Alma Mater Studiorum · Università di Bologna

**Scuola di Scienze**
**Dipartimento di Fisica e Astronomia**
**Corso di Laurea in Fisica**

# Topological data analysis with physical applications

**Relatore:**
**Prof. Roberto Zucchini**

**Presentata da:**
**Michele Busti**

Anno Accademico 2023/2024

*Dedicata a Renè, grazie per avermi insegnato tanto.*

# Abstract

Starting with basic concepts of topology, the main tools of topological data analysis are first introduced and then used in the context of physically relevant cases. In particular it is analyzed a set of experimental data obtained from a sampling of the cyclooctane molecule's conformations. The study reveals that, despite the high dimensional conformation space in exam, the geometry of the dataset is rather simple, it being an intersection of two bidimensional surfaces.

Partendo da concetti di base di topologia, vengono introdotti gli strumenti principali dell'analisi topologica di dati. Questi ultimi vengono poi applicati in esempi di rilevanza fisica, in particolare per analizzare un insieme di dati ottenuto da un campionamento delle conformazioni della molecola del cicloottano. Lo studio rivela che, nonostante lo spazio delle conformazioni abbia una dimensionalità elevata, la geometria dell'insieme dei dati è piuttosto semplice in quanto esso consiste in un'intersezione di due superfici bidimensionali.

# Contents

# Introduction

All kinds of experiments, ranging from physics to chemistry, biology and social sciences, have the common goal to collect as much data as possible from the object in study. Generally, as the complexity of the experiment grows, the amount of data that is collected get so large that any type of examination on single data items provides no useful information. This, however, does not rule out the possibility of a broader and less specific analysis that focuses on rougher traits of the data.

Topological data analysis (TDA for short) is a subarea of topology whose tools are crucial to get some kind of qualitative information on the collected data as a whole. Using the methods provided by TDA it is possible not only to reasonably distinguish between different datasets built over intrinsically different objects, but also to retrieve basic properties of what is being investigated despite the fact that the dataset can usually be noisy or inaccurate due to the effects of experimental errors.

A dataset can be viewed as a discrete and finite set of points embedded in a metric space of appropriate dimension. TDA aims to build a topological structure from this dataset that adequately encapsulates the main topological features of the space from which the data is gathered. Once this structure is assembled through one of several different procedures, it is possible to make use of standard topological tools to mathematically encode those properties of the structure that characterize it uniquely. Specifically, one of the most basic analysis that is done via TDA, is to measure the number of components and $k-$dimensional gaps of some topological space, in our case the dataset under consideration. Since one can only sample finitely many points, even though the object itself is continuous, the property extracted using TDA are employed to formulate an educated guess about the actual shape of the underlying dataset and are not to be intended as definitive and exact properties of it.

In molecular physics, molecules can be found to have many different conformations which presumably form a topological space. TDA can be implemented to study how the space is connected and its dimension and so whether it is composed of one or more components and if they have a manifold structure or not.

The applications of TDA cover many areas of science and more. In the field of medical imaging it has been used to classify MRI scans that indicates the presence of diseases and differentiate between different ones [1], [2]. In economics there have been instances where TDA has helped to point out recurrent signs of market instability and to study overall market dynamics [3].

# Chapter 1

# The conformation space of the cyclooctane molecule

## 1.1 Conformation space of a molecule

A molecule is a composite structure whose fundamental components are its atoms. The atoms in a molecule are linked via different types of bonds, which are not to be intended as stiff and firm connections, but rather as flexible ones. Consequently, once a set of atoms comes together to form the molecule, the relative distances and angles of rotation between them can vary in a wide range of values as figure 1.1.1 displays.

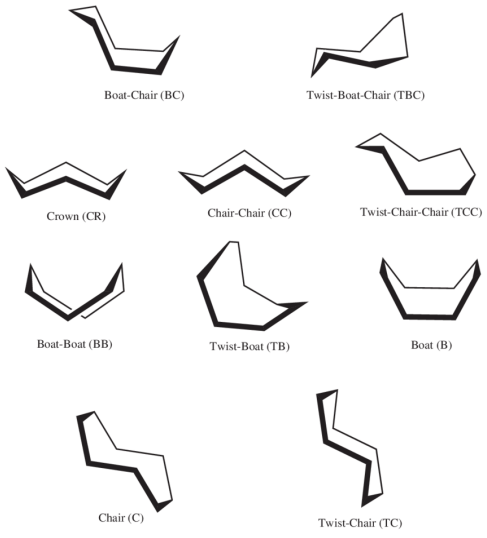


Figure 1.1.1: Few different conformations of the cyclooctane molecule. The graph vertices represents the eight carbon atoms in the molecule [4].

It is then possible define a *conformation* of a molecule as the set of coordinates of all of its atoms. A conformation of a molecule is effectively a list of number that, given a coordinate system, completely and uniquely specify those relative distances and angles between the molecule's atoms mentioned above.

As a ordered list of real numbers, a conformation of a molecule can be interpreted as a $d-$dimensional vector of $\mathbb{R}^d$ which in turn can be represented as a point in that very space. The collection of all the possible conformations of a molecule is then a subset of $\mathbb{R}^d$ which is called the *conformation space* of the molecule.

The molecule that will be used as a practical example from now on is the cyclooctane ($C_8H_{16}$). This molecule has a structure of a ring of eight carbon atoms, each bonded to a pair of hydrogen atoms reaching a total of 24 atoms. In 3D space each atom position is specified by three real numbers making a conformation of the cyclooctane a $24 \times 3 = 72-$dimensional vector of $\mathbb{R}^{72}$. It is relevant that once the conformation of the carbon ring is fixed, the positions of the hydrogen atoms can be determined through energy minimization techniques. Nonetheless the cyclooctane conformation still requires to specify all of the 24 atoms locations as per its definition.

## 1.2 The cyclooctane dataset

As a practical example it will be used a dataset $\mathcal{S}$ of 6400 conformation of the cyclooctane molecule [5], [6]. A geometric reconstruction of the shape the dataset $\mathcal{S}$ creates when embedded in $\mathbb{R}^{72}$ is performed. It is possible to visualize the result obtained from this reconstruction which is surprisingly a $2-$dimensional surface. Projecting the shape obtained in a 3D space through appropriate algorithms we get figure 1.2.1.

Figure 1.2.1: 3D embedded surface reconstruction of the dataset $\mathcal{S}$ [6].

It is immediately clear that there is a precise subset of $\mathbb{R}^{72}$ in which we can find

every single conformation that the cyclooctane molecule assumes. What it is shown is that the experimental points sampled from the conformation space of the cyclooctane do not cover in a somewhat uniform way the entire 72−dimensional space they live in as one might expect, but rather they arrange themselves to form a 2−sphere glued to a Klein bottle along two circumferences. Analyzing the surface in the vicinity of the two intersection rings, one finds a non-manifold geometry of two intersecting planes and thus a space that is locally non-Euclidean as shown in figure 1.2.2.

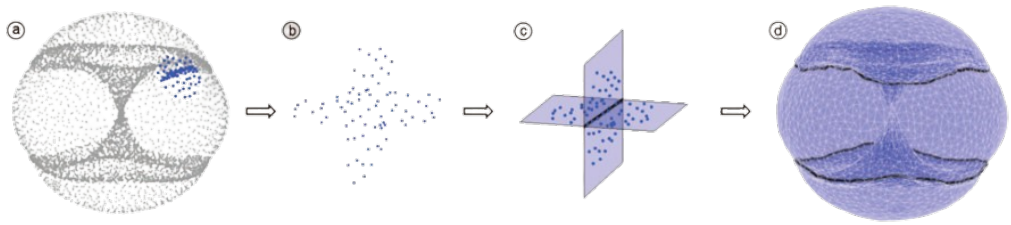


Figure 1.2.2: Local structure of the reconstructed surface near the intersection rings [7].

This kind of space is usually problematic when working with surface reconstruction techniques. On the other hand TDA does not require the assumption of manifold structure or smoothness which are frequently absent in real world cases, hence furnishing a more viable tool to study the cyclooctane conformation space.

## 1.3 The 2-sphere and the Klein bottle

The modelling of the surface reconstruction of the cyclooctane's conformation space is composed of two elements, namely a 2−sphere and a Klein bottle. It is interesting to study some of the geometric properties of those components and how the way they are connected have physical consequences on the way the molecule get its conformation.

The $n$−sphere is a really basic object defined in a metric space as the set of points at equal distance $R$ (usually set to 1 to define the unit $n$−sphere) from a common point $c$ called the center, in other words it is the set:

$$\mathbb{S}^n \;=\; \{x \in \mathbb{R}^{n+1} \;:\; ||x - c|| = R = 1\}$$

The $n$−sphere is then just the generalization to $n$ dimension of a circumference. In our case the 2−sphere is merely a spherical surface embedded in $\mathbb{R}^3$. One can intuitively already grasp some of the main topological characteristics of this object. Without necessarily give a precise definition, it seems plausible to say that the 2−sphere has 1 single connected component and 1 single 3−dimensional gap, respectively the surface and the

space encapsulated by it. The meaning of the words component and 3−dimensional gap of the 2−sphere (and any other topological space in general) will later be appropriately defined.

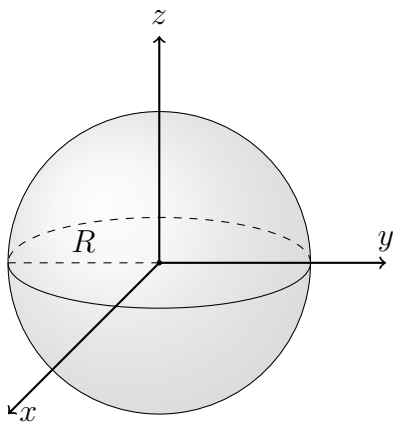


Figure 1.3.1: The 2−sphere in $\mathbb{R}^3$ centered in the origin.

The Klein bottle is a bit more articulated object as a non-orientable surface that has no *embedding* in $\mathbb{R}^3$. That would indeed require representing it without self intersections which is impossible in a 3−dimensional space. Relaxing that request however allows to visualize it through an *immersion* in $\mathbb{R}^3$ (figure 1.3.2), that is to our purpose basically equivalent. It should be noted that an embedding in a 4−dimensional space like $\mathbb{R}^4$ is possible.

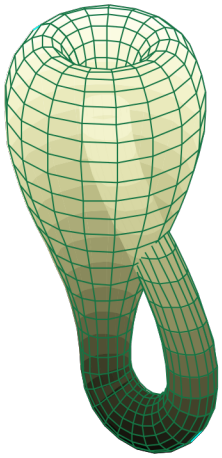

Figure 1.3.2: The Klein bottle immersion in $\mathbb{R}^3$ [8].

The Klein bottle construction can be better understood by the procedure shown in figure 1.3.3. Starting with a square whose sides orientation are represented by arrows, the opposite sides are glued first to form a cylinder and then to complete the Klein bottle. This gluing is performed in such a way that corresponding sides have matching orientations.

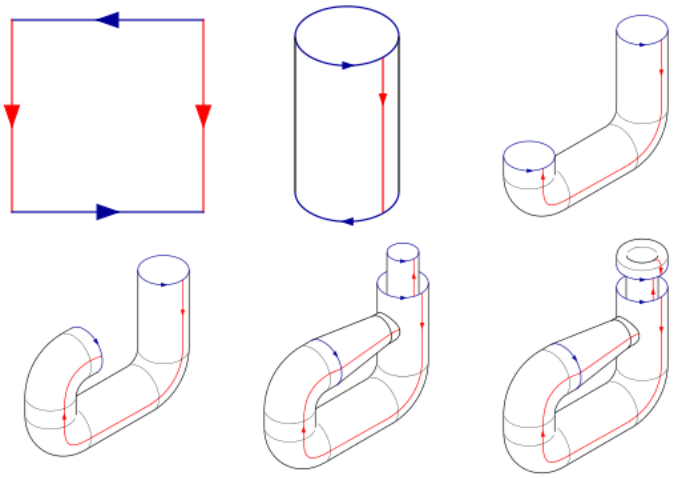

Figure 1.3.3: The procedure to construct a Klein bottle from a square.

In this case it already becomes more challenging to intuitively guess how many components and gaps the shape presents. Again, a more formal definition of what one means with those words will later help to answer the question.

As different conformations of the molecule corresponds to different energy states, it is reasonable to ask how transitions between different conformations behave when approached from the perspective of the conformation space. It is possible to show through appropriate calculations [7] that when the cyclooctane switch conformation near the intersection rings of the 2−sphere and the Klein bottle, the corresponding transitions are low energy ones. The rings could thus contain saddle points that favor transitions to conformation located either on the 2−sphere or on the Klein bottle.

# Chapter 2

# Homology and persistent homology

TDA has a pretty straightforward roadmap that can broadly be summarized by the following points:

1. Build a combinatorial structure over a discrete set $\mathcal{S}$ that is generally parameterized by a scale factor $\varepsilon$.
2. Calculate topological invariants on that structure.
3. Vary the scale factor $\varepsilon$ to build a family of combinatorial structures.
4. Study how the topological invariants evolve as a function of the scale factor $\varepsilon$ to find *persistent* invariants.

In this chapter all of these steps will be associated to topological objects used in *persistent homology*.

## 2.1 Simplices and simplicial complexes

We start by defining a basic topological structure over a set of points in $\mathbb{R}^n$ called a *simplex*.

**Definition 2.1.1** (Affine independece)**.** *Let $V = \{v_0, \ldots, v_k\}$ be a collection of points in $\mathbb{R}^n$. The points in $V$ are said to be affine independent if:*

$$\nexists \quad \alpha_j \ : \ v_i = \sum_{j=0, j\neq i}^{k} \alpha_j v_j \qquad \sum_{j=0, j\neq i}^{k} \alpha_j = 1$$

Equivalently, $v_0, \ldots, v_k$ are affine independent if the vectors $v_0 - v_i, \ldots, v_{i-1} - v_i, v_{i+1} - v_i, \ldots, v_k - v_i$ for any $v_i$ are linearly independent in the usual sense.

**Definition 2.1.2** (Convex hull)**.** *Let $V$ be a set of $k$ affine independent points in $\mathbb{R}^n$. The convex hull of $V$, denoted as $Conv(V)$, is the smallest convex set such that $V \subseteq Conv(V)$.*

**Definition 2.1.3** (Geometric $k-$simplex)**.** *Let $V$ be a set of $k+1$ affine independent points in $\mathbb{R}^n$ with $k \leq n$. $\sigma = Conv(V)$ is said to be a geometric $k-$simplex with the following properties:*

- *The dimension of $\sigma$ is $dim(\sigma) = k$.*
- *The vertices of $\sigma$ are the $k+1$ affine independent points in $V$.*
- *For every subset $V' \subseteq V$, the simplex $\tau = Conv(V')$ is a face of $\sigma$. Faces with 2 vertices are called edges.*
- *$\tau$ is a facet of $\sigma$ if it is a face such that $dim(\tau) = dim(\sigma) - 1$.*

It is easy to understand how simplices are a generalization of the $2-$dimensional triangle. The convex hull of 3 points is indeed a triangle, while one of 2 points is a straight line ($1-$dimensional triangle), the one of 4 points is a tetrahedron ($3-$dimensional triangle) and so on. This is shown in figure 2.1.1.

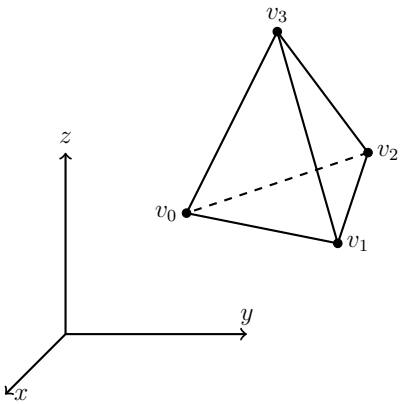


Figure 2.1.1: The convex hull of 4 affine independent point in $\mathbb{R}^3$ is a $3-$simplex which has the geometric shape of a tetrahedron.

In this example we can identify the properties of the simplex:

- The dimension of the simplex is 3 as it has been built over a set of 4 points.
- The set of vertices is the set of points $V = \{v_0, v_1, v_2, v_3\}$.
- The edges of the simplex are the edges of the tetrahedron, as they are the convex hull of pair of distinct vertices.

- The triangles formed by triplets of distinct vertices and edges therof are the faces of the simplex as they are the convex hull of subsets of $V$.
- The triangles formed by triplets of distinct vertices, but not the edges, are the facets of the simplex as they are faces of dimension 2.

It turns out that the objects of interest in TDA are not single simplices, but collections of them that satisfies certain conditions. These particular collections of simplices are called *simplicial complexes*. The way a $k-$simplex has been defined has a direct geometric interpretation since the set $V$ over which it has been built is represented as a set of points in $\mathbb{R}^n$. This approach, though very simple to visualize, is not the most general. From now on $V$ will simply be a discrete and finite set and its elements are not to be understood as points in some space anymore.

**Definition 2.1.4** (Abstract simplicial complex)**.** *Let $V$ be a discrete and finite set and $\mathcal{L}$ a family of non empty subsets of $V$. $\mathcal{L}$ is an abstract simplicial complex if it is closed under the inclusion relation, in other words:*

$$\varnothing \neq \tau \subseteq \sigma \in \mathcal{L} \implies \tau \in \mathcal{L}$$

*and it contains all the singlets:*

$$\forall\, v \in V \implies \{v\} \in \mathcal{L}$$

The properties defined for a single simplex extend naturally to a simplicial complex:

- The dimension of $\mathcal{L}$ is defined as $dim(\mathcal{L}) = \max_{\sigma \in \mathcal{L}} dim(\sigma)$.
- The set of vertices of $\mathcal{L}$ is defined as the union of the ones of all individual simplices.
- The edges of $\mathcal{L}$ are defined as the union of all the edges of the individual simplices.

As a more complex object it is convenient to give other additional definitions:

- $\mathcal{K} \subseteq \mathcal{L}$ with the structure of a simplicial complex is said to be a *subcomplex*.
- The subcomplex formed by the $m-$simplices in $\mathcal{L}$ with $m \leq n$ for a fixed n is said to be the $n-$*skeleton* of $\mathcal{L}$ denoted as $\mathcal{L}^{(n)}$.

Switching to more a more general set $V$ does not exclude the possibility of visualizing a simplicial complex since there exists generally a correspondence between the elements of $V$ and elements of $\mathbb{R}^n$. In that case it is formally said that there exists a *geometric realization* for the simplicial complex. This concept is better conceived through the use of the *body* of a simplicial complex.

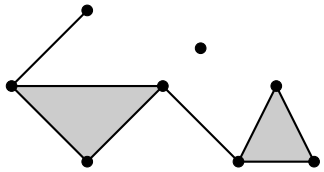

Figure 2.1.2: A simplicial complex in 2 dimension composed of two 2−simplex (the triangles), eight 1−simplex (the edges of the triangles and the two segments) and eight 0−simplex (the eight vertices)

**Definition 2.1.5** (Body of a simplicial complex)**.** *Let $\mathcal{L}$ be a simplicial complex, the union of the simplices of $\mathcal{L}$ as subsets of $\mathbb{R}^n$ through the convex hull of their vertices is called the body of $\mathcal{L}$ and it is denoted by $|\mathcal{L}|$.*

It is then possible to visualize a simplicial complex (obviously up to 3 dimension) thanks its body as show in figure 2.1.2.

## 2.2 Construction procedures

We now focus on effective procedures to systematically build a simplicial complex for a given dataset $\mathcal{S}$ which is itself a discrete and finite set. There are two main ways to start off, namely from a cover of an embedding of $\mathcal{S}$ in some topological space or from a graph whose set of vertices is $\mathcal{S}$.

Both of these strategy share the fundamental property that the simplicial complex they realize is parameterized by a scale factor $\varepsilon$ which is used in the cover based approach to define a cover and in the graph based approach to define a graph.

### 2.2.1 Cover based simplicial complexes

**Definition 2.2.1** (Nerve of a cover)**.** *Let $U = \{\mathcal{U}_i\}_{i\in I}$ be a cover of $\mathcal{S}$ and $N(U)$ a collection of subsets of indices $i \in I$ such that:*

1. $\varnothing \in N(U)$
2. $\bigcap_{j\in J} \mathcal{U}_j \neq \varnothing, J \subseteq I \implies J \in N(U)$

*$N(U)$ is called the Nerve of the cover $U$.*

The second condition in the definition of the nerve of a cover, being basically an inclusion relation, makes it immediately clear that $N(U)$ has the necessary property to be a simplicial complex whose simplices are the subsets of indices of $I$. The process to

build a simplicial complex is now as straightforward as defining a cover of $\mathcal{S}$ and using its nerve. We now review some basic examples.

### Čech complex

Using the standard euclidean metric in $\mathbb{R}^n$ the open ball cover of $\mathcal{S}$ denoted as $U_\varepsilon$ is defined as:

$$U_\varepsilon = \{B_\varepsilon(p) \ : \ p \in \mathcal{S}\} \quad B_\varepsilon(p) = \{x \in \mathbb{R}^n \ : \ d(x,p) < \varepsilon\}$$

The *Čech complex* over the dataset $\mathcal{S}$ is then just the nerve of its open ball cover:

$$\check{\text{C}}\text{ech}(\mathcal{S}, \varepsilon) = N(U_\varepsilon)$$

Note that the scale parameter $\varepsilon$ here is used to define the radius of the balls that cover the dataset. Intuitively, as the scale parameter grows, the balls that form the cover will intersect more and more, thus increasing the complexity of the nerve by adding more simplices. An example is given in figure 2.2.1.

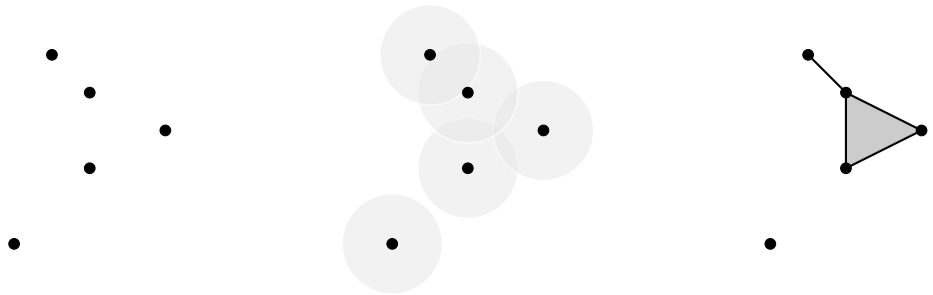

Figure 2.2.1: A Čech complex over a dataset of 5 points. The simplices are created based on the intersections of balls in the cover.

### Alpha complex

The Čech complex has the advantage of relying on a very basic procedure to build a simplicial complex, however the number of simplices generated is often so large that it is irrelevant due to the computational power it requires. It is possible to refine the open ball cover with the use of the Voronoi regions. The Voronoi region of a point in $\mathcal{S}$ is defined as follows:

$$R(p) = \{x \in \mathbb{R}^n \ : \ d(p,x) \leq d(p',x) \ \forall p' \in \mathcal{S}\}$$

An illustration of the Voronoi regions for a set of points is shown in figure 2.2.2.

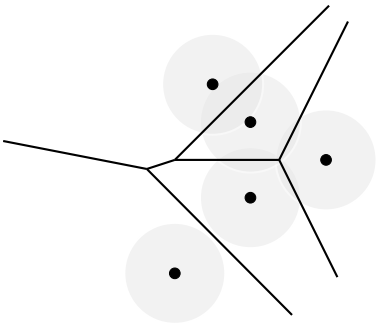

Figure 2.2.2: The Voronoi regions overlaid with the open ball cover of the same dataset used for the Čech complex example.

If the elements of $\mathcal{S}$ satisfies the condition that no $n+2$ elements lie on a $(n-1)-$sphere, intersecting the Voronoi regions of the points in the dataset $\mathcal{S}$ with their open ball cover results in a new cover, whose nerve is called the *Alpha complex*:

$$U'_\epsilon = \{B_\epsilon(p) \cap R(p) \mid p \in \mathcal{S}\} \quad A_\epsilon = N(U'_\epsilon)$$

The additional condition requires for example that in a plane no 4 points in $\mathcal{S}$ lie on a common circumference.

### 2.2.2 Graph based simplicial complexes

The simplicial complex that will be used in the analysis of the cyclooctane's conformation space isn't actually built off a cover of the dataset $\mathcal{S}$. Instead a particular graph called the $\varepsilon-$*neighborhood graph* is built using the points in $\mathcal{S}$ as the vertices.

**Definition 2.2.2** (Graph)**.** *A graph is a pair $(G, E)$ of a set of vertices $G$ and a set $E$ of subsets of $G$ which contain 2 distinct vertices. A graph is said to be complete if $E$ contains every possible pair of vertices.*

An example of a non-complete and a complete graph is given in figure 2.2.3.

Figure 2.2.3: The graph on the left is not complete while the one on the right is because every pair of vertices is connected.

**Definition 2.2.3** (Subgraph)**.** *Let $(G, E)$ be a graph and consider $G' \subseteq G$, $E' \subseteq E$. The pair $(G', E')$ is called a subgraph.*

**Definition 2.2.4** (Clique of a graph)**.** *A clique of a graph is a subgraph which is complete. A clique is said to be maximal if adding any other vertex $v \in G, v \notin G'$ results in a graph that is not complete, and thus it is not a clique.*

### Vietoris-Rips complex

**Definition 2.2.5** ($\varepsilon-$neighborhood graph)**.** *The $\varepsilon-$neighborhood graph, denoted as $G_\varepsilon$, is the graph built over the set of vertices $G$ whose edges are defined as follows:*

$$G_\varepsilon = (G, E_\varepsilon) \quad E_\varepsilon = \{(x, y) \ : \ d(x, y) \leq \varepsilon \ , \ x \neq y \in G\}$$

The *Vietoris-Rips complex* over a dataset $\mathcal{S}$ is the complex whose maximal simplices are the maximal cliques of the $\varepsilon-$neighborhood graph over the same dataset $\mathcal{S}$.

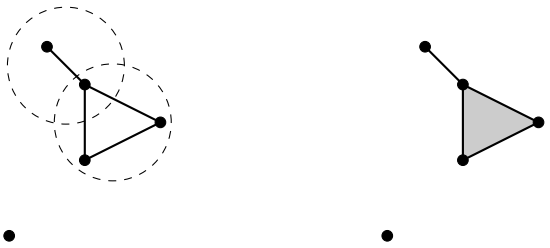

Figure 2.2.4: The Vietoris-Rips complex over the same dataset of 5 points used in previous examples. The scale $\varepsilon$ is the same as the previous complexes. The vertices encapsulated by the dashed circumferences represent the cliques in the graph which by definition translate to simplices.

It is useful to have a way to connect different simplicial complexes through *simplicial maps*.

**Definition 2.2.6** (Simplicial map)**.** *Let $\mathcal{L}, \mathcal{K}$ be two simplicial complexes. The map $f \ : \ \mathcal{L}^{(0)} \to \mathcal{K}^{(0)}$ is a simplicial map if every vertex of $\mathcal{L}$ is mapped to a vertex of $\mathcal{K}$:*

$$v_i \in \mathcal{L} \to f(v_i) \in \mathcal{K}$$

*Additionally, every simplex $\sigma \in \mathcal{L}$ must be mapped to a simplex $\tau \in \mathcal{K}$ through the mapping of its vertices.*

Note that a simplicial map may not preserve the dimension of a simplex since two distinct vertices in $\mathcal{L}$ could be mapped to the same vertex in $\mathcal{K}$.

## 2.3 Homology groups

Having now an overview on how a topological structure such as a simplicial complex can be assigned to a discrete and finite set, we turn our attention to the study of their invariants.

Formally a topological invariant between two topological spaces $\mathbb{X}, \mathbb{Y}$ is a map $f$ such that:

$$\mathbb{X} \simeq \mathbb{Y} \implies f(\mathbb{X}) = f(\mathbb{Y})$$

In the context relevant to simplicial complexes, the equivalence relation between the topological spaces is to be understood in terms of homotopic equivalence. That roughly means that there exist two maps between those topological spaces whose composition (both ways) can be continuously deformed into the identity map.

The relevance of a topological invariant actually lies in the opposite implication, in other words it is useful because it helps to rule out the possibility of two space being homotopically equivalent.

$$f(\mathbb{X}) \neq f(\mathbb{Y}) \implies \mathbb{X} \not\simeq \mathbb{Y}$$

Recalling that a simplicial complex can be viewed as a topological space through its body, the mathematical object that will be used to calculate invariants is the *homology group*, but a few more definition are needed before introducing it.

**Definition 2.3.1** (Oriented simplex)**.** *Let $\{v_0, \dots, v_k\}$ be the totally ordered set of vertices of a $k$ dimensional simplicial complex with the natural order $v_0 < \cdots < v_k$. An oriented simplex is a simplex $\sigma$ in the complex defined by the ordered subset of vertices:*

$$\sigma = \langle v_{i_0}, \dots, v_{i_p} \rangle, \quad v_{i_0} < \cdots < v_{i_p}$$

Figure 2.3.1 shows two different orientations of a $2-$simplex, while figure 2.3.2 gives an example of an oriented simplex and an orientation on one of its facets.

**Definition 2.3.2** (Chain group)**.** *Let $\mathcal{L}$ be a simplicial complex, $\mathcal{L}_p \subseteq \mathcal{L}$ the set of $p-$simplices in $\mathcal{L}$, $n_p$ the cardinality of $\mathcal{L}_p$ and $\mathbb{F}$ an abelian group, the set $C_p(\mathcal{L})$ whose elements are the formal sums:*

$$\sum_{i=1}^{n_p} c_i \sigma_i \quad c_i \in \mathbb{K},\ \sigma_i \in \mathcal{L}_p$$

*is called the p-th chain group.*

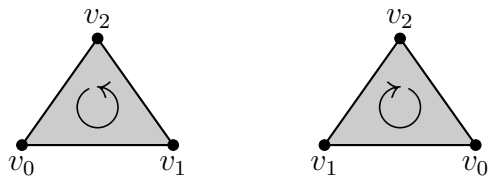


Figure 2.3.1: The circular arrow represent the orientation of the $2-$simplex. The orientation follows naturally the order of the vertices.

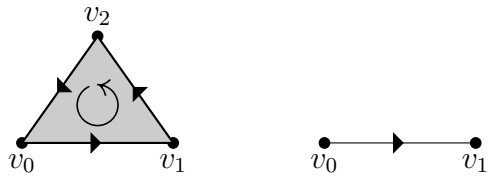


Figure 2.3.2: Removing $v_2$ from the oriented simplex $\langle v_0, v_1, v_2 \rangle$ gives the induced oriented facet $\langle v_0, v_1 \rangle$.

As the name implies $C_p(\mathcal{L})$ is a group, specifically it is a free abelian group that is generated exactly by the formal linear combination of $p-$simplices with coefficient in $\mathbb{F}$. The elements of the chain group are called *chains* and they can be understood as functions that map every $p-$simplex to an element in $\mathbb{F}$ which is just a reinterpretation of the formal linear combination used to define $C_p(\mathcal{L})$.

The last object needed to define the homology group serves as a way to connect different chain groups.

**Definition 2.3.3** (Boundary homomorphism)**.** *Let $\mathcal{L}$ be a simplicial complex. The following map between chain groups is linear and is thus defined just on the basis of the domain:*

$$\begin{aligned} \partial_p \ : \ C_p(\mathcal{L}) &\to C_{p-1}(\mathcal{L}) \\ \sigma &\to \sum_{i=0}^{p} (-1)^i < v_0, \cdots, \hat{v_i}, \cdots, v_p > \end{aligned}$$

*The notation $\hat{v_i}$ indicates that the $i-$th vertex has been removed.*

*For convenience it is also formally defined $\partial_0 \equiv 0$.*

The boundary homomorphism maps chains of $p-$simplices to chain of their facets. Reasoning as follows, that could serve as a way to identify $p-$dimensional gaps.

Consider the oriented simplicial complex in figure 2.3.3. Denoting the oriented $1-$simplices $\langle d,b\rangle, \langle b,c\rangle, \langle c,d\rangle$, and the chain $\gamma = m\langle d,b\rangle + l\langle b,c\rangle + n\langle c,d\rangle$ it is possible to apply the boundary homomorphism:

$$\partial_1(\gamma) = m\partial_1(\langle d,b\rangle) + l\partial_1(\langle b,c\rangle) + n\partial_1(\langle c,d\rangle) = m(d-b) + l(b-c) + n(c-d)$$

The result shows a linear combinations of vertices which, if we restrict the choice of coefficients to integers, can be interpreted as how many times the chain passes through one of it. Intuitively, if the chain enters $(+1)$ and leaves $(-1)$ an equal amount of time through each one of the vertices, resulting in 0 overall crosses, we conclude that it forms a closed loop. Expanding the linear combination:

$$m(d-b) + l(b-c) + n(c-d) = b(l-m) + c(n-l) + d(m-n)$$

$\gamma$ is a loop if $m = l = n$, in other words each edge is traversed an equal amount of time, which also means that the chain $\gamma$ is mapped into the null $0-$simplex. This is general and a $1-$chain is a loop if it is an element of $ker\partial_1$.

Applying the boundary homomorphism to the $2-$simplex present in the simplicial complex also gives the same result, but it seems reasonable to exclude it from the total number of loops since the inside of it is in a sense "filled". To tackle this types of cases we note that the loop $\gamma' = \langle b,c\rangle + \langle c,a\rangle + \langle a,b\rangle$ is itself the image of the boundary homomorphism $\partial_2$ applied to the $2-$simplex and so $\gamma' \in im\partial_2$.

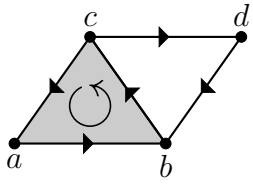


Figure 2.3.3: The boundary of the $2-$simplex (the filled triangle), forms a loop, but it should not be considered as such since it does not encapsulate a gap.

It seems now plausible to say that there exists a link between $ker\partial_1/im\partial_2$ and the number of $1-$dimensional loops, formally called *cycles*, in the simplicial complex. Though not exhaustive, this example can be directly generalized to define the homology group.

The boundary homomorphism has the following remarkable property:

**Theorem 2.3.1.** $\partial_p \circ \partial_{p+1} \equiv 0 \quad \forall p$

*Proof.* Writing $\partial_p \circ \partial_{p+1}$ as $\partial^2$, it is sufficient to show that $\partial^2\sigma = 0$ for every oriented $(p+1)-$simplex $\sigma$.

$\partial^2\sigma = \sigma'$ is a sum of $(p-1)-$dimensional faces of $\sigma$, written in terms of vertices one has:

$$\sigma' = \langle v_0, \dots, \hat{v_i}, \dots, \hat{v_j}, \dots, v_p \rangle \quad i < j$$

Recalling how an orientation is induced on a face of a simplex, we now take a look at the orientation sign of $\sigma'$ noting that it can be obtained in one of the following ways:

1. $v_j$ is removed first, giving a factor of $(-1)^j$. $v_i$ is removed next and, it being the $i-$th vertex, its contribution to the orientation of $\sigma'$ is a factor $(-1)^i$ giving a final orientation of $(-1)^j(-1)^i$.

2. $v_i$ is removed first, giving a factor of $(-1)^i$. $v_j$ is removed next and, it being now the $(j-1)-$th term, its contribution to the orientation of $\sigma'$ is a factor of $(-1)^{j-1}$ giving a final orientation of $(-1)^i(-1)^{j-1}$.

We get a sum over oriented simplex with opposite orientation and thus the result is 0 independently of the choice of $\sigma$. □

It immediately follows that $im\partial_{p+1} \subset ker\partial_p$. This guarantees that a quotient group can always be defined, leading to the concept of *homology group.*

**Definition 2.3.4** (Homology group)**.** *Let $\mathcal{L}$ be an oriented simplicial complex with dimension $n$. The $m-$th homology group is defined as the following quotient group:*

$$H_m(\mathcal{L}) = ker\partial_m / im\partial_{m+1} \quad 0 \leq m \leq n$$

It is now useful to restrict the choice of the abelian group $\mathbb{F}$, over which the chain groups are defined, to a commutative field. In this way the homology group obtains the structure of a vector space which turns out to be useful due to the possibility of defining its dimension.

**Definition 2.3.5** (Betti numbers)**.** *Let $\mathcal{L}$ be an oriented simplicial complex, the dimension of the $m-$th homology group is called the $m-$th Betti number $\beta_m$.*

Each unique $m-$dimensional cycle in the simplicial complex corresponds to an element of the $m-$th homology group's basis. Consequently, the Betti numbers quantify the number of unique *homology elements* ($m-$dimensional gaps) in a simplicial complex. It is now possible to give a meaningful definition of the terms used in section 1.3:

- The number of connected components of a simplicial complex is the Betti number $\beta_0$.

- For a simplicial complex in $\mathbb{R}^3$, the number of 3−dimensional voids (e.g. the space inside the 2−sphere) is the Betti number $\beta_2$.

The Betti numbers defined in this way are associated to simplicial complex, while in section 1.3 the objects in question were topological spaces. There is in fact no need to give an additional definition thanks to the following theorem.

**Theorem 2.3.2.** *Let $\mathcal{L}$ and $\mathcal{L}'$ be two simplicial complexes, if their bodies are homotopically equivalent then their homology groups are isomorphic:*

$$|\mathcal{L}| \simeq |\mathcal{L}'| \implies H_n(\mathcal{L}) \cong H_n(\mathcal{L}') \quad \forall\, n \leq dim(\mathcal{L}) - 1$$

*This result is independent of the choice of the group $\mathbb{F}$ over which the chain groups are built.*

As a body of a simplicial complex is itself a topological space, it is sufficient to work with one whose body is homotopically equivalent to the space of interest. A simplicial complex with this property is indeed called a *triangulation* of the topological space in study. An example is given in figure 2.3.4.

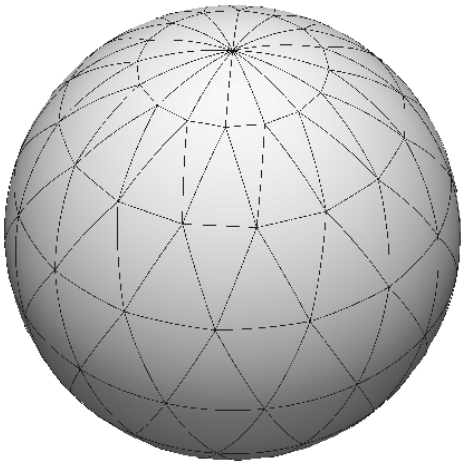

Figure 2.3.4: Triangulation of a $2-$sphere in $\mathbb{R}^3$ [9]. The topological space formed by the union of the triangles is homotopically equivalent to a $2-$sphere.

Lastly we turn our focus the way different homology groups are connected through appropriately defined maps, the following notation will be adopted:

- $\mathbf{Z}_q(\mathcal{L}) = \{\gamma_q : \gamma_q \in C_q(\mathcal{L}),\ \partial_q(\gamma_q) = 0\} \subseteq C_q(\mathcal{L})$ is the subgroup of the $q-$th chain group composed of $q-$cycles $\gamma_q$.
- $\mathbf{B}_q(\mathcal{L}) = \{\eta_q : \exists\ \zeta_{q+1} \in C_{q+1}(\mathcal{L}),\ \eta_q = \partial_{q+1}(\zeta_{q+1})\}$ is the subgroup of $\mathbf{Z}_q(\mathcal{L})$ composed of $q-$boundaries $\eta_q$.

Note that $\mathbf{Z}_q(\mathcal{L}) = ker\partial_q$ and $\mathbf{B}_q(\mathcal{L}) = im\partial_{q+1}$ so that $\mathbf{Z}_q(\mathcal{L})/\mathbf{B}_q(\mathcal{L}) = H_q(\mathcal{L})$.

**Definition 2.3.6** (Chain map)**.** *Let $\mathcal{L}, \mathcal{K}$ be two oriented simplicial complexes, the homomorphism between their $q-th$ chain groups:*

$$f_{\#}^{q} \ : \ C_q(\mathcal{L}) \to C_q(\mathcal{K})$$

*is said to be a chain map if*

$$f_{\#}^{q-1} \circ \partial_q = \partial_q \circ f_{\#}^{q}$$

A chain map is just a particular homomorphism between chain groups that preserves cycles and boundaries as stated by the following proposition:

**Proposition 2.3.1.** *Let $f_{\#}^{q} \ : \ C_q(\mathcal{L}) \to C_q(\mathcal{K})$ be a chain map, then:*

1. $f_{\#}^{q}(\mathbf{Z}_q(\mathcal{L})) \subseteq \mathbf{Z}_q(\mathcal{K})$
2. $f_{\#}^{q}(\mathbf{B}_q(\mathcal{L})) \subseteq \mathbf{B}_q(\mathcal{K})$

*Proof.* 1. Let $\gamma_q \in \mathbf{Z}_q(\mathcal{L}) \implies \partial_q \gamma_q = 0$, then by definition of chain map we get:

$$0 = f_{\#}^{q-1}(0) = (f_{\#}^{q-1} \circ \partial_q)(\gamma_q) = (\partial_q \circ f_{\#}^{q})(\gamma_q) = 0 \implies f_{\#}^{q}\gamma_q \in \mathbf{Z}_q(\mathcal{K})$$

which means that a cycle in $\mathcal{L}$ is mapped to a cycle in $\mathcal{K}$.

2. Let $\eta_q \in \mathbf{B}_q(\mathcal{L}) \implies \exists\, \zeta_{q+1} \in C_{q+1}(\mathcal{L}), \eta_q = \partial_{q+1}(\zeta_{q+1})$, then by definition of chain map we get:

$$f_{\#}^{q}(\eta_q) = (f_{\#}^{q} \circ \partial_{q+1})(\zeta_{q+1}) = (\partial_{q+1} \circ f_{\#}^{q+1})(\zeta_{q+1})$$

which by naming $f_{\#}^{q+1}(\zeta_{q+1}) = \theta_{q+1} \in C_{q+1}(\mathcal{K})$ implies that:

$$\exists\, \theta_{q+1} \in C_{q+1}(\mathcal{K}), f_{\#}^{q}(\eta_q) = \partial_{q+1}(\theta_{q+1})$$

which means that a boundary in $\mathcal{L}$ is mapped to a boundary in $\mathcal{K}$.

□

It is interesting to note that simplicial maps with certain properties naturally induce chain maps:

**Proposition 2.3.2** (Induced chain map)**.** *Let $f : \mathcal{L} \to \mathcal{K}$ be a simplicial map between oriented simplicial complexes that preserves both the dimension of the simplices and their orientation, that is given $v < w$ vertices in $\mathcal{L}$ then $f(v) < f(w)$ as vertices in $\mathcal{K}$. The following map is a chain map:*

$$\begin{aligned} f_{\#}^{q} \ : \ & C_q(\mathcal{L}) \to C_q(\mathcal{K}) \\ & \sum_i a_i \sigma_i \to \sum_i a_i f(\sigma_i) \end{aligned}$$

*Proof.* We need to check that $f_{\#}^{q-1} \circ \partial_q = \partial_q \circ f_{\#}^q$.

An oriented $q-$simplex is always of the form $\langle v_{i0}, \ldots, v_{iq} \rangle$ so that:

$$\partial_q \sum_i a_i \langle v_{i0}, \ldots, v_{iq} \rangle = \sum_i a_i \sum_{x=0}^{q} (-1)^x \langle v_{i0}, \ldots, \hat{v_{ix}}, \ldots, v_{iq} \rangle$$

applying $f_{\#}^{q-1}$ we get:

$$\begin{aligned} f_{\#}^{q-1} \circ \partial_q \left( \sum_i a_i \sigma_i \right) &= \sum_i a_i f_{\#}^{q-1} \left( \sum_{x=0}^{q} (-1)^x \langle v_{i0}, \ldots, \hat{v_{ix}}, \ldots, v_{iq} \rangle \right) \\ &= \sum_i a_i \sum_{x=0}^{q} (-1)^x \langle f(v_{i0}), \ldots, f(\hat{v_{ix}}), \ldots, f(v_{iq}) \rangle \end{aligned}$$

where in the last equivalence we used the fact that $f$ preserves the orientation so the image of the vertices are already ordered.

Applying the maps in the opposite order we get:

$$\begin{aligned} \partial_q \circ f_{\#}^q \left( \sum_i a_i \sigma_i \right) &= \partial_q \left( \sum_i a_i \sum_{x=0}^{q} (-1)^x \langle f(v_{i0}), \ldots, f(v_{iq}) \rangle \right) \\ &= \sum_i a_i \sum_{x=0}^{q} (-1)^x \langle f(v_{i0}), \ldots, f(\hat{v_{ix}}), \ldots, f(v_{iq}) \rangle \end{aligned}$$

where in the last equivalence we used the fact that $f$ preserves the dimension, so the boundary homomorphism of dimension $q$ can act on $f(\sigma_i)$. □

These maps in turn induce maps on homology groups in the following way:

**Definition 2.3.7** (Induced homology map)**.** *Let $\mathcal{L}, \mathcal{K}$ be an oriented simplicial complexes and $f : \mathcal{L} \to \mathcal{K}$ a simplicial map between them that preserves both the dimension and the orientation of the simplices. The following induced map between homology groups are defined:*

$$\begin{aligned} f_*^q \ : \ H_q(\mathcal{L}) &\to H_q(\mathcal{K}) \\ [\alpha] &\to [f_{\#}^q(\alpha)] \end{aligned}$$

Being a map between quotient groups, $f_*^q$ acts on equivalence classes, we thus need to check that it is well defined.

**Proposition 2.3.3.** *The induced map $f_*^q$ is well defined if $f_\#^q$ is the chain map naturally induced by $f$.*

*Proof.* We need to check that $f_*^q$ is independent of the representative's choice. Let $\gamma_q, \gamma_q'$ be two chains in the the same equivalence class, that is $\gamma_q - \gamma_q' \in \mathbf{B}_q(\mathcal{L}) \implies \gamma_q - \gamma_q' = \partial_{q+1}(\zeta_{q+1})$ for some $(q+1)-$chain $\zeta_{q+1}$. By definition of chain map we get the following:

$$f_\#^q(\gamma_q) - f_\#^q(\gamma_q') = f_\#^q(\gamma_q - \gamma_q') = (f_\#^q \circ \partial_{q+1})(\zeta_{q+1}) = (\partial_{q+1} \circ f_\#^{q+1})(\zeta_{q+1}) \in \mathbf{B}_q(\mathcal{K})$$

This shows that any two representatives in $H_q(\mathcal{L})$ are mapped to the same equivalence class in $H_q(\mathcal{K})$. The map $f_*^q$ is thus well defined. □

**Theorem 2.3.3.** *Let $f, g$ be simplicial maps, then for any given dimension $q$:*

- $(f \circ g)_\#^q = f_\#^q \circ g_\#^q$
- $(f \circ g)_*^q = f_*^q \circ g_*^q$

Given a family of simplicial complexes we now have all the instruments to characterize and connect them and their associated homology groups via appropriate maps. This is enough to study how a simplicial complex evolves when varying the scale parameter $\varepsilon$.

## 2.4 Variable scale analysis

The procedure consists in fixing a particular dimension, say 1, and keeping track of how the Betti number $\beta_1$ evolves. One can already imagine that by changing continuously the value of $\varepsilon$, some new $1-$dimensional loops will appear while other will be absorbed, typically in $2-$simplices.

### 2.4.1 Persistent homology

*Persistent homology* can better quantify the previous statement: a collection of the same type of simplicial complexes built over a range of scale parameters, called a *filtration*, can be shown to be intimately related to a collection of homology groups. By appropriately linking the various simplicial complexes, a connection between the homology groups naturally appears in the form of induced homology maps.

Scaling a simplicial complex often results in a new complex that is contained in the original one as the following example shows:

$$\check{\text{C}}\text{ech}(\mathcal{S}, \varepsilon_1) \subseteq \check{\text{C}}\text{ech}(\mathcal{S}, \varepsilon_2) \subseteq \check{\text{C}}\text{ech}(\mathcal{S}, \varepsilon_3) \quad \varepsilon_1 \leq \varepsilon_2 \leq \varepsilon_3$$

A visual representation of this chain of inclusions is given in figure 2.4.1.

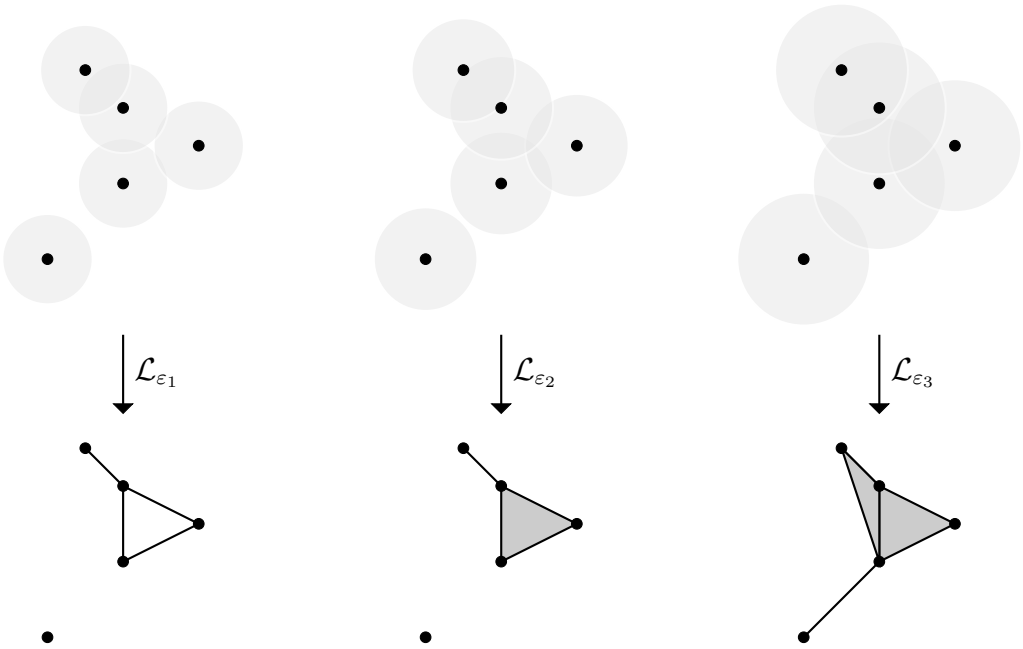


Figure 2.4.1: Different scale parameters results in different simplicial complexes forming a filtration.

**Definition 2.4.1** (Filtration)**.** *Let $\{\mathcal{L}_\varepsilon\}_{\varepsilon\geq 0}$ a collection of simplicial complexes parameterized by a scale factor $\varepsilon$ such that $\mathcal{L}_{\varepsilon_s} \subseteq \mathcal{L}_{\varepsilon_t}\ \forall\ \varepsilon_s \leq \varepsilon_t$. The collection of simplicial complexes together with the inclusion maps (i.e. simplicial maps) $i_{s,t} : \mathcal{L}_{\varepsilon_s} \hookrightarrow \mathcal{L}_{\varepsilon_t}$ is called a filtration of the simplicial complex $\mathcal{L}_\varepsilon$.*

Note that a filtration can be either continuous or discrete depending on how the scale parameter $\varepsilon$ vary.

It is now possible to connect two homology groups of two simplicial complexes in a filtration through the map induced by the filtration's inclusion map. An inclusion map is a map that leaves the simplices unaltered and thus preserve orientation and dimension, which as we have already shown is all it is needed to induce a map between homology.

Take as an example the following filtration:

$$\mathcal{L}_{\varepsilon_1} \xhookrightarrow{i_{1,2}} \cdots \xhookrightarrow{i_{m-1,m}} \mathcal{L}_{\varepsilon_m}$$

where one clearly has $i_{n,n+1} \circ i_{n+1,n+2} = i_{n,n+2}$, the following chain of maps follows:

$$H_q(\mathcal{L}_{\varepsilon_1}) \xrightarrow{i_{1,2*}} \cdots \xrightarrow{i_{m-1,m*}} H_q(\mathcal{L}_{\varepsilon_m})$$

where thanks to theorem 2.3.3 one has $(i_{s,t} \circ i_{t,u})_* = i_{s,u*}$.

**Definition 2.4.2** (Persistent homology)**.** *Let $\{\mathcal{L}_\varepsilon\}_{\varepsilon\geq 0}$ be a filtration of a simplicial complex ending at $\overline{\varepsilon}$ and $i_{s,t*}$ the induced map between the $q-th$ homology groups of the $s-th$ and $p-th$ simplicial complexes. The following objects are defined:*

$$H_q^{s,t} = im(i_{s,t*}) \quad \forall\, 0 \leq s \leq t \leq \overline{\varepsilon}$$
$$\beta_q^{s,t} = rank(i_{s,t*})$$

*The group $H_q^{s,t}$ is called the $q-$dimensional persistent homology group and the number $\beta_q^{s,t}$ is the $q-$dimensional persistent Betti number.*

The persistent Betti numbers are an indicator of how the non trivial homology elements evolve. Taking again figure 2.4.1 as an example, we note that $\beta_1 = 1$ for the first one and $\beta_1 = 0$ for the second one. We compute the persistent Betti number $\beta_1^{1,2} = 0$ meaning that every $1-$dimensional gap present in $\mathcal{L}_{\varepsilon_1}$ has collapsed in $\mathcal{L}_{\varepsilon_2}$.

### 2.4.2 Barcodes

The general description of variable scale analysis done to this point can be reviewed to get a more straightforward visual interpretation thanks to the concept of a filtration's *barcode*.

**Definition 2.4.3** (Barcode)**.** *Let $\{\mathcal{L}_\varepsilon\}_{\varepsilon\geq 0}$ be a filtration of a simplicial complex. Once a dimension $q$ has been fixed, each unique $q-$homology elements born at scale $\varepsilon = s$ and terminated ad scale $\varepsilon = t$ is assigned an interval $[s, t)$. The representation of all the intervals of every $q-$homology elements that appears at some point in the filtration is called the filtration's barcode.*

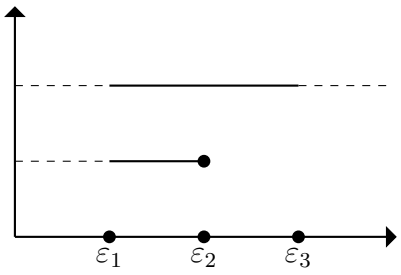


Figure 2.4.2: A basic barcode for the filtration in figure 2.4.1 representing the evolution of connected components ($0-$homology elements). Dashed intervals represent possible previous and successive complexes in the filtration. The lower interval is relative to the single vertex that gets absorbed in $\mathcal{L}_{\varepsilon_2}$, the upper one is relative to the triangle-edge component that *persists* throughout the steps of the filtration.

Adding a $k-$simplex to the simplicial complex has one of two possible outcomes: either creating a new $k-$dimensional homology class or destroying one. This is basically what an interval of the filtration's barcode represent, it is so customary to define the following.

**Definition 2.4.4** (Creator/destroyer simplices)**.** *Let $\{\mathcal{L}_\varepsilon\}_{\varepsilon\geq 0}$ be a filtration of a simplicial comlpex and $[s,t)$ be an interval of that filtration's barcode. There exists an associated pair of simplices to this interval:*

- *$\sigma \in \{\mathcal{L}_\varepsilon\}_{\varepsilon\geq 0}$ called the creator, which is the simplex that generates a new homology class associated to $[s,t)$.*
- *$\tau \in \{\mathcal{L}_\varepsilon\}_{\varepsilon\geq 0}$ called the destroyer, which is the simplex that merges the homology class associated to $[s,t)$ back into another class.*

The length of the interval associated to an homology element is known as its *lifetime*. The lifetime of an homology element can vary in a wide range of values, but the most interesting case is when it is said to be *semi-infinite*. A lifetime is semi-infinite when its associated barcode's segment is of the form $[s,+\infty)$, where the value $+\infty$ is not to be intended as actually infinite but indicates the largest value of the filtration's scale parameter. It should be noted that this kind of interval can be interpreted as those ones with no associated destroyer simplex.

This kind of homology elements are often an indications of actual properties of the dataset since they persist throughout the filtration and it is then plausible to say that they are not caused by noise or errors.

Since a barcode is just a direct visualization of persistent homology, we should expect a very close connection between it and persistent Betti numbers. That is indeed the case and this connection can be summarized in the following points:

- $\beta_q^{s,t}$ represents the number of intervals containing $s$ and extending at least through $t$.
- For a discrete filtration, $\beta_q^{s,s} - \beta_q^{s-1,s}$ represents the number of intervals starting precisely at $s$.
- For a discrete filtration, $\beta_q^{t-1,t-1} - \beta_q^{t-1,t}$ represents the number of intervals ending precisely at $t$.
- For a discrete filtration, $\beta_q^{s,t} - \beta_q^{s-1,t}$ represents the number of intervals starting at $s$ and extending at least through $t$.

And for a more precise quantification for the number of bars for a given interval:

- $\mathbf{n}_{s,t} = \beta_q^{s,t-1} - \beta_q^{s-1,t-1} - (\beta_q^{s,t} - \beta_q^{s-1,t})$ represents the number of intervals starting precisely at $s$ and terminating precisely at $t$.

- $\mathbf{n}_{s,\infty} = \beta_q^{s,m} - \beta_q^{s-1,m}$, where $m$ is the largest value of the scale parameter that characterize the filtration, is the number of intervals starting precisely at $s$ and extending to the end of the filtration, in other words the number of semi-infinite intervals.

# Chapter 3

# Persistent homology applied to the cyclooctane dataset

The broad overview of the analysis of the cyclooctane dataset $\mathcal{S}$ showed in chapter 1 can now be supported by the topological framework built in chapter 2. First a $4-$skeleton of the Vietoris-Rips complex is built and Betti numbers are calculated [6]. A more elaborated inspection is then performed by optimizing the computational complexity through the use of *tidy sets* [6] which are topological objects that generalize the concept of simplices and that will not be further examined.

## 3.1 Preliminary overview

### 3.1.1 Number of simplices in a generic dataset

Recalling how a Vietoris-Rips complex is built over a dataset $\mathcal{S}$, it is interesting to compare the upper bound of the number of simplices that can be generated for a given number of elements in $\mathcal{S}$ and the actual number that will be generated by the cyclooctane dataset.

**Proposition 3.1.1.** *Given a dataset $\mathcal{S}$ of $n$ elements, the maximal number of simplices of dimension $d \leq p$ in a Vietoris-Rips complex over $\mathcal{S}$ is:*

$$\nu_{\leq p} = -\binom{n}{p+1} {}_2F_1(1, -n+p+1; p+2; -1) + 2^n - 1$$

*Where ${}_2F_1(a, b; c; z)$ is the hypergeometric function.*

*Proof.* The proof is a straightforward calculation of all the possible combination of vertices. Since the Vietoris-Rips complex is based off a graph, a $p-$simplex will be generated

if $p$ vertices are connected independently of their order. That's just the combination of $n$ objects grouped in $p$, in other words the binomial coefficient $\binom{n}{p}$.

Summing the coefficients over values of $p$ ranging from 1 to a desired dimension of simplices concludes the proof:

$$\sum_{k=1}^{p} \binom{n}{p} = \nu_{\leq p}$$

□

For a 4−skeleton, as the one that will be constructed, the upper limit has a more manageable expression:

$$\nu_{\leq 4} = \frac{1}{24}(n^3 - 2n^2 + 11n + 14)$$

The upper bound scales up pretty rapidly as the number of vertices grows even for a 4−skeleton and, for the dataset $\mathcal{S}$ in exam composed of 6400 experimental points, we get an upper bound of $\nu^{\mathcal{S}}_{\leq 4} \approx 7 \cdot 10^{13}$.

It is clear that the enormous number we got is just an upper bound, but this shows why Vietoris-Rips, although it being a good starting point, will need some optimization that will permit to go beyond a basic 4−skeleton. This will result in the introduction of tidy sets in section 3.3.

### 3.1.2 Vietoris-Rips complex over $\mathcal{S}$

Appropriate algorithms for the construction of the Vietoris-Rips complex are applied to the cyclooctane's dataset $\mathcal{S}$. The $\varepsilon$−neighborhood graph is built with a value of $\hat{\varepsilon} = 0.4$ which has been chosen after evaluating the maximum interpoint distance between pairs of closest points at $\varepsilon_{max} = 0.18$. The graph has a total of 76 657 edges over the 6400 vertices. The maximal number of edges is the number of distinct 2−elements subsets of $\mathcal{S}$, which is just $N_{max} = \frac{6400 \cdot 6399}{2} = 20\ 476\ 800$. The graph built at $\hat{\varepsilon}$ contains only about 0.4% of the possible edges meaning that the data points are quite sparse at this scale.

Computing the maximal cliques of the graph results in the construction of the Vietoris-Rips complex. In this particular case the 4−skeleton is built resulting in a simplicial complex with over 3 million simplices.

## 3.2 4-skeleton analysis

### 3.2.1 Evolution of simplices

The variable scale analysis starts by building the filtration of the Vietoris-Rips complex over $\mathcal{S}$ in the range $0 \leq \varepsilon \leq 0.4$. Figure 3.2.1 shows how the number of simplices grows when increasing the value of $\varepsilon$.

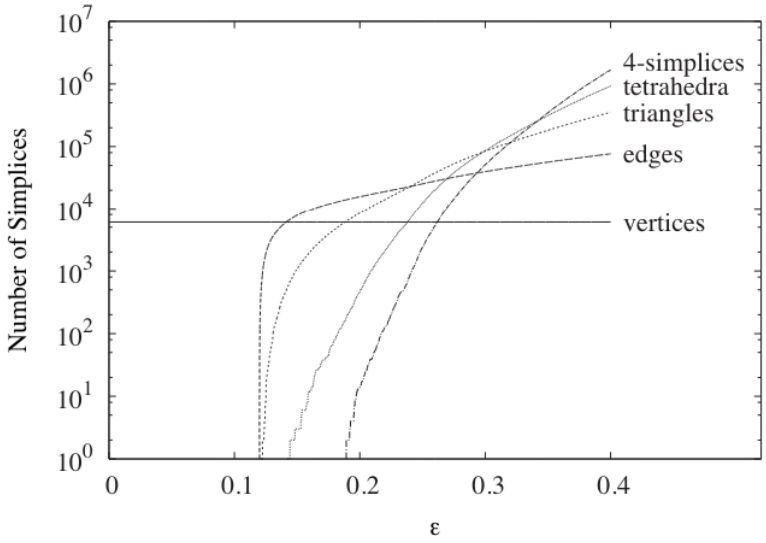


Figure 3.2.1: The number of simplices in the 4−skeleton of the Vietoris-Rips complex as a function of the scale parameter $\varepsilon$ [6].

The vertex line is horizontal since the number of them obviously doesn't depend on the scale parameter. The overall size of the complex has instead an approximately exponential growth.

### 3.2.2 Matrix based computation of Betti numbers

Homology groups calculation on simplicial complexes of these sizes are not trivial. To simplify the matter, in TDA it is customary to build chain groups over the field $\mathbb{Z}_2$. To calculate homology groups one needs to find the kernel and the image of two boundary homomorphisms which, them being linear maps on finite dimensional spaces, can be represented by matrices. This particular choice of field results in the entries of the matrices being either 1 or 0.

To better understand this concept, we explicitly calculate the Betti number $\beta_1$ on the simplicial complex $\mathcal{L}_{\varepsilon_1}$ in figure 2.4.1, expecting a value of 1. Figure 3.2.2 adds a label

to the vertices of the complex, the orientation of the edges are induced by alphabetical order (e.g. $\langle b, c\rangle$).

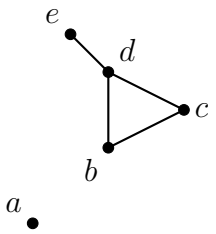


Figure 3.2.2: Simplicial complex $\mathcal{L}_{\varepsilon_1}$ in figure 2.4.1 with labelled vertices.

To compute $\beta_1$ we need to find the matrix forms of the operators $\partial_1, \partial_2$. We start by noting that $\partial_2$ is trivial since there are no $2-$simplices implying that $im\partial_2 \equiv 0$. The calculation of $ker\partial_1$ is a bit more involved, the notation for oriented simplices will be simplified in this section: $\langle b, c\rangle \implies bc,\ \langle c, b\rangle \implies cb$.

$\partial_1$ is a map from $C_1(\mathcal{L}_{\varepsilon_1})$ to $C_0(\mathcal{L}_{\varepsilon_1})$. The most simple basis of $C_1(\mathcal{L}_{\varepsilon_1})$ is composed of the $1-$chains $bc, cd, db, de$, giving a generic chain the vector form:

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = z \in C_1(\mathcal{L}_{\varepsilon_1}) \quad u_1, u_2, u_3, u_4 \in \mathbb{Z}_2$$

The coefficients are associated to the basis chains in the following way: $u_1 \to bc,\ u_2 \to cd,\ u_3 \to db,\ u_4 \to de$. This is completely equivalent to the way vectors in $\mathbb{R}^n$ are identified by specifying the $n$ coefficients that multiply the basis vectors $\vec{e}_i$. It is possible to arrange the vertices of the complex in a vector of $C_0(\mathcal{L}_{\varepsilon_1})$ in a similar way:

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{bmatrix} = s \in C_0(\mathcal{L}_{\varepsilon_1}) \quad w_1, w_2, w_3, w_4, w_5 \in \mathbb{Z}_2$$

The coefficients are associated to the basis vertices in the following way: $w_1 \to a,\ w_2 \to b,\ w_3 \to c,\ w_4 \to d,\ w_5 \to e$.

Applying $\partial_1$ to the generic chain $z = u_1 bc + u_2 cd + u_3 db + u_4 de$, recalling that the map is linear, we get the following:

$$\begin{aligned}\partial_1(z) &= u_1\partial_1(bc) + u_2\partial_1(cd) + u_3\partial_1(db) + u_4\partial_1(de)\\ &= u_1(c-b) + u_2(d-c) + u_3(b-d) + u_4(e-d)\\ &= a(0) + b(u_3-u_2) + c(u_1-u_2) + d(u_2-u_3-u_4) + e(u_4)\end{aligned}$$

which takes the vector form:

$$\partial_1 \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ u_3 - u_1 \\ u_1 - u_2 \\ u_2 - u_3 - u_4 \\ u_4 \end{bmatrix}$$

In $\mathbb{Z}_2$ we lose the distinction between $+$ and $-$ sign as $-1 \equiv 1\ mod(2)$, moreover $1+1=2\equiv 0\ mod(2)$. This leads to the following matrix form of $\partial_1$:

$$\partial_1(z) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ u_3 + u_1 \\ u_1 + u_2 \\ u_2 + u_3 + u_4 \\ u_4 \end{bmatrix}$$

In this form the kernel of $\partial_1$ is easily computed, finding it is the $1-$dimensional vector sub-space of $C_1(\mathcal{L}_{\varepsilon_1})$ spanned by the vector $k$ with the property:

$$\partial_1(k) = \partial_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} = 0$$

Recalling the definition of homology group and Betti number:

$$H_1(\mathcal{L}_{\varepsilon_1}) = ker\partial_1 \big/ im\partial_2 = ker\partial_1 \implies \beta_1 = dim(H_1(\mathcal{L}_{\varepsilon_1})) = dim(ker\partial_1) = 1$$

obtaining again the expected result as $\beta_1$ correctly identifies the only $1-$dimensional gap in the complex.

Note that it is no coincidence that the vector $k$ has those precise components. Taking a closer look at the chain associated to $k$ we find that $k = bc + cd + db$. That is exactly the chain that bounds the gap in the complex.

Applying this method to the Vietoris-Rips $4-$skeleton filtration over $\mathcal{S}$ gives the result shown in figure 3.2.3.

A few comments on the result:

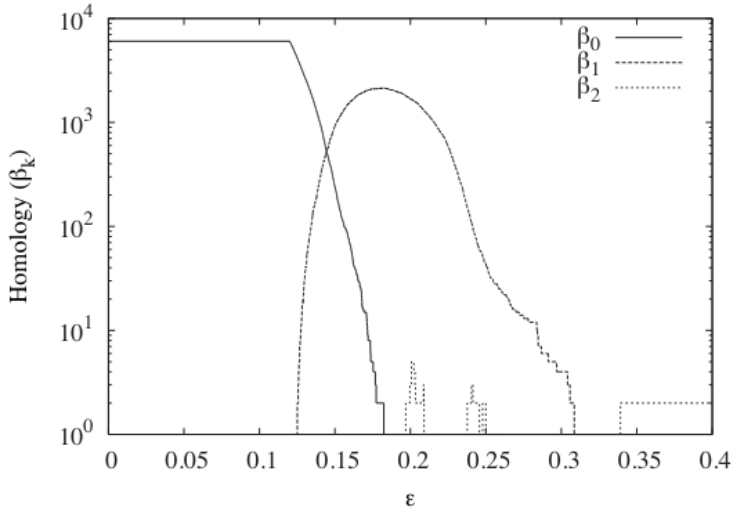


Figure 3.2.3: Betti numbers $\beta_n$ for $0 \leq n \leq 2$ of the Vietoris-Rips 4−skeleton filtration over $\mathcal{S}$ [6]. $\beta_3$ is not plotted as it is identically 0.

1. $\beta_3$ is identically 0
2. $\beta_0 = 0$ for $\varepsilon > 0.18 = \varepsilon_{max}$
3. $\beta_n$ for $\varepsilon \geq 0.3391 = \varepsilon_g$ are equal to the ones of the geometric reconstruction of the conformation space in figure 1.2.1.

Betti numbers of dimension greater of the topological space's dimension that they are measuring are always 0. (1) is then an early indication that the conformation space is arranged on a 2−dimensional surface in $\mathbb{R}^{24}$. (2) is an expected result as well, it indicates that the complex becomes completely connected at the maximum interpoint distance between pair closest of points. (3) gives the indication that $\varepsilon_g$ is the scale at which the complex approximate the conformation space in an optimal way.

The 4−skeleton analysis concludes with the last tool introduced in chapter 2. Figure 3.2.4 shows the Vietoris-Rips 4−skeleton filtration's barcode for $\beta_1$ [6].

The barcode has 3475 non empty intervals, only one of which is semi-infinite. This translates to an equal amount of 1−dimensional gaps in the complex throughout the filtration. The mean lifespan of the 1−dimensional gaps (i.e. the mean length of a segment) is $\bar{l} = 0.0382$ with a standard deviation $\delta l = 0.0260$. The only semi-infinite interval has a length of $l_\infty = 0.2545$ which is more than 8 standard deviation over the mean. This leads to the hypothesis that there exist a single 1−dimensional loop in the actual conformation space which is not a result of noise in the dataset $\mathcal{S}$.

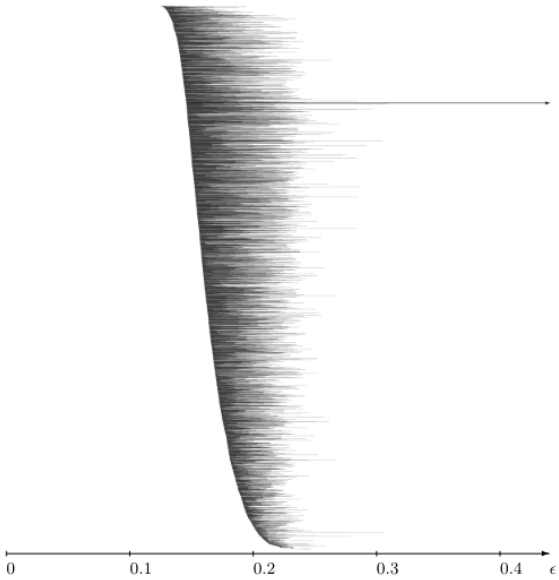


Figure 3.2.4: $\beta_1$ barcode. The arrowed segment indicates a semi-inifinite interval [6].

## 3.3 Deeper study using tidy sets

The approach used to get higher dimensional information on $\mathcal{S}$ is to optimize the topological structure built on it in such a way that it lowers considerably the amount of computational power needed. This is done through the use of generalized simplicial complex called *simplicial sets* and in particular the type derived by simplicial complexes through *collapse* procedures called *tidy sets*. Figure 3.3.1 shows the possible simplicial sets that can be built from a $2-$simplex. It is evident that only the first simplicial set, which is the standard $2-$simplex, has the usual structure of simplex.

Once this has been achieved, higher dimensional Betti numbers are computed without the need of an increased computational power.

What follows is a very broad overview of the collapse procedures used to build a tidy set from a simplicial complex. The scope of the next section is only to give a general idea of why a tidy sets is computationally more efficient than a simplicial complex and does not aim to give a precise and detailed description of them.

### 3.3.1 Tidy sets

A simplex is said to be *acyclic* if it does not contain any holes and thus every homology group of dimension greater than 0 is trivial, which means that the simplex has no $k-$dimensional gaps.

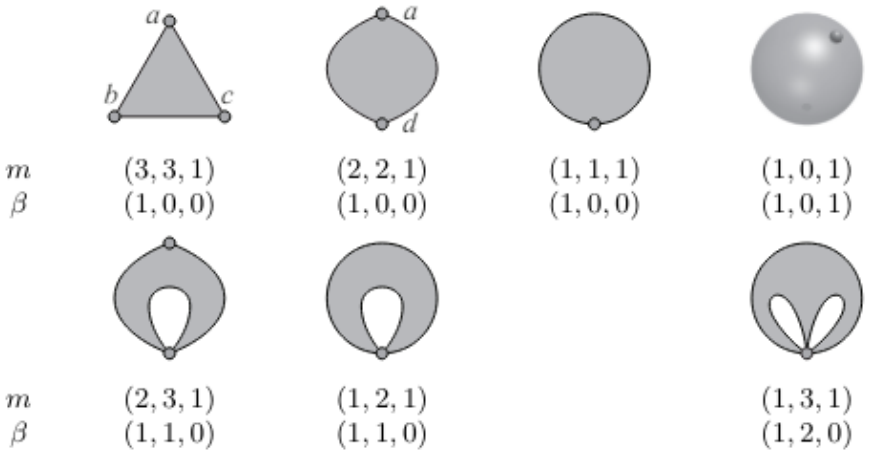


Figure 3.3.1: The vector $m$ counts the non-degenerate simplices of dimensions $0, 1, 2$ while the vector $\beta$ contains the Betti number $\beta_0, \beta_1, \beta_2$. [6].

Given a simplicial complex, consider the following collapse procedures:

- *Trimming*, which consists in removing all the simplices which have acyclic intersection with the rest of the complex.
- *Thinning*, which consists in collapsing to a single point all of the remaining acyclic simplices.

A simplicial complex which undergoes this collapse procedures is not in general a simplicial complex but a simplicial set.

**Definition 3.3.1** (Tidy set)**.** *Let $\mathcal{L}$ be a simplicial complex. The tidy set $T(\mathcal{L})$ is the simplicial set obtained through the trimming and thinning procedures that is minimal with respect to them.*

Figure 3.3.2 shows how a basic simplicial complex is turned into a tidy set through trimming and thinning procedures. First, the upper most $2-$simplex is removed since its intersection with the complex is an edge which is acyclic. The remaining $2-$simplex is itself acyclic and it is collapsed into one of remaining vertices. We remain with a complex composed of 6 acyclic edges. All of the edges are collapsed into a single common point. It is not possible to perform further trimming and thinning, we are left with a tidy set.

The key feature of these collapse procedures is that the tidy set obtained has the same exact characteristics, from an homology point of view, as the original simplicial complex despite the fact that it is, in a sense, a much *cleaner* and *simple* object. Taking a look back at figure 3.3.2 it is evident that the collapse procedures kept unchanged the only gap in the original simplicial complex.

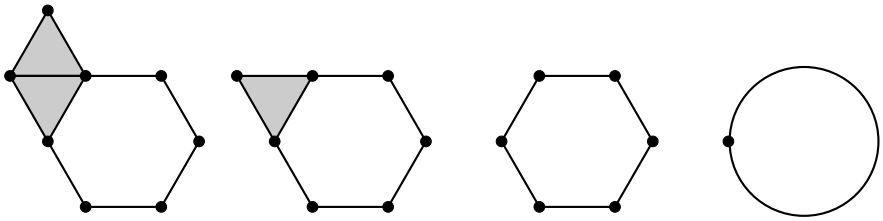

Figure 3.3.2: Collapse procedures applied to a simplicial complex. Note that, after the thinning procedure, what we are left with does not resemble a familiar simplicial complex, remarking the fact that a tidy set is indeed a more general simplicial set.

### 3.3.2 Higher dimensional Betti numbers

Revising the 4−skeleton of the Vietoris-Rips complex originally used, the collapse procedures trim 87% and thin 8% of the simplices. In terms of cliques the size of the tidy set is just about 5% of the original complex. A quick check on the Betti numbers confirms that no information has been lost as $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 2$.

This new approach can be exploited to extend the computation to the 11 dimensional tidy set, since the largest uncollapsed clique has size 12, and the relative higher dimensional Betti numbers. Once again we retrieve the familiar results for $\beta_{0,1,2}$. As hinted by the Betti number $\beta_3 = 0$ obtained from the old 4−skeleton, the results shows that the homology is trivial also for higher dimensions, that is $\beta_n = 0$ for $3 \leq n \leq 11$. This is a further evidence confirming the hypothesis that $\mathcal{S}$ is intrinsically a 2−dimensional surface.

# Bibliography


[1] Yashbir Singh et al. "Topological data analysis in medical imaging: current state of the art". In: *Insights into Imaging* (2023).

[2] Anuraag Bukkuri, Noemi Andor, and Isabel K. Darcy. "Application of Topological Data Analysis in Oncology". In: *Frontiers in Artificial Intelligence* (2021).

[3] Christopher Shultz. *Applications of Topological Data Analysis in Economics*. 2023. URL: https://ssrn.com/abstract=4378151.

[4] José Serrano Martínez et al. "Structure determination of di-$\mu$-hydroxo-bis[(2-(2-pyridyl)phenyl-$\kappa$ 2 N , C 1 )palladium(II)] by X-ray powder diffractometry". In: *Acta Crystallographica Section B-structural Science - ACTA CRYSTALLOGR B-STRUCT SCI* (2007).

[5] W. Michael Brown et al. "Algorithmic dimensionality reduction for molecular structure analysis". In: *The Journal of Chemical Physics* (2008).

[6] Afra Zomorodian. "Topological Data Analysis". In: *Proceedings of Symposia in Applied Mathematics* (2012).

[7] Shawn Martin et al. "Topology of cyclo-octane energy landscape". In: *The journal of chemical physics* (2010).

[8] Tttrung. *Klein bottle made with gnuplot*. URL: https://commons.wikimedia.org/w/index.php?curid=960446.

[9] J. Beckmann, Hrushikesh Mhaskar, and Jürgen Prestin. "Local numerical integration on the sphere". In: *GEM - International Journal on Geomathematics* 5 (Nov. 2014), pp. 143–162. DOI: 10.1007/s13137-014-0065-1.

[10] Žiga Virk. *Introduction to Persistent Homology*. Založba UL FRI, 2022.