Topological Machine Learning

Ankan Kar[a], Pranava Priyanshu[a], Guided by: Dr. Kuntal Roy[b]

[a]*Statistics & Mathematics Unit, Indian Statistical Institute, Bagalore, India*
[b] *Indian Institute of Science Education and Research, Bhopal, India*

Abstract

This report focuses on applications of Topological Machine Learning in real life. Here we provided an overview of Algebraic Topology's significance in Topological Machine Learning and Data Analysis. We explain Persistence Homology, chain complexes, and simplexes, highlighting their utility in analyzing data. Visualization techniques are employed to illustrate these concepts, along with corresponding analyses. We introduce Neural Networks and their role in Topological Machine Learning. We conclude by discussing a vectorization method for feeding input to Neural Networks in the context of Machine Learning and Topological Data Analysis.

Keywords: Topological Machine Learning, Topological Data Analysis, Persistent Homology, Chain Complex, Simplex, Neural Network

1. Introduction

Topological machine learning represents a burgeoning field situated at the intersection of topological data analysis (TDA) and machine learning. Topology, informally referred to as the encoding of the overall shape of data, provides a complementary approach to localized and rigid geometric features. Topological features possess the ability to effectively capture multiscale, global, and intrinsic properties present within datasets. The utility of such features has garnered significant recognition alongside the ascent of TDA, establishing topological information as generally accepted and pertinent in the context of data analysis.

Recent progress in TDA has witnessed the integration of topological methodologies into classical machine learning techniques and deep learning models, thereby affording opportunities for enhancement and augmentation. However, a notable challenge in various data analysis algorithms, including those employed in TDA, arises from the necessity to set multiple parameters. In the absence of prior domain knowledge, the correct selection of parameters for a given dataset becomes arduous.

The principal insight of persistent homology lies in its approach to harnessing information extracted from all parameter values, thereby encoding this extensive wealth of information into an intelligible and easily representable format. Within the realm of TDA, a mathematical interpretation emerges when this information is represented in the form of homology groups. Generally, the prevailing assumption posits that features persisting across a wide range of parameters hold the status of "true" features, bearing crucial significance in the data under analysis.

2. Some Concepts

In the latter sections of this paper, several key concepts will be extensively employed and elucidated to advance the research and enrich the discourse. The following brief explanations provide a foundational understanding of these concepts, serving as essential building blocks for deeper exploration and analysis within the context of the paper's subject matter.

2.1. Point Cloud

Point cloud refers to a set of data points in a high-dimensional space without any specific structure or connectivity. It serves as the raw input for topological data analysis (TDA) techniques. TDA methods analyze point clouds to extract essential topological features and structures, revealing hidden patterns and properties in complex datasets. By leveraging tools like persistent homology, point clouds enable the exploration and understanding of data from a topological perspective, making them crucial in various applications such as computer vision, robotics, and scientific research.

2.2. Simplex

Simplexes are fundamental geometric elements used to construct higher-dimensional spaces. A simplex of dimension "n" is a geometric object formed by connecting "n+1" points in a straight line (1-dimensional), a triangle (2-dimensional), a tetrahedron (3-dimensional), and so on. In simplicial complexes, a collection of simplexes are combined to represent the underlying data structure. Simplexes play a crucial role in topological data analysis (TDA) as they help identify and measure topological features, such as connected components and voids, enabling the understanding and analysis of complex datasets from a topological perspective.

2.3. Chain Complex

Chain complex is a fundamental mathematical construct used in algebraic topology to study topological spaces. It comprises a sequence of vector spaces (modules) and linear maps

called boundary operators that capture the boundaries of higher-dimensional simplices in a simplicial complex. Chain complexes are central to computing homology, which quantifies the topological features, such as connected components, holes, and voids, present in data represented as simplicial complexes. In topological machine learning, chain complexes play a crucial role in extracting topological features from data and facilitating their analysis and interpretation.

2.4. Nerve Complex

Nerve complex is a mathematical construct used to capture the topological relationships among subsets of a dataset. It is formed by considering the intersections of cover elements, where each element represents a region containing data points. The nerve complex helps extract topological features, providing valuable insights into the data's underlying structure and facilitating analysis in TDA applications.

2.5. Čech Complex

Čech complex is a fundamental mathematical construct used to capture topological features within a dataset. It is formed by considering the pairwise distances between data points and including simplices for sets of points that are within a certain distance threshold. The Čech complex helps uncover geometric and topological properties, such as connected components and voids, facilitating the analysis and understanding of complex datasets from a topological perspective.

2.6. Persistence Homology Group

Persistence Homology Group is a fundamental algebraic structure used to quantify and characterize topological features in a dataset. It represents the lifespan and connectivity of topological features, such as connected components and holes, as a parameter (e.g., radius) varies. Persistence Homology Groups provide valuable insights into the persistent and robust topological properties of data, allowing for a deeper understanding of complex datasets and aiding in data analysis and classification tasks. The persistent homology group PH of a point cloud is the persistence module defined as $PH_k(X) = \prod H_k(X_r)$ where X_r is the Čech complex of radius r of the point cloud X and H_k is the homology group.

2.7. Persistence Barcode

It is a multiset of intervals of $\mathbb{R}$. Persistence Barcode is a concise visual representation of the lifespan of topological features, such as connected components and holes, across different parameter values. It displays the birth and death of these features as horizontal bars, with the length of each bar corresponding to the duration of the feature's existence. Persistence Barcodes are a powerful tool for analyzing and comparing topological features in complex datasets, facilitating pattern recognition and data understanding in various applications.

2.8. Persistence Diagram

Persistence Diagram is a compact graphical representation of the birth and death of topological features, such as connected components and holes, as they persist across different parameter values. It consists of points in a 2D plane, where each point represents a topological feature, and its position indicates its birth and death parameters. Persistence Diagrams are essential for characterizing and comparing topological features in complex datasets, enabling efficient data analysis and pattern recognition.

2.9. Persistence Image

Persistence Image is a visual representation in topological machine learning that converts the output of persistent homology, which captures topological features in data, into a fixed-size image. Each pixel in the image represents a specific topological feature and encodes its "persistence" or lifetime during the analysis. It enables efficient analysis and processing of topological information using standard image processing techniques.

3. Algebraic Topology in Topological Machine Learning

In data analysis, the "manifold hypothesis" posits that data points are not randomly scattered but lie on a low-dimensional, smooth manifold embedded in a higher-dimensional space. Traditional statistical analysis focuses on summarizing data using statistical descriptors, but topological data analysis (TDA) takes a different approach by investigating the underlying manifold structure using algebraic methods.

Algebraic topology is a branch of mathematics that studies topological spaces through algebraic structures associated with them. In the context of TDA, the goal is to classify or distinguish between different underlying manifolds represented in the data. To achieve this, algebraic topology provides tools to associate computable algebraic structures, such as groups or vector spaces, to the manifold in a way that remains unchanged regardless of how the manifold is stretched, twisted, or bent (i.e., under homeomorphisms).

One crucial class of algebraic invariants used in TDA is called homology groups. These homology groups provide a way to encode essential topological features of the manifold in a computationally efficient manner. The idea is to represent the manifold combinatorially using chain complexes. A chain complex is a sequence of vector spaces (or groups) and linear maps that capture the connectivity and holes of the underlying manifold at different dimensions. The homology groups are then derived from these chain complexes and serve as stable and informative representations of the manifold's topological structure.

By utilizing the homology groups as algebraic invariants, TDA can robustly distinguish between different types of underlying manifolds, even when the data is noisy or incomplete. This allows TDA to offer valuable insights into the underlying shape and structure of complex data sets, making it a powerful tool in data analysis and machine learning applications.

Figure 1: Simplices in increasing dimensions of connected components. From a 0-dimensional point to a 3-dimensional tetrahedron.

Figure 2: Centre: Data, Left, Right: Potential Simplical Complexes

3.1. Simplex

Simplices are fundamental geometric objects used to build chain complexes and compute homological invariants. A k-simplex, denoted as Δ^k, can be understood as the standard k-simplex Δ^k. It is defined as the convex hull of the standard basis vectors in the (k+1)-dimensional Euclidean space, denoted as $\mathbb{R}^{k+1}$. Geometrically, Δ^k is a k-dimensional solid shape formed by k+1 vertices, where each vertex corresponds to one of the standard basis vectors (e.g., e_0 = (1, 0, ..., 0), e_1 = (0, 1, ..., 0), ..., e_k = (0, 0, ..., 1)). Mathematically, it is represented as follows: $\Delta^k := \{(x_0, x_1, ..., x_k) \in \mathbb{R}^{k+1} \mid \sum_{i=0}^{k} x_i = 1, x_i \geq 0 \; \forall \; i\}$.

We can also look at a general k-simplex as $[v_0, ..., v_k]$. This is defined as the convex hull of k+1 affinely independent points $v_0, v_1, ..., v_k$ in a k-dimensional Euclidean space. Affine independence ensures that the k+1 points do not lie on any hyperplane, and they are not collinear. This general k-simplex represents a broader class of shapes in the Euclidean space.

Simplices serve as the building blocks in algebraic topology, and they are used to construct chain complexes. A chain complex is a sequence of vector spaces (or groups) connected by boundary operators. These operators describe the boundary of each simplex in the chain complex and play a crucial role in computing homological invariants.

A singular k-simplex in a topological space X is a continuous map $\sigma : \Delta^k \to X$. It represents a way of mapping the geometric shape of Δ^k into the topological space X. For example, if X is a topological manifold, singular simplices allow us to model the manifold's features and shapes using these maps.

An important singular simplex in Δ^k is the inclusion of the i-th face of Δ^k, denoted as $F_i^k : \Delta^{k-1} \to \Delta^k$. The inclusion map represents how the (k-1)-dimensional face is embedded into the k-dimensional simplex. This inclusion is a fundamental concept in algebraic topology.

In this context, we are working in $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ for simplicity, which means we are considering algebraic structures over the field with two elements, making computations more manageable. Overall, the use of simplices, singular simplices, and chain complexes allows us to study and analyze the topological properties of spaces, including topological manifolds, in an algebraic manner.

We gave an example of simplices in Figure 1 where it displays simplices progressing from 0-dimensional to 3-dimensional connected components, arranged from left to right. It starts with a 0D point, followed by 1D two loops, 2D filled region, and ends with a 3D tetrahedron. In Figure 2 we can see the centre image is the data. The left complex obtained from the data is composed of various connected components in increasing dimensions: one 0-dimensional connected component (a single point), two 1-dimensional connected components (representing two separate loops), one 2-dimensional connected component (a filled region), and one 3-dimensional connected component (a solid tetrahedron). On the other hand, the right complex obtained from the same data depicts a 2-dimensional face that is not entirely connected, leading to the representation of a 2-dimensional loop, where certain parts of the face remain disjointed.

3.2. Chain Complex

Definition: A chain complex (A_*, d_*) is a sequence of abelian groups or modules $..., A_0, A_1, A_2, A_3, ...$ connected by homomorphisms (called boundary operators or differentials) $d_n : A_n \to A_{n-1}$, such that the compositions of any two consecutive maps is the zero map, i.e, $d_n \circ d_{n+1} = 0$.

In topological machine learning, the input data, often represented as a simplicial complex or a point cloud, is first transformed into a chain complex. The chain complex captures the topological features of the data at various dimensions. Each element in the chain complex is referred to as a chain, and it represents a certain collection of simplices or points in the dataset.

The boundary operators, d_i, play a crucial role in the chain complex. These operators describe how the chains in one dimension are related to the chains in the next lower dimension. Essentially, the boundary operator computes the boundary of each chain, and this boundary is represented as a linear combination of lower-dimensional chains. It describes how the simplices or points in one dimension are "glued together" to form the simplices in the next lower dimension.

The key property of the chain complex is that the composition of any two consecutive boundary operators results in a zero map. This property is known as the boundary operator's nilpotency. It ensures that the chains' boundaries close up, meaning that the boundaries of the boundaries are empty. This property is essential in algebraic topology as it allows the computation of topological invariants, such as homology groups.

In summary, chain complexes are vital algebraic structures used in topological machine learning to capture and represent the topological features of data. They enable the computation of topological invariants, which help analyze and understand the underlying manifold structure of the data in an algebraic fashion.

In any space X, its singular k-chains are elements of a vector space over the field $\mathbb{F}_2$, denoted as $C_k(X)$. These chains are generated by considering all singular k-simplices in X. The singular chain complex (C(X), ∂) of X is a sequence of vector

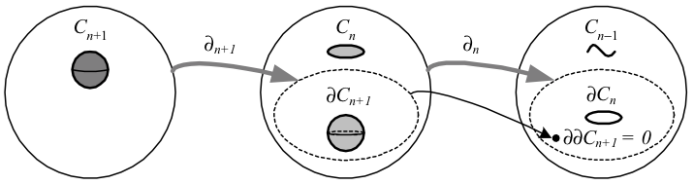


Figure 3: A demonstration of a Chain Complex with C_n as the n-chains

spaces with boundary maps. Each space $C_k(X)$ corresponds to the singular k-chains in X, and the boundary map ∂_k: $C_k(X) \to C_{k-1}(X)$ is defined as $\partial(\sigma) : \sum_i \sigma \circ F_i^k$. We have the elements of the group $Z_k(X) := \ker(\partial_k)$ as k-cycles in a given space X. These cycles are elements in the kernel of the boundary operator ∂_k, which means that they have no boundary themselves.On the other hand, the elements of the group $B_k(X) := \mathrm{im}(\partial_{k+1})$ are called k-boundaries. These boundaries are the images of the (k+1)-dimensional boundary operator ∂_{k+1}, representing k-cycles that are the boundary of some (k+1)-dimensional region.

The quotient group $H_k(X) := Z_k(X)/B_k(X)$ is defined as the k-th singular homology group of the space X. This group represents the equivalence classes of k-cycles modulo k-boundaries, providing a measure of the k-dimensional holes or topological features in X.Homology groups are topological invariants, meaning they remain unchanged under homeomorphisms. As a result, they encode intrinsic topological information about X. Utilizing homology groups and derived invariants, such as the Betti numbers $\beta_k := \dim(H_k(X))$, is valuable for studying classification problems. For instance, the 0-th Betti number β_0 counts the connected components of the space, reflecting its number of isolated pieces. Meanwhile, β_1 counts the number of 1-dimensional cycles, which corresponds to loops or nontrivial connected paths in the space. In short, the homology groups and their associated invariants are powerful tools in algebraic topology, allowing us to study the topological structure of a space and classify it based on its intrinsic topological properties.

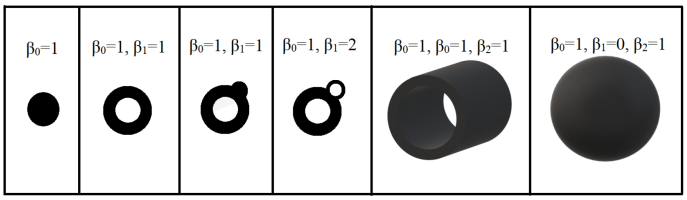


Figure 4: Betti numbers (β_k) characterizing topological holes in 0 to 3-dimensional space

4. Persistence Homology

Persistent homology is a method used to compute topological features of a space at different spatial resolutions. It aims to identify more persistent features, those that persist over a wide range of spatial scales, as they are considered to be more indicative of genuine underlying features rather than artifacts of sampling, noise, or parameter choices.

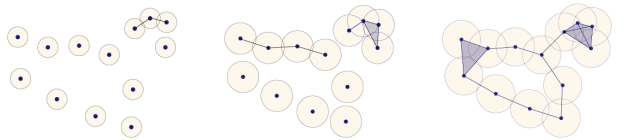
4.1. Filtration

Persistent homology extends the concept of homology to filtered chain complexes. A filtered chain complex is a sequence of chain complexes, denoted as $C^{\varepsilon_0} \subset C^{\varepsilon_1} \subset C^{\varepsilon_2} \subset C^{\varepsilon_3} \subset \ldots$, where inclusion maps $\iota^i : C^{\varepsilon_i} \to C^{\varepsilon_{i+1}}$ are present. Here, ε is a parameter that characterizes the spatial resolution.

In a filtered chain complex, the inclusion maps $\iota^{i,j} := \iota^j \circ \iota^{j-1} \circ \ldots \circ \iota^i : C^{\varepsilon_i} \to C^{\varepsilon_j}$ for $i < j$ describe the relationship between different chain complexes in the sequence. These filtered chain complexes naturally emerge in situations where there is a sequence of inclusions of spaces, such as $X^{\varepsilon_0} \subset X^{\varepsilon_1} \subset X^{\varepsilon_2} \subset X^{\varepsilon_3} \subset \ldots$.

By analyzing the evolution of homological features across different spatial scales in the filtered chain complex, persistent homology allows us to identify and quantify topological features that persist over a wide range of resolutions. These persistent features provide valuable insights into the underlying structure of the space and are robust indicators of genuine topological characteristics.

In certain scenarios, such as when considering sublevel sets $X_\varepsilon := f^{-1}(R_{<\varepsilon})$ of a filtration function $f : X \to \mathbb{R}$ or when examining a point cloud Y in a metric space (M, d) and defining $Y^\varepsilon := \bigcup_{y\in Y} B_\varepsilon(y) = g^{-1}(\mathbb{R}_{<\varepsilon})$ with filtration function $g : M \to \mathbb{R}$ given by $g(m) := \inf_{y\in Y}\{d(m, y)\}$, cases arise where we encounter similar structures. Here, $B_\varepsilon(y)$ represents the open ball of radius ε centered at y, and we implicitly consider $\varepsilon \simeq \varepsilon'$ if X^ε (or Y^ε) is canonically homeomorphic to X^δ (or Y^δ) for all $\delta \in [\varepsilon, \varepsilon']$.



(a) Rips complex at some scale (b) ..at a larger scale. (c) ...and at an even larger scale.

Figure 5: Example of a Vietoris-Rips filtration on point cloud data

An important property of (singular) homology is that it is functorial, meaning the inclusion maps $\iota^{i,j}$ induce well-defined maps on the respective homology groups, denoted as $H_k(\iota^{i,j}) : H_k(C^{\varepsilon_i}) \to H_k(C^{\varepsilon_j})$. This property ensures that the topological features are preserved under the inclusion of spaces within the filtration sequence. The Figure 6 provided below illustrates the construction of the Vietoris-Rips complex based on a distance filtration, a standard and widely used approach in topological data analysis (TDA). The Vietoris-Rips complex captures the topological structure of data points based on pairwise distance relationships, allowing for meaningful topological analysis in TDA applications.

Consider an illustrative example as in Figure 7 of a union-of-spheres where the Čech and Vietoris-Rips filtrations would yield different outcomes. In this scenario, if the space (denoted by red) is sufficiently small, the Vietoris-Rips complex would interpret these spheres as forming a 3-way connected component due to their close proximity. Conversely, the Čech filtration

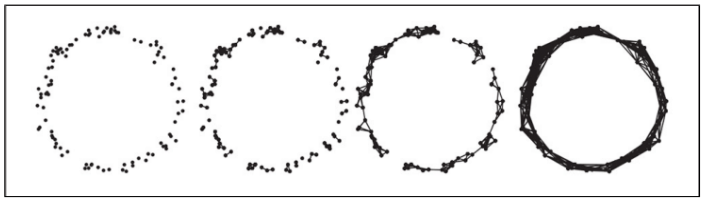


Figure 6: The Vietoris-Rips filtration for a "circle" point cloud shows various stages from left to right, where the connectivity of the underlying simplicial complex increases with the increment of ε

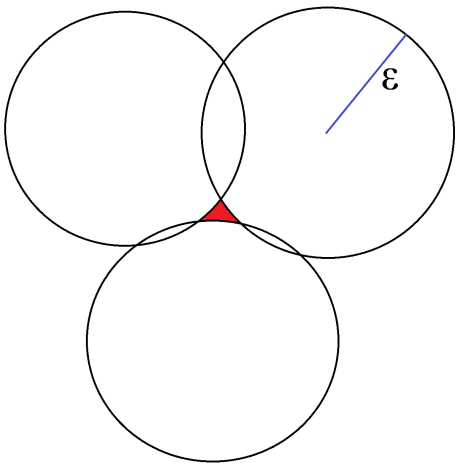


Figure 7: Example of a union-of-spheres at which the Čech and VR filtrations would differ.

would treat them as pairwise connected with a hole between them, capturing the absence of direct connections between the spheres despite their proximity. This disparity arises due to the different ways in which the two filtration methods account for proximity and connectivity in the point cloud data.

The k-th persistent homology groups are characterized as the images of the inclusion maps $\iota^{i,j}$, denoted as $H_k^{i,j} := im(H_k(\iota^{i,j})) = Z_k(C^{\varepsilon_i})/(B_k(C^{\varepsilon_i}) \cap Z_k(C^{\varepsilon_i}))$. These groups precisely consist of the k-th homology classes of C^{ε_i} that persist and still exist after applying the inclusion map $H_k(\iota^{i,j})$. A homology class $\alpha \in H_k(C^{\varepsilon_i})$ is considered "born" at C^{ε_i} if $\alpha \notin H_k^{i-1,i}$. If α is born at C^{ε_i}, it is said to "die" at C^{ε_j} if $H_k(\iota^{i,j-1})(\alpha) \notin H_k^{i-1,j-1}$ and $H_k(\iota^{i,j})(\alpha) \in H_k^{i-1,j}$. The persistence of α is given by $\varepsilon_j - \varepsilon_i$ and is set to infinity if it never dies. The persistent Betti numbers, defined as $\beta_k^{i,j} := dim(H_k^{i,j})$, carry valuable information on how the homology changes across the filtration. This information can be succinctly represented using a persistence diagram, which is a multiset $\overline{\mathbb{R}}^2 := \mathbb{R} \cup (\mathbb{R} \times \{\infty\})$.

Specifically, the persistence diagram of (homological) dimension k consists of points $(\varepsilon_i, \varepsilon_j) \in \overline{\mathbb{R}}^2$ with multiplicity $\mu_k^{i,j} := (\beta_k^{i,j-1} - \beta_k^{i,j}) - (\beta_k^{i-1,j-1} - \beta_k^{i-1,j})$ for all $i < j$. The multiplicity $\mu_k^{i,j}$ counts the number of k-th homology classes that are born at C^{ε_i} and die at C^{ε_j}. These persistence diagrams provide a concise and informative representation of the evolution of homological features across the filtration, aiding in the analysis and understanding of topological structures present in the data.

4.2. Persistence Diagram

The Figure 8 presents a persistent diagram derived from the Vietoris-Rips complex in Figure 6. In this diagram, the axes represent the ε values at which topological features are created and destroyed, respectively. Notably, the single point of high persistence corresponds to the primary topological feature of the point cloud, specifically its circular shape. Additionally, other topological features occur at smaller scales, denoting lower values of ε, forming a small dense cluster in the lower-left corner of the persistence diagram.

The persistent Betti-numbers, which provide information about the number of topological holes of different dimensions, can be recovered directly from the persistence diagram itself. Specifically, the number of connected components, loops, voids, and higher-dimensional features can be inferred from the pattern of points and their persistence in the diagram. This representation aids in comprehending the hierarchical evolution of topological features as the spatial resolution changes, offering valuable insights into the structure of the underlying data.

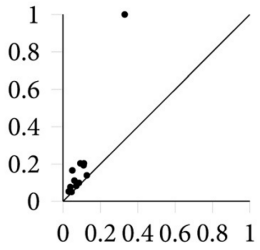


Figure 8: A persistence diagram contains 1-dimensional topological features (cycles) as points representing the births and deaths of these features, where each point corresponds to a cycle born at a certain value of ε and dying at another, providing information about their persistence across spatial resolutions

$\mathbf{d_c^p}$ Distance: Let D_x and D_y be two persistence diagrams with cardinalities n and m respectively such that $n \leq m$, and denote $D_x = \{x_1, ..., x_n\}$, $D_y = \{y_1, ..., y_m\}$. Let $c > 0$ and $1 \leq p \leq \infty$ be fixed parameters. The d_c^p distance between two persistence diagrams D_x and D_y is given by: where $\prod_m$

$$d_c^p(D_x, D_y) = \left(\frac{1}{m}\left(\min_{\pi \in \prod_m} \sum_{l=1}^{n} min(c, ||x_l - y_{\pi(l)}||_\infty)^p + c^p|m-n|\right)\right)^{\frac{1}{p}}$$

is the set of permutations of $(1, ..., m)$. If $m < n$, then define $d_c^p(D_x, D_y) := d_c^p(D_y, D_x)$.

There are several methods of comparing persistence homology outside of the persistence diagram comparisons. One of the common formulations, which has some convenient statistical results, is the landscape distance. In this method, one converts (birth, death) coordinates to a system of orthogonal lambda functions with coordinates (m, h) mapped as follows: $birth \rightarrow m, \quad death \rightarrow h$. This transformation results in a new space called the "persistence landscape," where each lambda

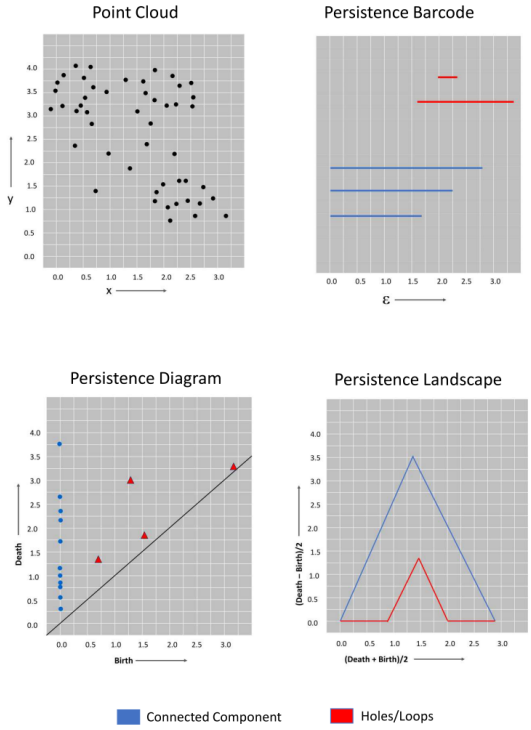


Figure 9: An example of a point cloud dataset, persistence barcode, persistence diagram, and persistence landscape, providing insights into the topological features, their durations, and evolutionary patterns.

function captures the topological persistence of a specific feature.

The persistence landscape resides in a Hilbert space, which allows for the definition of distances and angles, enabling convenient statistical analysis. Estimates of the mean in this space abide the central limit theorem, providing robust statistical properties.

The persistence landscape is stable, ensuring that small changes in the original persistence diagram result in small changes in the landscape representation, making it robust to noise and data perturbations.

Moreover, the landscape distance serves as a stable lower bound on the Wasserstein bottleneck distance, and it offers a straightforward way to calculate Z-scores, confidence intervals, etc., making it suitable for statistical applications.

Overall, the landscape distance provides a simple and clean formulation for statistical applications in the context of persistence homology, with desirable properties for analyzing and comparing topological features across datasets.

4.3. Persistence Landscape

After mapping the persistence points to the coordinates (m, h) using the transformations $m = \frac{death+birth}{2}$ and $h = \frac{death-birth}{2}$, the rescaled rank function $\lambda : \mathbb{R}^2 \to \mathbb{R}$ is defined to encode the topological persistence information in a compact and interpretable way.

The function $\lambda(m, h)$ is designed to capture the topological significance of each point in the persistence diagram. When $h \geq 0$, meaning the point corresponds to a valid persistence pair with a positive lifespan, $\lambda(m, h)$ takes the value $\beta^{t-m,t+m}$, which represents the Betti number at dimension k. This Betti number indicates the rank of the homology group of the topological space at the given dimension k and time t. Essentially, $\lambda(m, h)$ associates each point in the persistence diagram with the Betti number of the corresponding topological feature.

However, if h is negative, it implies that the point does not form a valid persistence pair. In this case, $\lambda(m, h)$ is set to 0 to indicate the absence of a meaningful topological feature at that point.

The persistence landscape $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ is then defined to further summarize the topological persistence information across various dimensions k and times t. For each dimension k and time t, $\lambda(k, t)$ is calculated by taking the supremum (or maximum) of m over all possible values, subject to the condition that the associated Betti number $\beta^{t-m,t+m}$ is greater than or equal to k. This function $\lambda(k, t)$ effectively identifies the longest persistence time for topological features with Betti numbers greater than or equal to k, offering insights into the existence and duration of such features.

The persistence landscape can be viewed as a collection of functions $\lambda_k : \mathbb{R} \to \mathbb{R}$, where each λ_k corresponds to a different value of k. By analyzing these functions, one can gain a comprehensive understanding of the evolution of topological features across various dimensions and times, making it a powerful tool for analyzing complex datasets.

In summary, the persistence landscape serves as a rich and informative representation of the topological persistence in the data. Its ability to capture the duration and significance of topological features makes it a valuable approach in various statistical applications of persistence homology, providing researchers with deeper insights into the underlying structure and dynamics of the data.

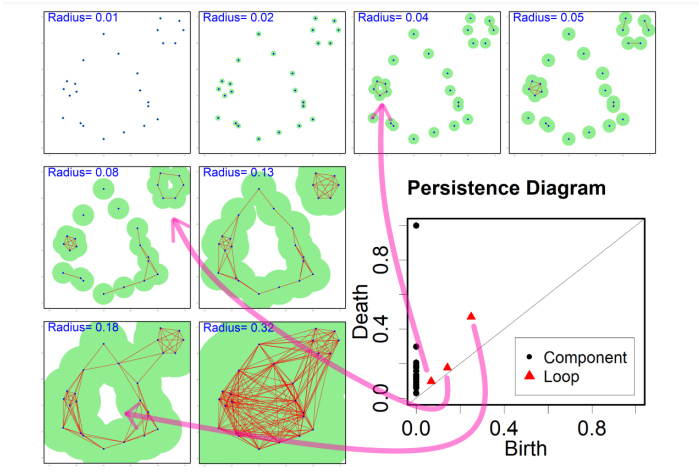


Figure 10: An illustrative example of a Persistence Diagram capturing the birth and death diameters of holes in a topological representation of a dataset. The diagram visually represents the lifespans of these topological features, providing insights into their creation and disappearance across different spatial resolutions.

4.4. Persistence Image

Persistence Image is a powerful and versatile technique in Topological Data Analysis (TDA) and Topological Machine Learning (TML) that allows us to transform the topological features of a dataset into a more amenable format for machine learning algorithms. It bridges the gap between topological features, typically represented by persistence diagrams or barcodes, and numerical data, enabling the application of traditional machine learning methods.

The process of creating a Persistence Image involves several steps:

(i) *Persistence Diagrams/Barcodes:* Given a dataset, TDA extracts its topological features using techniques like Persistent Homology, which provides a set of persistence diagrams or barcodes. These diagrams represent the birth and death times of topological features such as connected components, loops, and voids.

(ii) *Binning:* To convert the persistence diagram into a numerical format, the points in the diagram are binned into a grid. The bins are defined based on a grid resolution and cover the range of birth and death times in the diagram.

(iii) *Weighting Scheme:* Each point in the persistence diagram contributes to the corresponding bins in the grid, and a weighting scheme is applied to capture the significance or importance of each topological feature. Common weighting schemes include Gaussian or triangular kernels centered at the points.

(iv) *Integration:* The contributions from all points in the persistence diagram are integrated within their respective bins. This process generates a pixel intensity for each bin, representing the accumulated significance of topological features within that region of the grid.

(v) *Normalization:* Finally, the Persistence Image is normalized to ensure that it remains invariant to the scale of the input dataset. This normalization step makes the Persistence Image more robust and comparable across different datasets.

The resulting Persistence Image is a numerical representation of the topological features of the original dataset, suitable for various machine learning algorithms. It retains essential topological information while being amenable to standard image-based machine learning techniques.

Persistence Image has found applications in various domains, including object recognition, shape classification, and time series analysis. Its ability to capture the topological essence of complex datasets while leveraging established machine learning methodologies makes it a valuable tool in the intersection of TDA and TML.

5. Neural Network

Neural Networks (NNs) have been widely recognized for their ability to approximate arbitrary continuous functions in $\mathbb{R}^n$, making them popular for various pattern recognition tasks since the early 2000s. However, during that era, computational power was limited, which led researchers to seek efficient formulations, even for small networks.

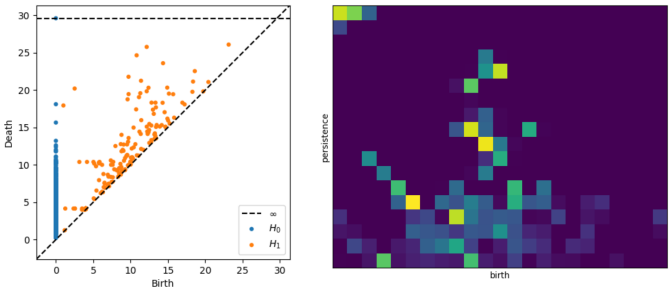


Figure 11: An example of a Persistence Image (right) derived from the Persistence Diagram (left) of a dataset where H_0 are connected components and H_1 are holes.

In the 1990s, a significant development emerged in constructing topologically optimal three-layer NNs. This approach drew upon foundational works from mathematical literature, demonstrating that it is possible to estimate the size of latent dimensions in the data space and limit an NN to the minimum number of nodes and edges required to effectively approximate this space. The key motivation was to find compact networks with reduced complexity to mitigate symmetries in the solution space and achieve more efficient approximations. While this procedure was suitable for small networks, it became computationally infeasible for larger networks and big data applications. As a result, finding computationally efficient topologies for neural networks remained a challenging problem, especially in modern high-dimensional settings.

Recent advancements have witnessed the application of persistent homology in exploring the topological optimization of neural networks. This approach aims to leverage topological data analysis techniques to identify and exploit essential topological features in the data and network structures, potentially leading to improved network design and more efficient approximations in large-scale applications. The integration of topological insights into neural network architectures presents exciting opportunities for advancing the understanding and optimization of complex models in the era of big data and high-performance computing.

5.1. Topological Machine Learning with Neural Network

The integration of persistence homology, a powerful topological data analysis technique, with neural networks has emerged as a promising approach to enhance their capabilities in handling complex data. The most commonly used methods for incorporating persistence homology into neural networks involve utilizing vectorized forms of topological filtration, such as persistence barcode, landscape, or persistence image. These representations enable the translation of topological features into numerical formats that can be efficiently processed by neural network architectures.

To incorporate persistence features into neural networks, we first need to discretize the continuous hierarchical signal produced by the persistence filtration. This continuous signal can be viewed as a point cloud in a high-dimensional space. The

discretization process involves choosing a fixed number of bins for the analysis. For persistence barcodes and landscapes, this entails selecting a single width for the bins, while for persistence landscapes, which have two-dimensional binning, independent choices for the width and height of the bins can be made.

The selection of appropriate binning parameters is a critical step, as it influences the resolution and granularity of the topological information captured by the persistence features. Too few bins might result in the loss of important topological features, leading to suboptimal neural network performance. Conversely, using too many bins could result in overfitting and increased computational complexity, making the model less interpretable and harder to generalize to unseen data.

Determining the optimal binning strategy becomes essential to effectively leverage persistence features for neural network tasks. Researchers and practitioners need to consider the trade-off between capturing essential topological information and maintaining computational efficiency and model interpretability. Additionally, the choice of binning can impact the robustness and generalization capabilities of the neural network models utilizing these topological features. By combining the expressive power of neural networks with the topological insights provided by persistence homology, researchers can design more sophisticated models that can extract meaningful features from complex data, leading to improved performance in various machine learning and pattern recognition tasks. The integration of persistence features into neural networks represents a promising direction for advancing the field of topological machine learning and exploring new avenues for understanding and analyzing complex data structures.

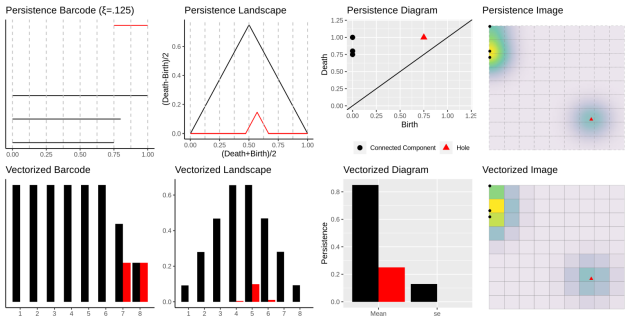


Figure 12: Demonstration of 1 and 2-dimensional binning methods for the vectorization of persistence barcodes, landscapes, diagrams, and images. These examples showcase commonly used techniques for converting persistent information into vectorized representations, retaining various aspects of the original persistence information.

After having the vectorized form of persistent information, we can easily use this as input for many algorithms. Support vector machines (SVMs), NNs, clustering algorithms (such as k-means), and others have all been shown effective in application. In NNs, the vectorized input is used as the input to the 0-th layer. If the data is already a single column, such as in the vectorized landscape, this can be input directly. If the vectorized data is greater than 1-dimensional, such as the persistence image, then they must first be unrolled into a single, "flat," column.

Also, we can use the signal attribute of some vectorization methods to input the data first to convolutional layers and treat the persistence information as an n-dimensional signal. By employing machine learning algorithms like SVMs, NNs, clustering with vectorized persistent information, researchers can effectively leverage topological features to improve the performance of their models in various applications. The utilization of these topological signatures in machine learning enables enhanced data analysis, classification, and pattern recognition, thereby expanding the applications of topological data analysis in diverse fields.

5.2. Topological Data Analysis in Neural Network

Various processes of Topological Data Analysis (TDA) have been successfully applied to Neural Networks (NNs) in diverse applications. A common approach is to directly incorporate persistence barcode data or persistence diagram data as inputs to an NN. Although these processes can vary widely in their specific applications, Convolutional Neural Networks (CNNs) are frequently utilized to optimize over persistence homologies.

However, a drawback to these studies, from a methodological standpoint, is that they often involve multiple choices in pre and post-processing steps, bootstrapping techniques, and other aspects. Additionally, NN learning is inherently a complex classifier with a multitude of parameterization choices, making it challenging to interpret and generalize results across different applications.

The combination of these numerous choices and the complexity of the classifier can hinder the generalization of results, potentially leading to non-robust outcomes. Furthermore, some of these methods, except for the vectorized landscape, rely on a vectorization of persistence information that is known to be unstable, despite its demonstrated success in empirical studies. In particular, the vectorization of the barcode and persistence diagram might encounter issues related to stability, while the vectorized landscape and persistence image have been shown to be more stable and reliable alternatives.

Hence, researchers should be cautious when applying TDA techniques to NNs and carefully consider the implications of various choices made during the process. Moreover, it is crucial to explore and identify more stable vectorization methods, such as the vectorized landscape and persistence image, to ensure the robustness and validity of the obtained results in real-world applications.

6. Results and Observations

We applied the concepts of topological machine learning to the "shapes.zip" dataset available in Python libraries. Specifically, we focused on four shapes from the dataset and performed the following steps:

(i) Data Preprocessing: We loaded the "shapes.zip" dataset and extracted the four shapes of interest.

(ii) We employed topological data analysis (TDA) techniques, specifically persistent homology, to analyze the dataset.

After applying persistent homology, we obtained a point cloud plot to visualize the distribution of one attribute in the dataset. This plot serves to clarify the spatial distribution and patterns of the data points with respect to the selected attribute and the whole dataset.

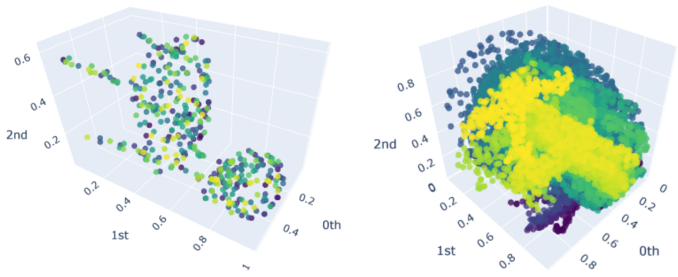


Figure 13: Left: Point cloud plot displaying the distribution of one specific attribute from the dataset. Right: Point cloud plot displaying the distribution of all attributes considered in the dataset.

(iii) We then made the persistence diagram for the considered attributes in the dataset.

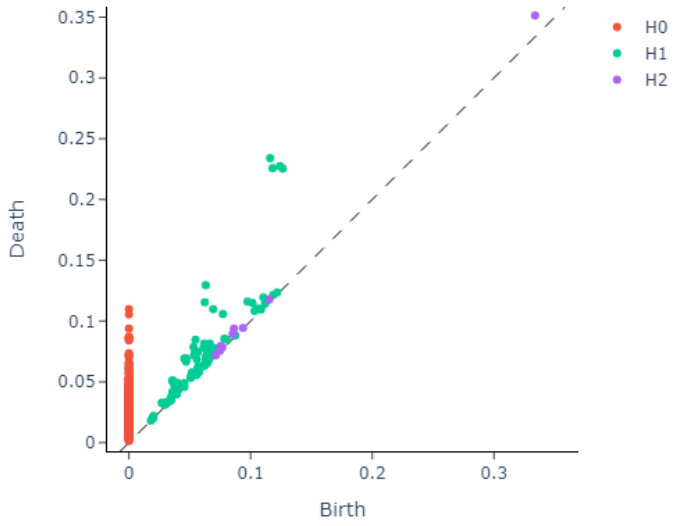


Figure 14: Persistence Diagram of the considered shapes

(iv) We convert each persistence diagram into a 3-dimensional vector using the persistent entropy method, which allows us to quantify the topological features. The resulting feature matrix is then plotted, providing a visual representation of the data's topological characteristics in a higher-dimensional space.

(v) Given the small sample size, we chose Random Forest Classification as the machine learning algorithm. We used the vectorized persistence information as input features (here we used feature matrix) and the labeled classes as target vectors to train the Random Forest model.

Link to the code is here: *Topological – machine – learning – using – Random – Forest.ipynb*

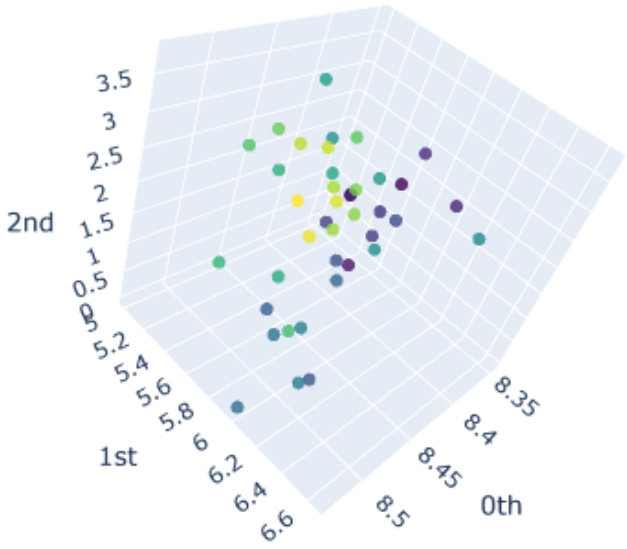


Figure 15: Point Cloud of Persistence Feature Matrix

7. Conclusions

TDA and TML have proven to be valuable tools for analyzing and understanding complex datasets. They provide a fresh lens to explore and interpret the underlying structures of data, uncovering hidden patterns and relationships. By capturing topological features, these approaches offer a more robust and meaningful representation of data, leading to improved machine learning models' performance. As the fields continue to advance, they hold great promise in various scientific, engineering, and real-world applications, contributing to the progress of data analysis and artificial intelligence.

Author Contribution

Ankan Kar: Conceptualization, Methodology, Validation, Inference, Implemantation, Writing – original draft, Revised manuscript. **Pranava Priyanshu**: Conceptualization, Methodology, Validation, Inference, Implemantation, Writing – original draft, Revised manuscript. **Dr. Kuntal Roy**: Supervision, Writing – review.

Acknowledgements

We would like to acknowledge all the researchers, scholars, and individuals whose work and contributions have paved the way for advancements in the field. Though there are no specific individuals to acknowledge, their collective efforts and the broader scientific community's dedication have undoubtedly influenced and inspired our research. We extend our gratitude to everyone who has contributed to the pool of knowledge, making this journey possible.

References

Hensel Felix, Moor Michael, Rieck Bastian, et al. 2021, “A Survey of Topological Machine Learning Methods”, Frontiers in Artificial Intelligence, Volume 4. DOI: 10.3389/frai.2021.681108, ISSN: 2624-8212

Love, Ephraim Robert, “Machine Learning with Topological Data Analysis.” PhD diss., University of Tennessee, 2021

Shafie Gholizadeh, Wlodek Zadrozny, “A Tutorial on Topological Data Analysis in Text Mining”, 2020

Conti, F.; Moroni, D.; Pascali, M.A. “A Topological Machine Learning Pipeline for Classification”. Mathematics 2022, 10, 3086.

Choudhary; Aruni, “Approximation algorithms for Vietoris-Rips and Čech filtrations”, Dissertation zur Erlangung des Grades des Doktors der Ingenieurwissenschaften (Dr.-Ing.) der Fakult¨at f¨ur Mathematik und Informatik der Universit¨at des Saarlande, doi:10.22028/D291-26959