Topological Data Analysis

Group Meeting Report

Haixuan Lin

Department of Physics, Fudan University

July 2, 2026

Summary

Introduction

What and why TDA?

TDA (Topological Data Analysis) refers to a set of techniques that use methods from **algebraic topology** to analyze data structures.

In Materials Data Science, an important property of materials is their "shape" or "structure," which can be effectively described using topological algebra.

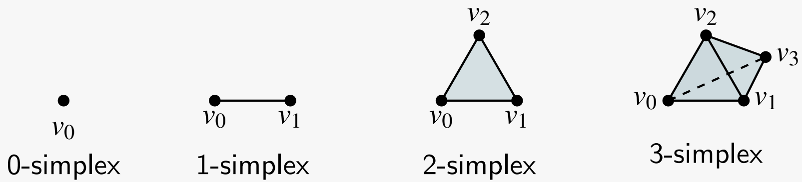
The Foundation of TDA

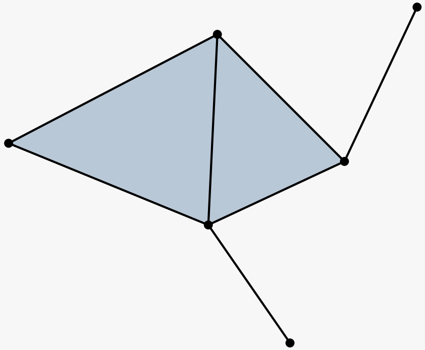
Summary

Definition (simplicial complex)

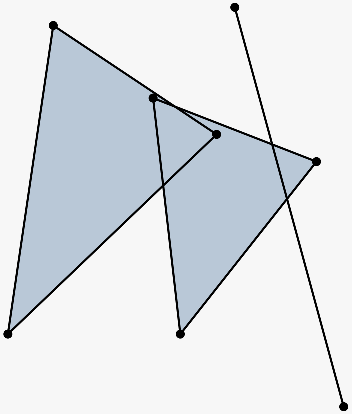
A 0-simplex is a point, denoted by $[v_0]$; a 1-simplex is an edge, denoted by $[v_0, v_1]$; a 2-simplex is a triangle, denoted by $[v_0, v_1, v_2]$; a 3-simplex is a tetrahedron, denoted by $[v_0, v_1, v_2, v_3]$; and so on for higher dimensions.

A **simplicial complex** is a collection of simplices that satisfy certain conditions, often described as "fitting together nicely".





fitting together nicely

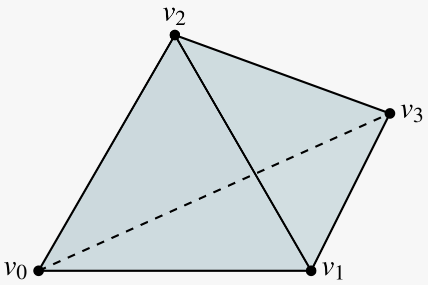


not fitting together nicely

We want to define an operator ∂ that enables us to obtain the "boundary" of a simplex or complex, for example:

$$\partial\left[v_1, v_2, v_3, v_4\right] = \left[v_1, v_2, v_3\right] - \left[v_1, v_2, v_4\right] + \left[v_1, v_3, v_4\right] - \left[v_2, v_3, v_4\right].$$

The minus signs ensure that every "face" is given a consistent "orientation".



3-simplex

Let K be a simplicial complex.

The k-**th chain group** of K, denoted $C_k(K)$, is the free abelian group whose generators correspond to the oriented k-simplices of K. That is,

$$C_k(K) = \left\{ \sum_i a_i \sigma_i \,\middle|\, a_i \in \mathbb{Z},\ \sigma_i \text{ is a } k\text{-simplex in } K \right\}.$$

The **boundary operator** $\partial_k : C_k(K) \to C_{k-1}(K)$ is a group homomorphism defined on a basis element $[v_0, v_1, \ldots, v_k]$ (an oriented k-simplex) by

$$\partial_k \left([v_0, v_1, \ldots, v_k] \right) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \widehat{v_i}, \ldots, v_k],$$

where $\widehat{v_i}$ indicates that v_i is omitted.

These boundary operators induce a sequence of abelian groups and homomorphisms, known as the **chain complex**:

$$\cdots \xrightarrow{\partial_{k+1}} C_k(K) \xrightarrow{\partial_k} C_{k-1}(K) \xrightarrow{\partial_{k-1}} \cdots \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} 0,$$

which is called the **simplicial chain complex**.

Theorem

For the simplicial chain complex of complex K, we have

$$\partial_k \circ \partial_{k+1} = 0$$

for all k.

Definition (homology)

The k-**th homology group** of K, denoted $H_k(K)$, is defined as the quotient group

$$H_k(K) = \ker \partial_k / \operatorname{im} \partial_{k+1},$$

where $\ker \partial_k$ is the group of k-cycles, and $\operatorname{im} \partial_{k+1}$ is the group of k-boundaries.

By the structure theorem of finitely generated abelian groups, we have:

Theorem

For k-th homology group $H_k(K)$, we have inner direct sum decomposition

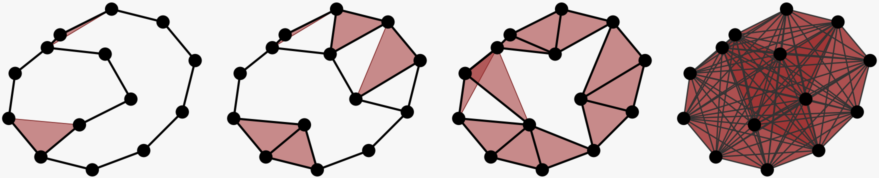
$$H_k(K) = F \oplus \mathrm{tor}(H_k(K)),$$

here $F \cong Z^{\beta_k}$, $\beta_k \in \mathbb{N}_+$ is called **Betti number**, *and $\mathrm{tor}(H_k(K))$ is called* **torsion subgroup**, *which is a finite group.*

Betti number is a topological invariant, which can be used to distinguish different topological spaces.

Example: The Betti number can be used to distinguish the shapes of Fermi surfaces in different materials, indicating its connection to physical properties such as conductivity.

Persistent Homology



The "holes" (simplices carrying information) birth and die as scale parameter increases.

Persistent homology is an extended homology theory, which can be used to better describe the topological structure of data when it comes to some scale-dependent topological features.

Definition (persistent homology)

A **filtration** of complex K is a sequence of nested simplicial complexes:

$$\varnothing \subseteq K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K.$$

Define the **persistent homology groups** of K as:

$$H_k^{ij} = Z_k(K_i) / \left(B_k(K_j) \cap Z_k(K_i)\right),$$

and its corresponding **persistent Betti numbers** as:

$$\beta_k^{ij} = \dim H_k^{ij}.$$

Summary

GLMY theory is a homology theory specifically developed for directed graphs, in contrast to ordinary homology theory, which applies only to undirected graphs. It was introduced by Grigoryan, Lin, Muranov, and Yau in 2012, and is named after its creators.

Example: If the directed edge 03 is not allowed in the given directed graph, then $\partial(013) = 13 - 03 + 01$ is not a valid boundary. Ordinary homology theory cannot handle this situation.

Definition (GLMY homology)

Let $G=(V,E)$ be a simple directed graph with vertices set $V=\{v_0,v_1,\cdots,v_n\}$ and directed edges set E. Denote

$$\mathscr{P}_n(G)=\{\langle v_{i_1}\cdots v_{i_n}\rangle, i_1\neq\cdots\neq i_n\}\cap E$$

as allowed n-paths and denoted $A_n(G)=\mathrm{span}_K\mathscr{P}_n(G)$ as the n-th **path space** of G on field K. Denote

$$\Omega_n(G)=A_n(G)\cap\partial_n^{-1}A_{n-1}(G)=\{x\in A_n(G):\partial_n x\in A_{n-1}(G)\}$$

as the maximal K-linear subspace of $A_n(G)$ such that ∂_n can be well-defined. Then the n-th GLMY homology group is defined as:

$$H_n(G,K)=\ker\partial_n/\operatorname{im}\partial_{n+1}.$$

We can also define hybrid homology by intersecting GLMY homology and persistent homology. This approach can be used to describe systems that involve both scale parameters and directed edges.

Summary

Euler characteristic is a topological invariant integer, which can be used to distinguish different topological spaces.

Definition (Euler characteristic)

The Euler characteristic of a simplicial complex K is defined as:

$$\chi(K) = \sum_{i=0}^{n} (-1)^i N_i.$$

where N_i is the number of i-simplices in K.

Example: When it comes to 3-dimensional Euclidean space, we have

$$\chi = V - E + F,$$

where V is the number of vertices, E is the number of edges, and F is the number of faces. We are familiar with this formula.

Theorem (Euler–Poincaré formula)

For a simplicial complex K, we have

$$\chi(K) = \sum_{i=0}^{n} (-1)^i \beta_i(K).$$

where $\beta_i(K)$ is the i-th Betti number of K.

Euler characteristic Curve (ECC)

The ECC (Euler Characteristic Curve) is an extension of the Euler characteristic that incorporates a filtration process. This allows it to describe simple directed graphs that are either edge-weighted or node-weighted.

Definition (filtration function)

A filtration function is a function $f : K \to \mathbb{R}$ that assigns a real number to each simplex in K. It induces a family of subspaces

$$X_t = f^{-1}\left((-\infty, t]\right) = \{x \in X : f(x) \leq t\}, \qquad t \in \mathbb{R}.$$

By $s \leq t \implies X_s \subseteq X_t$, we get a filtration of X:

$$\cdots \subseteq X_s \subseteq X_t \subseteq \cdots .$$

Under the filtration function f, a simplex cannot appear before any of its faces. Let χ be the Euler characteristic, define

$$\mathrm{ECC}_f : \mathbb{R} \to \mathbb{Z}, \qquad t \mapsto \chi(K_t).$$

This is the Euler characteristic curve.

Topological Learning and Applications in Materials Science

- Proposed the framework of **Topological Learning**, integrating TDA with machine learning for data-driven materials science.
- Introduced three representative paradigms of topological learning:
  - **Explicit Topological Representations**: use persistence diagrams/images, Betti numbers, and Euler characteristic curves as topological features for machine learning.
  - **Implicit Topological Regularization**: incorporate topological constraints into the training objective to preserve the intrinsic topology of data.
  - **Topology-Aware Architectures**: design neural network architectures that explicitly exploit topological structures.
- Established a general workflow:

$$\text{Material Structure} \rightarrow \text{Topological Descriptor} \rightarrow \text{Machine Learning} \rightarrow \text{Structure–Property Relationship.}$$

- Demonstrated that TDA provides robust, interpretable, and multiscale structural descriptors for complex material systems.

- Reviewed representative applications of TDA in materials science, including:
  - prediction of fundamental material properties;
  - polymer materials;
  - glassy materials (medium-range order);
  - heterogeneous catalysis;
  - porous crystalline materials (e.g., MOFs/COFs);
  - superionic conductors (ion transport networks).
- Highlighted the advantages of TDA over conventional descriptors:
  - robustness to noise and perturbations;
  - multiscale characterization of structural organization;
  - physically interpretable topological features;
  - seamless integration with modern machine learning methods.
- Discussed current challenges and future directions, including the incorporation of chemical information, more expressive filtration strategies, and topology-aware deep learning models.

References

[1] F. Lei, Z. Yang, and F. Li, *Basic Topology and Applications*. Beijing: Higher Education Press, 2025, ISBN: 9787040630428.

[2] S. Zheng et al., "Topological data analysis in materials science: Principles, machine learning integration, and application landscapes," *Chemical Reviews*, 2026, Advance online publication. DOI: 10.1021/acs.chemrev.5c01098.

[3] A. Hatcher, *Algebraic Topology*. Cambridge: Cambridge University Press, 2002, ISBN: 9780521795401.

The End