homology as a field in which I wanted to know more.

Then it was the time to extend my topology knowledge, thanks to [1, 2], and get the intuitive geometrical idea behind homology groups, to later on understand the more formal and algebraic definitions.

The next step was entering in the field of persistent homology, which was more complicated, since there was a lack of books and information due to the novelty of this theory. Thanks to my tutor who provided me many papers from which I could learn and come up with even more questions at the same time.

Then, I learned how LaTeX worked in order to write all my new knowledge down, and also using the softwares Google SketchUp and Inkscape to create most of the images that appear.

Finally, when my work seemed finished I was not quite satisfied since I had not the chance to compute persistent homology from any data set by myself. Fortunately, I could get a computer with Linux OS, and with more efforts and tries than expected I finally could install the software that I needed, GUDHI (Geometry Understanding in Higher Dimensions), and compute the persistent homology of a few point clouds.

Structure and Organization of This Work

This report is organised as follows. In chapter 2, a basic introduction to algebraic topology is given and the fundamental group is formally introduced. Chapter 3 has two sections, in the first one, the intuitive idea behind homology groups is given together with two practical examples of computing these. The second section defines simplicial homology and gives the formal algebraic definition of some tools used in the previous section. Chapter 4 explains persistence, and ways to go from data to topology to then make use of the framework set in the previous chapters to go a step beyond and explain persistent homology, its structure, stability and representation, to finally show some examples of persistent homology groups, computed with the help of [14], of two different point samples. Finally, chapter 5 summarizes the results of this work and it sketches possible further work.

Chapter 2

Algebraic Topology

Intuitively, we can say that Algebraic Topology is the study of shapes and properties of topological spaces independent of continuous deformations through algebra. It studies what remains constant when we continuously deform shapes.

That means that algebraic topology is the study of techniques for forming algebraic images of topological spaces.

Before explaining how Algebraic Topology studies this topological properties we first need to explain two concepts. Categories and Functors.

Definition 2.1. *A **Category** is a collection of objects that are linked by arrows. A category has two basic properties: the ability to compose the arrows associatively and the existence of an identity arrow for each object.*

Definition 2.2. *A **Functor** is a map between categories. Functors were first considered in algebraic topology, where algebraic objects (such as the fundamental group) are associated to topological spaces, and maps between these algebraic objects are associated to continuous maps between spaces.*

So, we can define one Category as different Topological Spaces and maps between them (like arrows), and another Category as different Algebraic Images and morphisms between them. Our functor will be a map between these two categories.

That is because functors have the characteristic feature that they form images not only of spaces but also of maps. Thus, continuous maps between spaces are projected onto homeomorphism between their algebraic images. Meaning that topologically related spaces have algebraically related images.

That means that with suitably constructed functors we may be able to form images with enough detail to reconstruct accurately the shapes of a large number of classes of topological spaces, which can be very interesting for us to study.

2.1 The Fundemental Group

One of the simplest and most important functors of algebraic topology is the fundamental group, which creates an algebraic image of a space out of the loops that it contains. In other words, it creates an algebraic image out of the paths from the space that start and end at the same point.

To define the fundamental group in a rigorous way we must remember a few things about Homotopies, since the fundamental group is the first group of homotopy.

Definition 2.3. *Let X and Y be Topological Spaces and $f_0, f_1 : X \to Y$ continuous maps. We will say that f_0 and f_1 are **homotopic** if a continuous map $F : X \times [0,1] \to Y$ such that $F(x,0) = f_0(x)$ and $F(x,1) = f_1(x)$ for all $x \in X$ exists.*

Note: if f_0 and f_1 are two homotopic maps, we will write $f_0 \sim f_1$, and F will be a homotopy between f_0 and f_1.

Definition 2.4. *Let X and Y be topological spaces, $A \subseteq X$ and $f_0, f_1 : X \to Y$ continuous maps such that $f_0(a) = f_1(a)$ for any $a \in A$. We will say that f_0 and f_1 are **relatively homotopic** in A if a map $F : X \times [0,1] \to Y$ such that $F(x,0) = f_0(x)$, $F(x,1) = f_1(x)$ and for any $a \in A$, $F(a,t) = f_0(a) = f_1(a)$ for $t \in [0,1]$. And we denote it as $f_0 \underset{A}{\sim} f_1$.*

Definition 2.5. *Let X be a Topological space. We will say that two paths $\gamma_0, \gamma_1 : [0,1] \to X$ such that $\gamma_0(0) = \gamma_1(0)$ and$\gamma_0(1) = \gamma_1(1)$ are **equivalent** if $\gamma_0 \underset{\{0,1\}}{\sim} \gamma_1$ and we will denote it as $\gamma_0 \sim \gamma_1$.*

More explicitly we have that $\gamma_0 \sim \gamma_1$ if there exists a continuous map $F : [0,1] \times [0,1] \to X$ such that $F(s,0) = \gamma_0(s)$, $F(s,1) = \gamma_1(s)$ and $F(0,t) = \gamma_0(0) = \gamma_1(0)$ and $F(1,t) = \gamma_0(1) = \gamma_1(1)$.

Note: $\sim$ is a equivalent relation both in functions and in paths.
Now we can define the fundamental group.

Definition 2.6. *Let X be a Topological Space, and $x \in X$ we denote*

$$C_x = \{\alpha : [0,1] \to X \text{ paths such that } \alpha(0) = \alpha(1) = x\}$$

*Now, let us consider in C_x the relationship of homotopies between paths. Then we can consider $C_x / \sim$, which we denote as $\Pi_1(X,x)$. This is the **Fundamental Group**.*

Theorem 2.7. *Let X be a topological space and $x \in X$, then $\Pi_1(X,x)$ is a group.*

Proof. Though to that all paths start and end in x and knowing the path is properties it is clear that for each two elements of $\Pi_1(X, x)$ you can define their product.

Note: Proving associativity needs you to realize that when concatenating paths, the "speed" at which you go through each of the paths does not change the homotopy type. □

Corollary 2.8. *If X is a path-connected topological space, then, for any two points $x, y \in X$ we have $\Pi_1(X, x) \cong \Pi_1(X, y)$.*

So, in a more informal way we can say that the elements of the fundamental group $\Pi_1(X, x)$ are all the not homotopic paths in X that start and end in x.

For example: In an annulus you have two kinds of paths that are not homotopic. Trivial loops that are like an identity function and other ones that surround the hole in the annulus, meaning that there are not homotopic to the trivial loops that we mentioned before. You can surround the hole as many times as you want, and we can define n as the number of times you surround the hole in one direction and $-n$ in the other one. So it is easy to see that the fundamental group of any point in an annulus (because an annulus is a path-connected space) is isomorphic to $\mathbb{Z}$. The fundamental group of S^2 is the trivial group $\{0\}$, since there will always exist a homotopy F between two loops in the continuous surface of a S^2, meaning that we only have one type of loop up to homotopy.

Definition 2.9. *A topological space is called **contractible** if it is homotopocaly equivalent to a point.*

Corollary 2.10. *If X is a contractible space and $x_0 \in X$ then $\Pi_1(X, x_0) = \{0\}$.*

Note: If X is a topological space that satisfies $\Pi_1(X, x_0) = \{0\}$ does not mean that X is contractible.

You can intuitively compute the fundamental group of some spaces like a torus (which has a fundamental group isomorphic to $\mathbb{Z} \times \mathbb{Z}$ like the fundamental group of $S^1 \times S^1$). But, it is certainly difficult to imagine the fundamental group of some other spaces, therefore it exists a very useful tool called the Seifert VanKampen theorem, which allows us to compute the fundamental group of a topological space X using the fundamental groups of open sub-spaces of X.

Definition 2.11. *A topological space X is **simply connected** if and only if it is path-connected and $\Pi_1(X, x_0) = \{0\}$ for one (therefore, for all) $x \in X$.*

This way, we can intuitively say that simply connected means that the space has no holes. Which will be interesting due to the content of the next chapter.

Chapter 3

Simplicial Homology

The fundamental Group is especially good for low-dimensional information (because it is all about loops). Therefore there are higher-dimensional analogues called homotopy groups $\Pi_n(X)$ to study higher-dimensional loops.

Example: $\Pi_2(S^2)$ is considering a two dimensional loop (a loop of loops), in a sphere S^2 (figure 3.1), and we see that $\Pi_2(S^2) = \mathbb{Z}$ owing to that the loop of loops surrounds the whole sphere and it only matters the number of times the sphere it surround and in which direction it does.

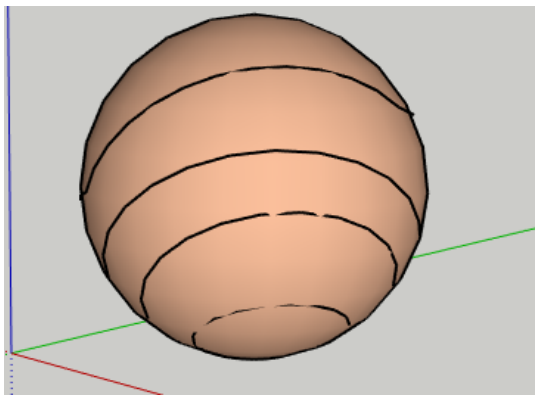

Figure 3.1: A 2-dimensional loop in a sphere.

And we can study these homotopy groups, but these are complicated and difficult to compute. And, even more, sometimes they can come out a little bit bizarre, for example, $\Pi_n(S^k)$ groups can be non trivial even when $n > k$, which is somehow measuring n-dimensional holes in k-dimensional spheres.

So what we want is an alternative to homotopy groups that is easier to compute. Those are the homology groups.

Definition 3.1. *Homology is a commutative alternative to homotopy. For a topological*

space X we have a functor to **homology groups** $H_n(X;\mathbb{K})$ *which are all commutative. Where* $\mathbb{K}$ *is a group that determine the coefficients of* $H_n(X;\mathbb{K})$*.*

The most common group to use is $\mathbb{Z}$, which makes from $H_n(X;\mathbb{Z})$ a $\mathbb{Z}$-module.

The homology group is the quotient of the group generated by their cycles and the group generated by their boundaries.

$$H_n(X;\mathbb{Z}) = \frac{< n - \text{dim cycles} >}{< n - \text{dim boundaries} >} \tag{3.1}$$

3.1 Intuitive Idea of Homology Groups

Next we give an intuitive idea of what *cycles* and *boundaries* are. In the next chapter we will go more in detail in this formula, and formally define the concepts *cycle* and *boundary*.

That is, as a result of the commutative properties of the homology group. A cycle it is similar to a loop, but it has not a starting point. If you look at the figure 3.2 of a topological space X, one can observe the loop $a + c + b$ which starts and ends at x, the loop $c + b + a$ which start and ends at y, and the loop $b + a + c$ which start and ends at t. But since we want $H_n(X;\mathbb{Z})$ to be commutative, $a + b + c = b + c + a = c + a + b$, which leads as to think that the starting and ending point does not really matter.

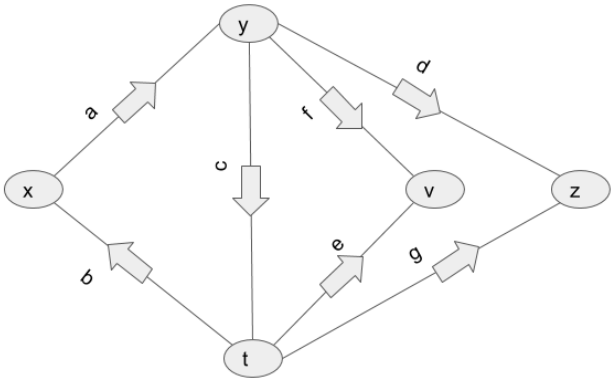

Figure 3.2: Directed Cell Complex.

Regarding the boundaries, a topological manifold with boundary M is a Hausdorff space in which every point has a neighborhood homeomorphic to an open subset of Euclidean half-space (for a fixed n): $\mathbb{R}^n_+ = \{(x_1, x_2, ..., x_n) \in \mathbb{R}^n : x_n \geq 0\}$.

We will call interior of M, the subspace formed by all the points $p \in M$ that have a neighborhood homeomorphic to $\mathbb{R}^n$. And we will call boundary of M, to the complement subspace of the interior of M (i.e. $M \setminus Int(M)$). The boundary of M will be an $(n-1)$-dimensional topological manifold.

Our next step, will be explaining how to compute the first homology group of the 1-dimensional skeleton (1-skeleton) of a Cell Complex (which until later explanation we can suppose it is like a graph) with the purpose to get the intuitive idea behind the first homology group and to get a little bit of contact with this machinery. But first we need to understand a couple of ideas.

3.1.1 Previous ideas we need to understand

A Cell Complex is a topological space formed by Cells of different dimensions. In order to describe a cell complex, one simply lists its cells, a n-dimensional Cell is homeomorphic to a n-dimensional closed ball, and how the boundary of those cells get attached to the rest of the complex.

In order to compute the first homology group, first of all we have to compute the 1-dimensional cycles (see equation 3.1).

Note: Remember that a principal ideal domain, or PID, is an integral domain in which every ideal is principal, meaning that it can be generated by a single element.

To compute an independent set of 1-dimensional cycles that generate all of them you need an oriented Cell Complex like in Figure 3.2. That means that, $a + c + b$ is a cycle, but $a + c - b$ is not one. Once we have the group generated by the cycles we will see that this group will be isomorphic to and n-dimensional abelian group, and since we normally work with $\mathbb{Z}$, our n-dimensional abelian group will be homeomorphic to $\mathbb{Z}^n$, or $\mathbb{Z}_2^n$, or more complex ones like $\mathbb{Z}^3 \bigoplus \mathbb{Z}_2 \bigoplus \mathbb{Z}_7^2 \bigoplus \mathbb{Z}_{13}^{n-6}$. Due to the fact that, if $\mathbb{K}$ is a PID and $\mathbb{M}$ a finitely generated $\mathbb{K}$-module, then there exists $r \geq 0$ and elements $d_1, d_2, \ldots, d_s \in \mathbb{K}$ such that:

$$\mathbb{M} \cong \mathbb{K}^r \oplus \mathbb{K}_{d_1} \oplus \cdots \oplus \mathbb{K}_{d_s}$$

Note: Remember that a principal ideal domain, or PID, is an integral domain in which every ideal is principal, meaning that it can be generated by a single element.

In order to avoid such types of groups that are not necessarily free , and to simplify later on computations that we want to realize with help of computers, we

will be working with $\mathbb{Z}_2$. And we will denote the homology group of dimension n as $H_n(X;\mathbb{Z}_2)$.

If we are working in $\mathbb{Z}_2$ we can see that no orientation is needed since $a = -a$. Moreover, $\mathbb{Z}_2$ is a field, which means that $H_n(X;\mathbb{Z}_2)$ is a $\mathbb{Z}_2$-vector field, which leads to an important simplification of the computation, owing to the fact that there is only one $\mathbb{Z}_2$-vector field of dimension n, which is $\mathbb{Z}_2^n$. Besides these two reasons, in later chapters we will see, that to compute persistent homology one needs a field of coefficients and it is usually the case that this field is used when an algorithm is implemented to run on a computer in order to compute homology groups.

That been said, we will for now on ignore the orientation of Simplicial and Cell Complexes. That means that we will now work with the representation in Figure 3.3 of our Cell Complex.

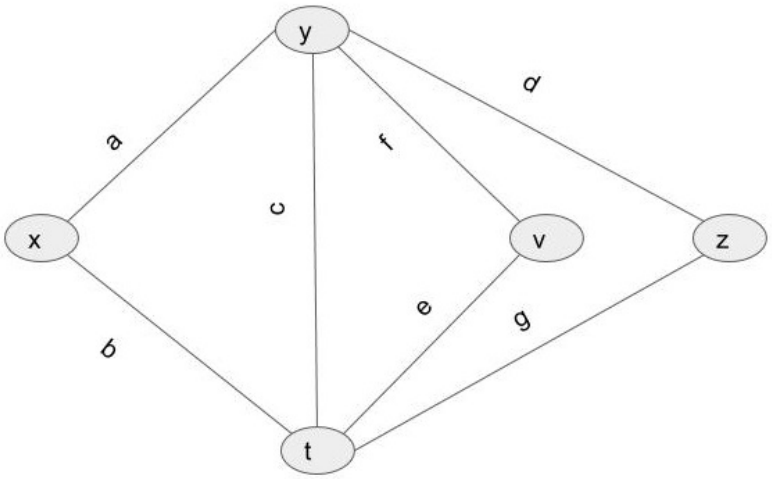

Figure 3.3: undirected Cell Complex.

Another idea that we need to understand are the Chain-Groups C_n and a specific morphism between them called the boundary operator.

Definition 3.2. *Let X be a Cell Complex. We will define the n-***Chain group** *as the the abelian group generated by the cells of dimension n. And we denote it as C_n.*

Meaning that, if we have a Cell Complex CW_A like in figure 3.3:

The C_0 would be the group generated by the vertices x, y, z, t, v And C_1 would be the group generated by the edges a, b, c, d, e, f, g.

And if we have a different Cell Complex CW_B like in figure 3.4. Where two 2-cells M, N have been attached along the cycle $a + b + c$.

Note: M and N are two-cells, meaning that there is an empty cavity between them.

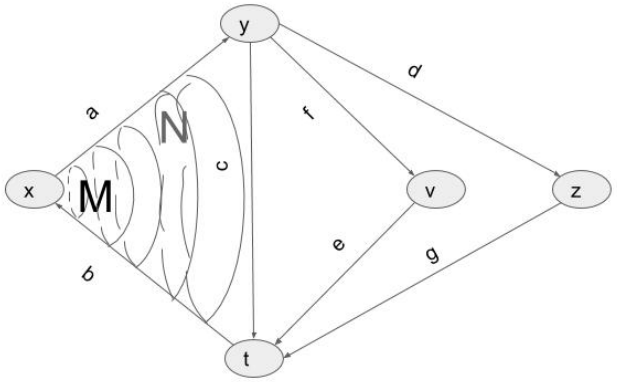

Figure 3.4: Another Cell Complex with two added 2-dimensional Cells. Here M and N are two-cells, meaning that there is an empty cavity between them.

We can see that C_0 and C_1 are the same but now we have C_2 generated by M and N.

Now we will define a morhpism between different chain groups called boundary operator, that sends every element to its boundary. And we will denote it with ∂_n. And we define it as $\partial_n : C_n \to C_{n-1}$.

So, for example, $\partial_1 : C_1 \to C_0$ will send each combination of edges to a combination of vertices.

In the next chapter, we will define this morphism algebraically in a more formal way, but for now let us geometrically see it this way.

Example 3.3. $\partial_1(a) = x + y$ (the signs of x and y would depend on the direction of the edge, but due to the fact that we are working on $\mathbb{Z}_2$ and x=-x that will be irrelevant).

If we apply this function to all the generators from C_1 we have a set of generators for $\partial_1(C_1)$ and we see that it is a morphism. Moreover, since we are working on $\mathbb{Z}_2$, which is a field, C_n is a $\mathbb{Z}_2$-vector field for each n. And the boundary operator is a linear map.

Other examples: $\partial_1(f) = y + v, \partial_1(c + f) = t + v, \partial_2(M) = a + b + c$, $\partial_0(z) = 0$.

Intuitively we can think that a 1-dimensional cycle will have a boundary equal to zero in C_0, owing to that there are no vertex that delimit the end of a 1-dimensional cycle. If we check this idea with all the 1-dimensional cycles of A we see that $\partial(a + b + c) = 0, \partial(a + b + e + f) = 0$. ($\partial$ is a linear map, that means that $\partial(a + b + c) = \partial(a) + \partial(b) + \partial(c) = y + x + x + t + t + t = 2x + 2t + 2y = 0$ since we are in $\mathbb{Z}_2$. The same goes for $\partial(a + b + e + f) = 0$.)

So an element $c \in C_1$ will be a cycle iff (if and only if) $\partial_1(c) = 0$. Which implies that $<$ 1-dim cycles $>= \text{Ker}(\partial_1)$. The same happens for every dimension. If we look at CW_B on figure 3.4 we can see that $\partial_2(M + N) = \partial(M) + \partial(N) = (a + b + c) + (a + b + c) = 2a + 2b + 2c = 0$. So we can say that $<$ n-dim cycles$>= ker(\partial_n)$.

3.1.2 Computing homology groups

So let us compute the first homology group of CW_A. $H_1(CW_A; \mathbb{Z}_2) = \frac{<1-\text{dim cycles}>}{<1-\text{dim boundaries}>}$. Which will be equal to $< 1 - \text{dim cycles} >= \text{Ker}(\partial_1)$ due to it has no 1-dimensional boundaries. (remember that, with our geometric intuition so far, a 1-dimensional boundary has to be a 2-dimensional manifold minus its interior, and in A there are not any 2-dimensional cells).

We want to compute the *Ker* of $\partial_1 : C_1 \to C_0$, and we know that any element of C_1 is a lineal combination of a, b, c, d, e, f, g so for $\forall c \in C_1$ there $\exists \lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6 \in \mathbb{Z}_2$ such that $c = \lambda_0 \cdot a + \lambda_1 \cdot b + \lambda_2 \cdot c + \lambda_3 \cdot d + \lambda_4 \cdot e + \lambda_5 \cdot f + \lambda_6 \cdot g$ (Remember we are working on $\mathbb{Z}_2$ so λ_i only can be equal to 0 and 1 for $i = 1, 2, 3, 4, 5, 6$) Meaning that:

$$\partial_1(c) = \partial(\lambda_0 \cdot a + \lambda_1 \cdot b + \lambda_2 \cdot c + \lambda_3 \cdot d + \lambda_4 \cdot e + \lambda_5 \cdot f + \lambda_6 \cdot g)$$

And, due to the fact that ∂_1 is a linear map:

$$\partial_1(\lambda_0 \cdot a + \lambda_1 \cdot b + \lambda_2 \cdot c + \lambda_3 \cdot d + \lambda_4 \cdot e + \lambda_5 \cdot f + \lambda_6 \cdot g) =$$

$$= \lambda_0 \cdot \partial_1(a) + \lambda_1 \cdot \partial_1(b) + \lambda_2 \cdot \partial_1(c) + \lambda_3 \cdot \partial_1(d) + \lambda_4 \cdot \partial_1(e) + \lambda_5 \cdot \partial_1(f) + \lambda_6 \cdot \partial_1(g)$$

And, thanks to the image 3.3 we see that:

$$\partial_1(a) = x + y$$

$$\partial_1(b) = x + t$$

$$\partial_1(c) = t + y$$

$$\partial_1(d) = z + y$$

$$\partial_1(e) = v + t$$

$$\partial_1(f) = y + v$$

$$\partial_1(f) = z + t$$

which leads us to the equality:

$$\lambda_0 \cdot \partial_1(a) + \lambda_1 \cdot \partial_1(b) + \lambda_2 \cdot \partial_1(c) + \lambda_3 \cdot \partial_1(d) + \lambda_4 \cdot \partial_1(e) + \lambda_5 \cdot \partial_1(f) + \lambda_6 \cdot \partial_1(g) =$$

$$= \lambda_0 \cdot (x+y) + \lambda_1 \cdot (x+t) + \lambda_2 \cdot (t+y) + \lambda_3 \cdot (z+y) + \lambda_4 \cdot (v+t) + \lambda_5 \cdot (v+y) + \lambda_6 \cdot (z+t) =$$

$$= x \cdot (\lambda_0 + \lambda_1) + y \cdot (\lambda_0 + \lambda_2 + \lambda_3 + \lambda_5) + z \cdot (\lambda_3 + \lambda_6) + t \cdot (\lambda_1 + \lambda_2 + \lambda_4 + \lambda_6) + v \cdot (\lambda_4 + \lambda_5)$$

Which leads us to the conclusion:

$$c \in \text{Ker}(\partial_1) \Leftrightarrow \partial_1(c) = 0 \Leftrightarrow x \cdot (\lambda_0 + \lambda_1) + y \cdot (\lambda_0 + \lambda_2 \lambda_3 + \lambda_5) + z \cdot (\lambda_3 + \lambda_6) + t \cdot (\lambda_1 + \lambda_2 + \lambda_4 + \lambda_6) + v \cdot$$

$$\Leftrightarrow \begin{cases} \lambda_0 + \lambda_1 = 0 \\ \lambda_0 + \lambda_2 + \lambda_3 + \lambda_5 = 0 \\ \lambda_3 + \lambda_6 = 0 \\ \lambda_1 + \lambda_2 + \lambda_4 + \lambda_6 = 0 \\ \lambda_4 + \lambda_5 = 0 \end{cases}$$

Now, we could solve this equation system, but we will transform it in his matrix form in order to see how we would handle this computation with code.

Notice that we can transform this equation system in a matrix only because of the fact that we work in $\mathbb{Z}_2$ which is a field.

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \underset{\text{reduced form}}{\sim} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Which gives us following reduced equation system:

$$\begin{cases} \lambda_0 + \lambda_1 = 0 \\ \lambda_1 + \lambda_2 + \lambda_3 + \lambda_5 = 0 \\ \lambda_3 + \lambda_6 = 0 \\ \lambda_4 + \lambda_5 = 0 \end{cases}$$

Since we have a system with 4 equations and 7 variables we will set λ_1, λ_5 and λ_6 as the parameters α, β, γ respectively.

So the solution will be:

$$\begin{cases} \lambda_0 = \alpha \\ \lambda_1 = \alpha \\ \lambda_2 = \alpha + \beta + \gamma \\ \lambda_3 = \gamma \\ \lambda_4 = \beta \\ \lambda_5 = \beta \\ \lambda_6 = \gamma \end{cases}$$

And those are all the solutions. Let us write it in vector form:

$$\begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \end{pmatrix} = \alpha \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \beta \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} + \gamma \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

In other words, the vector formed by λ_i for $i = 0, 1, 2, 3, 4, 5, 6$ is formed by the linear combination of the vectors (1,1,1,0,0,0,0,),(0,0,1,1,0,0,1,),(0,0,1,0,1,1,0). Meaning that the kernel has dimension 3, and these 3 vectors form a basis for all the cycles. That means that all the one dimensional cycles in CW_A are generated by those

$1-$ dimensional Cycles $=< (a+b+c), (c+d+g), (c+e+f) >$

Which leads us to the conclusion that the cycles form a group isomorphic to $\mathbb{Z}_2^3$.

Finally, since there is no non-trivial 1-dimimensional boundary, we conclude:

$$H_1(CW_A; \mathbb{Z}_2) =< 1 - \text{dimensional Cycles} >= \mathbb{Z}_2^3$$

.

For a more general graph X with v number of vertices and e edges, we can apply a well known theorem from graph theory.

Theorem 3.4. *Every connected graph X contains a Spanning Tree. (A tree which is a subgraph of X which includes all the vertices.)*

Note: A Tree is a graph without loops or double paths.

Proof. Let X be a connected graph, if X does not have a cycle it is a spanning tree. If it has any cycle, you just have to remove edges without disconnecting the graph.

When you can not remove any edge without disconecting the graph, you will have a subraph X', that will be, by definition, a Tree. □

Using that theorem, we conclude that for every graph X that represents a topological space we have a Spanning Tree. A tree with v vertices has $v-1$ edges. So there are $e-(v-1)$ edges that are not included in the spanning tree. And all of these edges that are not included in the spanning tree represents a cycle each.

And each one is independent from the other because it uses a new edge. With linear algebra you can determine which cycles are generators of the graph.

So the first homology group of a Graph X is $H_1(X;\mathbb{Z}_2) \cong \mathbb{Z}_2^{e-v+1}$.

Example 3.5. If we look at the graph in the figure 3.5, where the red subgraph represent a spanning tree we see that it has 7 vertices and 17 edges. Its spanning tree has 6 edges, so there are 11 different edges that generate a cycle, and the first homology group of this graph G is $H_1(G;\mathbb{Z}_2) = \mathbb{Z}_2^{11}$.

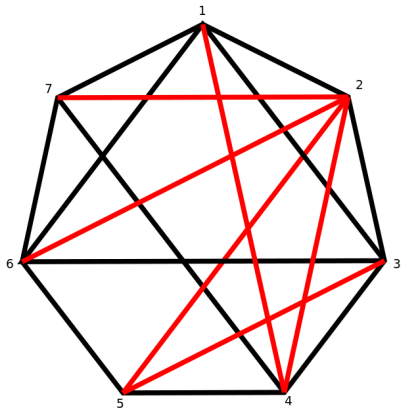

Figure 3.5: A graph and a spanning tree of it in red. This is not a planar graph. The vertices of the outer polygon marked with a number are the only vertices of the graph. When lines cross within the interior of the polygon, they are actually not touching.

In some way, we can say that $H_n(X;\mathbb{Z}_2)$ somehow measures the number of n-dimensional holes in X.

To make more clear and to extend the definition a little bit more accurately we well go up on dimensions and calculate H_1, H_2 of our Cell Complex CW_B represented in Figure 3.4.

In CW_B, C_0 and C_1 are the same since CW_A and CW_B have the same points and edges. But, $C_2 =< M, N >$ in CW_B. And $\partial_2(M) = a+b+c = \partial_2(N)$. That

will change the first homology group. Owing to the fact that the cycle $a+b+c$ does not surround a hole anymore, it is now homotopocaly trivial since we can contract it to a point sliding by M or by N. So, the presence of the 2-cells modifies our idea of what a cycle is. $a+b+c$ was a cycle that represented a hole in A. In B, obviously $a+b+c$ still represents a cycle, but now it is homotopic to 0. Algebraically that implies that the cycle $a+b+c$ should not count anymore as far as measuring 1-dimensional holes. This suggests that we form a quotient of the group of cycles that do not enclose holes (boundaries). In order to do so we will quotient the group of 1-dimensional cycles by the subgroup of 1-dimensional boundaries. Which is how we have defined homology groups.

$$H_1(CW_B;\mathbb{Z}_2) = \frac{< 1-\text{dim cycles} >}{< 1-\text{dim boundaries} >}$$

We already know that n-dimensional cycles $= \text{Ker}(\partial_n)$. What the n-dimensional boundaries concerns, in order to compute the group generated by the n-dimensional boundaries we have to find all the cycles that enclose $n+1$ dimensional cells. And that is easy, we just have to compute all the boundaries of the $n+1$ cells, i.e. $\partial_{n+1}(C_{n+1})$.

As we see $\partial_2(M) = a+b+c$, and $a+b+c$ is a boundary.

So, $c \in C_n$ is a boundary $\Leftrightarrow \exists c' \in C_{n+1}$ such that $\partial_{n+1}(c') = c$ for a given n. Therefore: $< n$-dimensional boundary $>= \partial_{n+1}(C_{n+1}) = \text{Im}(\partial_{n+1})$.

In conclusion we have that for every n:

$$H_n(X;\mathbb{Z}_2) = \frac{< n-\text{dim cycles} >}{< n-\text{dim boundaries} >} = \frac{\text{Ker}(\partial_n)}{\text{Im}(\partial_{n+1})}$$

In particular:

$$H_1(CW_B;\mathbb{Z}_2) = \frac{< 1-\text{dim cycles} >}{< 1-\text{dim boundaries } >} = \frac{\text{Ker}(\partial_1)}{\text{Im}(\partial_2)} =$$

$$= \frac{< (a+b+c),(c+d+g),(c+e+f) >}{< a+b+c >} =< (c+d+g),(c+e+f) >\cong \mathbb{Z}_2^2$$

Let us now compute $H_2(CW_B;\mathbb{Z}_2)$.

With this information we now know that

$$H_2(CW_B;\mathbb{Z}_2) = \frac{< 2-\text{dim cycles} >}{< 2-\text{dim boundaries} >} = \frac{\text{Ker}(\partial_2)}{\text{Im}(\partial_3)}$$

$\text{Im}(\partial_3) = \{0\}$ since $C_3 = 0$.

If we repeat the process that we have done earlier, for $c' \in C_2$ there $\exists \lambda_0, \lambda_1 \in \mathbb{Z}_2$ such that $c' = \lambda_0 \cdot M + \lambda_1 \cdot N$ which implies that $\partial_2(c') = \lambda_0 \cdot \partial_2(M) + \lambda_1 \cdot \partial_2(N) = \lambda_0 \cdot (a+b+c) + \lambda_1 \cdot (a+b+c)$.

As a result of this we have:

$$c' \in \text{Ker}(\partial_2) \Leftrightarrow \partial_2(c') = 0 \Leftrightarrow a \cdot (\lambda_0 + \lambda_1) + b \cdot (\lambda_0 + \lambda_1) + c \cdot (\lambda_0 + \lambda_1) = 0 \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \lambda_0 + \lambda_1 = 0 \\ \lambda_0 + \lambda_1 = 0 \\ \lambda_0 + \lambda_1 = 0 \end{cases} \sim \lambda_0 + \lambda_1 = 0$$

And, as we did before:

$$\begin{pmatrix} \lambda_0 \\ \lambda_1 \end{pmatrix} = \alpha \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Which leaves us with the fact that the vector (λ_0, λ_1) is formed by the linear combination of $< (1,1) >$. Meaning that all the 2-dimensional cycles in CW_B are generated by the cycle $M + N$. And ultimately

$$H_2(CW_B; \mathbb{Z}_2) = \frac{< 2 - \text{dim cycles} >}{< 2 - \text{dim boundaries} >} = \frac{\text{Ker}(\partial_2)}{\text{Im}(\partial_3)} =< (M+N) > \cong \mathbb{Z}_2$$

Now that we have computed a few homology groups and we have had a first contact with the basic formulas and hopefully have got an intuitive idea of what this functor does, we want to make a framework in which we can set this theory up in a more formal way and not just by drawing pictures.

There basically two approaches to setting up a theory of homology. The first one is the initial one that Poincaré introduced in 1895, and it is what we call Simplicial homology where we work with spaces build with simplices.

There are also alternatives to simplicial complexes, such as cell complexes or Δ-complexes that follow a similar logic of glueing elements but work with spaces built in a different way or with different elements.

And then we have the singular homology, which allows us a more flexible approach to compute the homology groups of every kind of topological space, but it is significantly more complicated.

In this work we will focus on the simplicial homology which is easier to understand but it is powerful enough to compute the homology group of the spaces that are interesting for us, moreover it is the one used to compute homology groups via computers and coding meaning that it is an indispensable idea in order to understant persistent homology.

3.2 Simplicial Homology

3.2.1 Simplicial Complexes

First of all we have to define what a Simplicial complex is.

Definition 3.6. *A **simplex** is a topological manifold of dimension n determined by $n+1$ points in a space of dimension equal to or greater than n which satisfies: n-simplex $= \{(t_0, t_1, ..., t_n) \in \mathbb{R}^n | \sum_i t_i = 1$ and $t_i \geq 0$ for all $i\}$.*

Example 3.7. A triangle together with its interior determined by its three vertices is a two-dimensional simplex in the plane or any space of higher dimension.

To get the intuitive idea behind it, a simplex is a point, segment, triangle or its higher-dimensional analogues (tetrahedron, pentachoron, etc.).

Note: We denote a simplex S as $S = [V_0, V_1, \ldots, V_n]$, where V_i for $i \in 0, \ldots, n$ are the vertices of the given simplex.

Definition 3.8. *A **face** of a simplex S is another simplex $P \subseteq S$ whose vertices are also vertices of S.*

Example 3.9. If we have a 2-simplex (a triangle) it has three different 1-dimensional faces (the three different segments) and three different 0-dimensional faces (the three points).

Now that we have the basic building blocks we will use these to build up the topological spaces we want called Simplicial Complexes, and with this kind of framework, homology will be much easier to determine and compute.

Definition 3.10. *A **Simplicial Complex** is a topological space formed by different simplices not necessarily of the same dimension which have to satisfy that any two simplices are either disjoint or meet in a common face.*

Example 3.11. See Figure 3.6

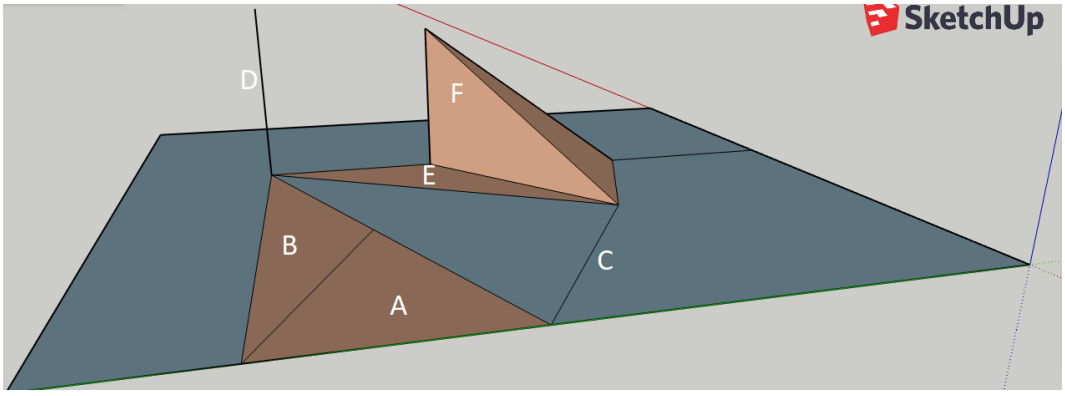

Figure 3.6: An example of a simplicial complex.

As you can see the intersection of A and B is a 1-simplex, the intersection between A and C is a 0-simplex such as the intersection between B and D or D

and E, or B and E. The intersection of E and F is also a 1-simplex, and we could have another tetrahedron which could share a triangle with F and it would still have been a Simplicial Complex.

CW_B, the topological space with we have worked previously (see Figure 3.4) is a Cell Complex, which is a generalization of a Simplicial Complexes. CW_B is not a Simplicial Complex because M and N share three 1-simplices, so they have 3 faces in common, not one or zero.

So, if we want to determine the homology group of a topological space X, we will look for a simplicial complex S, such that $X \cong S$, and calculate the homology group of S. These types of homeomorphisms are called triangulations.

Definition 3.12. *A **triangulation** of a topological space X is a homeomorphismus $|S| \to X$ where S is an abstract simplicial complex. We will say that X is triangulable if it admits some triangulation. In other words, a triangulation of X is a decomposition $X = \bigcup_{j \in J} T_j$ where each T_j is a closed from X homeomorphic to a simplex such that for all pair $i, j \in J$, either the intersection $T_i \cap T_j$ is empty or $T_i \cup T_j$ is a common face of T_i and T_j.*

Note: Two triangulations are called equivalent if there exists another triangulation which is a refinement of both.

Example 3.13. Let us show an example: Let T be a Torus, the easiest way to triangulate it is by looking at it in his planar form. See Figure 3.7.

The first idea that someone would come up with to make the planar form of a torus homeomorphic to a set of triangles would be this one shown in Figure 3.8. Nevertheless, $T_1 \cup T_2$ is not a simplicial complex due to $T_1 \cap T_2 = a \cup b \cup c$. So, T_1 and T_2 are not neither disjoint, and do not meet in a common face, they meet in three faces, which means that $T_1 \cup T_2$ is not a simplicial complex.

The idea is to triangulate T_1 and T_2 until the union of all the simplices form a simplicial complex.

And that is exactly what we have in Figure 3.9.

3.2.2 Algebraic definition of homology theory tools

Now that we have explained how to treat our topological spaces in order to easily compute their homology groups we will define in a more formally algebraic way a couple of ideas in terms of simplicial complexes that we have mentioned before to see how it worked out geometrically.

Note: Remember:

$$H_n(X; \mathbb{Z}_2) = \frac{< n - \text{dim cycles} >}{< n - \text{dim boundaries} >} = \frac{\text{Ker}(\partial_n)}{\text{Im}(\partial_{n+1})}$$

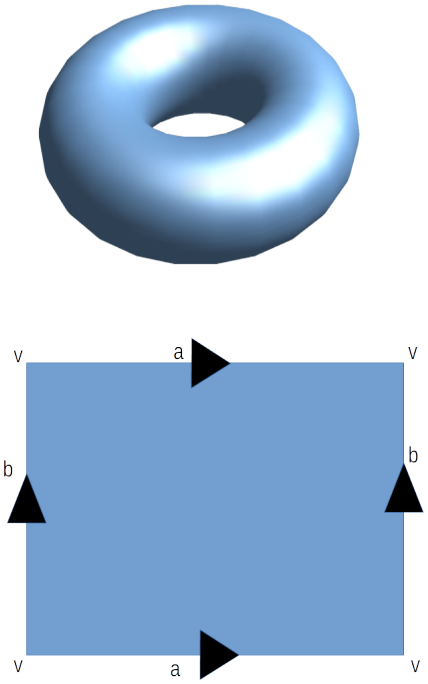

Figure 3.7: A Torus in his 3D and his planar form.

First of all, let us explain the boundary and the boundary operator for a n-simplex S homeomorphic to a topological space. $\partial_p : C_p(S) \to C_{p-1}(S)$.

Definition 3.14. *If $S = [V_0, \ldots, V_p]$ is an oriented simplex with $p > 0$, we define the **boundary operator** as:*

$$\partial_p(S) = \sum_{i=0}^{p} (-1)^i \cdot [V_0, \ldots, \overset{\wedge}{V_i}, \ldots, V_p],$$

Where the symbol $\overset{\wedge}{V_i}$ stands for not counting the vertex V_i.

Note: Thanks to the right side of the equation it is easy to see that $\partial_p(S)$ lies in C_{p-1}. Every term of the sumatory $[V_0, \ldots, \overset{\wedge}{V_i}, \ldots, V_p]$ is a $(p-1)$-simplex.

Since we are working in $\mathbb{Z}_2$ and we do not care about orientation and $-v = v$. We will have:

$$\partial_p(S) = \sum_{i=0}^{p} [V_0, \ldots, \overset{\wedge}{V_i}, \ldots, V_p],$$

Example 3.15. If we have a segment $[V_0, V_1], \partial_1([V_0, V_1]) = V_0 + V_1$.

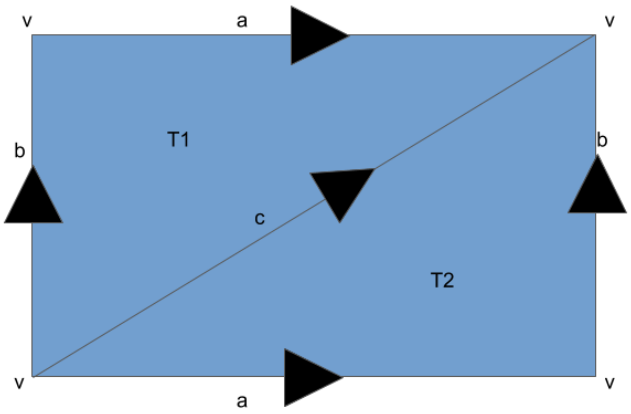

Figure 3.8:

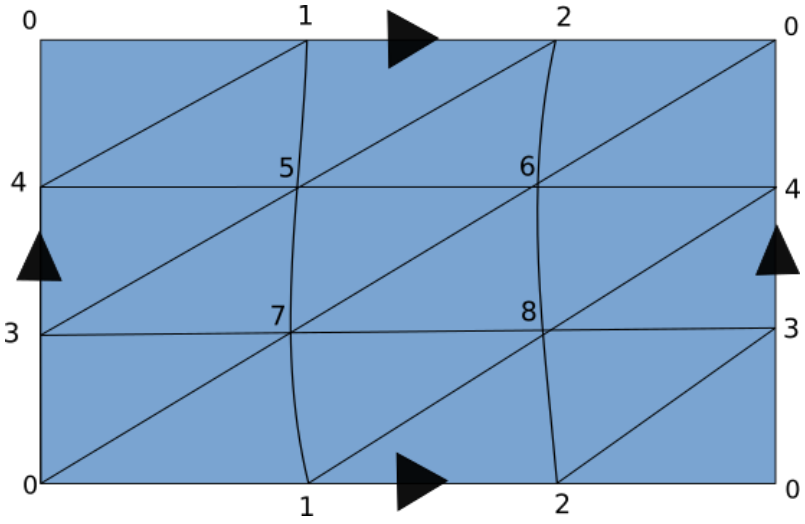

Figure 3.9: Triangulate Torus

For now on, owing to that we will work always with $\mathbb{Z}_2$ since is the most common field used in persistent homology we will ignore the signs, and the boundary operator will be used like in the previous equation.

Now, we have to see that this morphism is well-defined. Meaning that for a Simplex S, one can choose whatever way he wants to assign the vertices and they boundary should remain the same. If we have that $S = [V_0, \ldots, V_i, \ldots, V_j, \ldots, V_n]$,

you can choose to switch V_i and V_j. This leads us to the equality that we want:

$$\partial_n([V_0,\dots,V_i,\dots,V_j,\dots,V_n]) = V_0+\cdots+V_i+\cdots+V_j+\cdots+V_n = (*)$$

$$= V_0+\cdots+V_j+\cdots+V_i+\cdots+V_n = \partial_n([V_0,\dots,V_j,\dots,V_i,\dots,V_n])$$

* (That equality will be always true because we work in abelian groups, not only because of the fact that we work on $\mathbb{Z}_2$).

Meaning that for a concrete simplex S you will always have the same boundary.

Now, let us study in a more formal way the homology groups. We have defined the homology group as a quotient. But to make sure that this group is well defined, we have to see that $\mathrm{Im}(\partial_p) \subseteq \mathrm{Ker}(\partial_{p-1})$. And since $c \in \mathrm{Ker}(\partial_{p-1}) \Leftrightarrow \partial_{p-1}(c) = 0$ if we prove this next lemma true we will have demonstrated that our quotient is well defined.

Lemma 3.16. $\partial_{p-1} \circ \partial_p = 0$.

Proof.

$$\partial_{p-1}(\partial_p([V_0,\dots,V_p])) = \sum_{i=0}^{p}(-1)^i \cdot \partial_{p-1}([V_0,\dots,\hat{V_i},\dots,V_p]) = \sum_{i=0}^{p}\partial_{p-1}([V_0,\dots,\hat{V_i},\dots,V_p])$$

(remember we are working on $\mathbb{Z}_2$ and $v = -v$)

$$= \sum_{j<i}^{p}\partial_{p-1}([\dots,\hat{V_j},\dots,\hat{V_i},\dots]) + \sum_{i<j}^{p}\partial_{p-1}([\dots,\hat{V_i},\dots,\hat{V_j},\dots]) =$$

$$= 2\cdot\sum_{j<i}^{p}\partial_{p-1}([\dots,\hat{V_j},\dots,\hat{V_i},\dots]) = 0$$

It is almost trivial to see that the two terms that we are adding are the same since you only have to switch V_i and V_j and we already have seen that the boundary operator is well defined. And once we have that they are the same it is easy to see that our expression is equal to zero since $\forall z \in \mathbb{Z}_2, 2\cdot z = 0$. □

A good question could be, why should we triangulate this torus in a simplicial complex? What Figure 3.8 shows us is a Δ-complex, which is another way to approach homology, where the boundary operator and homology in general can be defined in a similar way as in simplicial complexes.If you compute the homology groups of this two spaces homeomorphic to the Torus, they will isomorphic. But, with the Δ-complex it is much easier. It still only has one vertex, meaning that $\partial_1(a) = \partial_1(b) = \partial_1(c) = v - v = 0$. Which implies that $\mathrm{Ker}(\partial_1) \cong \mathbb{Z}_2^3$. And, since $\partial_2(T_1) = \partial_2(T_2) = a + b + c$. $\mathrm{Im}(\partial_2) \cong \mathbb{Z}_2$. i.e. $H_1(T;\mathbb{Z}_2) \cong \mathbb{Z}_2^2$.

Computing it with the triangulation shown in 3.9 you come up with the exact same result but with much more computational cost. So why would we chose the Simplicial Complex triangulation instead of a Δ-Complex triangulation? That is because in a Simplicial Complex is uniquely defined by its boundary. And you can uniquely define a simplex in there, for example, the 2-simplex in the left upper corner could be defined by its edges [(0,1),(0,4),(1,4)]. In the Δ-Complex, T_1 and T_2 have the exact same boundary. The fact that the simplices that form the complex are uniquely defined by its boundary, make it much easier to compute homology using algorithms.

Analogous to the Seifer-Van Kampen to compute Homotopy groups, in homology theory we have the Mayer-Vietoris sequences, which are a powerful tool to compute homology groups by splitting a topological space into subspaces where homology is easier to compute.

Definition 3.17. *We define the $n-th$* ***Betti number*** *of a Simplicial Complex C with coefficients from a field $\mathbb{K}$ like $\beta_n(C;\mathbb{K}) = dim_{\mathbb{K}}(H_n(C;\mathbb{K}))$.*

Note: Due to we work always with $\mathbb{Z}_2$ which is a field, we do not have torsion in mind.
If we have a topological space X and we are working with a field $\mathbb{K}$ (like we are doing with $\mathbb{Z}_2$), $H_n(X;\mathbb{K})$ is a $\mathbb{K}$-vector field. which means that, if the homology groups are finite-dimensional, then:
$H_n(X;\mathbb{K}) \cong \mathbb{K}^m$ for some $m \in \mathbb{Z}$.

And from that, it can be deduced that...

$$\beta_n = dim_{\mathbb{K}}(H_n(X;\mathbb{K})) = m$$

Betti numbers are a topological tool that measures the dimension of the homology group meaning that in some way, Betti numbers estimate the connectivity, by measuring the number of n-dimensional holes when the group over we work is a field.

$\beta_0(C;\mathbb{K})$ measures the connected components of a simplicial complex C.

$$\beta_0(C;\mathbb{K}) = dim(\frac{\text{Ker}(\partial_0)}{\text{Im}(\partial_1)})$$

Thanks to the algebraic definition of the boundary operator, we see that $\partial_0(c) = 0$ for every $c \in C_0$. the 0-simplices are defined by one single vertex, meaning that if v_0 is that vertex for c. $c = [v_0]$. And $\partial_0(c) = (-1)^0 \cdot \sum_0^1 [\overset{\wedge}{V_0}] = 0$. Meaning that dim(ker($\partial_0$))=dim($C_0$). And since we are making the quotient with Im(∂_1), i.e. we are making the quotient with the boundary of the edges, all the vertices that are

connected by edges fall in the same class, which means that we have a different class for every connected component.

$\beta_1(C;\mathbb{K})$ measures the number of 1-dimensional holes, that is straightforward, you are looking for 1-dimensional cycles that do not surround a 2-dimensional manifold, which is a definition for a 1-dimensional hole. Following the same logic, $\beta_2(C;\mathbb{K})$ measures the number of 2-dimensional holes, (bubbles or cavities). You measure the number of 2-dimensional cycles that are not the boundary of a 3 dimensional manifold.

And, so on, β_n measures in some way the number of n-dimensional holes.

Note: From now on, in order to simplify notation, and since we are working in $\mathbb{Z}_2$, $H_n(X)$ and $\beta_n(X)$ will refer to $H_n(X;\mathbb{Z}_2)$ and $\beta_n(X;\mathbb{Z}_2)$ respectively.

Example 3.18. Our Cell Complex CW_A from Figure: 3.3 has the following Betti numbers:

$$\beta_0(CW_A) = dim_{\mathbb{Z}_2}(H_0(CW_A)) = 1,$$
$$\beta_1(CW_A) = dim_{\mathbb{Z}_2}(H_1(CW_A)) = 3,$$
$$\beta_i(CW_A) = dim_{\mathbb{Z}_2}(H_i(CW_A)) = 0 \text{ for } i \geqslant 2.$$

And for our Cell Complex CW_B, in Figure 3.4 we have that:

$$\beta_0(CW_B) = dim_{\mathbb{Z}_2}(H_0(CW_B)) = 1,$$
$$\beta_1(CW_B) = dim_{\mathbb{Z}_2}(H_1(CW_B)) = 2,$$
$$\beta_2(CW_B) = dim_{\mathbb{Z}_2}(H_2(CW_B)) = 1,$$
$$\beta_i(CW_B) = dim_{\mathbb{Z}_2}(H_i(CW_B)) = 0 \text{ for } i \geqslant 3.$$

$H_0(CW_A) \cong H_0(CW_B) \cong \mathbb{Z}_2$ due to the complexes are fully connected and the quotient is $\text{Im}(\partial_1)$, we have that $x \sim y \sim z \sim v \sim t$. Which leaves us only with one vertex. (You can also see that $\text{Ker}(\partial_0(CW_A)) \cong \text{Ker}(\partial_0(CW_B)) \cong \mathbb{Z}_2^5$ and $\text{Im}(\partial_1(CW_A)) \cong \text{Im}(\partial_1(CW_B)) \cong \mathbb{Z}_2^4$.)

Note: Notice, that changing our field, different Betti numbers can come out.

A slight modification of the Betti numbers will be our main tool to determine the persistence of topological features in our next chapters which goes deeper in a new technique to analyze data via algebraic topology using persistent homology.

Chapter 4

Persistent Homology

Persistent homology is a new technique, first defined in [15] and whose first effective algorithm was given in [11]. This technique consists on applying algebra to obtain topological features (such as components or holes) of data (a set of discreet points with a metric).

It consists in transforming a cloud of points into a family of topological spaces (Simplicial Complexes to be precise) paramatrized by a variable to see which topological features persists in different values of that given parameter and are more likely to represent true features of the underlying space rather than noise.

4.1 From Data to Topology

In first place, we have to convert our cloud of data to a simplicial complex. To do that, we set all our data points in our metric space to vertices (0-simplices). Then, a distance $\delta > 0$ is chosen. Two vertices will be connected by an edge if the distance between them is smaller or equal to δ. Now we have a graph, that captures the connectivity of our data (clustering), but do not give any further information. So we want to fill this graph with simplices. And to do so we have few options, for instance:

Definition 4.1. *Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the **Čech complex**, C_δ, is the abstract simplicial complex whose k-simplices are determined by unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ whose closed $\delta/2$-ball neighborhoods have a point of common intersection.*

Definition 4.2. *Given a collection of points$\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the **Rips complex**, R_δ, is the abstract simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ that are pairwise within distance δ.*

Example 4.3. In figure 4.1 you can see three points and they respective closed $\delta/2$-balls. In the Rips Complex, the three points are pairwise less than δ apart, so they form a 2-simplex. In the other side, we have the Čech complex. Since the intersection of the three balls is empty, the points do not form a 2-simplex, but they are pairwise intersected, so they form three 1-simplex (an empty triangle).

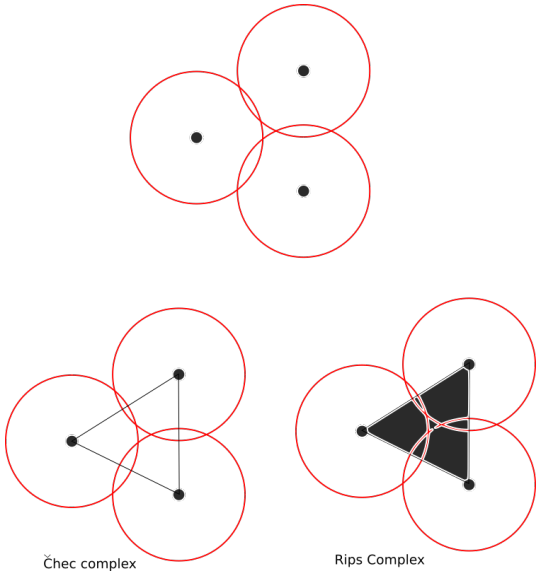

Figure 4.1: Čech Complex vs Rips Complex.

Note: As one can clearly see in this image, the Čech Complex and the Rips Complex can have different homotopy groups.

Theorem 4.4. *([6]). The Čech theorem (or 'nerve' theorem) states that C_δ has the same homotopy type as the union of closed balls with radius $\delta/2$ that have as they center each point of the given cloud.*

Meaning that our Complex behaves like a subs set of $\mathbb{E}^n$. In order to get a more illustrative idea, what this theorem is stating is that both picture in Figure 4.2 have the same homotopy type (or are homotopically equivalent).

One can see that both have the same number of components and the same number of one-dimensional holes, and also have an isomorphic fundamental group ($\mathbb{Z}_2$). Indeed, one is a deformation retract of the other.

To sum everything up, thanks to this theorem it is clear that the Čech Complex is a good enough topological approximation to what the underling object could be, nevertheless, the Čech complex and various topologically equivalent subcomplexes, have a very expensive computational cost.

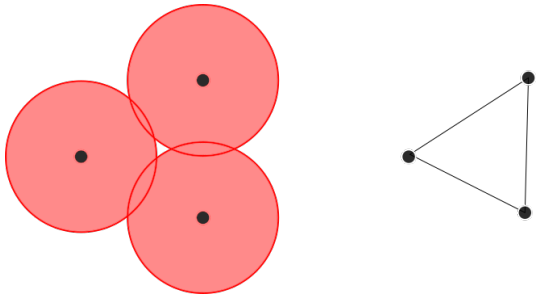

Figure 4.2: The union of three closed balls and its respective Čech Complex.

However, the Rips Complex is much easier to compute than the corresponding Čech Complex despite having at least, the same amount of simplices. That is because of the fact that the Rips Complex is a *flag* complex (it is maximal among all the simplicial complexes with the same 1-skeleton). Therefore, the 1-skeleton determines univocally the complex. That implies that the Rips Complex can be stored as a graph and later be reconstructed, meaning that the entire boundary operator needed for a Čech complex do not need to be stored. That been said, this computational simplification comes not without any handicap. The disadvantage is that in general, a Rips complex R_δ do not necessarily behave like a subset of $\mathbb{E}^n$ nor an n-dimensional space at all. But, once we have introduced the idea of persistence, we will see that the Rips complex will also be a good approximation if one handles the parameter in a concrete way.

Logically, the next step would be to ask which distance δ to chose, if such distance exist, to capture these true features from our underlying object from our data. If δ is sufficiently small, the complex will be a discrete set of points. Contrarily, if δ is too large, the Simplicial Complex will be a single $(n-1)$-simplex, where n is the number of points of our data, which would have a trivial homology.

In [3] there is a perfect example of a sampling of points on a planar annulus (See Figure 4.3) where one can observe that when δ is large enough to have removed all the small holes within the annulus, the characteristic hole that discerns it from a disk is already filled in.

So, does actually the optimal δ exists?

The answer is that, if such ideal choice of δ exist, it is rare.

It is insufficient to know the number of components and different dimensional holes (Betti numbers) of a single simplicial complex created from clouds of data (regardless one chooses the Čech Complex or the Rips Complex) with a particular δ. It is a mistake to ask which value of δ is optimal. What we need is some kind of tool to declare which holes are essential and which can be ignored (considering them noise). But, the problem is than neither homology non homotopy gives

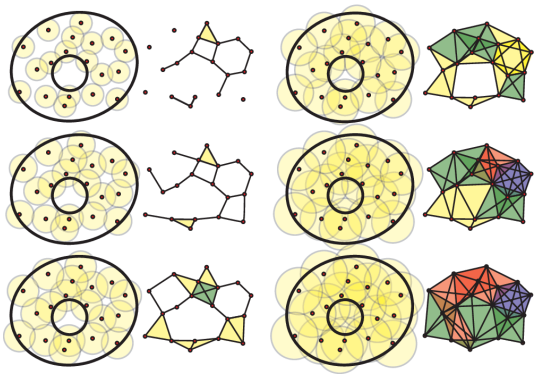

Figure 4.3: A sequence of Rips complexes for a point cloud data set representing an annulus extracted from [3].

any information of this kind, these functors compute if there is a hole, with no regard to how fragile this hole can be. And here is where we explain the idea of persistence, which is a novelty that is applied to topology with the aim of discernoing the significant topological features from simple noise.

4.2 Persistence

Persistence is the rigorous answer to this problem given by Edelsbrunner, Letscher, and Zomorodian [16] and later refined by Carlsson and Zomorodian [12]. Given a parameterized family of spaces C, those topological features which persist over a significant parameter range are to be considered as significant and those that are short-lived features should be considered noise.

Given values $\delta_0, \delta_1 \in \{\delta_i\}_{i\in I}$, if a new topological feature of a space appear at δ_0, and at δ_1 this feature disappears, we will say that δ_0 is its birth-time and that δ_1 is its dead-time, and that this feature lived from δ_0 to δ_1.

We can represent the *persistence* of this feature as a pair (δ_0,δ_1). The distance between δ_0 and δ_1 measures how long this feature has lived.

Example 4.5. Let us supose we have a sequence of Čech Complexes C=$\{C_\delta\}_{\delta\in I}$ (See Figure 4.4), parametrized by a distance δ, and where I is increasing sequence of values $\{\delta_i\}_{i\in\mathbb{N}}$. This sequence is associated to a fixed cloud of points (in this case three points), and we can look at it as our parameterized family of spaces. By lookinkg at the picture one can observe that for $\delta_i = 5$, there is a hole, and that at $\delta_i + 1 = 6$ there is not one, so the persistence of this hole is $(\delta_i, \delta_i + 1) = (5, 6)$.

Note: Notice that if the sequence $\{\delta_i\}_{i\in\mathbb{N}}$ is increasing, $R_{\delta_i} \hookrightarrow R_{\delta_j}$ for $i < j$,

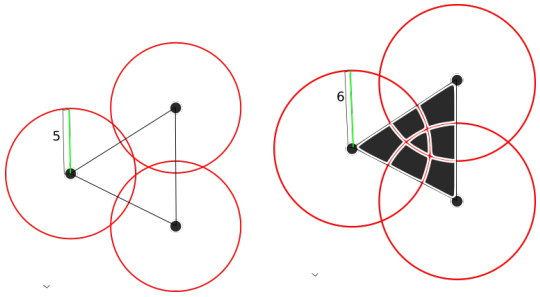

Figure 4.4: A simple example of the persistence of a topological feature (in this case, a hole).

meaning that after a feature has died it can not appear again (but one of the same kind, can).

So, to determine the features of the underlying space, instead of computing the homology group for a single complex R_{δ_i}, one examines the homology of the iterated inclusions $\iota : H_*(R_{\delta_i}) \to H_*(R_{\delta_j})$ for all $\delta_i < \delta_j$. These maps reveal which features persist.

That explains why Rips Complexes are acceptable approximations to Čech complexes. Although no single Rips complex is a good approximation to a single Čech complex, pairs of Rips complexes can capture the appropriate Čech complex. For any δ, it is trivial to see that $C_\delta \hookrightarrow R_\delta$ due to that the restrictions from the Čech complexes are much stronger that the ones of the Rips complexes. That means that if for every δ one can find a δ' that satisfies $R_{\delta'} \hookrightarrow C_\delta$. We would have a chain of inclusion maps:

$$R_{\delta'} \hookrightarrow C_\delta \hookrightarrow R_\delta \tag{4.1}$$

Which leads us to the conclusion that any topological feature which persists under the inclusion $R'_\delta \hookrightarrow R_\delta$ is in fact a topological feature of C_δ.

Lemma 4.6. *For any $\delta > 0$, there is a chain of inclusion maps:*

$$R_\delta \hookrightarrow C_{\delta\cdot\sqrt{2}} \hookrightarrow R_{\delta\cdot\sqrt{2}}$$

Proof. For a cloud of points P, we will prove that $R_\delta \hookrightarrow C_{\delta\cdot\sqrt{2}}$. In other words, if a collection n points are pairwise at distance $2 \cdot \delta$ or closer, then the balls of radius $\delta \cdot \sqrt{2}$ centered on these points have a non-empty intersection.

We will demonstrate it for the worst of the cases (when the Čech Complex differs at most from the Rips complex). When $d(x, y) = \delta$ for each two points $x, y \in$P, where d represents the euclidean distance. This is the distribution where the closed balls are more far away from each other, meaning that the probability

that a simplex is added in the Rips Complex and not added in the Čech complex is maximal.

Once demonstrated the worst case, that would imply that if the distance between any two points of the cloud are closer than δ, the Čech complex would be at least equal to the Rips, what simplices reefer.

Let us suppose that we have n points distributed like we said before in $\mathbb{R}^m$, where $m \geq n-1$. We can set up a coordinate system where each point is included in a different axis of $\mathbb{R}^m$, meaning that each point p, can be represented as $p_i = (0, 0, \dots, 0, \underset{i}{a}, 0, \dots)$. The value of a is easy computable. The distance between every two points will be δ, so: $d(p_i, p_j) = \sqrt{0 + \dots + 0 + a^2 + 0 + \dots + 0 + a^2 + 0 \dots} = \sqrt{2} \cdot a = \delta \longrightarrow a = \frac{\delta}{\sqrt{2}}$.

What we want to see is that the $\bigcap_{i \leq n} D_{\delta \cdot \sqrt{2}}(p_i) \neq \emptyset$ where $D_{\delta \cdot \sqrt{2}}(p_i)$ is the closed ball with radius $\delta \cdot \sqrt{2}$ and center p_i. To do so, we will see that the barycenter, is included in all the balls. Let B be the barycenter, $B = \sum_{i=0}^{n+1} \frac{p_i}{n} = (\frac{\delta}{n \cdot \sqrt{2}}, \frac{\delta}{n \cdot \sqrt{2}}, \dots, \underset{n}{\frac{\delta}{n \cdot \sqrt{2}}}, 0, \dots)$.

We want to see that for every i, $d(B, p_i) \leq \frac{\delta \cdot \sqrt{2}}{2}$. Owing to that $\delta > 0$ and $d(x, y) \geq 0$. To simplify everything we will multiply each side by itself. Meaning that we want to demonstrate: $d(B, p_i)^2 \leq \frac{\delta^2}{2}$.

And it is easy to see, that with this distribution, the distances for all the points to B will be the same, since $B - p_i = (\frac{\delta}{n \cdot \sqrt{2}}, \frac{\delta}{n \cdot \sqrt{2}}, \dots, \underset{i}{\frac{\delta}{n \cdot \sqrt{2}} - \frac{\delta}{\sqrt{2}}}, \dots, \underset{n}{\frac{\delta}{n \cdot \sqrt{2}}}, 0, \dots)$. And since $\frac{\delta}{n \cdot \sqrt{2}} - \frac{\delta}{\sqrt{2}} = \frac{\delta \cdot (1-n)}{n \cdot \sqrt{2}}$, we have that $B - p_i = (\frac{\delta}{n \cdot \sqrt{2}}, \frac{\delta}{n \cdot \sqrt{2}}, \dots, \underset{i}{\frac{\delta \cdot (1-n)}{n \cdot \sqrt{2}}}, \dots, \underset{n}{\frac{\delta}{n \cdot \sqrt{2}}}, 0, \dots)$.

Which implies that:

$$d(B, p_i)^2 = |B - p_i|^2 = \sum_{j \neq i} (\frac{\delta}{n \cdot \sqrt{2}})^2 + (\frac{\delta \cdot (1-n)}{n \cdot \sqrt{2}})^2 = \frac{(n-1) \cdot \delta^2}{2 \cdot n^2} + \frac{(1-n)^2 \cdot \delta^2}{2 \cdot n^2} =$$

$$= \frac{\delta^2 \cdot ((n-1) + (1-n)^2)}{2 \cdot n^2} = \frac{\delta^2 \cdot (n^2 - n)}{2 \cdot n^2} = \frac{\delta^2 \cdot (n-1)}{2 \cdot n}$$

That last expression is the distance between B and any point of the point cloud. And, to finalize:

$$\frac{\delta^2 \cdot (n-1)}{2 \cdot n} \leq \frac{\delta^2}{2},$$

due to $\frac{n-1}{n} \leq 1$. Meaning that the barycenter will be in every ball of center p_i and radiuses $\delta \cdot \sqrt{2}$. Which implies that $R_\delta \hookrightarrow C_{\delta \cdot \sqrt{2}}$. As said before, the implication $C_\delta \hookrightarrow R_\delta$ is trivial, since in order to have an non empty intersection of closed balls of radius δ, the points must be a at most at distance δ pairwise.

And with this two implications we have that:

$$R_\delta \hookrightarrow C_{\delta\cdot\sqrt{2}} \hookrightarrow R_{\delta\cdot\sqrt{2}}$$

□

This lemma implies that the δ' mentioned before in the inclusions of 4.1 exists if $\frac{\delta}{\delta'} \geqslant \sqrt{2}$.

4.3 Persistence in Homology

First of all, we need to define the family of topological spaces in which we will work.

Definition 4.7. *A **persistence complex** C=$\{C_*^{\delta^i}\}_{i\in I}$ is a sequence of chain complexes together with their chain maps $x : C_*^{\delta_i} \to C_*^{\delta_{i+1}}$. This is motivated by having a sequence of Rips or Čech complexes of increasing parameter δ from an increasing sequence of values $\{\delta_i\}_{i\in I}$. Since Rips or Čech complexes grow with δ, the chain maps x are naturally identified with inclusions.*

Definition 4.8. *For $i < j$, the (i,j)-**persistent homology** of C, denoted $H_*^{i\to j}(C)$, is defined to be the image of the induced homomorphism $x : H_*(C_*^i) \to H_*(C_*^j)$.*

Continuing with the point of before, consider a filtration of Rips Complexes R= $\{R_i\}_{i\in I}$ parametrized by proximities $\{\delta_i\}_{i\in I}$ as our family of topological spaces. The previous lemma 4.6 implies that if $\frac{\delta_i}{\delta_j} \geq \sqrt{2}$, then $H_n^{i\to j}(R) \neq 0$ implies that $H_n(C_{\delta_j}) \neq 0$. Meaning that properties in the Čech complex are deduced by the persistent homology of the Rips filtration.

Let us a choose a PID of coefficients R and place a graded $R[x]$-module structure on C with x acting as a map between complexes that can be composed in the following form $x^m \in R[x]$ and sends $x^m : C_*^i \to C_*^{i+m}$ via m applications of x. We also will assume that each complex from C is finitely generated as an $R[x]$, since it clearly exist a parameter $M \in R$, such that, C_n^M is a n-simplex, and $H_p(C_n^M; R) = 0$ for $p \geq n$.

The problem is, that, although C is a free $R[x]$-module, meaning it has unique basis and a unique rank, (due to it is a filtering via chain maps x,) $H_*(C)$ is not necessarily free, it is certainly a $R[x]$-module, but, like with the non-persistent homology, if one takes coefficients over a ring other than a field, there are many possible forms for the homology groups.

That is why we are working all the time with coefficients $\mathbb{Z}_2$. It is a field. If we choose our coefficients from a field $\mathbb{K}$, the classification of $\mathbb{K}[x]$-modules is much

easier, if one takes in consideration the Structure Theorem for PIDs which tells us that the only graded ideals of $\mathbb{K}[x]$ have following form: $x^n \cdot \mathbb{K}[x]$. If one also assumes that all Betti numbers are finite, which is the case for the most reasonable finite-dimensional spaces, that implies that:

$$H_*(C, \mathbb{K}) \cong \bigoplus_{\mathrm{i}} x^{t_i} \cdot \mathbb{K}[x] \oplus \left(\bigoplus_{\mathrm{j}} x^{r_j} \cdot \left(\frac{\mathbb{K}[x]}{x^{s_j} \cdot \mathbb{K}[x]} \right) \right)$$

This classification has a natural interpretation. The free part of the equation has a bijective correspondence with those topological features which appear at the parameter t_i and persist for all other values of $\{t_i\}_{i \in I}$ for indices greater than i. These are the features that persist to infinity. The torsion elements correspond to those topological features that come into existence at r_j and dissapear at $r_j + s_j$. So, this theorem provides a pair with the birth-time and dead-time for each feature, except those that persist to infinity.

4.4 Representation

Now that we have explained some of the algebra behind *Persistent Homology*, we want to represent the persistence of the topological features in a more graphical way with the aim of making it a little bit more understandable. We have defined the persistence of a feature as a pair (x, y) where x is the parameter where this feature is born (birth-time) and y is the parameter where the feature dies (death-time).

There are two common used ways to represent the persistence of this features. One is the *persistent diagram*, where each feature is described as a point in a plane where the x axis represent the birth-time of the feature and the y axis represent the death-time. So, logically, all the points will be find over the diagonal $x = y$ since a feature can not die before it has been born. In this diagram the most important feature would be the ones furthest from the diagonal, hence they will be the more persistent ones.

Example 4.9. See figure 4.5.

The other way to represent persistence, in which one we will go more indeed, are the *persistent barcodes*.

This way, consists in the fact that each feature has an interval (bar) in the barcode, and all the intervals are set on top of another. The x axis will represent the parameter δ that parametrizes our family of Complexes. And each interval of the barcode will begin in its birth-time and end in its dead-time.

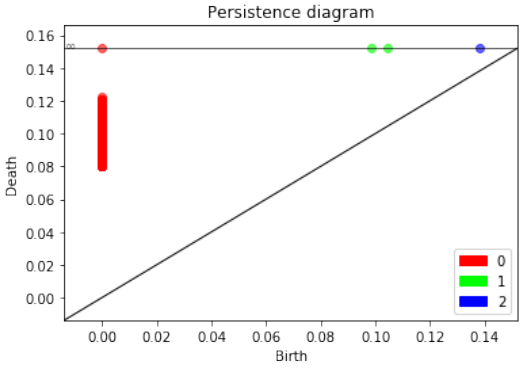

Figure 4.5: Example of a persistence diagram, where the color of the points tells us the dimension of the features.

Theorem 4.10. *([12]). The dimension $dim_{\mathbb{K}}$ of the persistent homology group of a filtration R=$\{R_i\}_{i\in I}$, $H_n^{i\to j}(R)$ is equal to the number of intervals in the barcode of $H_n(R)$ between i and j. And in particular $dim_{\mathbb{K}}(H_n(R_i))$ is equal to the number of intervals in the complex R_i for each n.*

Example 4.11. If we take a look at figure 4.6, extracted from [3]) one can see the different Rips Complexes from a given filtration R from a sample of points of an annulus. In every Rips Complex one can see that it has an interval in the barcode for each of their features in that value of the parameter. For example, in the fourth image, on can see that at this time it has one interval of $H_0(R)$, three intervals of $H_1(R)$ and none of them of $H_2(R)$, and, at this time, it certainly has one connected component, three one dimensional holes, and not a single cavity or two-dimensional hole. If one looks at the whole barcode , it is easy to see which features are more persistent. There is only one big interval of $H_0(R)$, and one of $H_1(R)$, meaning that it has one connected component and one 1-dimensional hole, meaning that the underlying figure of our data cloud has isomorphic homology groups to the homology groups of S^1 (which is a retract of an annulus).

So, in a more intuitive way of though, a barcode is the persistence analogue of a Betti number. Remember, $\beta_k = dim_{\mathbb{Z}_2}(H_K) = rank(H_K)$, and a barcode do not give any further information of the structure. It is only a continuously paramatritzed rank. The reason why one chooses barcodes to represent persistence is the capacity to distinguish noise and significant topological features by measuring the different intervals lengths.

Example 4.12. Let us see an example of the persistence of a topological feature represented in a persistence diagram and in a persistence barcode.

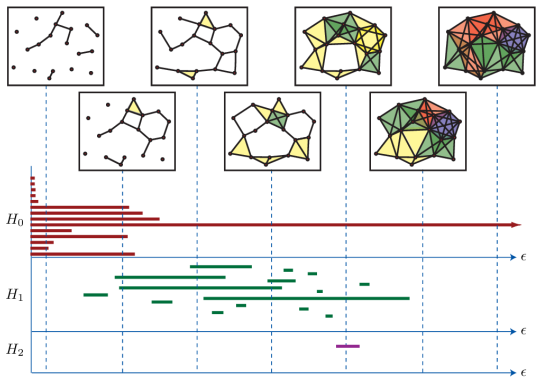

Figure 4.6: Example of barcodes for $H_0(R), H_1(R), H_2(R)$ for a given filtration R from a data cloud. Exctracted from [3].

The hole of Figure 4.4 which had birth-time 5 and dead-time 6, and is represented in Figure 4.1 in a persistence diagram.

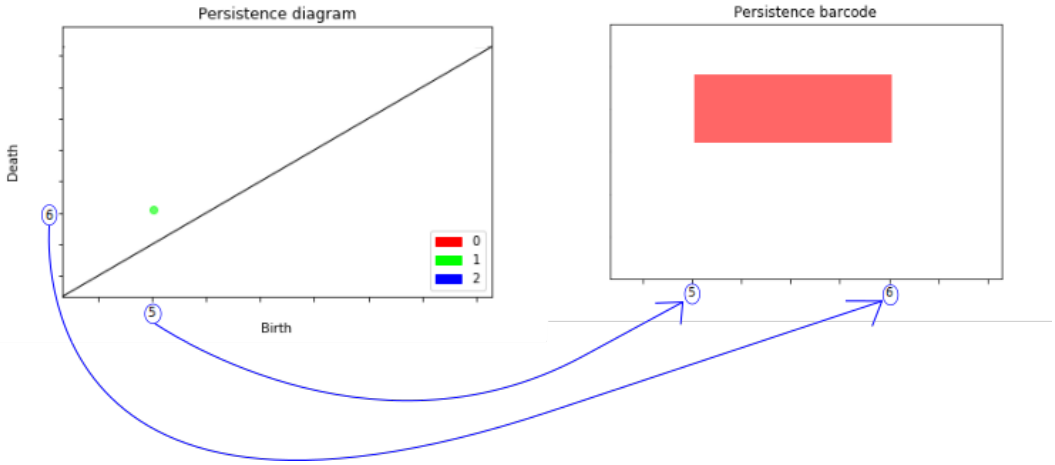

Figure 4.7: The persistence of the feature in Figure 4.4 represented in a persistence diagram and a persistence barcode.

As we will see in the next section, barcodes and persistence diagrams are stable in the presence of noise.

4.5 Stability

Let us recapitulate. We have now, Persistent Homology is the homology of a filtration, and we set up a theory that allows us to genarate a filtration of Simplicial

Complexes parametrized by the distance of the points. That, leads us to a way to calculate topological features starting from discrete data.

4.5.1 Homology Inference

One of the most fundemental applications of **Persistent Homology** is the *Homolgy Inference*. Which tackles the following problem.

Given a finite sample of points $P \subset \Omega$ of an unkown shape $\Omega \subset \mathbb{R}^m$, we want to determinate $H_*(\Omega)$.

In order to do so, we want to approximate the homology of Ω via barcodes. But, we first, need to explain a couple of ideas.

Let $X \subseteq \mathbb{R}^m$ be a closed set, we define $d^X : \mathbb{R}^m \to \mathbb{R}$ as a function that gives the euclidean distance to the nearest point in X.
$d^X(p) = \inf_x ||p - x||_2$, for $x \in X$ and $p \in \mathbb{R}^m$.

Definition 4.13. *The* **Parellel Body** X_0^ϵ *is defined as* $X^\epsilon = d^{-1}[0, \epsilon]$.

Meaning that the paralel body, are all the points that are at distance ϵ or less from any point of X.

Definition 4.14. *Let Y be another closed set in $\mathbb{R}^m$ the* **Hausdorff distance** *between two sets $d_H(X, Y)$ is defined as the infimum $\epsilon > 0$ for which $X \subseteq Y^\epsilon$ and $Y \subseteq X^\epsilon$.*

Which implies that if $d_H(X, Y) = \epsilon$, all the points from X are at a distance ϵ or less from a point of Y and viceversa.

Definition 4.15. *Let f be a real function on X. A* **Homological Critical Value** *of f is a real number a for which there exists an integer such that for all $\epsilon > 0$ suffiecent small the map $H_k(f^{-1}(-\infty, a - \epsilon)) \to H_k(f^{-1}(-\infty, a + \epsilon))$ induced by inclusions is not an isomporhism.*

If $f = d^X$, it is the distance in which a betti number B_k changes.

Definition 4.16. *The* ***Homological feature size*** *of X, denoted as $hfs(X)$ is the infimum positive homological critical value of d^X.*

Now, we can use a theorem to handle our initial problem.

Construct X such that $H_*(X) \cong H_*(\Omega)$ from a finite sample of points $P \subset \Omega \subset \mathbb{R}^m$.

Theorem 4.17. Homology Inference Theorem. *([5]). For all $\delta \in \mathbb{R}$ such that $d_H(\Omega, P) < \delta < hsf(\Omega)/4$, and all suffecently small $\epsilon > 0$, $rank(H_p(\Omega^\epsilon)) = rank(H_p(Im(f_\delta^{3\delta})))$, where $f_\delta^{3\delta} : H_p(P^\delta) \to H_p(P^{3\delta})$.*

In other words, if $P^\delta = D_\delta(P) = \bigcup_{p\in P} D_\delta(p)$ covers Ω, (being $D_\delta(p)$, the closed ball with radius δ and center p,) which is a direct implication of the inequality $d_H(\Omega, P) < \delta$, and the inclusions $D_\delta(P) \hookrightarrow D_{2\delta}(P) \hookrightarrow D_{3\delta}(P)$ preserve homology, which is an implication of the other inequality $\delta < hsf(\Omega)/4$, since $3\delta + d_H(\Omega, P) < hsf(\Omega)$, meaning that, $H_*(D_\delta(P)) \hookrightarrow H_*(D_{2\delta}(P)) \hookrightarrow H_*(D_{3\cdot\delta}(P))$ are isomorphisms. Which implies that they do preserve homology, meaning that $dim(H_p(\mathrm{Im}(f_\delta^{3\delta}))) = dim(H_p(D_\delta(P))) = dim(H_p(D_{2\delta}(P))) = dim(H_p(D_{3\delta}(P))) = dim(H_p(\Omega^\delta))$ for any p.

Since we work with fields, that imply that $H_p(\Omega^\delta)$ is isomorphic to all the union of closed balls that we listed, meaning that we have almost found the shape we wanted, except for one detail. We have now determinated the homology for Ω^δ, but not for Ω. So, in order to determine the homology of Ω, one has to assume a certain regularity of the shape Ω, (meaning that we are not working with fractals, or other irregular structures,) so, that it exists a small enough δ, such that the $\Omega \cong \Omega^\delta$.

Now, we can determinate how to compute the homology of a shape with a sample of points by computing the persistent homology of the thickening of this points by closed balls. And thanks to the **Čech Theorem** (theorem 4.4), we know, that this thickening has the same kind of homotopy as its corresponding Čech Complex. And since a filtration of Čech Complexes and Rips Complexes have similar (although not the same) persistent features, we can compute the homology of the underlying shape by calculating the persistent homology of the Rips filtration.

In this subsection, one can deduce, that for any sufficiently well distributed set of points that satisfyes the inequalities, one will obtain the same homology, meanin that there exists some kind of stability of persistence homology for near point clouds. Let us see, that this stability is also present in persistent diagrams and persistent barcodes.

4.5.2 Stability of persistence diagrams

In the first place, one needs to know that there is a way to define persistent homology given a function, rather than given a thickening parameter. But I will not focus on that, so I will consider only the particular case in which the function is d^P for some point cloud, since it is the example I have covered in this report.

Definition 4.18. *The* **bottleneck distance** *measures the similiraty between two persistence diagrams. Given functions f, g that create filtrations, (such as d^P, d^Q, for two point clouds, P, Q), and $Dgm_p(f), Dgm_p(g)$ the persistence diagrams for dimension p from the filtrations derivated from f and g respectively. And η being a bijection between these two*

diagrams, the Bottleneck distance is defined as:

$$d_B(Dgm_p(f), Dgm_p(g)) = \inf_{\eta} \sup_{x} ||x - \eta(x)||_\infty \tag{4.2}$$

Which implies that, the bottleneck distance between two diagrams, $d_B(Dgm_p(f)$ and $Dgm_p(g) = d$, is the shortest distance d such that it exists a perfect matching between the points of the two diagrams (completed with all the points in the diagonal so it do not has cardinality mismatches) such that any couple of mathced points are at L_∞-distance d or less.

And the equation 4.2 surely does that, it is equal to the supremum of the L_∞-distance of all the points with its match, and then takes the infimum of this distance between all the possible matchings.

Theorem 4.19. *([5]). Given two functions f, g as before, for each dimension p the bottleneck distance between the the persistence diagrams of dimension p is bounded from above by the L_∞-distance between the two funcions.*

$$d_B(Dgm_p(f), Dgm_p(g)) \leq ||f - g||_\infty \tag{4.3}$$

And, if f and g, are the distance functions to two set of points P, Q. ($f = d^P, g = d^Q$). Theorem 4.3 implies that if any point of P is less than d (in L_∞-distance) apart from a point of Q and viceversa, the points from the persistence diagrams for both point clouds will be also less than d apart from its matching point, since:

$$d_B(Dgm_p(f), Dgm_p(g))) \leq ||f - g||_\infty = ||d^P - d^Q||_\infty = d_H(P, Q). \tag{4.4}$$

4.5.3 Stability of persistence barcodes

Knowing that persistence diagrams have this stability, it is easy to see that persistence barcodes also have it.

Theorem 4.20. *If $||f - g||_\infty = d$ there exists a matching between the intervals of the persistence barcode of f and g such that:*

- *matched intervals have endpoints with distance equal or less than d, and*
- *unmatched intervals have length equal or less than $2 \cdot d$.*

Proof. (Remember that for each interval in the persistence barcode there is a point in the persistent diagram if they came from the same filtration.) From equation 4.4 we know that there exists a matching between the points of the two diagrams such

that all the points are at L_∞-distance d or less. That means that these two points differ by d in birth-time and dead-time at most. This implies that for each couple of mathced points of the diagram that are not in the diagonal there exist a couple of matched intervals of the barcode that also differ at most in d in both birth-time and dead-time which, in other words means that the ending a starting point of the mathced intervals differ by d or less. There are points in the persistence diagram that are matched with the diagonal, and that leaves us with unmatched barcodes, but since the L_∞-distance from those points to the diagonal is also equal or less than d: Let us suppose the worst case, where the point with coordinates (a, b) is at distance d in both birth-time and dead-time from the diagonal. That means that the nearest point from (a, b) is $(a + d, b - d)$ since the points are only above the diagonal. And given that the diagonal is $x = y$, $a + d = b - d \rightarrow b - a = 2 \cdot d$. And, $b - a$ is the dead-time minus the birth-time, the length of the respective barcode. And since this was the worst case, it clear that the unmatched intervals have length equal or less than $2 \cdot d$. □

Which implies that the stability of persistent homology is also present in the persistent diagrams and persistent barcodes. This stability gives the tools to perform *Homology inference* as it has shown that the persistent homology will determine the homology from the underlyng space if a good enough sample is given. And it also means that persistent diagrams and persistent barcodes are stable with respect to data perturbation. So noise will not affect the homology of our filtration.

4.6 Persistent Homology Examples

Now, we will see, two examples from *Persistent Homology* computed with the help from the Gudhi library [14] [1].

The data from the first example are 200 points from a sphere S^2. I took a sample of points of a sphere from the library and then selected 200 random points. The seconds example are 300 points from a Torus, which I also extracted from [14].

Let us see this points in a more graphical way in Figure 4.8.

Setting up parameters (like maximum dimension of simplices, or maximum distance that parametrizes the Rips Complex) that we have to have in mind in order to have a reasonable computational time, one can calculate the persistence barcodes and diagrams from the sphere (see Figure 4.9), and from the torus (see Figure 4.10).

[1] The code used in those examples can be found at `https://github.com/fritzpere/Gudhi_examples.git`

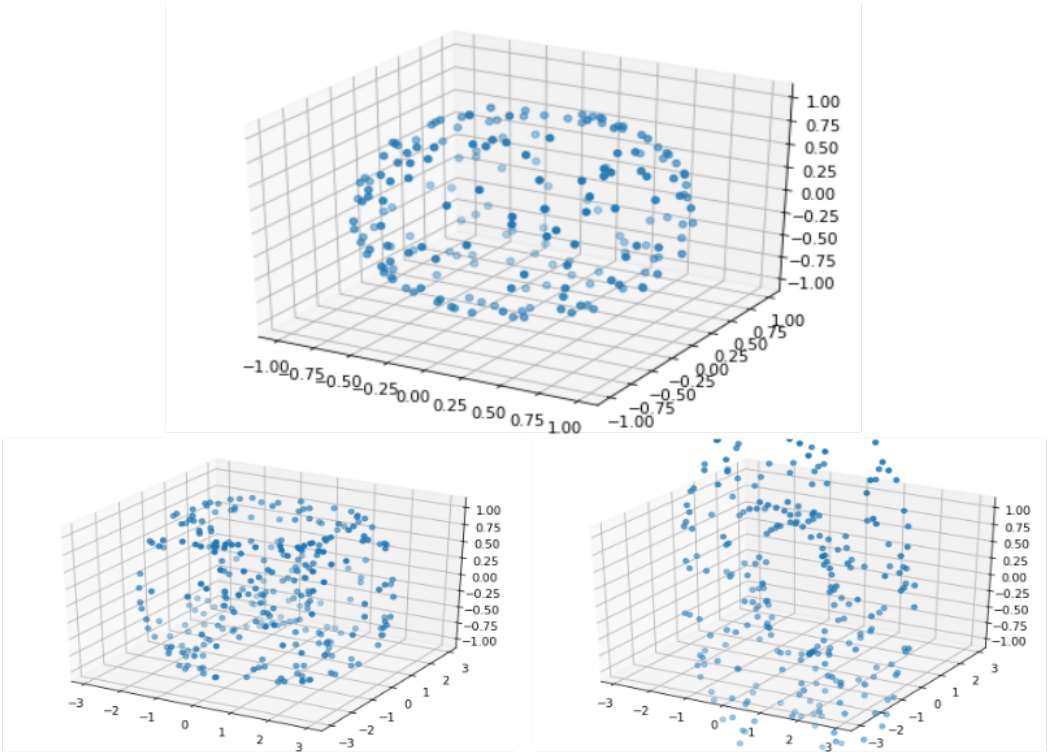

Figure 4.8: Sample points of a sphere(Up).Sample points of a Torus (Down).

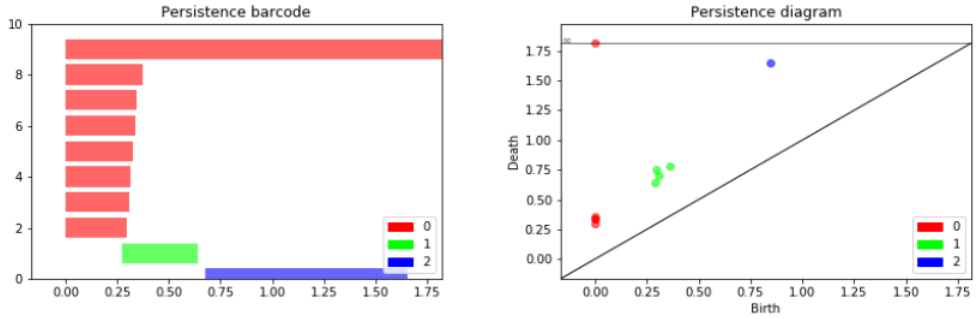

Figure 4.9: Persistence barcode of the sphere (Left).
Persistence diagram of the sphere (Right).

Then, I added noise to check the stability of those persistence. I added a random number between 0 and 0.1 to each coordinate of each points of both sphere and torus and recalculated the persistent barcodes and persistence diagrams with the same parameter. See Figure 4.11 for the sphere and Figure 4.12 for the torus.

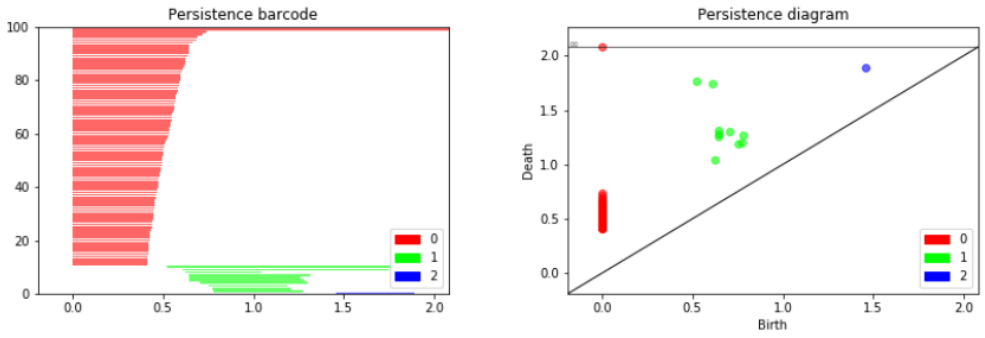

Figure 4.10: Persistence barcode of the torus (Left).
Persistence diagram of the torus (Right).

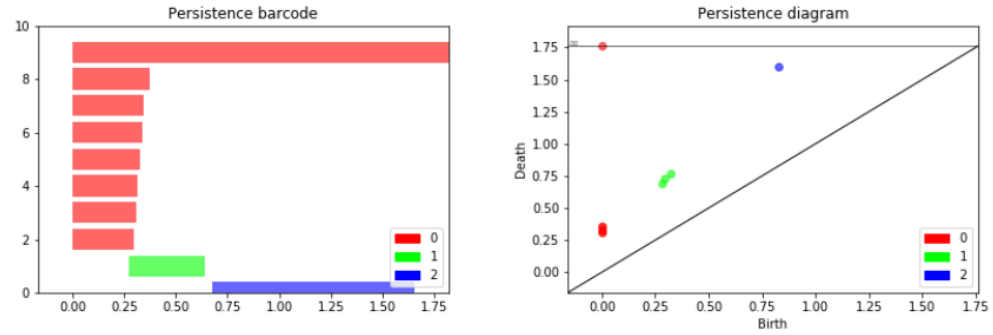

Figure 4.11: Persistence barcode of the sphere with noise (Left).
Persistence diagram of the sphere with noise (Right).

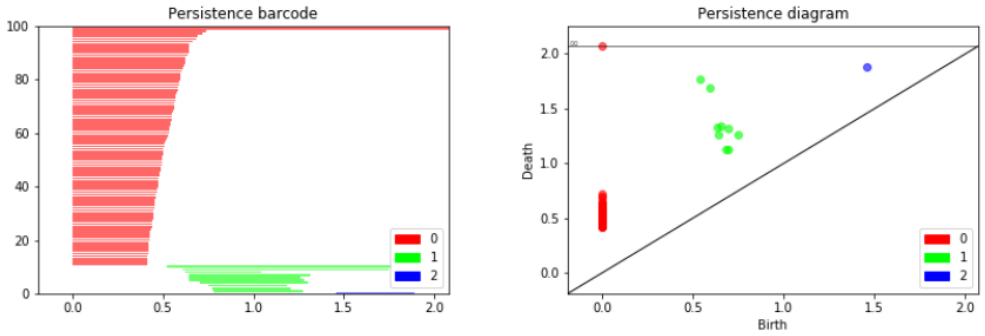

Figure 4.12: Persistence barcode of the torus with noise (Left).
Persistence diagram of the torus with noise (Right).

Chapter 5

Conclusions

In this work we have provided a formal introduction to algebraic topology and the fundamental group. Then we have explained the intuitive idea behind homology groups and formally defined simplicial homology and all the algebraic tools needed to understand it.

Next, we have entered in persistent homology, explaining how to represent persistent homology and the remaining ideas needed in order to explain the structure theorem and the stability theorem, both theorems set the basis for persistent homology to be a robust mathematical theory.

Finally, two examples of computing persistent homology are given, where one can appreciate that, persistent barcodes and persistence diagrams are indeed stable against noise.

Persistent Homology is a new technique in topological data analysis with many applications that allows us to denoise and discover other features from high-dimensional data by studying its shape. This two applications evoke many questions that can work as a source for future works. For example, a more statistical research on how short-living a feature must be in order to be considered noise. Or, since persistent homology only gives us information about Betti numbers, going beyond only persistence to find tools able to distinguish between spaces with the same Betti numbers.

These are just two examples of the immense amount of possibilities to cover, this field will be studied in the near future and hopefully extended to give us even more information starting from noisy and high-dimensional data.

Bibliography

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

[2] Munkres, James R. *Elements Of Algebraic Topology*. CRC PRESS, 2019.

[3] Ghrist, Robert. "Barcodes: the persistent topology of data." *Bulletin of the American Mathematical Society* 45.1 (2008): 61-75.

[4] Gebhart, Thomas, and Paul Schrater. "Adversary detection in neural networks via persistent homology." arXiv preprint arXiv:1711.10056 (2017).

[5] Cohen-Steiner, David, Herbert Edelsbrunner, and John Harer. "Stability of persistence diagrams." *Discrete '&' Computational Geometry* 37.1 (2007): 103-120.

[6] Borsuk, Karol. "On the imbedding of systems of compacta in simplicial complexes." *Fundamenta Mathematicae* 35.1 (1948): 217-234.

[7] Adcock, Aaron, Daniel Rubin, and Gunnar Carlsson. "Classification of hepatic lesions using the matching metric." *Computer vision and image understanding* 121 (2014): 36-42.

[8] Xia, Kelin, Zhixiong Zhao, and Guo-Wei Wei. "Multiresolution persistent homology for excessively large biomolecular datasets." *The Journal of chemical physics* 143.13 (2015): 10B603-1.

[9] Hiraoka, Yasuaki, et al. "Hierarchical structures of amorphous solids characterized by persistent homology." *Proceedings of the National Academy of Sciences* 113.26 (2016): 7035-7040.

[10] Lawson, Peter, et al. "persistent Homology for the Quantitative evaluation of Architectural Features in prostate Cancer Histology." *Scientific reports* 9 (2019).

[11] Edelsbrunner, Herbert, and John Harer. "Persistent homology-a survey." *Contemporary mathematics* 453 (2008): 257-282.

[12] Zomorodian, Afra, and Gunnar Carlsson. "Computing persistent homology." *Discrete '&' Computational Geometry* 33.2 (2005): 249-274.

[13] Nomitch, Michael. *Teaching Tree*. Teachingtree.Co, 2020, http://www.teachingtree.co/teachers/N

[14] GUDHI project, "GUDHI, Geometry Understanding in Higher DImensions." GUDHI, *software available at:* https://gudhi.inria.fr/interfaces/.

[15] Robins, Vanessa. "Towards computing homology from finite approximations." *Topology proceedings*. Vol. 24. No. 1. 1999.

[16] Edelsbrunner, Herbert, David Letscher, and Afra Zomorodian. "Topological persistence and simplification." *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE, 2000.