Wasserstein convergence of Čech persistence diagrams for samplings of submanifolds

Charles Arnal, David Cohen-Steiner, Vincent Divol

November 12, 2024

Topological data analysis studies the shape of data, i.e. its geometric or topological structure.

The Čech persistence diagrams $\mathrm{dgm}_i(\mathsf{S})$ of a set S capture some aspects of its topology.

Persistence diagrams can be vectorized and used as inputs to various machine learning methods for classification or regression.

Common task in TDA: recover information regarding the shape of some set S using a finite sampling $A \subset S$.

Idea: use the persistence diagrams $\mathrm{dgm}_i(A)$ as estimates of $\mathrm{dgm}_i(S)$.

Problem: $\mathrm{dgm}_i(A)$ is only known to converge weakly to $\mathrm{dgm}_i(S)$ as $A \to S$. In particular, machine learning-friendly features might not converge.

In our article, we show that if

- S is a generic submanifold, and
- $A_n \subset S$ is a random finite sampling of cardinality n,

then $\mathrm{dgm}_i(A_n) \xrightarrow{n\to\infty} \mathrm{dgm}_i(S)$ for the stronger Wasserstein metric.

This guarantees the stability of many machine learning features.

Further details for the specialists

More specifically, we show:

- A stronger version of the bottleneck theorem,
- Finiteness in expectation of the number of points in some regions of the diagram,
- Asymptotic expression for the total persistence of $\mathrm{dgm}_i(\mathsf{A}_n)$,
- Convergence of $\mathrm{dgm}_i(\mathsf{A}_n)$ to $\mathrm{dgm}_i(\mathsf{S})$ for the Wasserstein distances.