Lecture 1: From Data to Simplicial Complexes

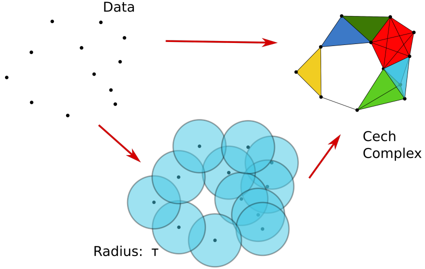
Claudia Landi

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Italian-Japanese summer school in Mathematics - Pisa, 2026

Motivation

- The idea of Topological Data Analysis (TDA) is to assign topological invariants to data
- Data however is typically presented as a discrete sample whose topology is rather uninteresting
- we must transform the data into a "continuous" object which (hopefully) is topologically similar to the shape of the underlying geometric object from which the data is sampled.



- The resulting continuous spaces need to be discretized in order to study them on a computer, and this discretization process is typically done by means of simplicial complexes.

Abstract simplicial complexes

An abstract simplicial complex is a generalization of a graph.

Definition

An **abstract simplicial complex** is a finite collection A of finite non-empty sets, such that if α is an element of A, then so is every nonempty subset of α.

- We call the elements of A **simplices** and the **dimension** of $\alpha \in A$ is $\dim \alpha = |\alpha| - 1$. The dimension of A is given by $\dim A = \max_{\alpha \in A} \dim \alpha$.
- Singletons are called **vertices**, doubletons are colled **edges**, triads are called **triangles** etc.
- $B \subseteq A$ is a **subcomplex** of A if B is an abstract simplicial complex.

Simplicial maps

Let A and B be abstract simplicial complexes with vertex sets V_A and V_B, respectively.

Definition

A **simplicial map** $f\colon A \to B$ is a function such that for every simplex $\{p_0, \ldots, p_n\}$ in A its image $\{f(p_0), \ldots, f(p_n)\}$ is a simplex in B.

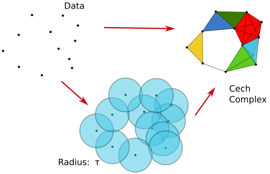
Note that f can decrease the dimension of simplices.

Definition

Two simplical complexes A and B are called **isomorphic** if there is a bijective correspondence $f\colon V_A \to V_B$ satisfying $\{a_0, \ldots, a_n\} \in A$ if and only if $\{f(a_0), \ldots, f(a_n)\} \in B$.

The Čech Simplicial Complex

For $p \in \mathbb{R}^d$, let $B_r(p) := \{q \in \mathbb{R}^d : \|q - p\| \leq r\}$ be the **closed ball** of radius r centered at p.



Definition Let $P = \{p_0, \ldots, p_n\} \subset \mathbb{R}^d$ be a set of points (a.k.a. point cloud).
The **Čech complex** of P at **scale** r is defined as

$$\check{C}_r(P) = \{\sigma \subseteq P : \cap_{p \in \sigma} B_r(p) \neq \emptyset\}$$

- $\check{C}_r(P)$ is an abstract simplicial complex: if $\sigma \in \check{C}_r(P)$ and $\tau \subseteq \sigma$, then $\emptyset \neq \cap_{p \in \sigma} B_r(p) \subseteq \cap_{p \in \tau} B_r(p)$, implying $\tau \in \check{C}_r(P)$
- $r \leq r'$ implies that $\check{C}_r(P)$ is a subcomplex of $\check{C}_r(P)$.
- For sufficiently large r, $\dim \check{C}_r(P) = n$

Nerve Theorem

$\check{C}_r(P)$ can be geometrically realized as a polyhedron $|\check{C}_r(P)| \subseteq \mathbb{R}^n$, with $n = |P| - 1$:

For $n \geq 0$, let Δ^n be the convex hull in $\mathbb{R}^n$ of 0, $e_1, \ldots, e_n$. If P has cardinality $n+1$, there is a bijection γ between the vertices of $\check{C}_r(P)$ and the set $\{0, e_1, \ldots, e_n\}$. We define

$$|\check{C}_r(P)| := \bigcup_{\sigma \in \check{C}_r(P)} \mathrm{conv}(\{\gamma(p) | p \in \sigma\}) \subseteq \Delta^n$$

Nerve Theorem (without proof)

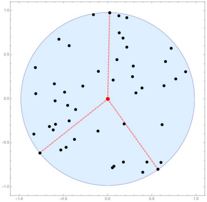
$|\check{C}_r(P)|$ is homotopy equivalent to $\cup_{p \in P} B_r(P)$.

Cost of deciding simplexes in the Čech complex

Let σ contain $k+1$ points of $P \subseteq \mathbb{R}^d$ $(k \geq d)$.

Does $\sigma \in \check{C}_r(P)$?

- $\sigma \in \check{C}_r(P)$ if and only if there exists a ball of radius $\leq r$ in $\mathbb{R}^d$ containing all the elements of σ.
- Such ball is unique and determined by at most $d+1$ points on its boundary (a $(d-1)$-dimensional sphere)



- Among the $k+1$ points of σ, there are $\binom{k+1}{d+1} \sim k^{d+1}$ subsets with $d+1$ points.

- Deciding whether there is a center c for a sphere of dimension $d-1$ passing through $d+1$ given points has a fixed cost because it corresponds to decide if a linear system of d equations in d unknowns (the coordinates of c) admits a solution: cost of computing the determinant of a $d \times d$ matrix.
- For each such sphere, we must check if it encloses the remaining $k-d$ points.
- The total cost in the worst case is of order $(k-d)\binom{k+1}{d+1} \sim k^{d+2}$.

This cost makes the construction of the Čech complex computationally untractable for large k.

The Vietoris-Rips complex

Instead of checking all subcollections, we may just check pairs and add 2- and higher-dimensional simplices whenever we can.

Definition

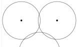
The **Vietoris-Rips complex** of P at **scale** r is the simplicial complex

$$VR_r(P) = \{\sigma \subseteq P | diam(\sigma) \leq 2r\} = \{\sigma \subseteq P | d(p,q) \leq 2r \; \forall p,q \in \sigma\}$$

with the **diameter** of a subset $\sigma \subseteq P$ given by

$$\mathrm{diam}(\sigma) = \max_{p,q \in P} \|p - q\|.$$

- The Vietoris-Rips complex is completely determined by its **1-skeleton** (i.e., the subcomplex of all simplices of dim ≤ 1)
- The cost of verifying if $\sigma \in VR_r(P)$: $O(k^2)$ ($k = \dim \sigma$).
- $VR_r(P)$ and $\check{C}_r(P)$ have the same 1-skeleton.
- $VR_r(P)$ does not enjoy the Nerve Theorem



VR- and Cech-complexes are interleaved

$$\check{C}_r(P) \subseteq VR_r(P) \subseteq \check{C}_{2r}(P)$$

The latter bound can however be significantly sharpened:

- **Theorem** (Jung's Thm., without proof)

  Let Q be a finite point set in $\mathbb{R}^d$. Then Q is contained in a closed ball with radius $r \leq \delta \operatorname{diam}(Q)$ where $\delta = \sqrt{\frac{d}{2(d+1)}}$.

- **Corollary**

$$\check{C}_r(P) \subseteq VR_r(P) \subseteq \check{C}_{2\delta r}(P)$$

where $2\delta = 2\sqrt{\frac{d}{2(d+1)}} = \sqrt{\frac{2d}{d+1}}$.

It follows that, for $d = 1$, $\check{C}_r(P) = VR_r(P)$.