A Brief Introduction to Persistent Homology

Nehal Chigurupati

February 2024

1 Homology Without Topology

Definition: Consider X $\subseteq \mathbb{R}^n$, linearly independent. The **simplex** σ spanned by vertices in X is the convex hull of the vectors in X. Explained another way, if we think of each vector in X as a point in Euclidean space, then σ is the smallest set that contains the points in X and all the straight lines between points in X. An m-simplex, then, is the m-dimensional equivalent of a triangle in $\mathbb{R}^2$.

Remark: If X $\subseteq \mathbb{R}^n$ has dimension m $\leq$ n, then X is referred to as an m-simplex.

Definition: A **face** of a simplex σ is another lower dimensional simplex spanned by a subset of the vertices of σ. For instance, if σ is a 2-simplex (i.e. a triangle), then the faces are the sides of the triangle.

Definition: A **simplicial complex K** is a collection of simplices such that:

- If K contains a simplex σ, then K also contains every face of σ
- If two simplices in K intersect, then their intersection is a face of both simplices.

Definition: For X $\subseteq \mathbb{R}^n$ and r $\in \mathbb{R}$, r $\geq$ 0, the **Cech simplicial complex** with vertex set X and resolution r consists of vertices X, subsets of which form a simplex {x_0, ..., x_k} when $\cap_{i=0}^{k} B(x_i, r) \neq \varnothing$. Such a simplicial complex is denoted **Cech**(X;r).

Informally, we can think of the Cech simplicial complex as representing a dataset X at some spatial resolution r. When the bubbles of radius r about each element of a set of points overlap at a common point, the points are grouped together into a simplex. As the spatial resolution increases, more points are absorbed into a few large simplices. If every point in X is contained in a ball of radius r, then the Cech complex would consist of a single simplex. On the other hand, as the spatial resolution decreases, there are a lot of lower-dimensional simplices. If r is sufficiently small such that no two points have overlapping balls (i.e. a

very precise resolution), then the only simplexes in the Cech complex are 0-dimensional (points).

Informally, in algebraic topology, homology is a tool to determine the shape of a space, primarily by determining "holes" in the space of various dimension. Traditionally, for some space X, homology would give you a set of 1-dimensional holes, 2-dimensional holes, 3-dimensional holes and so on.

Persistent homology attempts to identify these holes in Cech complexes on a dataset at different spatial resolutions. Computationally, persistent homology begins at an extremely fine spatial resolution, counting the number of lower-dimensional simplices. The spatial resolution is then gradually increased, with the computer tracking the resolutions at which lower dimensional simplices "die" by being subsumed by a larger simplex. If the points in the space are extremely far apart, then the resolution r would need to be extremely large for all vertices to be subsumed into a single simplex. Taking the average of these death times, then, should give a measure of the sparsity of the dataset. If the average is large, then the points in the dataset are spread far apart. On the other hand, if the average is small, then the points are close together.

An interesting, unique characteristic of persistent homology is that it allows you to determine the "dimensionality" of the open spaces between points, i.e. it is not just giving you a raw measure of how far apart the points are, but rather a measure of how far apart points are and in which dimensions they differ substantially.

2 Using persistent homology to determine range of talent in a roster

Consider an NBA team G with players P = $\{p_1, p_2, ..., p_m\}$, and play types T = $\{t_1, ..., t_k\}$.

For each player p_i, consider the function ϕ_i: T $\to \mathbb{R}$ that takes some play type, and returns a real number corresponding to their skill at the input play type. Ideally, for fixed play type t_k, if $\phi_i(t) \geq \phi_j(t)$, then p_i should be "better" at play t_k than player p_j.

For each player p_i, let v_i be the point $(\phi_i(t_1), ..., \phi_i(t_k)) \subseteq \mathbb{R}^k$, i.e. a point with the j-th component equal to player p_i's proficiency at play type t_j. For each team, we then have a dataset V = $\{v_i$: i $\in \{1, ...m\}\}$.

If these points are extremely far apart, then the team G has great diversity of talent in the play tuples t_i. To determine such sparsity of V, we can use persistent homology, taking the mean of the death resolutions for each lower-

dimensional simplex. A higher mean death time means that the team has players proficient in a wide variety of play types, while a lower mean death time might suggest that the players on the team all have similar skillsets.

In practice, we can create offensive/defensive play types on the following actions:

- Isolation
- Transition
- Pick and roll ball handling
- Pick and roll rolling
- Off-ball screen
- Cuts
- Putbacks
- Post-ups
- Spot-up shooting

Our proficiency rating can map offensive play types to the z-score of a player's points-per-possession, so ϕ_i(isolation offense) would map player p_i to their points-per-possession when isolating on offense relative to season average. On the other hand, the proficiency rating should map defensive play types to the z-score of

$$1/[\text{opponent PPP when performing a given play type}]$$

to ensure that higher z-scores correspond to better defensive performances. So, for instance, ϕ_i(isolation defense) would map player p_i to the z-score of $1/[\text{PPP of isolation plays with } p_i \text{ defending}]$.

For each team, we can run three different variations of persistent homology. To determine offensive range, we can let the points associated to each player strictly correspond to talent on offensive play types. To determine defensive range, we can let the points correspond to defensive skills. Finally, for net range, we can fuse both offensive and defensive play types into a single vector.

Performing persistent homology on the three datasets gives three different metrics: offensive range, defensive range, and net range. Offensive range corresponds to a roster's diversity of skill on offense in the above play types, while defensive range corresponds to proficiency on defensive varieties of the above play types. Finally, net rating combines the two, measuring the range of skill in a roster performing and defending against the above play types.