Wasserstein Stability for Persistence Landscapes

Wanchen Zhao   Advisor: Peter Bubenik

Filtration / Homology / Vectorization

Data → **Chain Complexes** → **Persistence Modules** → **Vector Summaries**

$X_0 \subset X_1 \subset X_2 \subset X_3 \subset X_4 \subset \ldots \in Ch^{\mathbb{N}}$

$V_0 \to V_1 \to V_2 \to V_3 \to V_4 \to \ldots \in Vec^{\mathbb{N}}$

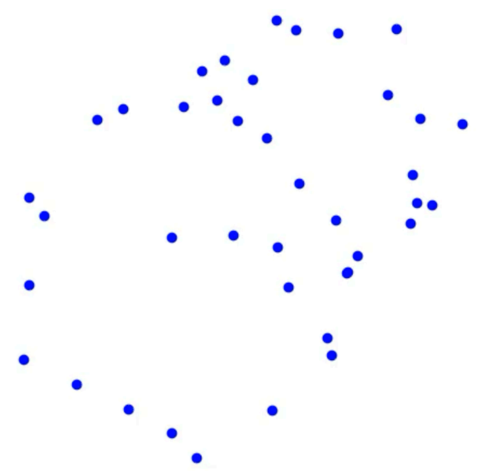


Fig. 1: Point cloud in metric space. More generally, data can come in many other forms and be preprocessed to obtain a filtration of topological spaces.

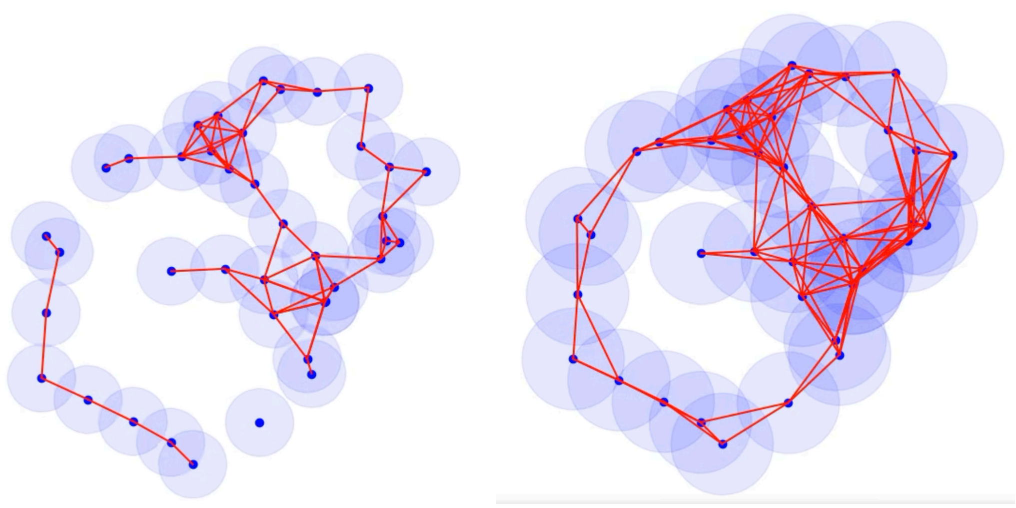


Fig. 2: Vietoris-Rips Filtration. There're also other ways to construct a simplicial complex on the set of vertices, such as the Alpha Complex.

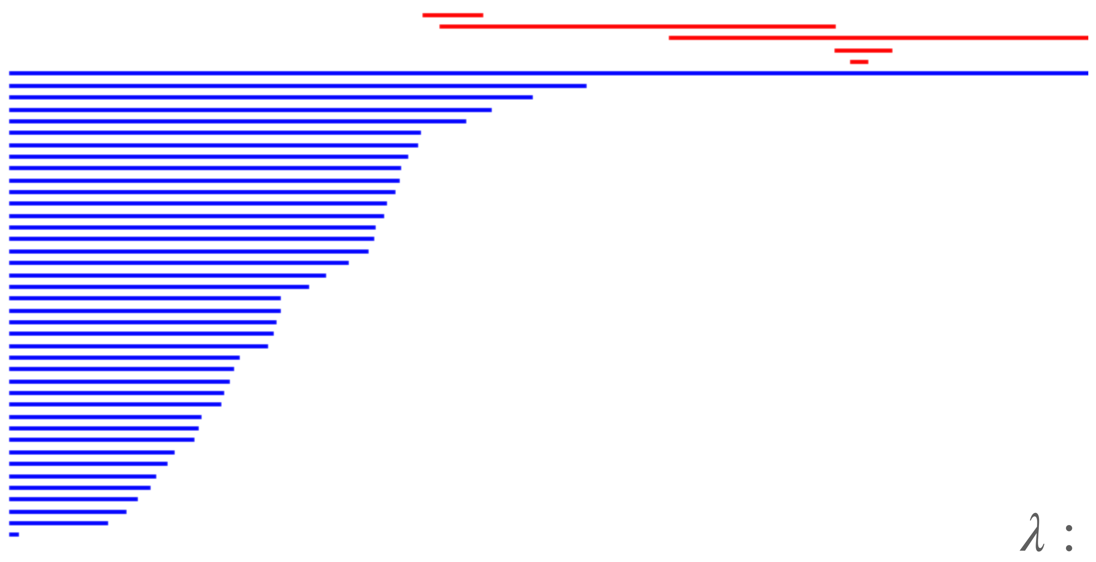
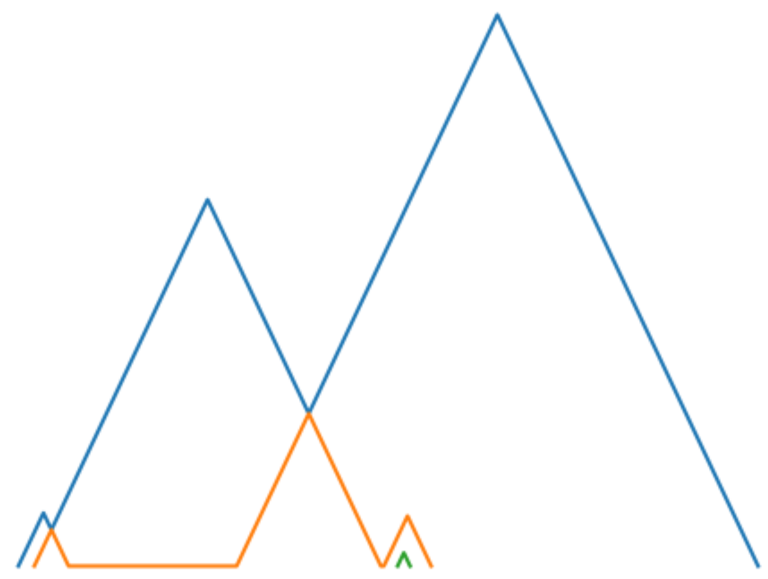


Fig. 3: Persistence Barcodes: $M \cong \bigoplus_j I_j$.



$\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$

$(k, t) \mapsto sup\{h > 0 \,|\, rankM(t-h \leq t+h) \geq k\}$

Fig. 4: Persistence Landscape

Topological Data Analysis

Topological data analysis(TDA) provides tools for analyzing topological features of various kinds of data. One of them is persistent homology. The persistence pipeline is described in the figure above. From data one can construct a filtration of chain complexes and apply the homology functor. The resulting object is called a persistence module, which can be represented as a persistence barcode under the following condition.

Structure Theorem in TDA: If $M \in Vec^{\mathbb{R}}$ is pointwise finite dimensional, then $M \cong \bigoplus_j I_j$, where I_j is an interval module.

Each bar in the barcode is an interval module that represents the lifespan of a homology class, which encodes the topological feature of the dataset.

Wasserstein Distance

The persistence barcodes form a metric space with p-Wasserstein metric for $1 \leq p \leq \infty$ defined as follows:

$$W_p^d(\alpha, \beta) := \inf_{\phi} \{ \sum_j d(x_j, \phi(x_j))^p \}^{1/p},$$

where d is a metric on bars and ϕ is a partial matching between bars.

Lemma(W. Zhao 2024): W_1^d satisfies the triangle inequality.

The metric d measures distance between interval modules in $\mathbb{R}$, so what's a good choice?

(Carlsson et al. 2005) The convention is to use $d(I, J) = \|(\inf I, \sup I) - (\inf J, \sup J)\|_1$. Observe equivalently $d(I, J) = \int |\dim I - \dim J|$. We denote this classical choice of metric by d_{dim} and we propose a new metric that uses the rank invariant instead.

Definition(W. Zhao 2024): $d_{rank}(I, J) := \int |rank\, I - rank\, J|$

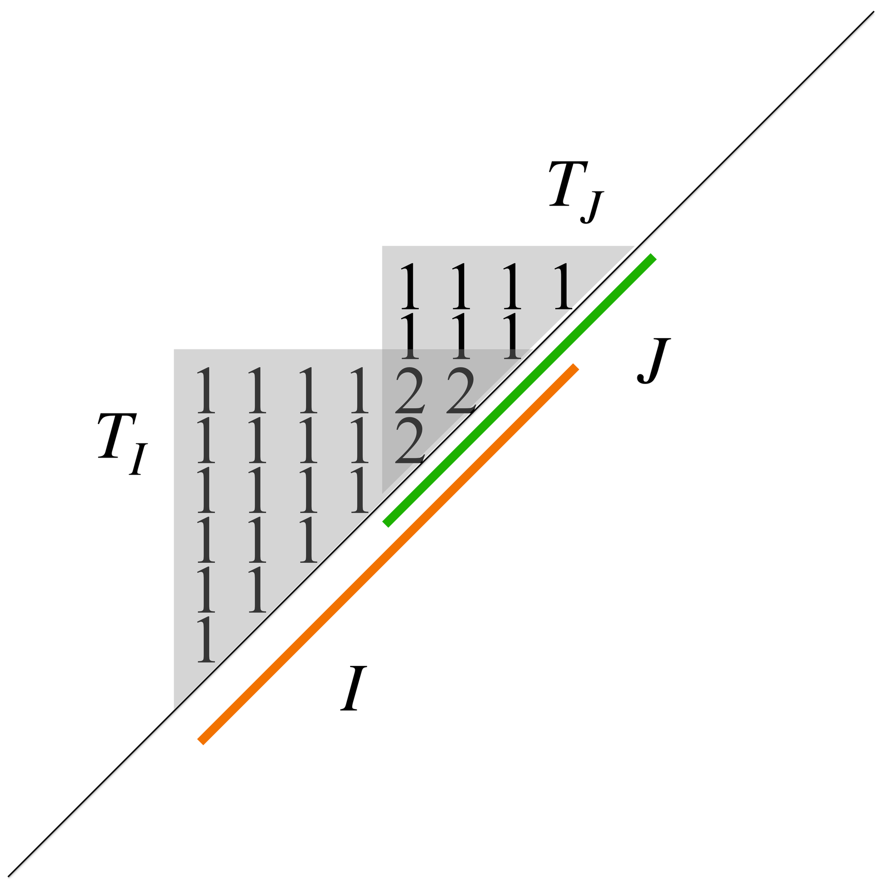


Fig. 5: $rank(I \oplus J) : Int(\mathbb{R}) \to \mathbb{Z}$. The shaded triangles T_I, T_J denote the support of rank(I), rank(J) for interval modules I, J.

Let W_1^{rank} denote the 1-Wasserstein distance using d_{rank} as the underlying metric.

Metric for Interval Modules

Interpretation of our choice of metric

Consider the following invariants in the Abelian category $Vec^{\mathbb{R}}$.

- $dim : \mathbb{R} \to \mathbb{Z}$
- $rank : Int(\mathbb{R}) \to \mathbb{Z}$

The rank invariants form a complete set of invariants. It keeps tracks of the number of homology classes that persist on an interval, which is the key idea of persistent homology. The *dim* invariants count the number of homology classes alive at a given filtration value and do not form a complete set of invariants. Therefore, We choose to use the rank invariants in d. Below is an equivalent definition of d_{rank}.

$d_{rank}(I, J) = m(T_I \Delta T_J)$, where $I, J \in Int(\mathbb{R})$, $T_I = \{K \in Int(\mathbb{R}) \,|\, K \subset I\}$, m is the Lebesgue measure on $\mathbb{R}^2$.

On the other hand, $d_{dim}(I, J) = \mu(I \Delta J)$ where μ is the Lebesgue measure on $\mathbb{R}$.

Vectorization and Stability Results

After obtaining a topological summary, one often needs to conduct further analysis on it, using machine learning and statistical tools. Thus we'd like to vectorize the persistence barcodes. Persistence landscapes and persistence images are often used as vector summaries for persistence barcodes.

Let Λ be the vectorization map that sends barcodes to their corresponding persistence landscapes. We have the following stability result using 1-Wasserstein distance with our choice of d.

Theorem(W. Zhao 2024):

For persistence barcodes α, β,

$$\|\Lambda\alpha - \Lambda\beta\|_1 \leq \frac{1}{2} W_1^{rank}(\alpha, \beta).$$

And this bound is sharp.

Future Work

- Stability from data to persistence barcodes using the new distance
- Generalize the new distance to multiparameter persistence modules $M : P \to Vec$ where P is some poset category.
- Metric for indecomposables of multiparameter persistence modules

Acknowledgments:

Special thanks to Professor Peter Bubenik for helpful discussions. This research project has been funded by NSF grants No. DMS-1764406, No. DMS-2324353, No. 594594, No. DMS-2113506

References:

Bubenik, P. Statistical Topological Data Analysis Using Persistence Landscapes. *Journal of Machine Learning Research 16* 77-102 (2015).

Carlsson, G; Zomorodian, A; Collins, A; Guibas, L. "Persistence barcodes for shapes". *International Journal of Shape Modeling.* **11** (2): 149–187. (2005)