Topological Data Analysis

and Persistence Theory

NSF/CBMS Conference

Valdosta State University

August 8-12, 2022

Peter Bubenik, University of Florida

Lecture 4 : TDA and Statistics

Outline:
1. TDA and Hilbert Space
2. The Persistence Landscape
3. Statistics with Persistence Landscapes
4.

Please interrupt me !!!

1. TDA and Hilbert Space

1.1 Why Hilbert Space?

Recall the TDA pipeline:

$$\text{Data} \longrightarrow \text{Geometry} \xrightarrow{\text{Homology}} \text{Algebra} \longrightarrow \text{Summary} \longrightarrow \text{Analysis}$$

Persistence Diagrams and Wasserstein distance

Statistics Machine Learning

Statistics and Machine Learning depend on Linear Algebra.

Want: a vector space and inner product (ie inner product space).

We want summaries that lie in a complete inner product space (ie. a Hilbert space).

1.2 Nonembeddability

Theorem For any $p \in [1, \infty]$, the metric space of persistence diagrams with the Wasserstein distance W_p does not embed into a Hilbert space.

Theorem For any $p \in (2, \infty]$, the metric space of persistence diagrams with the Wasserstein distance W_p does not coarsely embed into a Hilbert space.

$p \in [1, 2]$ open

1.3 Feature maps and kernels

A <u>feature map</u> is a map $\Phi : \underset{\text{set}}{X} \longrightarrow \underset{\text{Hilbert space}}{H}$.

Given a feature map Φ, we may define $k : X \times X \longrightarrow \mathbb{R}$

by $k(x, x') = \langle \Phi x, \Phi x' \rangle$, called a <u>kernel</u>.

inner product on H.

<u>Theorem</u> $k : X \times X \to \mathbb{R}$ is a kernel iff it is symmetric and positive definite.

$\hookrightarrow$ $\forall n \in \mathbb{N}$, $\forall x_1, \ldots x_n \in X$, the $n \times n$ matrix $(k(x_i, x_j))$ has non-negative eigenvalues.

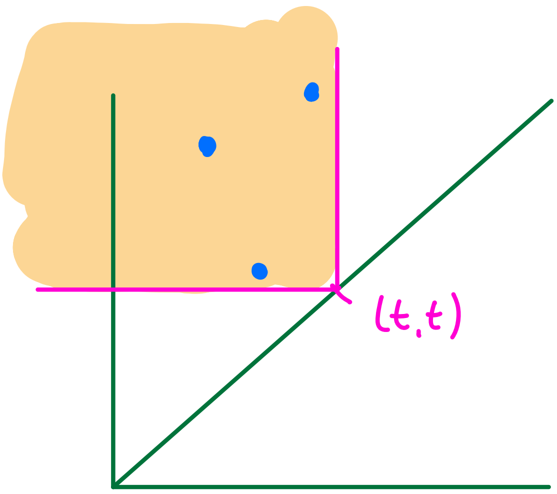
We want a Hilbert space H and a feature map

$$\Phi : \mathrm{Dgm} \longrightarrow H.$$

2. The Persistence Landscape

2.1 Erosion of Persistence Diagrams

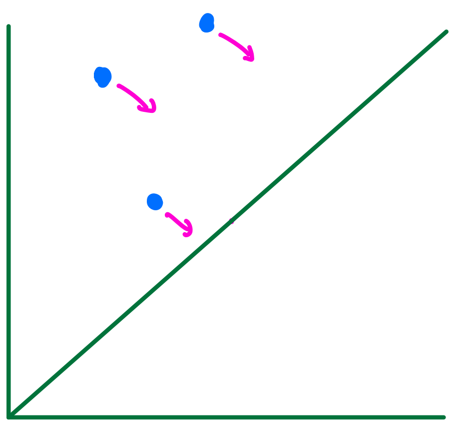
Let $M: \mathbb{R} \to \mathrm{Vect}$ be a persistence module with persistence diagram $\mathrm{Dgm}\, M = \{(b_i, d_i)\}_{i \in I}$



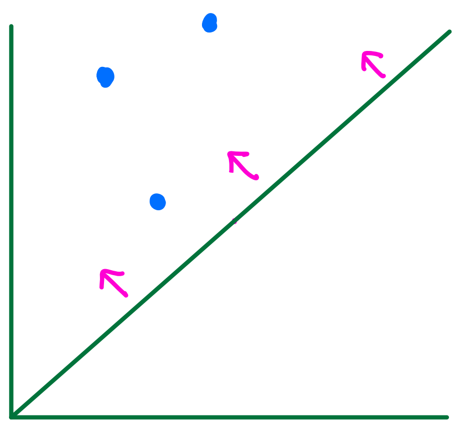
Lemma $\dim M_t = \#$ of points in $\mathrm{Dgm}\, M$ in the upper left quadrant Q_t.

Given $\varepsilon \geq 0$, let the *ε-erosion* of $\mathrm{Dgm}\, M$ be given by $(\mathrm{Dgm}\, M)_\varepsilon = \{(b_i+\varepsilon, d_i-\varepsilon)\}_{i \in I}$

{ remove if $b_i+\varepsilon > d_i-\varepsilon$

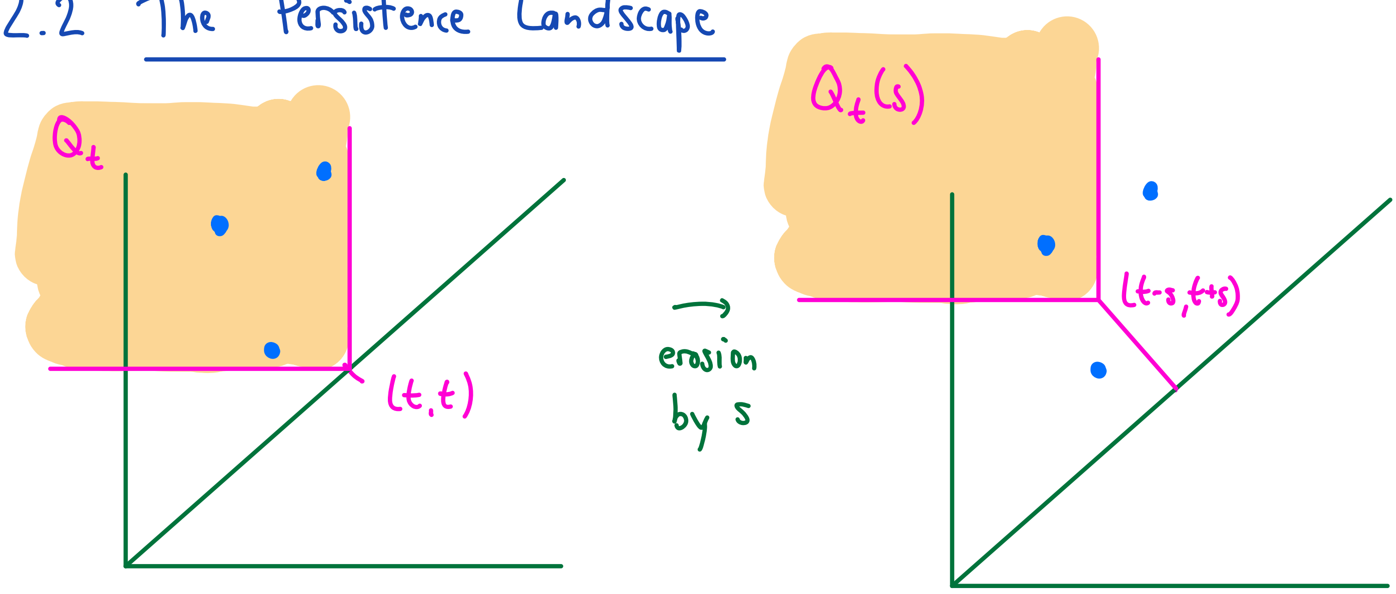


OR



ε confidence

2.2 The Persistence Landscape



Consider the # of points in the quadrant $Q_t(s)$.

For $k \in \mathbb{N}$,

Define $\lambda_k(t) = \max(s \mid \#\text{ points in } Q_t(s) \geq k)$

We obtain a sequence of functions $\lambda = (\lambda_1, \lambda_2, \lambda_3, \dots)$ called the Persistence Landscape.

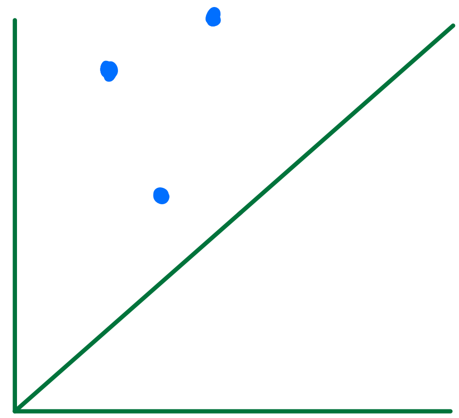
It has inner product $\langle \lambda, \rho \rangle = \sum_k \int \lambda_k(t)\, \rho_k(t)\, dt$

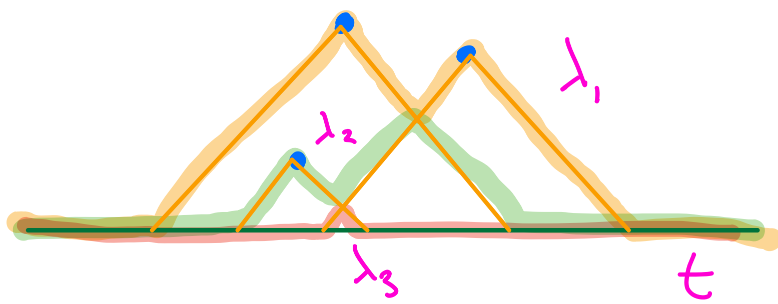
Equivalently, have $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ given by $\lambda(k,t) = \lambda_k(t)$.
$\lambda \in L^2(\mathbb{N} \times \mathbb{R})$.

We have a feature map $\Lambda : \mathrm{Dgm} \longrightarrow L^2(\mathbb{N} \times \mathbb{R})$

$$D \longmapsto \lambda$$

Graphing the persistence landscape:





Properties:

(a) PL. Each λ_k is piecewise linear with slope ± 1 on its support

(b) Lossless. Pers Diag $\to$ Pers Landscape is invertible

(c) Stable. Point cloud $\to$ Pers Landscape is nonexpansive

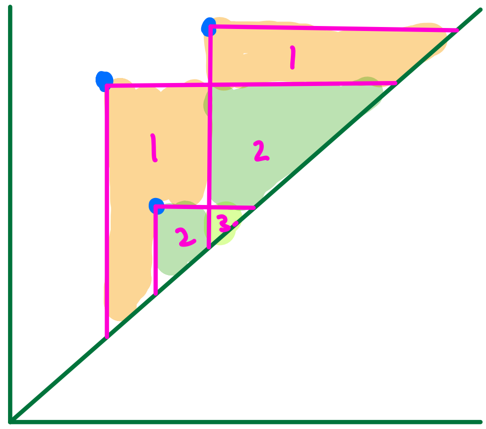
Hausdorff dist   supremum norm

2.3 Graded Persistence Diagram and Persistence Landscape

Rank function, Rank

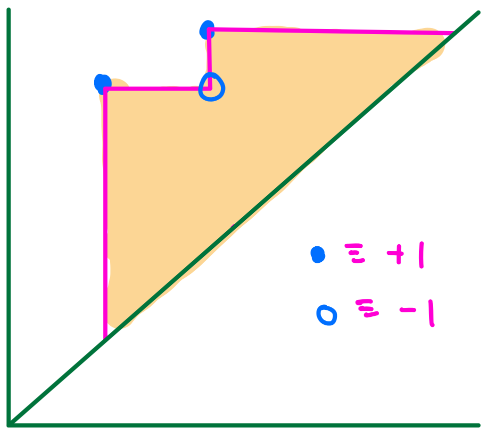
and

Persistence Diagram, PD

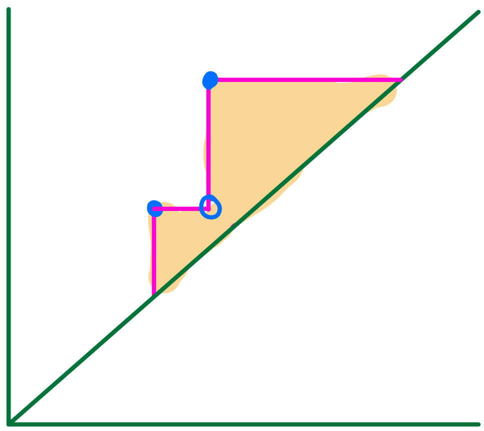


Graded Rank functions and Graded Persistence Diagrams:

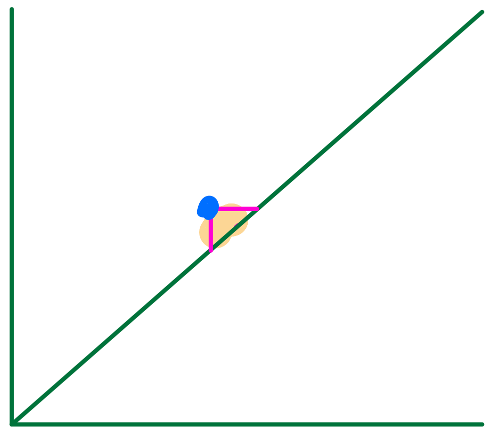
Rank_1 and PD_1

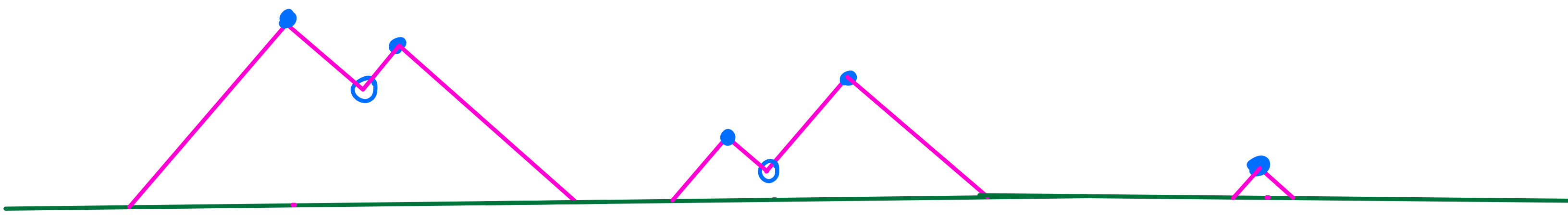


Rank_2 and PD_2



Rank_3 and PD_3





λ_1 λ_2 λ_3

Theorem

The positive points of PD_k are the local maxima of λ_k.

The negative points of PD_k are the local minima of λ_k.

3. Statistics with Persistence Landscapes

3.1 Average Persistence Landscape

Data $\longrightarrow$ Persistence Landscape

X $\qquad\qquad$ λ

Data $\longrightarrow$ Persistence Landscapes

$X_1, \dots, X_N$ $\qquad\qquad$ $\lambda^{(1)}, \dots \lambda^{(N)}$

Let $\bar{\lambda}(k,t) = \frac{1}{N} \lambda(k,t)$

3.2 Hypothesis Testing

If we have two experimental conditions

Data $\qquad\qquad$ Average Persistence Landscapes

$X_1, \dots X_N \longmapsto \bar{\lambda}$

$Y_1, \dots, Y_N \longmapsto \bar{\rho}$

Difference of Average PL : $\bar{\lambda} - \bar{\rho}$

Is this difference significant?

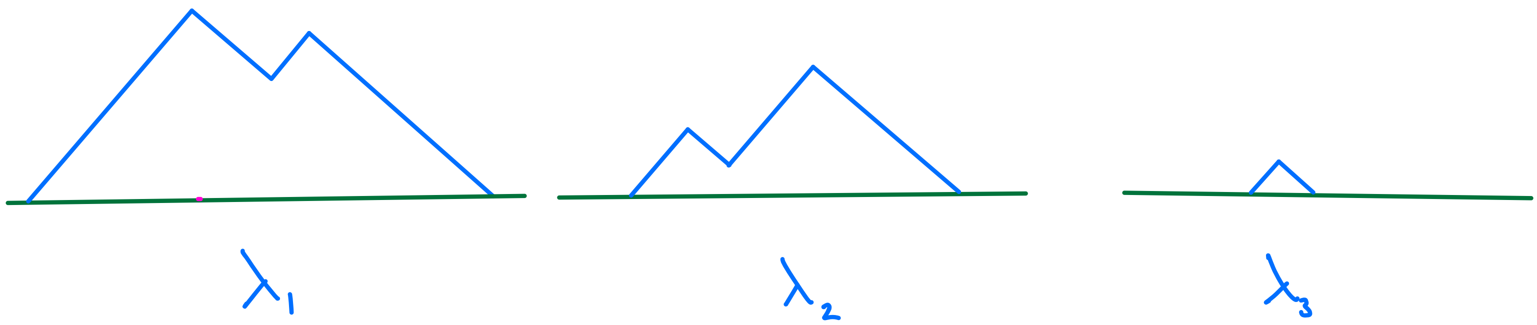
Test statistic: $\| \bar{\lambda} - \bar{\rho} \|$

We can obtain a p value for this statistic using a permutation test.

3.3 Discretizing the Persistence Landscape

$\lambda \in L^2(\mathbb{N} \times \mathbb{R})$ may be approximated by a point in $\mathbb{R}^D$ for some large D as follows:

1. Discretize the support.
2. Evaluate $\lambda_1, \lambda_2, \ldots, \lambda_K$ on this grid.
3. Concatenate the numbers.



3.4 Using Average Persistence Landscapes

If one is a situation where data is cheap and abundant then it is advised to repeat and use Average PL instead of PL.

If one is in a situation where the data is too large to compute persistent homology then subsample many times and compute the average persistence landscape.

<u>Theorem</u> In certain generic situations, not only can we use a persistence landscape to reconstruct a persistence diagram, we can use an average persistence landscape to reconstruct all of the persistence diagrams used to compute it.