



AIMS

**African Institute for
Mathematical Sciences
CAMEROON**

Topological Data Analysis: Fundamental Aspects for Traditional Data Science Approaches

Aurelie Jodelle Kemme (jodelle.kemme@aims-cameroon.org)
African Institute for Mathematical Sciences (AIMS)
Cameroon

Supervised by: Prof Franck Kalala Mutombo
AIMS Senegal and University of Lubumbashi, DRC

07 May 2020

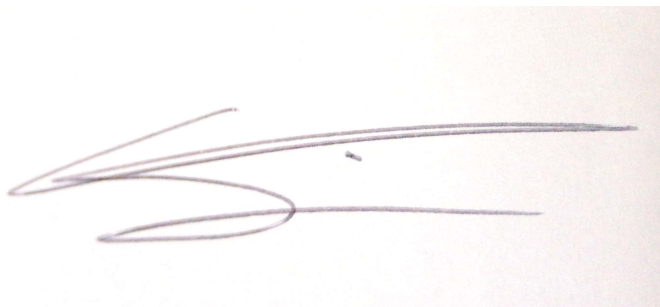
Submitted in Partial Fulfillment of a Structured Masters Degree at AIMS-Cameroon

Abstract

Topological data analysis (TDA) is a new and flourishing category of artificial intelligence, a method of data science for the analysis, and understanding of very high dimensional data using tools derived from algebraic topology and combinatorial analysis. The failure experienced by other methods such as machine learning in presenting the geometric structure of information of data, has given rise to this new approach to data analysis. The objection to this work is to lift a veil on this new approach to data analysis specializing in the study of very high dimensional data by describing the shapes of this data. In this project, we present a general introduction to the tools needed for the application of TDA and a description of some methods it uses such as the Mapper Algorithm and the Persistence Homology method.

Declaration

I, the undersigned, hereby declare that the work contained in this essay is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

A handwritten signature in purple ink, appearing to read 'Aurelie Jodelle Kemme', is written on a light-colored background.

Aurelie Jodelle Kemme, 07 May 2020.

Contents

Abstract	i
1 Introduction	1
1.1 Structure of study	5
1.2 Literature Review	5
2 Preliminaries	7
2.1 Metric spaces	7
2.2 Topological Spaces	13
2.3 Covers and simplicial complexes	15
2.4 Geometric and abstract simplicial complexes.	17
3 TDA Methods	21
3.1 Mapper Algorithm	21
3.2 Geometric Reconstruction and Homological Inference	22
3.3 Persistence Homology	26
3.4 Some Examples of Persistence Homology	28
4 Conclusion	30
Acknowledgements	31
References	32

1. Introduction

Topological Data Analysis, also abbreviated (TDA), is a recent growing field in data science that emerged from several works in applied algebraic topology and computational geometry [13] to extract relevant pieces of information from complex data. It aims at providing well-founded mathematical tool from pure and applied mathematics, statistical and algorithmic methods to exploit the topological and underlying geometric structures for exploratory data analysis and machine learning [10, 13]. The understanding of high dimensional data by using topological tools make TDA new and innovative tools for complex data to be used in complementarity to other machine learning and artificial intelligence tools[5]. Its constantly growing influence in machine-learning is the result of the variety of theories defined in topology.

Topology is a very exciting field of mathematics which study the shape of objects and has many applications in many different real world problems where one of them is the analysis and understanding high dimension of complex data set.

The idea of topology was introduced in 18th century by the great Swiss mathematician Leonhard Euler [17] and the challenge started with the famous Königsberg bridge and the Figure [1.1a] is showing the schematic diagram of the problem. Which was about the German city of Königsberg (now it is Russian Kaliningrad) was situated on the river Pre-gel. It had a park situated on the banks of the river and two islands. Mainland and islands were joined by seven bridges. A problem was whether it was possible to take a walk through the town in such a way as to cross over every bridge once, and only once [17].

To solve the problem, the great Swiss mathematician Leonhard Euler took all the information about the bridges, the river and converted it into a simple network. We can see from the Figure [1.1b] the simple representation of the problem into the network.

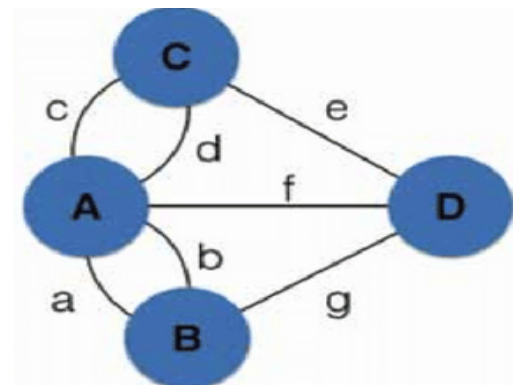
By visualizing the network graph as is shown in the Figure [1.1b], he realized that it was impossible to take a walk through all the edges in such a way that to cross an edge only once. Therefore, impossible to take a walk through the town in such a way as to cross over every bridge once, and only once.

The general idea behind the Königsberg problem was to show how a difficult can be turn out into a simple network to get the solution.

The interest of studying and understanding of high dimensional data is the fact that industries nowadays perceive data as an essential commodity and fuel for the sustainability of their businesses. They churn



(a) A schematic diagram of the seven Königsberg bridges.



(b) Network of the problem in figure

Figure 1.1: Model of the Königsberg problem

raw data into a meaningful product and uses it to draw insights for decision making, for better functioning of the industry. They do this by using both heavy and simple computational tools which have proven to be very effective over time. But In general, modern data are complex and often come as (sampling of) metric spaces or sets, and those spaces are endowed with a similarity measure with, possibly complex topological or geometric structure [5]. We say that data is complex whenever the underlying phenomena that the data seeks to capture are complex. But data can also be complex because of big data (one billion rows or more than five hundred features) or because it has rich features (features are what statisticians call variables or columns and mathematicians dimension) or because of both reasons. We can take an example of genetic data from cancer patients with more than five hundred thousand features.

The general aim of the study of complex data is to capture that information from the many connections hidden in this complex data set. The starting point was the fact that data has shape and shape has meaning, therefore, the challenge is how to extract meaning from complex data. Generally, within the data, only some features are relevant depends on the study. Then, the question is how can we get the features that one expects us to highlight from the data? That is where we started thinking about how to analyze the data by using topological tools and we called this method the Topological Data Analysis method (TDA). The main idea was data has shape and shape has meaning, and since topology study the shape, TDA is going to provide a technical sense, geometry summaries of the data problem.

Therefore, we can define TDA in simple words as a method from data science that uses some methods and algorithms from algebraic topology and geometric tools to study and better understand complex data.

As machine learning method uses some methods such as supervised learning and unsupervised learning, TDA also uses some methods [5] such as the Mapper method, whose goal is to give the summary of data, the persistent homology method, where the aim is to give the geometric structure of the information, the persistent landscapes which are sequences of piecewise-linear functions, whose goal is to bring a new topological summary approach for data that could solve the issue of combining the main tool of the subject, the barcode or persistence diagram with statistics and machine learning. Where the barcode and the persistence diagram are the two standard topological summaries of data.

But among the most widespread theories used by TDA, the most popular methods are the Mapper method and the persistent homology method which can be considered as an unsupervised machine learning technique to analyze high-dimensional data.

The Mapper algorithm works in general as for a given data set, it will use a map to provide a summary of the data and a cover that will provide an easy way to visualize the summary of data.

The persistence homology is a method aimed at computing topological features of space at different spatial resolutions [13].

The idea behind the summary of data by TDA was to bring another way of summarizing the data to improve what the other methods have proposed because, In any field, we are often interested in summarizing and comparing data sets when faced with a complex data analysis task. And often, it is cumbersome to work with the data directly because of the higher dimension. This is where persistent homology comes into play.

Persistence homology will take a data set and encode topological multi-scale features into a point cloud in a plane called persistence diagram, that will allow us to easily visualize and compare the (topological) features present in a data set [8].

We saw in the previous paragraph that TDA is using some methods where the aim is to infer relevant qualitative and quantitative topological structures directly from complex data. Now how can we be trustful about the coherence of these methods? or how can we get the confidence regions for topological features? and how can we deal with outliers and providing robust methods?. For answers at these questions, we will use a statistical deterministic approach of TDA.

Deterministic approaches are just the opposite of random approaches. It gives us information that some future event can be calculated exactly without the involvement of randomness.

Deterministic approaches have mostly relied on to TDA and topological inference. But the challenge is these deterministic approaches do not take into account the random nature of data and the intrinsic variability of the topological quantity they infer [5].

So with a statistical approach of TDA, data are considering to be generated from unknown distribution and the inferred topological features by TDA methods are seen as estimators of topological quantities describing an underlying object so the aim of the statistical approach is to

1. Provide confidence regions for TDA and discourse significance of the estimated topological quantities [5];
2. Central tendency for persistent homology [14];
3. Dealing with outliers and providing Robust methods fo TDA [5];
4. Representations of persistence in Euclidean spaces [14];
5. Consistency and convergence of TDA methods [5];
6. Statistical analysis of Mapper [14];
7. Develop kernels for topological descriptors[14];
8. selecting relevant scales at which the topological phenomenon should be considered, as a function of observed data [5].

Although a statistical approach has provided some solutions to the confidence of TDA method, but the powerful of TDA among other methods such as machine learning method is still to demonstrate.

Machine learning is a collection of techniques using some methods for visualization, prediction, classification and some others relevant tasks in order of better understanding data. In visualization techniques, projections of the data points are produced in two or sometimes three dimensions and the plots on these coordinates are made in the usual way. Therefore, the visualization techniques will reveal the scatter plot methods [2]. In the projection techniques processes, we have principal components analysis, multi-dimensional scaling and projection pursuit. However, with the extremely high dimension data, the scatter plot provided by machine learning techniques will be very hard to visualize and understand so, the TDA will enable us to construct a topological network by using the machine learning techniques result where the aim is to move from extremely high dimension data to its reduce dimension one. It will, therefore, transform the representation of the scatter plot produced by machine learning techniques to a representation of scatter plot easier to understand and interact with. For the representation of this scatter plot by using TDA, similar data will be grouped into nodes and connection between nodes which have data points in common will be represented with an edge. Where each node represents multiple data points.

The Figure [1.2] shows on the left the scatter plot produced by machine learning techniques, and we can see that the visualization of similarity between data points is still hard although this scatter plot is showing different clusters. In the middle, we have a network where each node represents a data point and it looks like a spider-web so it's very difficult to understand. On the right, it is a topological network of the scatter plot provided by machine learning. Here, each node represents a collection of similar data

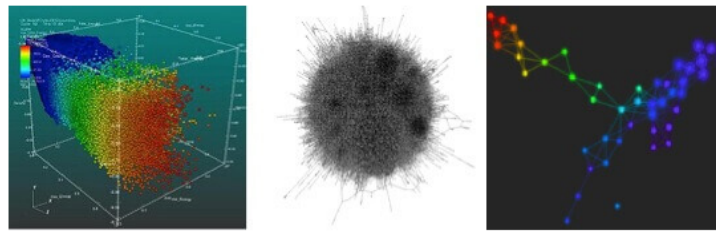


Figure 1.2: Different networks of a data points with different techniques [2].

points and that we can easily visualize the connections between those nodes which are given us the relationship between them.

Other challenges with machine learning models are although there have achieved tremendous success in text analysis, image analysis, computer vision, speech recognition, etc. Their application in high-dimensional complex systems has been obstructed significantly by proper feature representation.

We could solve this problem by using features from geometric analysis because it can characterize very well the local structure information. But this method tends to be inundated with details and will result in data complexity. Or features generated from traditional topological models because they preserve the global intrinsic structure information but they have disadvantages to reduce too much structure information and are rarely used in quantitative characterization [15].

So the performance key of machine learning models is based on a suitable features representation that can both keep the data intrinsic information and reduce data complexity and dimensionality. Deeply based in algebraic topology, persistent homology (PH) offer a delicate equilibrium between intrinsic structure characterization and data simplification and has been applied to different areas successfully [15]. But the combination of persistent homology and machine learning has been hindered greatly by three challenges

1. namely topological representation of data [15];
2. PH-based distance measurements or metrics [15];
3. PH-based feature representation [15].

Progress has been made in these problems with the growth up of topological data analysis. Deducing topological and geometrical information from data in the form of summary representations of topological features, by using topological data analysis methods such as persistent homology, offer another approach in machine learning problems. However, such topological signatures are highly impracticable for most machine learning techniques because information often comes with an unusual structure. And to map these topological signatures into machine learning compatible representations, many strategies have been proposed and unfortunately, there are still suffering from being agnostic to the target learning task.

Since machine learning techniques don't give an answer about training data, they give only information, from TDA methods, we propose a technique that enables us to input topological signatures in deep neural networks and to learn an optimal mapping of the task during training.

1.1 Structure of study

The aim of this study is to explore theoretical background to an alternative approach to data analysis called Topological Data Analysis. This approach is about analysis of extremely high dimensional data using tools from algebraic topology. For this purpose, we

1. Review the mathematics of distance measurement, shape formation and associated theory.
2. Using the mathematical foundations from (1), investigate how they form the basis for the analysis of complex data sets in comparison.
3. Explore the on a sample data set the different algorithms of Topological Data Analysis.

In order to highlight the TDA methods, we will divide this work into four different chapters as follows; Chapter one, with two sub-sections, will give a general introduction to TDA in the first sub-section and a literature review in the second sub-section. Following this, chapter two will highlight all the mathematical tools necessary for the application of the TDA and will give us a description of them and their links with the TDA. The chapter three divided into three sub-sections, is giving a descriptive analysis of the methods used in TDA. In the first sub-section, the description of the Mapper method will be offered and follow in the second sub-section by the description of the persistent homology method and in the last sub-section, some examples of the persistence homology. Chapter four will give us a general conclusion of the work. In this study, we will focus much more on the method of persistent homology.

1.2 Literature Review

1.2.1 Previous Study and Results.

1.2.2 New Text classification for Natural Language Processing. This study was carried out in China by Lei and Yumiao [12].

a. Context: Text classification is currently one of the most difficult research topics, and one of its difficulties is the high dimension of feature space. In a high dimension space, features can redundant or without relationship among them, which is responsible for spacial treatment inconveniences of high dimension, prone to over-learning, time and space overhead, It is important to reduce the dimension of features but without impacting the classification accuracy.

b. Methodology: The authors have chosen to classify the textual data from Persian poems of two Iranian poets Ferdowsi and Hafez and they used two R packages TDA and TDA staus implemented by persistent homology. After data preprocessing, they made document tern matrix by using $TF-IDF$ (Term Frequency-Inverse Document Frequency) algorithm and the next step is to sketch a persistent diagram, barcode and persistent landscapes for a sample of Ferdowsi poems using persistence homology. And compute at the end Wasserstein distances between persistent Diagrams of correspondence parts of Hafez's and Ferdowsi's poems [12].

c. Result: The accuracy was improved and we get a better classification effect.

1.2.3 Using Topological Data Analysis for diagnosis of pulmonary embolism. This study was carried out in Camerino (Italy) by Rucco, Matteo and Falsetti, Lorenzo and Herman, Damir and Petrossian, Tanya and Merelli, Emanuela and Nitti, Cinzia and Salvi, Aldo [16].

a. Context: Pulmonary embolism (PE) is a common and very deadly infection. Patients who usually have this infection usually die within the first few hours.

Several methods have been developed to improve the performance of pulmonary embolism diagnostics, such as the Wells and Geneva Revised Scores, one of the most validated and widely used clinical decision rules. However, these methods have limitations because only 10 out of every 100 people diagnosed with suspected pulmonary embolism confirm the results.

b. Methodology: The authors bring a new approach of clinical prediction rule-based of a new approach of the selection of features based on topological concepts to perform pulmonary embolism diagnostics. They used TDA to overcome the obstacle of missing null values in the data set. By studying patients using their topology, they were able to identify two distinct groups of pulmonary embolism patients. And they also used the topology network from the Iris software (Ayasdi, Inc., Palo Alto) to detect several subgroups of patients susceptible but not infected with pulmonary embolism for future correct diagnostic.

c. Result: Topological Data Analysis of the data allowed them to identify the major characteristics best associated with diagnostic factors for pulmonary embolism. This information was used to detect the entry space of the Back propagating artificial neural network (Bp-ANN). This dataset helped to increase the clinical performance of about 20%.

1.2.4 Chatter Classification in Turning using Machine Learning and Topological Data Analysis.

This study was carried out in the United States of America by Firas A. Khasawneh, Elizabeth Munch, and Jose A. Perea [11].

a. Context: Since two decades, the identification and detection of chatter in the machining process have been a frequent field of research. And these chatter are often described using delay differential equations. But the usual equations describing the occurrence of these vibrations are non-linear delay differential equations, which presents a challenge for the study of these chatter. Even the use of one of the classification algorithms for supervised machine learning techniques has proven ineffective in determining the metrics that can capture the chatter characteristics and also defining a threshold that signals its occurrence.

b. Methodology: They authors used a new approach to capture very these chatter. They used supervised machine learning methods and Topological Data Analysis methods to obtain a process descriptor that can detect chatter. The features that they used were derived from the persistence diagram of an attractor reconstructed from time series. The approach was tested using deterministic and stochastic rotation models.

c. Result: The result shows that they obtained a 97% success rate on the deterministic model labeled by the stability diagram obtained using spectral elevation.

2. Preliminaries

Topological Data Analysis is a method of data science that took its essence from pure mathematics and combinatorial tools so it needs some pieces of knowledge background such as metric space, topological space, and graph theory before starting applying it.

2.1 Metric spaces

2.1.1 Definition (Metric Space). A metric space (M, ρ) is a couple of a nonempty set M with a real-valued function ρ on the cartesian product $M \times M$ of M with itself such define: $\rho : M \times M \rightarrow \mathbb{R}_+$ and such that the metric axioms as follow below, are verified $\forall x, y, z \in M$ [7].

$$\begin{aligned}\rho(x, y) &\geq 0 \\ \rho(x, x) &= 0 \\ \rho(x, y) &= 0 \Leftrightarrow x = y \\ \rho(x, y) &= \rho(y, x) \text{ (symmetry)} \\ \rho(x, y) &\leq \rho(x, z) + \rho(z, y) \text{ (triangle inequality)}.\end{aligned}\tag{2.1.1}$$

If ρ satisfied all the metric axioms above, then ρ is called a metric in M or a distance in M . If x, y, z are belong to the same line then, $\rho(x, z) = \rho(x, y) + \rho(y, z)$.

If the third axiom is not satisfied then ρ is called a pseudo-metric in M

2.1.2 Example (Metric space). For $M = \mathbb{R}$, the metric is given by $\rho(x, y) = |x - y|$.

Let's prove that $|\cdot|$ is a distance in \mathbb{R} . $\forall x, y \in \mathbb{R}$

$$\begin{aligned}\rho(x, y) &= |x - y| \geq 0 \\ \rho(x, x) &= |x - x| = 0 \\ \rho(x, y) &= 0 \iff |x - y| = 0 \Leftrightarrow x = y \\ \rho(x, y) &= |x - y| = |-(y - x)| = |y - x| = \rho(y, x) \text{ (symmetry)}. \\ \rho(x, z) &= |x - z| = |x - y + y - z| \leq |x - y| + |y - z| = \rho(x, y) + \rho(y, z).\end{aligned}$$

2.1.3 Definition (Compactness of Metric Space). A space metric X is compact if every cover of X has a finite sub-cover

The particularity of matrix space in TDA is that it provides some distances very useful to quantify the proximity between dataset. Among all those distances, we will talk about specifically the Hausdorff distance and the Gromov-distance.

2.1.4 Definition (Hausdorff distance). Let $\mathcal{K}(M)$ be the set of compact subsets of (M, ρ) and let A, B be two compact subsets of M . The Hausdorff distance noted $d_H(A, B)$ is

$$d_H(A, B) = \inf\{\delta \geq 0 \mid \forall a \in A \exists b \in B \mid \rho(a, b) \leq \delta \text{ and } \forall b \in B, \exists a \in A \mid \rho(a, b) \leq \delta\}.\tag{2.1.2}$$

The Figure [2.1] is the construction of the Hausdorff distance between the compact sets A and B which is the distance from a point a on A farthest to B to a point b on B nearest to a .

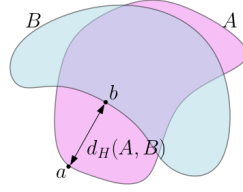


Figure 2.1: Hausdorff Distance [5].

Let's define a distance function from a point $x \in M$ to a nonempty compact set A .

$$\begin{aligned} d(\cdot, A) : M &\longrightarrow \mathbb{R}_+ \\ x &\longmapsto d(x, A) = \inf_{x \in M} d(x, a) = \min_{x \in M} d(x, a) \quad \forall a \in A. \end{aligned} \quad (2.1.3)$$

If $x \in A$ then $d_H(x, A) = 0$.

2.1.5 Theorem. For given two compacts subsets A and B of M , and for all $x \in M$, The equalities (2.1.4), (2.1.5), (2.1.2) and (2.1.6) are equivalent

$$d_H(A, B) = \max\{\sup_{b \in B} d(b, A), \sup_{a \in A} d(a, B)\} \quad (2.1.4)$$

$$= \sup_{x \in M} |d(x, A) - d(x, B)| \quad (2.1.5)$$

$$= \|d(\cdot, A) - d(\cdot, B)\|_\infty. \quad (2.1.6)$$

Proof. First Let's prove that (2.1.2) and (2.1.4) are equivalents.

$$\forall a \in A \exists b \in B \mid d(a, b) \leq \delta \text{ we have } \inf_{b \in B} d(a, b) \leq \delta, \quad (2.1.7)$$

then,

$$d(a, B) \leq \delta \quad \forall a \in A$$

so,

$$\sup_{a \in A} d(a, B) \leq \delta \quad (2.1.8)$$

symmetrically,

$$\forall b \in B \exists a \in A \mid d(a, b) \leq \delta \text{ then, } \sup_{b \in B} d(A, b) \leq \delta. \quad (2.1.9)$$

From (2.1.8) and (2.1.9) we obtain

$$\max\{\sup_{a \in A} d(a, B), \sup_{b \in B} d(A, b)\} \leq \delta. \quad (2.1.10)$$

Since $d_H(A, B)$ is the smallest non negative δ such that the inequation (2.1.12) is satisfied, then

$$d_H(A, B) = \max\{\sup_{a \in A} d(a, B), \sup_{b \in B} d(A, b)\}. \quad (2.1.11)$$

Secondly Let's prove that (2.1.4) and (2.1.5) are equivalents.

By using (2.1.1), we have

$$d(a, x) \leq d(a, b) + d(b, x) \text{ which implies that } \inf_{a \in A} d(a, x) \leq \inf_{a \in A} d(a, b) + d(b, x).$$

Then,

$$\begin{aligned} \inf_{a \in A} d(a, x) &\leq \sup_{b \in B} \inf_{a \in A} d(a, b) + \inf_{b \in B} d(b, x) \\ d(A, x) &\leq \sup_{b \in B} \inf_{a \in A} d(a, b) + d(B, x). \end{aligned}$$

We get

$$d(A, x) - d(B, x) \leq \sup_{b \in B} \inf_{a \in A} d(a, b). \quad (2.1.12)$$

Symmetrically, we have

$$\begin{aligned} d(B, x) &\leq \sup_{a \in A} \inf_{b \in B} d(a, b) + d(A, x) \\ -(d(A, x) - d(B, x)) &\leq \sup_{a \in A} \inf_{b \in B} d(a, b). \end{aligned} \quad (2.1.13)$$

Then from the inequalities (2.1.12) and (2.1.13), we get

$$|d(a, x) - d(B, x)| \leq \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\}.$$

So,

$$\sup_{x \in M} |d(a, x) - d(B, x)| \leq \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\}. \quad (2.1.14)$$

In particular for $a = x \in A$, we have

$$\begin{aligned} \sup_{a \in A} \inf_{b \in B} d(a, b) &= \sup_{x=a \in A} d(x, B) \\ &\leq \sup_{x \in M} |d(x, A) - d(x, B)|. \end{aligned} \quad (2.1.15)$$

Symmetrically, by setting $b = x \in B$, we get

$$\begin{aligned} \sup_{b \in B} \inf_{a \in A} d(a, b) &= \sup_{x=b \in B} d(x, A) \\ &\leq \sup_{x \in M} |d(x, A) - d(x, B)|. \end{aligned} \quad (2.1.16)$$

From (2.1.15) and (2.1.16), we have

$$\max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\} \leq \sup_{x \in M} |d(a, x) - d(B, x)|. \quad (2.1.17)$$

Therefore, the inequalities (2.1.14) and (2.1.17) give

$$\max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\} = \sup_{x \in M} |d(a, x) - d(B, x)|.$$

□

2.1.6 Hausdorff distance in TDA. One of the main areas of Artificial Intelligence (AI) research is face detection. As human identification features, facial features have the advantage of making it easier to obtain image samples compared to fingers. Face detection research focuses on static faces and the object of research is often a static image of the face without depth rotation [12]. In order to adapt to the closing of some sports fields, it will be necessary to make effective use of continuous motion image sequences to improve the efficiency of facial recognition and minimize the decline in recognition of the effect caused by the blurred image. A measure of similarity between a general face model and possible instances of the object within the image using the Hausdorff Distance (HD) can tolerate disturbances in the locations of the points. Many researchers helped to improve the performance of conventional measurement in terms of speed [12]. So the Hausdorff distance will provide a convenient way to quantify the proximity between data sets from the same ambient metric space.

The compact sets A and B can be rotated to reduce the Hausdorff distance and give rise to another distance called The Gromov-Hausdorff distance. The Gromov-Hausdorff distance minimizes the Hausdorff distance and bring a solution to how to study the similarity between the data sets when they do not belong to the same metric space.

2.1.7 Definition (Gromov-Hausdorff distance). The Gromov-Hausdorff distance between two compact metric spaces (M_1, ρ_1) , (M_2, ρ_2) noted $d_{GH}(M_1, M_2)$ is the infimum of the real numbers $r \geq 0$ such that there exists a metric space (M, ρ) and two compact sub-spaces $C_1, C_2 \subset M$ that are isometric to M_1 and M_2 and such that $d_H(C_1, C_2) \leq r$, [12].

The Figure [2.2] is the construction of the Gromov-Hausdorff between the compact sets A and B . Which is the distance from a point a on A nearest to B to a point b on B nearest to a .

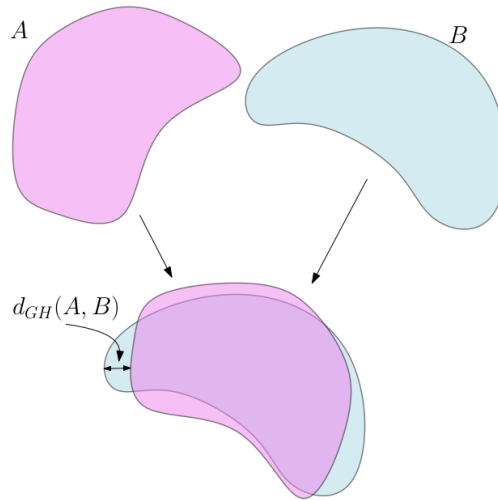


Figure 2.2: Gromov-Hausdorff distance [5]

2.1.8 Gromov-Hausdorff distance in TDA. The Gromov-Hausdorff distance is useful in TDA to study the stability properties persistence diagrams.

The Gromov- Hausdorff distance is able to detect the metric similarity between the shapes as it operates on their metric structure, which means that, shapes are viewed as metric spaces. It also compares the full metric information contained in the shapes, as opposed to other distances that may only compare

simple (incomplete) invariants. Therefore two shapes will be declared equal if and only if they are isometric.

2.1.9 Definition. (Isometric metric spaces)

Two compact metric spaces (M_1, ρ_1) and (M_2, ρ_2) are isometric if there exists a bijection $\phi : M_1 \rightarrow M_2$ that preserves distances.

$$\text{i.e. } \rho_2(\phi(x), \phi(y)) = \rho_1(x, y) \text{ for any } x, y \in M_1.$$

2.1.10 Definition. (Continuity metric spaces)

Let (X, d) and (Y, d') be two metric spaces. The map $f : X \rightarrow Y$ is continuous in $x_0 \in X$ if

$$\forall \epsilon > 0 \exists \delta > 0 \text{ such that } d_X(x, x_0) \leq \delta \implies d_Y(f(x), f(x_0)) \leq \epsilon. \quad (2.1.18)$$

2.1.11 Proposition. For a given $f : X \rightarrow Y$ and an open set U of Y , f is continuous if and only if the inverse image of all open cover U of Y by f noted $f^{-1}(U)$ is an open set of X . And similar, for a given closed set V of Y , f is continuous if and only the inverse image of V by f noted $f^{-1}(V)$ is a closed set of X .

Proof. First we assume f is continuous and U is an open set of Y and let's prove that $f^{-1}(U)$ is an open set of X .

Let's $x \in X$ by definition we have $f(x) = y \in Y$ $x \in f^{-1}(U)$ it follows $f(x) \in U$ moreover U is an open set of Y then for all $\epsilon > 0$ there exists $r > 0$ such that $B_Y(f(x), r) \subset U$. Furthermore f is continuous in X then there exists $\delta > 0$ such that $f(B_X(x, \delta)) \subset B_Y(y, \epsilon) \subset U$ so, $B_X(x, \delta) \subset f^{-1}(U)$ then $f^{-1}(U)$ is an open set of X .

Secondly we assume now that U is an open set of Y and $f^{-1}(U)$ is an open set of X let's prove that f is continuous in X . For all $x \in X$ and $\epsilon > 0$ by setting $U = B_Y(y, \epsilon)$ we have $f^{-1}(B_Y(y, \epsilon))$ is an open set of X then there exists $\delta > 0$ such that $B_X(x, \delta) \subset f^{-1}(B_Y(y, \epsilon))$ then, $f(B_X(x, \delta)) \subset B_Y(y, \epsilon)$ so f is continuous. \square

2.1.12 Definition. (Convexity)

Let's A be a subset of \mathbb{R}^n , A is convex for a given points $x, y \in A$ If for every $t \in [0, 1]$, we have $(1-t)x + ty \in A$. In other words, If the ligne segment joining any two points in A is wholly contained in A .

2.1.13 Example. (Convex sets)

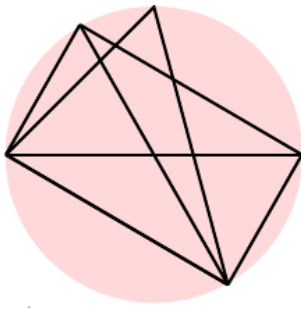
$$\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \text{ such that } x^2 + y^2 \leq r^2\}. \quad (2.1.19)$$

The Figure [2.3b] is a graphical representation of the example (2.1.19). And we realize that for each two points belong to \mathcal{C} the ligne segment joining these points is wholly contained in \mathcal{C} .

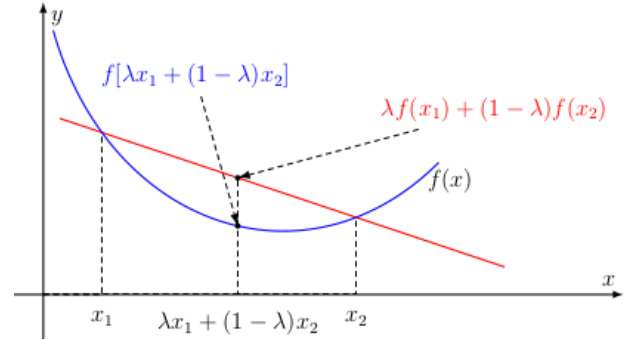
Let E be a nonempty set and $\forall x, y \in E$, the function f is convex if and only If we can find $\lambda \in [0, 1]$ such that $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$. In other words, If its value is below the interpolation formed between any two points.

2.1.14 Example. (Convex functions)

1. Exponential function : e^{ax}



(a) Graph of convex function



(b) Graph of convex function

Figure 2.3: Examples of a convex set in the left and a convex function in the right

2. Norm function : $\|x\|_{\infty} = \max_k |x_k|$

3. The Figure [2.3b] is showing an example of convex function and we can see that the interpolation line is above the convex combination of f

2.1.15 Definition. (Homeomorphism)

A given function f is a homeomorphism if f is continuous, bijective and it has an inverse which is also continuous.

2.1.16 Example (Homeomorphism functions).

$$\begin{aligned} f : \mathbb{R} &\longrightarrow (a, \infty) \\ x &\longmapsto a + e^x \end{aligned}$$

and

$$\begin{aligned} f : \mathbb{R} &\longrightarrow (-\infty, b) \\ x &\longmapsto b - e^x \end{aligned}$$

are the homeomorphism functions.

2.1.17 Definition. (Homeomorphic spaces)

Two sets A and B are homeomorphic if there exists a homeomorphism $g : A \rightarrow B$ is a homeomorphism.

2.1.18 Example (Homeomorphic spaces). The spaces $(0, 1)$ and $(1, \infty)$ (with their topologies being the unions of open balls resulting from the usual Euclidean metric on these subsets of \mathbb{R}) are homeomorphic.

Proof. (Example 2.1.18) Let f be a map such that

$$\begin{aligned} f : (0, 1) &\longrightarrow (1, \infty) \\ x &\longmapsto \frac{1}{x} \end{aligned}$$

Let's show that f is a homeomorphism.

- (1) Bijection : Let $x, y \in (0, 1)$ and with the assumption $f(x) = f(y)$. $f(x) = f(y)$ is equivalent to write $\frac{1}{x} = \frac{1}{y}$. Then, $x = y$. So f is a subjection.

Let $b \in (1, \infty)$. This means $b > 1$. So $0 < \frac{1}{b} < 1$. If we set $a = \frac{1}{b}$ we obtain

$$f(a) = \frac{1}{a} = b \text{ with } 0 < a < 1.$$

So f is an injection. Then f is bijective.

- (2) Continuity : f is continuous as a rational function continuous in its domain.

- (3) Continuity of the map f^{-1} : The map f^{-1} is defined by

$$\begin{aligned} f^{-1} : (1, \infty) &\longrightarrow (0, 1) \\ x &\longmapsto \frac{1}{x}, \end{aligned}$$

f^{-1} such define is also continuous. So, f^{-1} is a homeomorphism between $(0, 1)$ and $(1, \infty)$. Then these spaces are homeomorphics. \square

2.2 Topological Spaces

2.2.1 Definition (Topological Space). A topological space (X, τ) is a set X with a topology τ where τ is a collection of open subsets such that

1. $\emptyset \in \tau$ and $X \in \tau$;
2. τ is stable of a finite intersection of its subsets;
3. τ is stable for arbitrary union of its subsets.

2.2.2 Definition (Compactness of Topological Space). A topological space X is compact if it is separated and for any overlap of openings of X a finished underlap of X can be extracted. ie. $\forall \{U_i\}_{i \in I}$ open subset of X such that $X = \cup_{i \in I} U_i$, we can find $J \subset I$ finite such that $X = \cup_{i \in J} U_i$

For a given two topological spaces X and Y , they have the same topology if they are homeomorphics. i.e If there exist two continuous bijective maps $g : X \rightarrow Y$ and $f : Y \rightarrow X$ such that the map $f \circ g$ and $g \circ f$ are respectively the identity map of X respectively Y . In general, we will require the topological spaces X and Y to being homeomorphic to ensure that they will have the same topological features of interest of TDA.

There exist another notion which study different shapes to verify if they have the same topology. That new notion is called homotopy, will provide the homotopy equivalence between functions and spaces.

2.2.3 Definition. (Homopoty equivalence of functions)

Two maps $f, g : X \rightarrow Y$ are homotopics if they are continuous and there exists a continuous map $H : X \times [0, 1] \rightarrow Y$ such that for any $x \in X$ we have $h(x, 0) = f(x)$ and $H(x, 1) = g(x)$.

2.2.4 Definition. (Homotopy equivalence of spaces)

Two spaces X and Y are homotopy equivalent if there exists two maps $f: X \rightarrow Y$ and $g: Y \rightarrow X$ such that $f \circ g$ and $g \circ f$ are homotopic to the identity map of Y and X respectively.

The Torus whose equation is $(x^2 + y^2 - a^2)^2 + z^2 = b^2$ is homotopic to a coffee cup. In the figure [2.4],

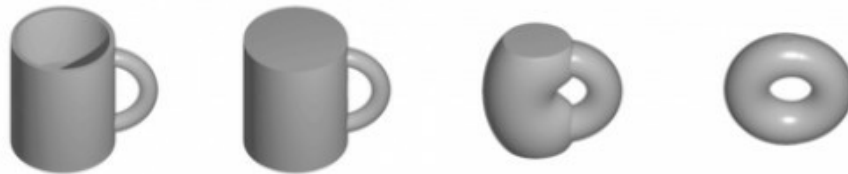


Figure 2.4: Homotopy equivalent between torus and coffee cup. (Source: Internet)

we see that if we quash a coffee cup we will get a torus which has the same properties with the coffee cup. So the deformation of variance of an object affect the shape but the properties are remaining the same.

If the functions f and g are homeomorphic, then they are homotopic and any set which is homotopy equivalent to a point is contractible.

2.2.5 Example (Homotopic functions). Let $f, g: \mathbb{R} \rightarrow \mathbb{R}$ be two continuous real-functions. By setting

$$F: \mathbb{R} \times [0, 1] \longrightarrow \mathbb{R}$$

$$(x, t) \longmapsto (1 - t)f(x) + tg(x)$$

1. We have $F(x, t)$ is a sum of two continuous functions so $F(x, t)$ is continuous.
2. $F(x, 0) = (1 - 0).f(x) + 0.g(x) = f(x) \quad \forall x \in \mathbb{R}$
3. $F(x, 1) = (1 - 1).f(x) + 1.g(x) = g(x) \quad \forall x \in \mathbb{R}$

With F such defined it clear that f and g are homotopic.

2.2.6 Definition (Cover). Let X, τ be a topological space. An open cover \mathcal{U} of X is a collection $\{U_i\}, i \in I$ of open subsets of X such that $\cup_I U_i = X$.

2.2.7 Example (Open cover). The set $A = \{(1/n, 1), n \in \mathbb{N} - \{0\}\}$ is an open cover of the open interval $(0, 1)$.

2.2.8 Definition (Pullback cover). Let's $f: \mathcal{X} \rightarrow \mathbb{R}^d; d \geq 1$ be continuous map. The pullback cover is defined as a collection of open sets $(f^{-1}(U_i))_{i \in I}$ for a given cover $\mathcal{U} = (U_i)_{i \in I}$.

2.2.9 Definition (Refined pullback). The refined pullback or connected components of the open sets $(f^{-1}(U_i))_{i \in I}$ is the collection of graph where any two vertices are connected to each other by an edge.

2.2.10 Example (Connected components). The Figure [2.5] is an example of connected components in the left, is a connected component with twelve vertices and the right is a connected component with four vertices.

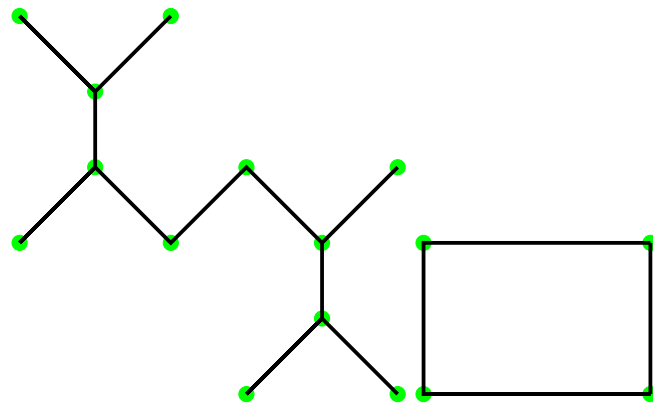


Figure 2.5: Graph of two components

2.3 Covers and simplicial complexes

2.3.1 Graph Theory. A graph is a collection of nodes or vertices connected between them by edges. We have two type of graphs which are a Directed graph and an Undirected graph.

2.3.2 Definition. (directed graph) A directed graph is a graph where the edges have a direction associated with them.

2.3.3 Example.

1. Directed Analytic Graph (DAG) there is no direct edge starting and ending on the same vertex.
2. A Tree. It's a restriction of DAG. In this case, a node is connected to only one other node.

2.3.4 Definition. (Undirected graph) An Undirected graph is a graph where the edges don't have a direction associated with them.

2.3.5 Example.

1. Connected graph : In this graph, there is no unreachable node.
2. Biconnected graph : In this graph, there is no articulation points. So it can not be broken dow into any pieces.
3. Complete graph : In this graph, each node is connected to another node. So, A complete graph is a graph where each vertex is connected to other vertices by an edge. Then, a complete graph will show all the nodes and the edges and a complete graph with n nodes, will have exactly $n(n-1)/2$ edges.

2.3.6 Simplicial complexes. The higher dimensional generalization of triangles is called Smplicial complexes. From its relevant tools from topology (using open covers) to combinatorial, it is very helpful for TDA. Therefor, how can we construct a simplicial complexes where given a data set or a topological space or a metric space? We will introduce the notion of Vietoris-Rips Complex noted $\text{Rips}_\alpha(X)$ and Čech complex noted $\text{Cech}_\alpha(X)$.

For a given metric space (M, d) , and a $\alpha \in \mathbb{R}$ the Vietoris-Rips complex characterizes the topology of

a points set and its particularity in TDA is that it is easy to extend its construction to high dimension and it's defined by

$$Rips_{\alpha}(X) = \{[x_0, x_1, \dots, x_k] \text{ of } (k+1) \text{ simplices such that } d(x_i, x_j) \leq \alpha \forall (i, j)\} \quad (2.3.1)$$

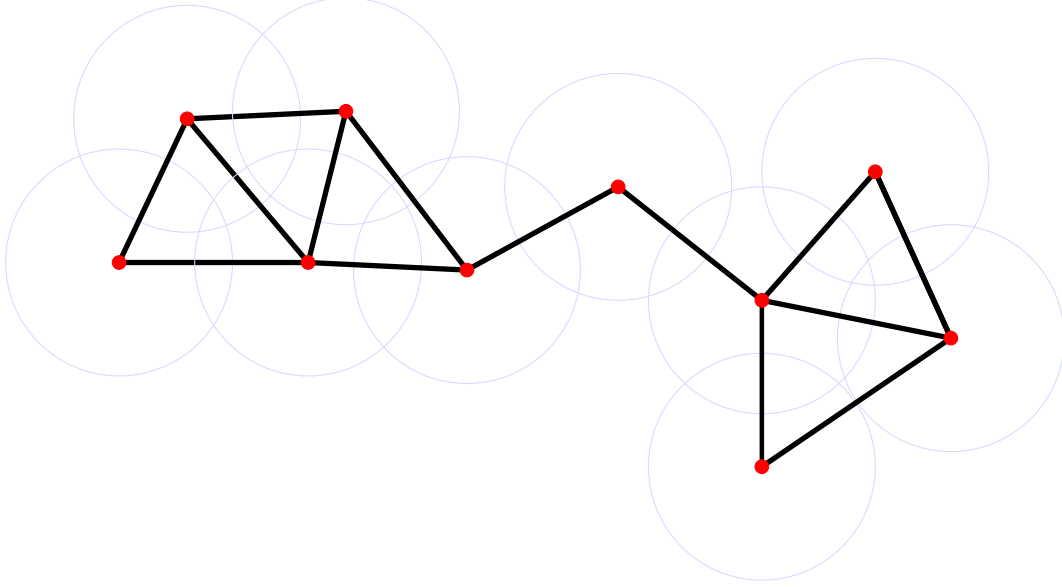


Figure 2.6: Vietoris-Rips complex

The Figure [2.6] is given an example of vietoris $Rips_{2\alpha}$. So for each two connected nodes, the distance between those nodes is less or equal than 2α .

The more close to vietoris-Rips is the Čech Complex denoted by

$$Cech_{\alpha}(X) = \{[x_0, x_1, \dots, x_k] \text{ of } (k+1) \text{ simplices such that } \bigcap_{x_i, \forall i} B(x_i, \alpha) \neq \emptyset\}. \quad (2.3.2)$$

The Figure [2.7] is an example of Čech Complex and we can see that all the circles which are sequent, their centers are connected among them.

1. **0-Simplex** is a single vertex;
2. **1-Simplex** is two vertices connected by an edge;
3. **2-Simplex** is three vertices connected pair wise by edges with a single face;
4. **3-Simplex or tetrahedron** is four vertices connected pair wise by edges joined by four faces;
5. **4-Simplex or orpentachoron** has five vertices, ten edges, ten faces, five cells, and one 4-element

Then, in general, k -simplex will have $(k+1)$ vertices.

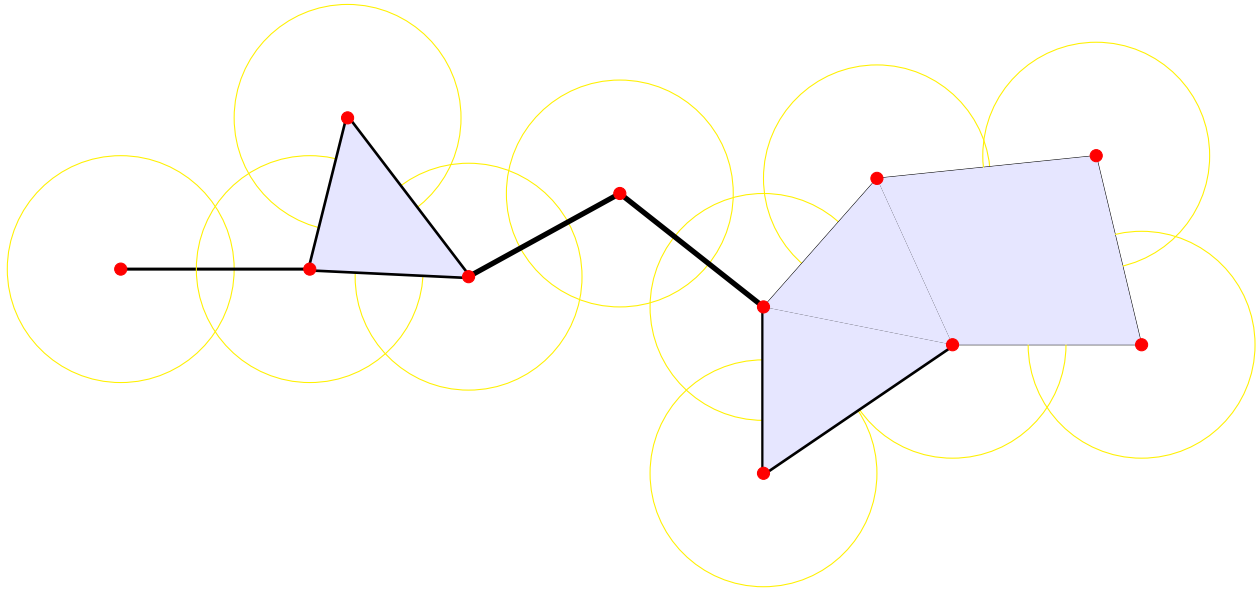
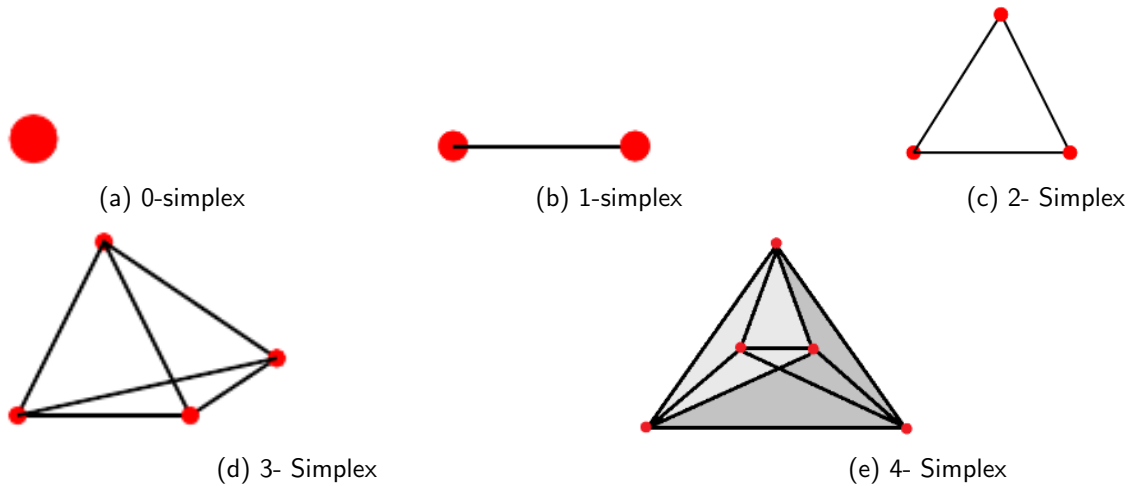


Figure 2.7: Čech Complex

Figure 2.8: Example of k -simplex.

2.4 Geometric and abstract simplicial complexes.

2.4.1 Definition (Geometric simplicial complex). A geometric simplicial complex $K \in \mathbb{R}^d$ is a collection of simplices such that

1. For any face F of a simplex of K , F is also a simplex of K
2. For any two simplices of K , the intersection is either equal to the empty set or is a common face of both.

For a given finite set $P = \{p_0, p_1, \dots, p_k\} \subset \mathbb{R}^d$, the points of P are called affinely independent points and respectively affinely dependent points, if the k vectors $p_0\vec{p}_1, p_0\vec{p}_2, \dots, p_0\vec{p}_k$ are linearly independent (respectively linearly dependent) in \mathbb{R}^d .

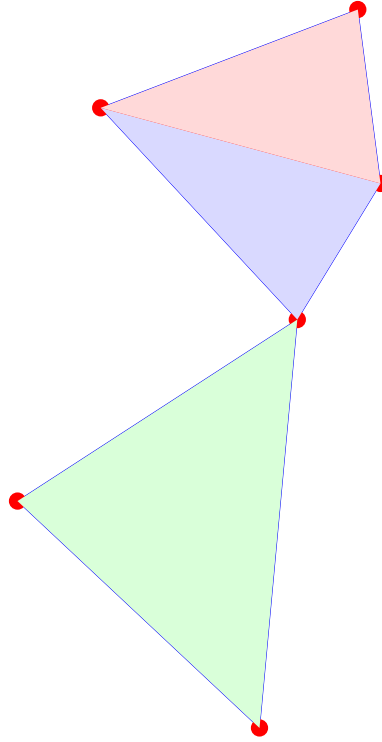


Figure 2.9: Example of simplicial complex

2.4.2 Definition (k -simplex). The k -dimensional simplex σ spanned by a family of $k + 1$ affinely independent points $P = \{p_0, p_1, \dots, p_k\} \subset \mathbb{R}^d$ is the set of convex combination $\sum_{i=0}^k \lambda_i p_i$ such that

$$\sum_{i=0}^k \lambda_i = 1 \text{ and } 0 \leq \lambda_i \leq 1, \quad (2.4.1)$$

where the elements of P are called vertices of σ .

The Figure [2.9] is an example of simplicial complex with six vertices.

2.4.3 Definition (Abstract simplicial complexes). Let $V = \{v_1, v_2, \dots, v_k\}$ be a nonempty set. An abstract simplicial complex with a vertex V is a set \tilde{K} of finite subsets of V such that the properties below are verified:

1. All elements of V belong to \tilde{K} .
2. For any $\sigma \in \tilde{K}$, any subset of σ belongs to \tilde{K} .

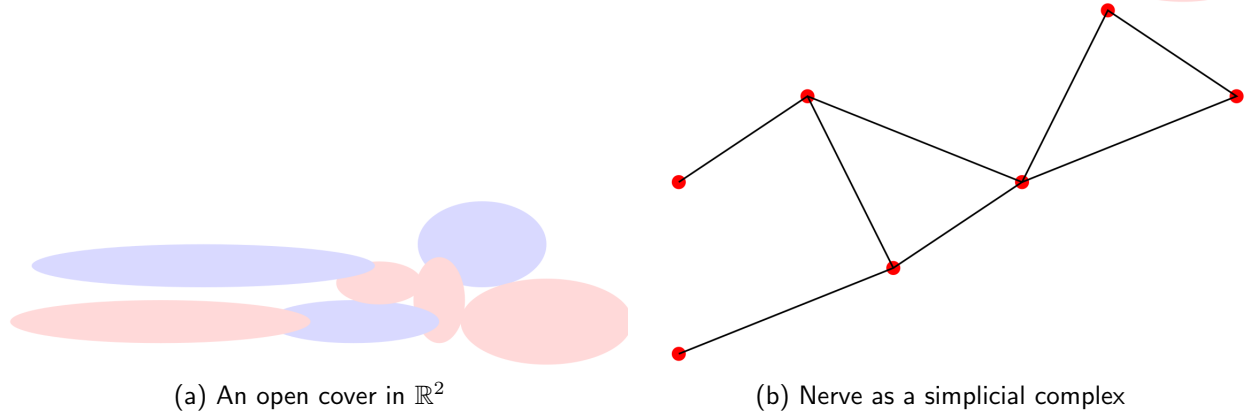
Then, an abstract simplicial complexes \tilde{K} can be seen as a topological space $|\tilde{K}|$.

A complete graph with K vertices will form a $(K - 1)$ -Simplex.

2.4.4 Definition (Nerve of a cover). Given a cover $\mathcal{U} = (U_i)_{i \in I}$ of a metric space (M, ρ) , the nerve of a cover \mathcal{U} is the abstract simplicial complex $C(\mathcal{U})$ whose vertices are the U_i such that

$$\sigma = \{[U_{i_0}, \dots, U_{i_k}] \in C(\mathcal{U}) \text{ iff } \cap_{j=0}^k U_{i_j} \neq \emptyset\}, \quad (2.4.2)$$

where σ is the k -dimensional simplex[5].

Figure 2.10: A cover in \mathbb{R}^2 and its Nerve.

The Figure [2.10a] is a cover formed by the union of seven open subset of \mathbb{R}^2 and the Figure [2.10b] shows its nerve.

2.4.5 Definition (Persistence Modules [9]). A persistence Module M is a couple formed by a vector space M_a with $a \in \mathbb{R}$ and a linear map $M(a \leq b) : M_a \rightarrow M_b$ for all $a \leq b$ with $b \in \mathbb{R}$ such that the linear map $M(a \leq a) : M_a \rightarrow M_a$ is the identity map and for another point $c \in \mathbb{R}$ such that $a \leq b \leq c$ we have

$$M(b \leq c) = M_c^b \circ M(a \leq b) = M(a \leq c)[5]. \quad (2.4.3)$$

Persistence Module is the key of algebraic study in TDA.

2.4.6 Definition. Let V and W be two persistence Modules. We define an homeomorphism of degree r between V and W as a collection of linear maps $\Phi_a : V_a \rightarrow W_{a+r} \forall a \in \mathbb{R}$ such that

$$a \leq s, \quad \Phi_s \circ v_s^a = w_{s+r}^{a+r} \circ \phi_a, \quad (2.4.4)$$

where w_{s+r}^{a+r} respectively v_s^a is a map of W and V respectively.

We can decompose the persistence module V in sum of intervals modules. The set of these intervals is independent to the decomposition of V and its called persistence barcode of V . We can represent each interval by a couple $(x, y) \in \mathbb{R}^2$.

The disjoint union of these points is called multi-set. If we take the union of multi-set with the diagonal $\Delta = \{x = y\}$, we obtain the persistence diagram of V .

Let dgm_1 and dgm_2 be two persistence diagrams of V . A matching between dgm_1 and dgm_2 is a subset m of $dgm_1 \times dgm_2$ such that every point in $(dgm_1 \setminus \Delta) \cap (dgm_2 \setminus \Delta)$ appears exactly once in m . So for $p \in dgm_1 \setminus \Delta$ and $q \in dgm_2 \setminus \Delta$, each of the sets $(\{p\} \times dgm_2) \cap m$ and $(\{q\} \times dgm_1) \cap m$ contains a single point[5].

The bottleneck distance between dgm_1 and dgm_2 is defined by

$$d_b(dgm_1, dgm_2) = \inf_{\text{matching } m} \max_{(p,q) \in m} \|p - q\|_\infty.$$

The determination of bottleneck is based on the largest distance among the pairs and do not take into account the closeness and the remaining pairs of points[5]. So to solve this issue we use the Wasserstein distance defined by:

$$W_p(dgm_1, dgm_2)^p = \inf_{\text{matching } m} \sum_{(p,q) \in m} \|p - q\|_\infty^p.$$

If M is a persistence module, then for $a \leq b$ the corresponding Betti number of M noted $\beta^{a,b}$ is the dimension of image of its linear map M_b^a . i.e $\beta^{a,b} = \dim(\text{im}(M_b^a))$. Moreover, if $a \leq b \leq c \leq d$ then, $\beta^{b,c} \geq \beta^{a,d}$.

The rank function also called simplest function is the function

$$\begin{aligned} \lambda : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ (b, d) &\longmapsto \begin{cases} \beta^{b,d} & \text{if } b \leq d, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (2.4.5)$$

If we set $m = \frac{b+d}{2}$ and $h = \frac{d-b}{2}$ the map (2.4.5) becomes

$$\begin{aligned} \lambda : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ (m, h) &\longmapsto \begin{cases} \beta^{m-h, m+h} & \text{if } h \geq 0 \leq d, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (2.4.6)$$

2.4.7 Definition (Persistence Landscapes). The persistence landscape is a function defined by

$$\begin{aligned} \lambda : \mathbb{N} \times \mathbb{R} &\longrightarrow \mathbb{R} \cup \{-\infty, +\infty\} \\ (k, t) &\longmapsto \lambda(k, t). \end{aligned} \quad (2.4.7)$$

The aim is to map persistence diagrams into a function space[1], and sometimes it can be considered as a Banach space or even a Hilbert space. This new summary persistence landscapes has the advantages to allow one to apply tools from statistics and machine learning. Moreover, the mapping from persistence diagrams to persistence landscapes is stable and invertible[1].

It had been seen that these preliminaries are very useful in TDA because from the metric space we saw relevant tools such as the Hausdorff distance and the Gromov-Hausdorff distance to quantify the proximity between data set whose are either belong or not into the same metric space. In topological space we to study the similarities between the shapes by using homotopy between data set and at the end how to represent the relationship between data point or a set of data points.

3. TDA Methods

The methods used by Topological Data Analysis are essentially based on the application of algebraic topology in particular and some other fields from pure mathematics. Among several methods used by TDA, we will focus in a particular two which are the Mapper method and the Persistent Homology method.

3.1 Mapper Algorithm

The Mapper algorithm is one of the most popular methods used in TDA . It is used for exploratory data analysis and visualization by providing the summaries of the data for a given data set X by using a function f and a cover \mathcal{U} of $f(X)$.

To use the mapper algorithm for a given high dimensional data set X , the users must first select the real-valued function f often called lens or filter defined by $f: X \rightarrow \mathbb{R}^d$ that strongly fit with the relevant features needed for the data analysis. And with the Mapper method, the first thing will be the mapping of the original data set into lower-dimensional data set through its filter function f .

In general, the Mapper algorithms has the structure as following below:

Input[6]

1. A data set X endowed with a metric;
2. A function $f : X \rightarrow \mathbb{R}^d$ with an integer number $d \geq 1$;
3. A cover $\mathcal{U} = \{U_i\}_{i \in I}$ of $f(X)$.

Methodology[6]

1. For any element $U \in \mathcal{U}$, decompose $f^{-1}(U)$ into clusters C_{U_1}, \dots, C_{U_k} where $C_{U_i} \in \mathcal{U}$;
2. Derive the nerve of the cover of X defined by those clusters C_{U_i} .

Output[6]

1. A vertex v_{U_i} for any cluster C_{U_i} ;
2. An edge between the vertices v_{U_i} and v_{U_j} if and only if $C_{U_i} \cap C_{U_j} \neq \emptyset$. Where v_{U_i} represents the vertex of the cluster C_{U_i} and v_{U_j} the vertex of the cluster C_{U_j} .

The Mapper method could be an easy method but the choice of the filter, the cover, and the cluster make it widely difficult. But there are some standard functions[6] which something work and are strongly recommended are:

1. The Principal Component Analysis function (PCA). Its principal characteristic is to reduce the dimensionality of functional imaging data and often using for visualization;

2. Non Linear Dimension Reduction coordinates (NLDR);
3. Eigen functions of Laplace and density estimates function;
4. The centrality function $f(x) = \sum_{y \in X} d(x, y)$;
5. The eccentricity function $f(x) = \max_{y \in X} d(x, y)$.

Because they have advantages that they don't require any particular knowledge from the data.

Secondly, the choice of the cover \mathcal{U} of $f(X)$.

This choice is very sensible because it has a high impact on the output of the Mapp. The best way to choose the cover \mathcal{U} of $f(X)$ is to take \mathcal{U} as a set of regularly spaced intervals of equal length ϵ covering the set $f(X)$ where ϵ is sometimes called the resolution of the cover. But between two consecutive intervals, we generally have an overlap that we can estimate in percentage g , and this percentage is called the gain of the cover.

The parameters ϵ and g are the keys to the choice of cover because small modification of these parameters will bring high change in the output. So, the best way to choose the right parameters is to take a range of parameters and after investigation, choose the ones whose output provides the most information from the user's expectations.

The way of clustering $f^{-1}(U)$ depends essentially on the choice of the cluster of U_i . So to choose the cluster it is recommended to take the connected components of the sub-graph spanned by the nodes in the set $f^{-1}(U)$ [6].

Since the nerve of covers is a way to summarize, visualize and explore data, therefore, with the good choice of the real valued-function f and a cover \mathcal{U} of $f(X)$ the Mapper algorithm will give the summary of the data X through the nerve of the refined pullback of the cover \mathcal{U} of $f(X)$. And the nerve graph objectives are to provide a simple convenient way to visualize the summary of the data. But taking the gain of the cover less than fifty percent will make every point in the real line be covered with at most two open sets of \mathcal{U} .

Two of the standard application of the Mapp algorithm are the clustering and the features selection. The challenge encountered is the choice of parameters of the graph to build a Mapper graph.

3.2 Geometric Reconstruction and Homological Inference

3.2.1 Geometric reconstruction. Another approach to build a cover for a given data set $X_n = \{x_0, x_1, \dots, x_n\} \subset \mathbb{R}^d$ and use the nerve of this cover to extract the topological structure of data set is to choose the union of balls of fixed radius centered on the x_i [5]. Moreover if K is compact subset of \mathbb{R}^d , then, for a given real positive number r , the union of balls of radius r centered on K and noted

$$K^r = \cup_{x \in K} B(x, r) \quad (3.2.1)$$

and called r -offset of K is by definition the r -sublevel set of the distance function

$$\begin{aligned} d_K : \mathbb{R}^d &\longrightarrow \mathbb{R} \\ x &\longmapsto \inf_{y \in K} \|x - y\| \end{aligned}$$

Then we have $K^r = d_K^{-1}([0, r])$. (3.2.2)

Proof. Let prove that the equalities (3.2.1) and (3.2.2) are equivalent.

Let $y \in \mathbb{R}^d$, $y \in K^r$ implies $\exists x \in K \mid y \in B(x, r)$, which is equivalent to

$$d_K(y) \leq d(x, y) \leq r, \quad (3.2.3)$$

so,

$$d_K(y) \in [0, r]$$

Composing by the pre-image we have

$$y \in d_K^{-1}([0, r]),$$

and then,

$$K^r = d_K^{-1}([0, r]).$$

□

If K is a smooth compact submanifold, for some good choice of r , the offset of X_n are homotopy equivalent to K .

The estimation of topology and geometry structure of shapes sampled with a lower amount of noise is made by using in different context the offset-based inferences.

The notion of distance function has three fundamental properties as following[3].

1. If the space of continuous functions is equipped with the infinite norm $\|f\|_\infty = \sup_{x \in \mathbb{R}^d}$, Then the map $d_K : K \rightarrow \mathbb{R}_+$ is Lipschitz. This means that, there exist a constant $k \geq 0$ such that

$$\forall x, y \in \mathbb{R}^d, \|d_K(x) - d_K(y)\|_\infty \leq k \|x - y\|_\infty.$$

2. The distance function d_K is 1-Lipschitz. This means that

$$\forall x, y \in \mathbb{R}^d, \|d_K(x) - d_K(y)\|_\infty \leq \|x - y\|_\infty.$$

3. The map d_K^2 is 1-semi-concave. ie. The function $\|x\|^2 - d_K^2(x)$ is convex.

The distance function endowed with these properties has successfully answered the question of recovering geometric and topological features of an unknown compact subset of Euclidean space from the approximating point cloud data. But one of its limitations is the fact that it does not deal effectively with outliers nor with noise[3]. Therefore, to overcome this issue, it has been proposed in [4] to redefine the way point clouds are interpreted. And for doing that, the class of compact subsets of \mathbb{R}^d has been replaced by the class of probability measures on \mathbb{R}^d . Where finite point clouds have been replaced by the sum of Dirac masses and k -dimensional manifolds by their volume form.

So the notion of the distance function has been introduced to a probability measure to bring a new notion of distance which is the distance like-function.

3.2.2 Definition (Distance like-function). The distance like-function is a non negative function ϕ defined by:

$$\begin{aligned}\phi : \mathbb{R}^d &\longrightarrow \mathbb{R}_+ \\ x &\longmapsto \phi(x)\end{aligned}$$

such that $\phi(x)$ satisfy the conditions (2) and (3).

Let ϕ be a distance like-function and ϕ^r the r -sublevel of ϕ .

1. For $x \in \mathbb{R}^d$, and $\alpha > 0$, x is said α - critical If the norm of the gradient of ϕ respect to x is less or equal to α , i.e., $\|\nabla_x \phi\| \leq \alpha$.
2. The α -reach of ϕ is: $\alpha\text{-reach} = \max\{r > 0 \text{ such that } \forall \alpha \in (0, 1) \ x \in \phi^{-1}((0, r]) \text{ is not } \alpha\text{-critical}\}$.
3. The weak feature size of ϕ at r noted $Wfs_\phi(r)$ is:

$$Wfs_\phi(r) = \min\{r' > 0 \text{ such that } \phi \text{ does not have any critical value in the interval } [r, r+r']\}$$

4. For two positive numbers r_1, r_2 such that $r_1 \leq r_2$, the sub-level sets $\phi^{-1}([0, r])$ are isotopic for $r \in [r_1, r_2]$ if $\forall x \in \phi^{-1}((r_1, r_2]), \nabla \phi(x) \neq 0$. Means that ϕ has no critical points in the subset $\phi^{-1}((r_1, r_2])$.

To facilitate the boundary computations, let restrict to homology with coefficients in \mathbb{Z}_2 and we are going to introduce the notion of chains complexes. let K be a simplicial complex and k a non negative integer.

3.2.3 Definition (Space of k -chains). The space of k -chains of K and over \mathbb{Z}_2 noted $C_k(K)$ is the set whose elements are the formal finite sum of k -simplices of K with the form $\sum_{i=1}^p \epsilon_i \delta_i$. where each δ_i is a unique k -simplices of K and $\epsilon_i \in \mathbb{Z}_2$. So $c \in C_k(K)$ means that $c = \sum_{i=1}^p \epsilon_i \delta_i$.

3.2.4 Proposition. The space of simplicial k -chain $C_k(K)$ is a vector space or an abelian group over \mathbb{Z}_2 .

Proof. For another k - chain $c' = \sum_{i=1}^p \epsilon'_i \delta_i$, we have:

$$\begin{aligned}c + c' &= \sum_{i=1}^p \epsilon'_i \delta_i + \sum_{i=1}^p \epsilon_i \delta_i \\ &= \sum_{i=1}^p (\epsilon'_i + \epsilon_i) \delta_i \in C_k(K), \quad \text{because } (\epsilon'_i + \epsilon_i) \in \mathbb{Z}_2.\end{aligned}$$

For $\lambda \in \mathbb{Z}_2$, we have:

$$\begin{aligned}\lambda c &= \lambda \sum_{i=1}^p \epsilon_i \delta_i \\ &= \sum_{i=1}^p (\lambda \epsilon_i) \delta_i \in C_k(K), \quad \text{because } (\lambda \epsilon_i) \in \mathbb{Z}_2;\end{aligned}$$

$$\begin{aligned}
c + c' &= \sum_{i=1}^p \epsilon'_i \delta_i + \sum_{i=1}^p \epsilon_i \delta_i \\
&= \sum_{i=1}^p (\epsilon'_i + \epsilon_i) \delta_i \\
&= \sum_{i=1}^p (\epsilon_i + \epsilon'_i) \delta_i = c' + c \in C_k(K),
\end{aligned}$$

because $(\mathbb{Z}_2, +)$ is an abelian group so $(\epsilon'_i + \epsilon_i) = (\epsilon_i + \epsilon'_i) \in \mathbb{Z}_2$. □

For each k we can define a linear mapp δ_k on a k -simplex $\sigma = [v_0, \dots, v_k]$ such that:

$$\begin{aligned}
\delta_k : C_k(K) &\longrightarrow C_{k-1}(K) \\
\sigma &\longmapsto \delta_k(\sigma) = \sum_{i=0}^k (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k],
\end{aligned}$$

where δ_k is called the boundary operator and $[v_0, \dots, \hat{v}_i, \dots, v_k]$ is the simplex spanned by the nodes of σ with \hat{v}_i removed.

3.2.5 Theorem. *The boundary operator of the boundary operator is null. ie., $\delta_k \delta_k = \delta_k \circ \delta_k = 0$.*

Proof.

$$\begin{aligned}
\delta_k(\delta_k) &= \delta_k\left(\sum_{i=0}^k (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k]\right) \\
&= \sum_{i=0}^k (-1)^i \delta_k([v_0, \dots, \hat{v}_i, \dots, v_k]) \quad (\text{linearity of } \delta_k) \\
&= \sum_{i=0}^k (-1)^i \left(\sum_{j=0}^k (-1)^j [v_0, \dots, \hat{v}_i, \dots, \hat{v}_j, \dots, v_k]\right) \\
&= \sum_{i < j} (-1)^i [v_0, \dots, \hat{v}_i, \dots, \hat{v}_j, \dots, v_k] + \sum_{i > j} (-1)^i [v_0, \dots, \hat{v}_j, \dots, \hat{v}_i, \dots, v_k] \\
&= 0
\end{aligned}$$

□

3.2.6 Definition (k -Boundaries). The set of k -Boundaries of K noted $B_k(K)$ is defined by

$$B_k(K) = \{c \in C_k(K) \text{ such that } \exists c' \in C_k(K), \delta_k(c') = c\}. \quad (3.2.4)$$

$B_k(K)$ such defined is the image of δ_k .

3.2.7 Definition (k -cycle). The set of k -cycle of K noted $Z_k(K)$ is defined by

$$Z_k(K) = \{c \in C_k(K) \text{ such that } \delta_k(c) = 0\}. \quad (3.2.5)$$

$Z_k(K)$ such defined is the kernel of δ_k .

The homology group of K or the k -dimensional simplicial homology group is noted $H_k(K) = Z_k(K)/B_k(K)$. The rank of this groups is the number of linearly independent k -dimensional holes in the space.

Two spaces homotopic equivalent have their corresponding chain maps chain homotopic (means that there is an algebraic connection from one to the other).

Despite the name chain, the k -chain doesn't need to be connected.

3.3 Persistence Homology

A persistent homology is a powerful tool used in the analysis of topological data. It is used to efficiently code multi-scale topological features of nested families of simplicial complexes and topological spaces as a set of points in a plane called a persistence diagram. The points in the plane, also called persistence points, where each point represents a topological feature, are each one associated with persistence (or life).

The general idea[9] of persistent homology is to describes the changes in homology that occur to an object which evolves with respect to a parameter, and it uses filtration procedure such that each topological generator is endowed with a geometric measurement. During the filtration, nested simplicial complexes encoded with structural topological information from different scales are products. We realized that some topological invariants have a long life duration (topological signal) in these simplicial complexes when others die quickly (topological noise) when filtration value changes.

Since short-live noise is considered as a characteristic with topological noise, and long-live noise is considered as a topological signal, the use of statistical knowledge as a persistence homology whose goal is to create a set of confidence that will allow us to separate the topological signal from the topological noise.

It also provides an efficient algorithm to compute Betti number (which is a sequence of numbers with indicating how many rows of each complex in the considered families have) and encodes the evolution of the homology groups of the nested complexes across the scales.

We can visualize the results from persistence homology by many methods such as a persistent diagram, persistent barcode, the persistence landscape, and the persistence image.

Choosing persistence diagrams as opposed to the original raw data set has the advantages that persistence diagrams capture patterns such as connected components and loops. It also allows us to compute a meaningful distance between persistence diagrams. Finally, persistence diagrams are stable.

3.3.1 Definition (Filtration of topological space [5]). Let (M, τ) be a topological space and T a finite or infinite subset of \mathbb{R} . A filtration of a topological space M is a nested family of subspaces $(M_r)_{r \in T}$ such that it is non-decreasing for the inclusion, i.e., for another $r' \in T$, if $r \leq r'$ then

$$M_r \subset M_{r'} \quad \text{and} \quad M = \bigcup_{r \in T} M_r,$$

where r and r' are called scale parameters.

In other words, for a given simplicial complex Σ , a filtration of Σ [9] is a finite sequence of subcomplexes

$$\Sigma^f := \{\Sigma^p \mid 0 \leq p \leq m\} \quad (3.3.1)$$

of Σ such that:

$$\emptyset = \Sigma^0 \subset \Sigma^1 \subset \dots \subset \Sigma^m = \Sigma. \quad (3.3.2)$$

Let define the map f by: $f : M \rightarrow \mathbb{R}$. $M_r = f^{-1}((-\infty, r])$ is a filtration called the sub-level set filtration of f .

Despite strong stability properties, the persistent homology of filtrations classically used in Topological Data Analysis, such as the Rips-Vietoris complexes $Rips_r$ and the Čech complexes $Cech_r$

In general, let K be a simplicial complexes with a vertex V and $\sigma = [v_0, \dots, v_k]$ a simplex of K . A function $f : v \rightarrow \mathbb{R}$ can be extended to all simplices of K by:

$$f([v_0, \dots, v_k]) = \max\{f(v_i), i \in \{0, \dots, k\}\}. \quad (3.3.3)$$

A family of subcomplex $K_r = \{\sigma \in K \text{ such that } f(\sigma) \leq r\}$ is a filtration called the sublevel filtration set of f . The homology of a filtration $(K_r)_{r \in T}$ of a simplicial complexes changes when r increases. That change can be due to the new appearing of the connected components or the merging of the existing connected components or the appearing of loops and cavities. The goal of persistent homology is to tracks these changes, by identifies the appearing features and associates a life time to them. The information getting at the end is encoded as a set of intervals called barcode.

3.3.2 Statistic and persistent homology. (Confidence region for persistent homology)

When point clouds do not have a geometric shape, it is very complex to analyze the persistence diagram. Usually, in this case, many topological features are closed on the diagonal. Since they correspond to topological structures that die off very quickly after they appear in the filter, they are called topological noise.

But with the confidence regions of the persistence diagram, we can perfectly solve the problem of distinguishing between topological noise and topological signal.

For the determination of this confidence region, we can use the Bottleneck distance or the Wasserstein distance.

When we want to estimate a persistence diagram dgm with an estimator \hat{dgm} , we typically look for some values ν_α such that:

$$P(d_b(\hat{dgm}, dgm) \geq \nu_\alpha) \leq \alpha.$$

If we take $\alpha \in (0, 1)$ and B_α a closed ball of radius α and centered at \hat{dgm} in the space of persistence diagrams, we can visualize the topological signature of the points in this ball in different ways.

We can visualize the confidence set by adding a band at (vertical) distance $\nu_\alpha/2$ starting from the diagonal (the Bottleneck distance is defined for the ℓ_∞ norm). Therefore the points outside the band are considered significant topological features.

Persistence homology gives more informations about the shape than the classical homology. The difference between both classical homology and persistence homology is that, classical homology captures cycles in a shape by ignoring the boundary cycles, and the persistence homology authorize the retrieval of these cycles that are non-boundary elements in a certain level of filtration and will turn into boundaries cycles in some subsequent step[9].

3.3.3 Example. The Figure [3.1] is given the representation of the map $f : [0, 3] \rightarrow \mathbb{R}$, its persistence bar-code and its persistence diagram.

If $F_r = f^{-1}((-\infty, r])$ is the sub-level set filtration of f then, if $r = a_1$ we will see the first appearing

of a connected component and the persistence diagram will save the value a_1 as a birth time of a connected component and will start tracking this value by creating an interval starting by a_1 . The same process will restart when we will have $r \in \{a_2, a_3, a_4, a_5, a_6, a_7\}$. But some components will merge together by pairwise to give rise to a single one.

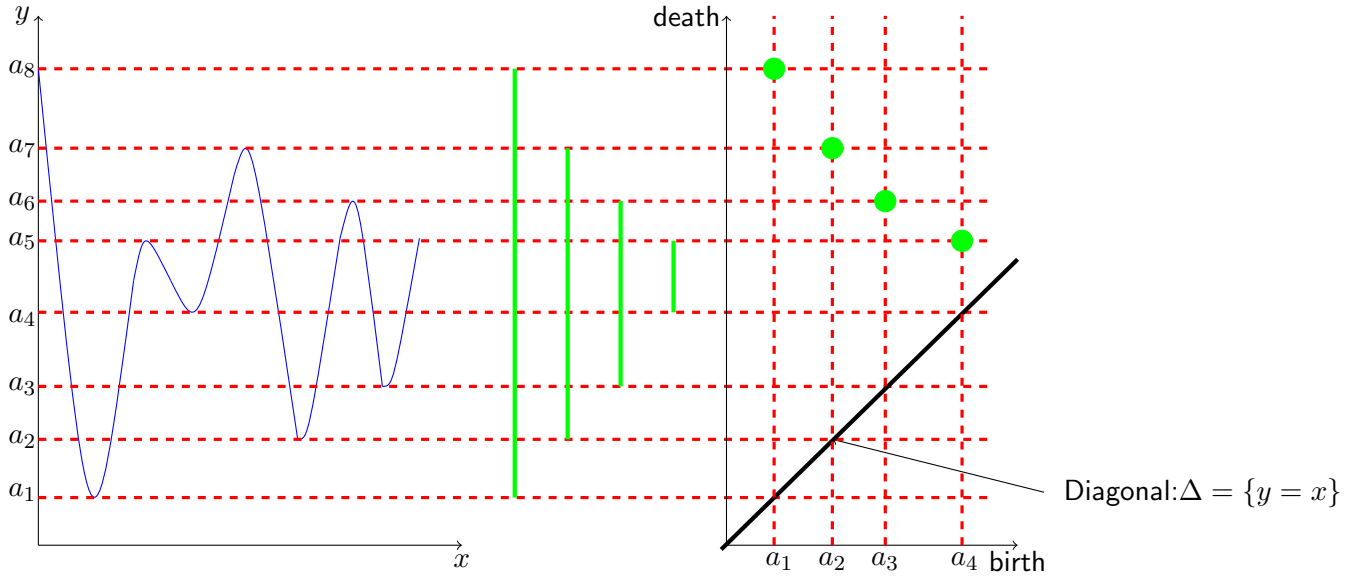


Figure 3.1: The persistence barcode (in the middle) and the persistence diagram (in the right) of the function $f : [0, 3] \rightarrow \mathbb{R}$

Topological Data Analysis methods are the most powerful methods among all machine learning method because their focusing in the representation of the shape of data by giving the geometry summaries through a map, a nerve through a cover, and its success to dealing with outliers using distance like-function, encoding multi-scale topological features through the persistence homology and it amelioration of machine learning result has brought successfully a new and better understanding of high dimensional data.

3.4 Some Examples of Persistence Homology

In this section, some examples of applications of the persistence homology are shown. The scatter plot from machine learning method is shown with the persistence landscape from TDA by using Gudhi library. The data used are from the protein binding dataset.

3.4.1 TDA with Python using the Gudhi Library. Protein binding can enhance or detract from a drug's performance. As a general rule, agents that are minimally protein bound penetrate tissue better than those that are highly bound, but they are excreted much faster.

The Figure [3.2] is the persistence diagram of the BarCodes-Rips0 (Barcode-Rips by taken only the first line of the distance matrix) and it is showing the correlation between the protein0. In fact the data from Protein binding dataset are coming as a sample of points in a space. The distance matrix was used to represent protein structures in a coordinate-independent and to get the BarCode, we computed

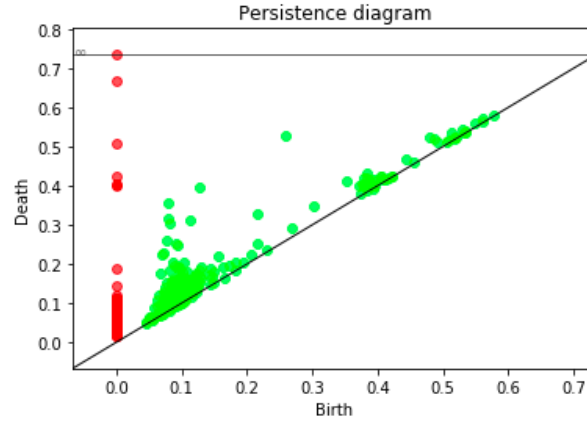


Figure 3.2: Persistence diagram

the 0-dimension homology which is their number of connected component, and from the BarCode, we get the persistence diagram.

3.4.2 Machine Learning and Topological Data Analysis: Visualization of Graphs.. First, we will generate persistence diagrams with orbits of dynamical systems. This dataset is very common in TDA. We use the following system, which depends on a parameter $r > 0$

$$\begin{cases} x_{n+1} = x_n + ry_n(1 - y_n) \\ y_{n+1} = y_n + rx_{n+1}(1 - x_{n+1}) \end{cases}$$

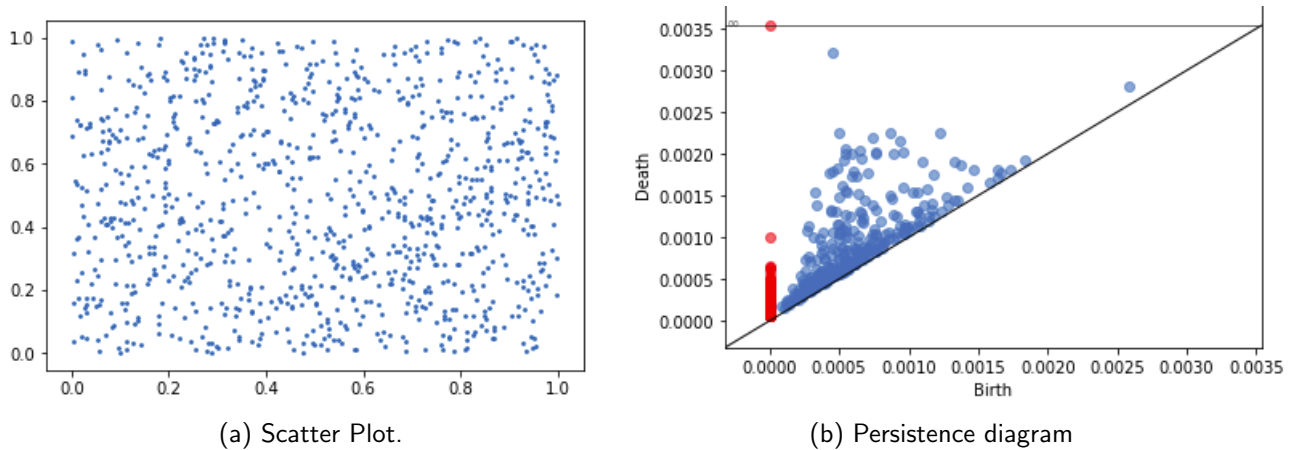


Figure 3.3: persistence diagrams with orbits of dynamical systems.

The Figure [3.3a] shows the scatter plot of the points cloud and by visualizing it, we can think that these points are random. The Figure [3.3b] is the persistence diagram generated from these points cloud by using the α -Complex filtrations. Moreover, we can see that the persistence diagram provides a better dimension reduction of high dimension data more than the scatter plot.

4. Conclusion

Topological Data Analysis has been seen to be a powerful method in data science to study, analyze, and exploratory a very high dimensional data set. Its great success using persistence homology method to code multi-scale features of nested families of simplicial complexes and topological spaces as points in a plane has helped to represent the geometry summaries of the data which come as a sample of metric space by covering this sample with the union of balls with fixed radius and centered in those points using the Mapper method. It has been seen that, when come to the representation of geometry structure of informations, machine learning is not able to give such representation.

Those methods from TDA have proved to be very useful to machine learning techniques to improve the geometry representation of the information. But although all the strategies that TDA has made to encode the representation of the topological information into machine learning compatible representation, they still have difficulties in experimental the target learning task.

In the future work we will improve the Mapper algorithms, especially the choice of the function, the cover and the clusters using by the Mapper algorithms.

Acknowledgements

I would like first of all to thank my parents for the continuous support they have provided to me. A great thank you to the entire AIMS community and especially to AIMS Cameroon for giving me the opportunity to take part in their prestigious program, to build a network with people from all over the world. Thank you to Prof. Mama Foupouagnigni president of AIMS Cameroon for providing an excellent working environment and a caring staff. A special thank you goes to Prof. Marco Garuti, Academic Director of AIMS Cameroon, for offering me a measurable follow-up, motivation in the difficult days. I would also like to thank Professor Franck Kalala Mutombo for providing the support for this master thesis, interesting discussions and beneficial comments are acknowledged and very much appreciated. And a particular thanks to my classmates and my assigned tutor for their support.

References

- [1] Peter Bubenik. The persistence landscape and some of its properties. *arXiv preprint arXiv:1810.04963*, 2018.
- [2] Gunnar Carlsson. Topological data analysis: A framework for machine learning. 2013.
- [3] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference using distance-like functions.
- [4] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for measures based on distance functions. 2011.
- [5] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*, 2017.
- [6] Frederic Chazal and Bertrand Michell. Covers and nerves: union of balls,geometric inference and mapper. 2016.
- [7] AURA Conci and CS Kubrusly. Distance between sets-a survey. *arXiv preprint arXiv:1808.02574*, 2018.
- [8] René Corbet, Ulderico Fugacci, Michael Kerber, Claudia Landi, and Bei Wang. A kernel for multi-parameter persistent homology. *Computers & Graphics: X*, 2:100005, 2019.
- [9] Ulderico Fugacci, Sara Scaramuccia, Federico Iuricich, and Leila De Floriani. Persistent homology: a step-by-step introduction for newcomers. In *Eurographics Italian Chapter Conference*, pages 1–10, 2016.
- [10] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. In *Advances in Neural Information Processing Systems*, pages 1634–1644, 2017.
- [11] Firas A Khasawneh, Elizabeth Munch, and Jose A Perea. Chatter classification in turning using machine learning and topological data analysis. *IFAC-PapersOnLine*, 51(14):195–200, 2018.
- [12] Yumiao Lei. Topological methods for the analysis of applications. In *International Conference on Modern Educational Technology and Innovation and Entrepreneurship (ICMETIE 2020)*, pages 6–9. Atlantis Press, 2020.
- [13] Dindin Mery. From topological data analysis to deep learning: No pain no gain. 2018.
- [14] Bertrand Michel. Statistics and topological data analysis. *Data Rich Phenomena (Modelling, Analysing and Simulation Using Partial Differential Equations)*, 5, 2015.
- [15] Chi Seng Pun, Kelin Xia, and Si Xian Lee. Persistent-homology-based machine learning and its applications—a survey. *arXiv preprint arXiv:1811.00252*, 2018.
- [16] Matteo Rucco, Lorenzo Falsetti, Damir Herman, Tanya Petrossian, Emanuela Merelli, Cinzia Nitti, and Aldo Salvi. Using topological data analysis for diagnosis pulmonary embolism. *arXiv preprint arXiv:1409.5020*, 2014.
- [17] Eric W Weisstein. Königsberg bridge problem. 2002.