Topological Data Analysis

Naheed Anjum Arafat
Phd Student (3rd Year)
School of Computing

Supervisor: Stéphane Bressan

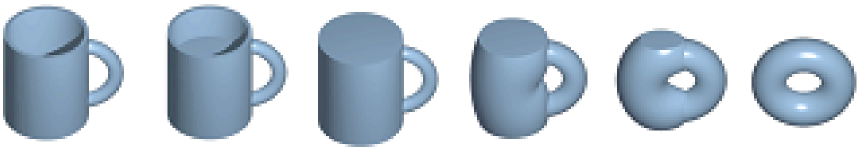
November 25, 2020

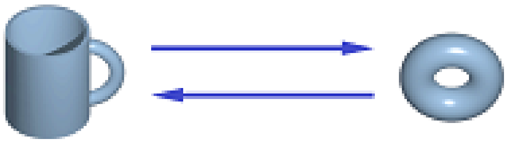
Outline

Topology: Coffee Mug-Donut Story

The Game of **Continuous Deformation**:-

- Allowed: Stretch, Squeeze, Bend, Twist
- Not allowed: Tearing
- Goal: Try to transform one into another.





Topology

Topology is the study of **objects** *which are* **invariant** *under certain kinds of transformation.*

Topology

Topology is the study of **objects** *which are* **invariant** *under certain kinds of transformation.*

- Objects:- Polyhedrons, Complexes, Manifolds etc.
- Invariants:- Euler Characteristics, Betti Numbers, Fundamental groups etc.

Topology

Topology is the study of **objects** *which are* **invariant** *under certain kinds of transformation.*

- Objects:- Polyhedrons, Complexes, Manifolds etc.
- Invariants:- Euler Characteristics, Betti Numbers, Fundamental groups etc.

How to study?

- The Objects as Groups (Algebra).
- The invariants as *Ranks*.

Topology

Topology is the study of **objects** *which are* **invariant** *under certain kinds of transformation.*

- Objects:- Polyhedrons, Complexes, Manifolds etc.
- Invariants:- Euler Characteristics, Betti Numbers, Fundamental groups etc.

How to study?

- The Objects as Groups (Algebra).
- The invariants as *Ranks*.

Why its important?

- To characterize, classify, compare and distinguish objects.

Topology before 20th century

Leonard Euler [1707-1783]

- Object: Polyhedrons
- Invariant: Euler Characteristics, $\chi = V - E + F$

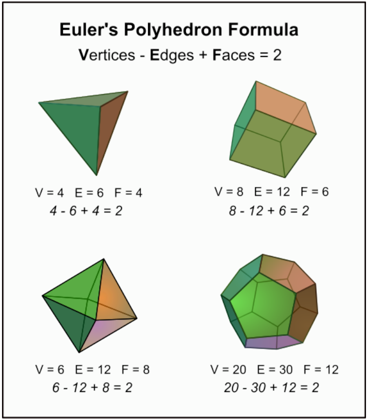
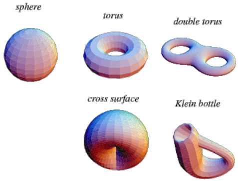


Figure 1: **Eulers polyhedron formula:** Euler characteristic of any convex polyhedron is 2

Topology in 20th century: Birth of Algebraic Topology

Henri Poincaré [1854 - 1912]

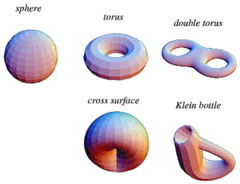
- Object: 3D Manifolds
- Invariant: Betti numbers, Fundamental groups



Henri Poincaré [1854 - 1912]

- Object: 3D Manifolds
- Invariant: Betti numbers, Fundamental groups

Application in Dynamical systems:
Poincaré-Hopf theorem For a vector field on a compact manifold with a finite number of fixed points, the sum of indices at the fixed points is equal to the Euler characteristic of the manifold.
Special case: Hairy ball theorem- the sum of indices = 2



Hairy ball theorem: You can't comb a hairy ball flat without creating a cowlick

Hairy ball theorem: You can't comb a hairy ball flat without creating a cowlick

Simplicial Complex, Manifold Triangulation

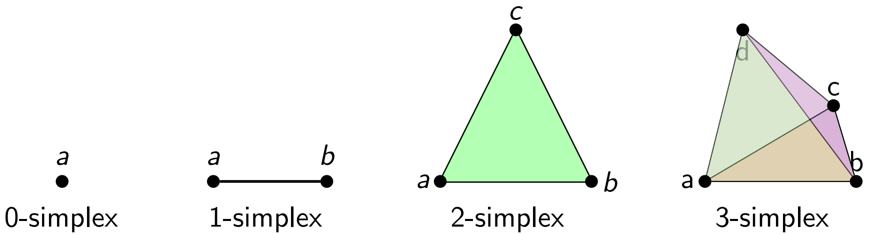


Figure 2: k-simplices

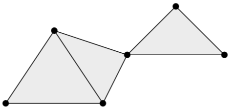


Figure 3: Simplicial Complex

Simplicial Complex, Manifold Triangulation

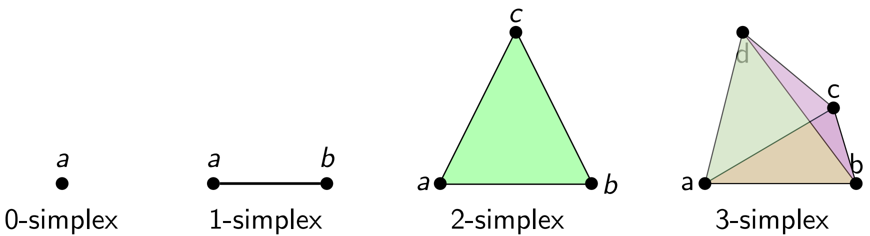


Figure 2: k-simplices

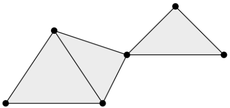


Figure 3: Simplicial Complex

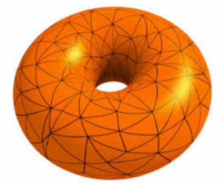


Figure 4: Triangulation of a torus

Simplicial Homology - Emmy Noether (1926)

Homology group: Let, $\mathcal{K}$ be the simplicial complex and $S_k(\mathcal{K})$ be the set of its k-simplices. The power set of S_k is an abelian group (termed as the k-th chain group C_k) under the operation of $+_2$(symmetric set difference). The chain groups are connected by boundary map $\partial_k : C_k \rightarrow C_{k-1}$.

k-**th Homology group**, $H_k \triangleq Ker(\partial_k)/Im(\partial_{k+1})$

k-**th Betti Number**, $\beta_k \triangleq rank(H_k)$

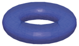
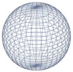
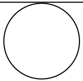
Simplicial Homology - Emmy Noether (1926)

Homology group: Let, $\mathcal{K}$ be the simplicial complex and $S_k(\mathcal{K})$ be the set of its k-simplices. The power set of S_k is an abelian group (termed as the k-th chain group C_k) under the operation of $+_2$(symmetric set difference). The chain groups are connected by boundary map $\partial_k : C_k \to C_{k-1}$.

k-**th Homology group**, $H_k \triangleq Ker(\partial_k)/Im(\partial_{k+1})$

k-**th Betti Number**, $\beta_k \triangleq rank(H_k)$

β_k **counts the number of K-dimensional holes.**

				**SOC**
β_0(Connected Components)	1	1	1	3
β_1(Cycles/Loops)	2	0	1	1
β_2(Voids)	1	1	0	0

Topological Data Analysis (TDA)

Preamble:

- Scientific data is Noisy, high dimensional and often in the form of Point Cloud from a metric space, e.g. output of 3D scanners, structure of biomolecules, sensor readings etc.
- We may not know the intrinsic dimension of the data.
- We do not have a unique simplicial representation of the data.

Topological Data Analysis (TDA)

Preamble:

- Scientific data is Noisy, high dimensional and often in the form of Point Cloud from a metric space, e.g. output of 3D scanners, structure of biomolecules, sensor readings etc.
- We may not know the intrinsic dimension of the data.
- We do not have a unique simplicial representation of the data.

What TDA does

- **Detects** and **Visualizes Global** features of the data.
    - Global features: Clusters/Connectivity, Cycles, Voids etc.

Topological Data Analysis (TDA)

Preamble:

- Scientific data is Noisy, high dimensional and often in the form of Point Cloud from a metric space, e.g. output of 3D scanners, structure of biomolecules, sensor readings etc.
- We may not know the intrinsic dimension of the data.
- We do not have a unique simplicial representation of the data.

What TDA does

- **Detects** and **Visualizes Global** features of the data.
    - Global features: Clusters/Connectivity, Cycles, Voids etc.

How?

- **Detection:** TDA Looks into the simplicial representation at **all scale**, in all dimension.
- **Visualization:** Provides a **topological summary** of the data, in all dimensions.

Persistent Homology (PH): mathematical tool for TDA

TDA workflow: PH constructs a nested sequence of Simplicial complexes (**Filtration)**. Then tracks the persistence (birth and death) of Homology groups across the filtration. Then computes Topological summaries (barcodes,persistent diagrams) from the birth,death pairs.

Point cloud $\Longrightarrow$ Nested complexes (Filtration) $\Longrightarrow$ barcodes.

Persistent Homology (PH): mathematical tool for TDA

TDA workflow: PH constructs a nested sequence of Simplicial complexes (**Filtration)**. Then tracks the persistence (birth and death) of Homology groups across the filtration. Then computes Topological summaries (barcodes,persistent diagrams) from the birth,death pairs.

Point cloud $\Longrightarrow$ Nested complexes (Filtration) $\Longrightarrow$ barcodes.

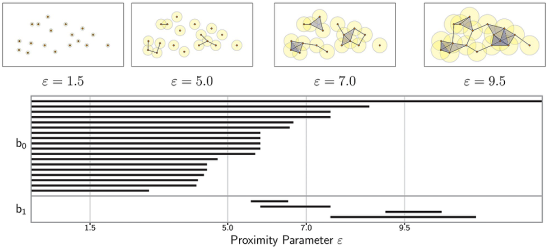


Figure 5: b_0 represents persistence in dimension 0, b_1 represents persistence in dimension 1.

Interpretation: Long bars are significant features, short bars are noise.

Another example

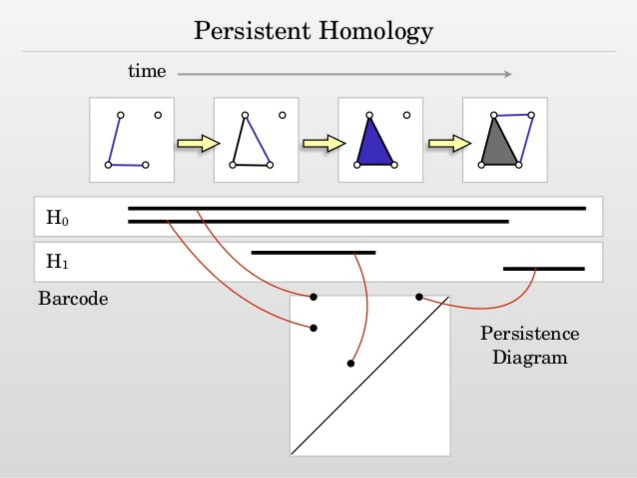


Figure 6: Barcodes in dimension 0,1 and Persistent Diagram.

Applications: Natural image

- A natural image (e.g. digital photograph) can be viewed as a vector in a very high-dimensional space P. The dimension d_p of P is the number of pixels in the camera resolution. A collection of such image constructs thus lies in some unknown manifold M embedded in P. **Can we find a dense sub-manifold of M?**

Applications: Natural image

- A natural image (e.g. digital photograph) can be viewed as a vector in a very high-dimensional space P. The dimension d_p of P is the number of pixels in the camera resolution. A collection of such image constructs thus lies in some unknown manifold M embedded in P. **Can we find a dense sub-manifold of M?**
- The space of 3x3 high-contrast patches from natural images has the topology of a Klein bottle [1], which is a 2-manifold.

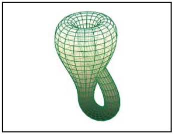


Figure 7: Klein bottle can not be embedded in $\mathbb{R}^3$, but can be embedded in $\mathbb{R}^4$ and higher.

- This implies, a natural image patch can be represented as a vector in a klein bottle embedded in dimension d_p. This can be exploited to design a compression algorithm.

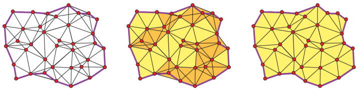
Applications: Coverage in Sensor network

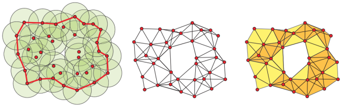
- **Problem Statement:** Consider a region $D \subset \mathbb{R}^2$, a set of sensors X and a subset $F \subset X$ of *fence* sensors on the boundary of D. Each sensor can identify any other sensor within radius r_b. Each sensor has a coverage of r_c. Fence nodes at all times know the identity of their neighbors.

  Can we determine coverage of sensors only from local information? [2]

Applications: Coverage in Sensor network

- **Problem Statement:** Consider a region $D \subset \mathbb{R}^2$, a set of sensors X and a subset $F \subset X$ of *fence* sensors on the boundary of D. Each sensor can identify any other sensor within radius r_b. Each sensor has a coverage of r_c. Fence nodes at all times know the identity of their neighbors.
  Can we determine coverage of sensors only from local information? [2]
- The **co-ordinates**, **orientations** and **distances** of neighboring nodes — all are unknown.
- Using relative homology of the simplicial complex built from the network of X, they gave a sufficient condition of when all the sensors covers the domain D.





Applications: Molecular biology

Study of Protein structure and folding [4] — First application of TDA in molecular biology.

- They proposed **Molecular topological fingerprint** (barcode constructed from PH of protein configuration) for protein characterization, identification and classification.
- They Proposed **slicing method** to analyze the topological fingerprints of alpha helices and beta sheets.
- They proposed **accumulated bar length** measure to quantitatively model protein flexibility.

Applications: Complex network

TDA of Brain functional network from fMRI [3]

- They studied 15 patients repeatedly under *Placebo* and *Psilocybin* drugs. They collected segmented fMRI images and computed the time series correlation from them. Finally, they constructed a weighted network from the regions (vertices) and computed correlation coefficient (edge weight) between regions.
- They proposed a new representations (**Homological scaffolds**) of the topological features of the brain functional network.
- Experiment shows clear difference between the effects of these drugs on the patients.

TDA research in our Lab

- **Complex Network:** We are trying to characterize different real world networks using persistent homology.
  - Challenges: Real world networks are large. Computing filtration and PH is memory inefficient. We are working on memory efficient solutions to the existing algorithms.
- **Climate Network:** We have analyzed historical sea surface temperature data of the eastern pacific peninsula with a goal to separate the spatial topology of el-nino years from normal years.
  - We built an interactive map visualization of homological cycles across different years.
  - We are trying to interpret the cycles (Climate scientists!)

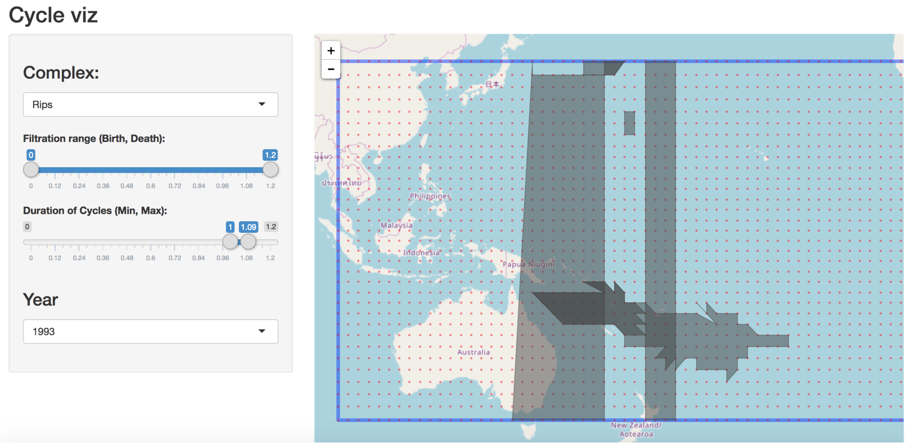


Figure 8: https://naheedcse.shinyapps.io/rips/

Summary

- Topology studies objects (**simplicial complexes, manifolds**) under continuous deformation.
- Topology **characterizes** objects by computing **invariants** (betti numbers)
- Homology **detects** global features of the data in one scale.
- Persistent homology **detects** and **represents** global features of the data in all scale.
- TDA has applications in dynamical systems,complex networks, biology, image analysis, 3D shape analysis, signal analysis and many more.

Thank You!

References

Carlsson, Gunnar et al. *On the local behavior of spaces of natural images.*
International journal of computer vision, 2008

De Silva, Vin and Ghrist, Robert. *Homological sensor networks*
Notices of the American mathematical society, 2007

Petri et al. *Homological scaffolds of brain functional networks*
Journal of The Royal Society Interface, 2014

Xia, Kelin and Wei, Guo-Wei. *Persistent homology analysis of protein structure, flexibility, and folding*,
International journal for numerical methods in biomedical engineering, 2014,