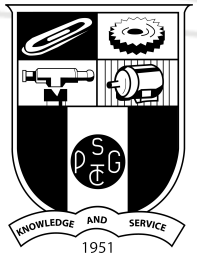



# Approximation of Persistence Homology For Large Data-sets

Karthikeya Subramanian

*Institution Guide: Dr. Sai Sundara Krishnan*

*Project Guide: Dr. Amit Chattopadhyay*

# Abstract

Topological data analysis (TDA) is a powerful approach to handle complex data structures that has emerged in recent years. Persistent homology is a successful TDA tool that generates summaries of a data-set's shape and size using the theory of homology, resulting in a persistence diagram that captures all of the topological information. However, persistent homology's computational expense makes it challenging to apply to large data-sets, which are now readily obtainable with modern data acquisition techniques. This project focuses on a proposed method that uses bootstrapping to compute persistent homology while taking the overall mean and their practical performance will be explored. This approach aims to explore new avenues for computing persistent homology in a computationally efficient manner.

# Basics of TDA

**Simplex:** A simplex is a geometric object defined as the convex hull of a set of affinely independent points in Euclidean space.
In formal notation, an n-simplex is defined as:

$$\Delta^n = \{\sum_{i=1}^{n} \lambda_i v_i \mid \sum_{i=1}^{n} \lambda_i = 1\,, \lambda_i \geq 0\}$$

where $v_0, \ldots, v_n$ are affinely independent points in Euclidean space. The points $v_0, \ldots, v_n$ are called the vertices of the simplex $\Delta^n$, and the coefficients $\lambda_0, \ldots, \lambda_n$ are called the barycentric coordinates of a point in the simplex. The dimension of the simplex is $n$, and it has $n + 1$ vertices. For example, a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron.

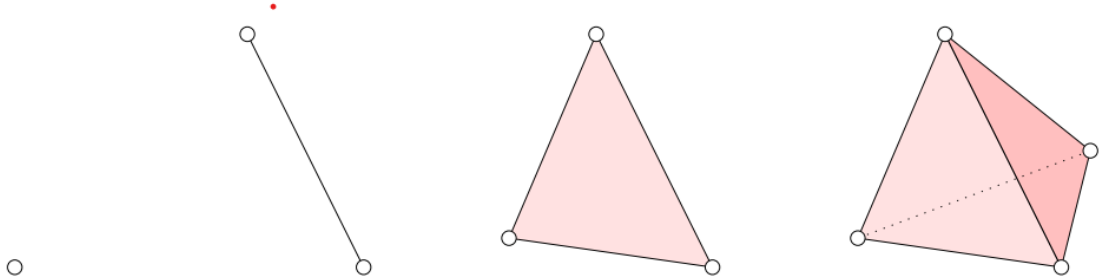


Figure: Simplex

**Simplicial Complex:** A simplicial complex is a finite collection of simplices $K$ such that $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$, and $\sigma, \sigma_0 \in K$ implies $\sigma \cap \sigma_0$ is either empty or a face of both.

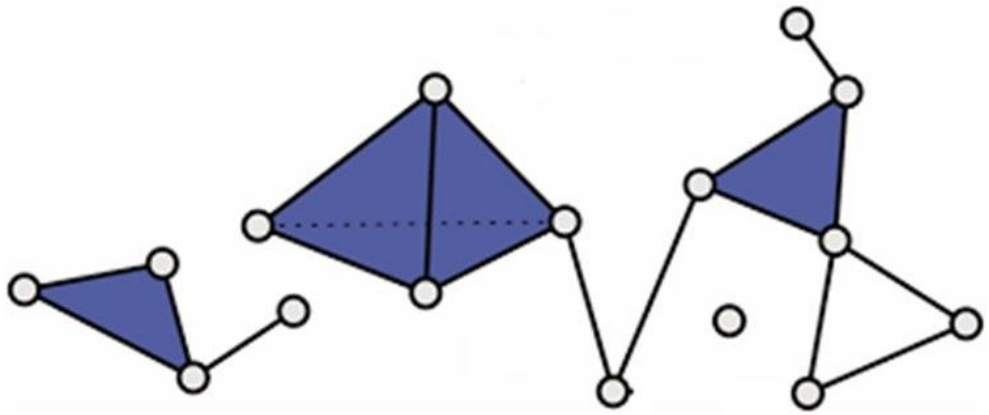


Figure: Simplicial Complex

# Homology

Homology groups provide a mathematical language for the holes in a topological space. Perhaps surprisingly, they capture holes indirectly, by focusing on what surrounds them. Their main ingredients are group operations and maps that relate topologically meaningful subsets of a space with each other. This section focuses on the various groups involved in the setup.

**Chain Complex:**

Let $K$ be a simplicial complex and $p$ a dimension. A $p$-chain is a formal sum of $p$-simplices in $K$, denoted by $c = \sum_i a_i \sigma_i$, where the $\sigma_i$ are the $p$-simplices and the $a_i$ are coefficients that can be integers, rational numbers, real numbers, elements of a field, or elements of a ring. In computational topology, we mostly work with coefficients that are either 0 or 1 (modulo 2 coefficients). This means that we can think of a chain as a set of $p$-simplices, specifically those $\sigma_i$ with $a_i = 1$. However, this way of thinking is not always convenient when working with other coefficient groups.

Two $p$-chains are added component-wise, like polynomials. Specifically, if $c = \sum_i a_i \sigma_i$ and $c' = \sum_i b_i \sigma_i$ then $c + c' = \sum_i (a_i + b_i) \sigma_i$, where the coefficients satisfy $1 + 1 = 0$. In set notation, the sum of two $p$-chains is their symmetric difference. The $p$-chains together with the addition operation form the group of $p$-chains denoted as $(C_p, +)$, or simply $C_p = C_p(K)$ if the operation is understood. Associativity follows from associativity of addition modulo 2. The neutral element is $0 = \sum_i 0\sigma_i$. The inverse of $c$ is $-c = c$ since $c + c = 0$. Finally, $C_p$ is abelian because addition modulo 2 is abelian.

**For a $p$-chain**, $c = \sum a_i \sigma_i$, the boundary is the sum of the boundaries of its simplices, $\partial_p c = \sum a_i \partial_p \sigma_i$. Hence, taking the boundary maps a $p$-chain to a $(p-1)$-chain, and we write $\partial_p : C_p \longrightarrow C_{p-1}$. Notice also that taking the boundary commutes with addition, that is, $\partial_p(c + c') = \partial_p c + \partial_p c'$. This is the defining property of a homomorphism, a map between groups that commutes with the group operation. We will therefore refer to $\partial_p$ as the boundary homomorphism or, shorter, the boundary map for chains. The chain complex is the sequence of chain groups connected by boundary homomorphisms,

$$\cdots \xrightarrow{\partial_{p+2}} C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p-1}} \cdots$$

It will often be convenient to drop the index from the boundary homomorphism since it is implied by the dimension of the chain it applies to.

**Cycles and boundaries:**

We distinguish two particular types of chains and use them to define homology groups. A $p$-cycle is a $p$-chain with empty boundary, $\partial c = 0$. Since $\partial$ commutes with addition, we have a group of $p$-cycles, denoted as $Z_p = Z_p(K)$, which is a subgroup of the group of $p$-chains. In other words, the group of $p$-cycles is the kernel of the $p$-th boundary homomorphism, $Z_p = \ker \partial_p$. Since the chain groups are abelian so are their cycle subgroups. Consider $p = 0$ as an example. The boundary of every vertex is zero ($C_{-1} = 0$), hence, $Z_0 = \ker \partial_0 = C_0$. For $p > 0$, however, $Z_p$ is usually not all of $C_p$.

## Fundamental Lemma of Homology

$\partial_p \partial_{p+1} d = 0$ for every integer $p$ and every $(p+1)$-chain $d$.

**Proof.** We just need to show that $\partial_p \partial_{p+1} \tau = 0$ for a $(p+1)$-simplex $\tau$. The boundary, $\partial_{p+1}\tau$, consists of all $p$-faces of $\tau$. Every $(p-1)$-face of $\tau$ belongs to exactly two $p$-faces, so $\partial_p(\partial_{p+1}\tau) = 0$.
It follows that every $p$-boundary is also a $p$-cycle or, equivalently, that $B_p$ is a subgroup of $Z_p$. Figure 3 illustrates the subgroup relations among the three types of groups and their connection across dimensions established by the boundary homomorphisms.

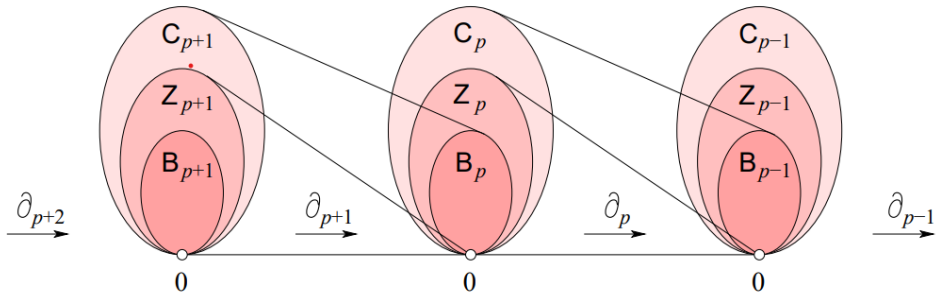



Figure: Chain Complex Homomorphisms

# Filtration

A filtration is a sequence of topological spaces that tracks the evolution of a space over a parameter like distance, scale, or density. It begins from an initial space, progresses to a final space, and intermediate spaces represent intermediate stages of evolution. Filtrations can be made through nested subsets of a space or geometric shapes around data points. These capture the topological features at various resolutions and how they change as parameters change. Filtrations are usually used with persistent homology to identify and quantify topological features that remain persistent across multiple scales or parameters.

# Cech complexes

Consider a case in which the convex sets are closed geometric balls, all of the same radius $r$. Let $S$ be a finite set of points in $\mathbb{R}^d$ and write $B_x(r) = x + rB_d$ for the closed ball with center $x$ and radius $r$. The Cech complex $C(S, r)$ of $S$ and $r$ is isomorphic to the nerve of this collection of balls.

$$Cech(r) = \{\sigma \subseteq S \mid \bigcap_{x \in \sigma} B_x(r) \neq \emptyset\}$$

Clearly, a set of balls has a non-empty intersection iff their centers lie inside a common ball of the same radius. Indeed, a point $y$ belongs to all balls iff $\|x - y\| \leq r$ for all centers $x$. An easy consequence of Helly's Theorem is therefore that every $d + 1$ points in $S$ are contained in a common ball of radius $r$ iff all points in $S$ are. This is Jung's Theorem which predates the more general theorem by Helly.

# Persistent Homology

Persistent homology adapts classical homology form algebraic topology to finite metrics spaces. The construction of persistent homology starts with a *filteration* which is a nested sequence of topological spaces: $X_0 \subseteq X_1 ... \subseteq X_n = X$. In particular, We focus on $Vietoris - Rips$ (VR) filtration for finite metric spaces $(\mathcal{X}, d_{\mathcal{X}})$. Let $\epsilon_1 \leq \epsilon_2 \leq ... \leq \epsilon_n$ be an increasing sequence of parameters. The $Vietoris - Rips$ Complex $\text{VR}(\mathcal{X}, \epsilon_i)$ at scale $\epsilon_i$ is constructed by adding a node for each $x_j \in \mathcal{X}$ and a $k-$simplex for each set $x_{j_i}$ with diameter less than $\epsilon_i$. Thus,

$$VR(\mathcal{X}, \epsilon_1) \longrightarrow VR(\mathcal{X}, \epsilon_2) \longrightarrow ... \longrightarrow VR(\mathcal{X}, \epsilon_n)$$

The sequence of inclusions induces maps in homology for any fixed dimension $r$. Let $H_r(\mathcal{X}, \epsilon_i)$ be the homology group of $\text{VR}(\mathcal{X}, \epsilon_i)$ with coefficients in a field .

Then we have the following sequence of vector spaces:

$$H_r(\mathcal{X}, \epsilon_1) \longrightarrow H_r(\mathcal{X}, \epsilon_2) \longrightarrow \ldots \longrightarrow H_r(\mathcal{X}, \epsilon_n)$$

The collection of $H_r(\mathcal{X}, \epsilon_i)$, together with vector space homomorphisms $H_r(\mathcal{X}, \epsilon_i) \longrightarrow H_r(\mathcal{X}, \epsilon_j)$ is called a persistence module.
When each $H_r(\mathcal{X}, \epsilon_i)$ is finite dimensional, the persistence module can be decomposed into rank one summands which corresponds to birth and death times of homology classes. Let $\alpha \in H_r(\mathcal{X}, \epsilon_i)$ be a non-trivial homology class $\alpha$ is born at $\epsilon_i$ if it is not in the image of $H_r(\mathcal{X}, \epsilon_{i-1}) \longrightarrow H_r(\mathcal{X}, \epsilon_i)$ it is dead entering $\epsilon_j$ if the image of $\alpha$ via $H_r(\mathcal{X}, \epsilon_i) \longrightarrow H_r(\mathcal{X}, \epsilon_{j-1})$ is not in the image $H_r(\mathcal{X}, \epsilon_{i-1}) \longrightarrow H_r(\mathcal{X}, \epsilon_{j-1})$, but the image of $\alpha$ via $H_r(\mathcal{X}, \epsilon_i) \longrightarrow H_r(\mathcal{X}, \epsilon_j)$ is in the image of $H_r(\mathcal{X}, \epsilon_{i-1}) \longrightarrow H_r(\mathcal{X}, \epsilon_j)$.The collection of birth-death intervals $[\epsilon_i, \epsilon_j)$ is called a *barcode* and it represents the persistent homology of VR filteration of $\mathcal{X}$.

## Persistence Diagram

Equivalently, we can regard each interval as an ordered pair of birth–death as coordinates and plot each in a plane $\mathbb{R}^2$ , which provides an alternate representation of a *barcode* as a *persistence diagram*.

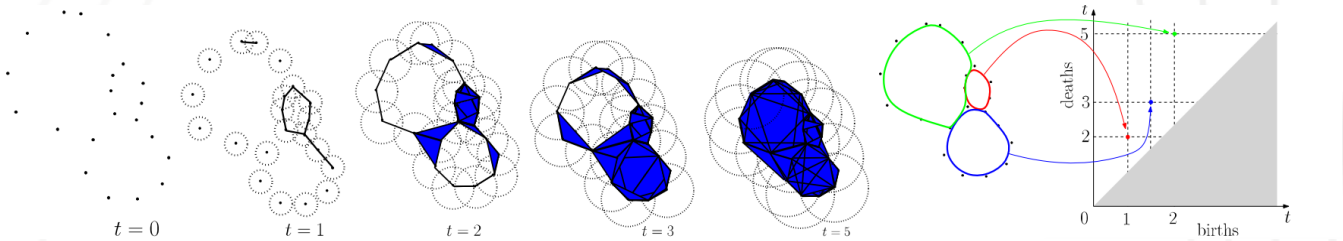



Figure: Filtration

**Definition** A *persistence diagram D* is a locally finite multi-set of points in the half-plane $\Omega = \{(x, y) \in \mathbb{R}^2 \mid x \leq y\}$ together with points on the diagonal $\partial\Omega = \{(x, x) \in \mathbb{R}^2\}$ counted with infinite multiplicity. Points in $\Omega$ are called $off - diagonal$ points. The persistence diagram with no off-diagonal points is called *empty persistence diagram* by $D_\emptyset$.

# Metrics on Persistence Diagrams

## Wasserstein distance

The $p$-th Wasserstein distance between two persistence diagrams, $D_1$ and $D_2$, is defined as

$$W_{p,q}(D_1, D_2) = \inf_{\gamma} \left( \sum_{x \in D_1} ||x - \gamma(x)||_q^p \right)^{\frac{1}{p}}$$

where $\gamma$ ranges over all bijections from $D_1$ to $D_2$. The set of bijections is nonempty because of the diagonal.

**Persistence Measure:**

An alternative approach is to define persistent diagrams as measures on $\Omega$ as a form of Dirac's measure. Let $\mu$ be Radon measure supported on $\Omega$ the $p-$ *total persistence* is defined as

$$pers_p(\mu) = \int_{\Omega} ||x - x^T|| d\mu(x)$$

Any Radon measure with finite $p-$ *total persistence* is called a persistence measure. Space of all persistence measure is denoted by $\mathcal{M}_p$. The optimal transport distance between measures $\mu$ and $\nu$ is

$$OT_{p,q}(\mu, \nu) = \inf_{\pi} \left( \int ||x - y||_q^p d\Pi(x, y) \right)^{\frac{1}{p}}$$

$(\mathcal{M}_p, OT_p)$ forms a polish space which is more advantageous as in measures are linear objects, this can be used for statistical purposes.

# Two Measures of Centrality

**Frechet Mean of Persistent diagrams** is the diagram which minimizes the Frechet function

$$Fr_\rho(D) = \frac{1}{B}\sum_{i=1}^{B} W_p^p(D_i, D)$$

The previous works prove that this function is not convex and there is no guarantee to find the global minimizer, there is a greedy algorithm to find the local minimizer. which makes the mean not unique

**Mean Persistence Measure** Let $D = \{D_1, D_2, ..., D_B\}$ be persistent diagram then the empirical mean of is simply $\mathbb{D} = \sum_{i=1}^{B} D_i$, So for any Borel set A in the first quadrant such that $A \subset \Omega$. The expectation is the average number of points in the persistent diagrams within the set

# Experimental Results

Experiments ran using mean persistence measures and Fréchet means of persistence diagrams to estimate the persistent homology of large data sets using sub-sampling. In particular, let $\mathcal{X}$ be a large point cloud with a predefined probability distribution satisfying some standard assumptions.We take $B$ number of $i.i.d's$ consisting of $n$ points from $\mathcal{X}$ are chosen. Then compute the mean persistence measure and Fréchet mean which can be regarded as two types of averages of persistence diagrams of sub-sample sets.

**Step 1 : A large point cloud** We initially start with 5000 point, point cloud of an annulus. We denote it by $\mathcal{X}$

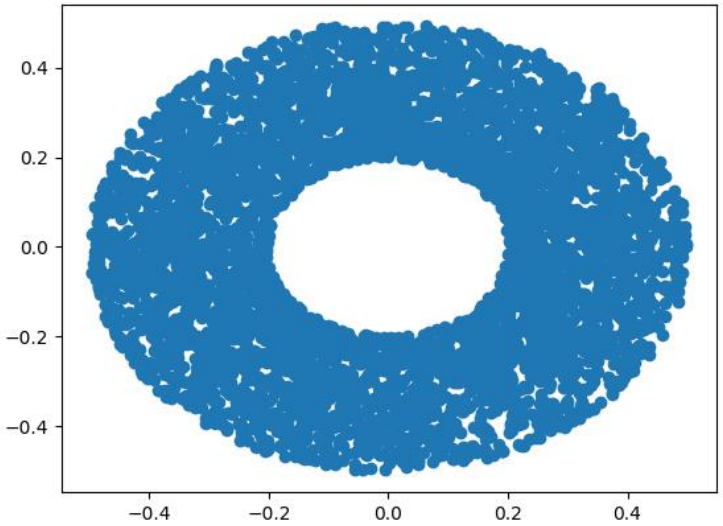
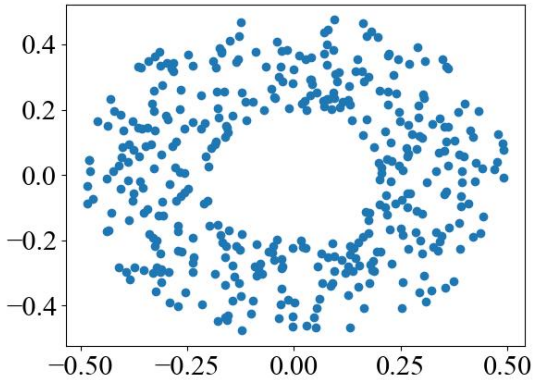
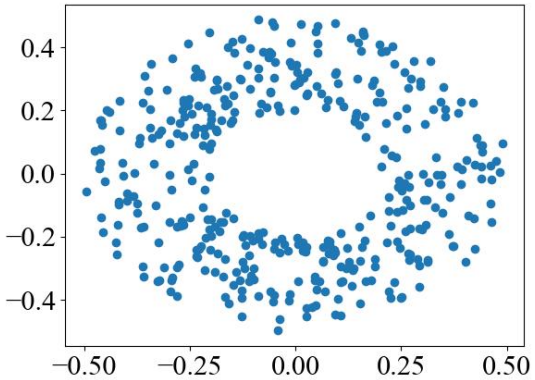



Figure: A point cloud of Annulus

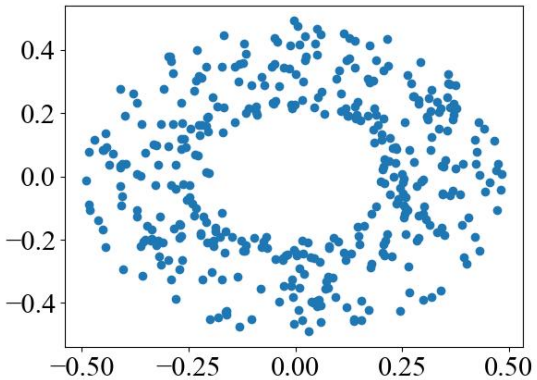
**Step 2 : Sub-Sampling** We further take $B - i.i.d$ (independent identically distributed)$\{X_1, X_2, ..., X_B\}$ from $\mathcal{X}$, each of size $n$. In this example, let $n = 400$ *and* $B = 3$




(a) Sub-Sample 1

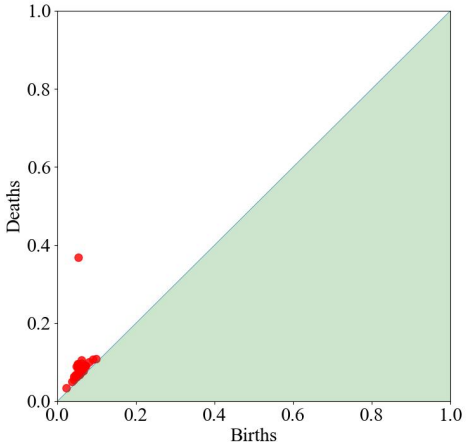



(b) Sub-Sample 2

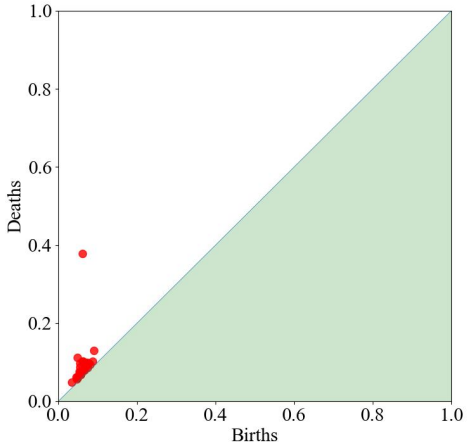



(c) Sub-Sample 3

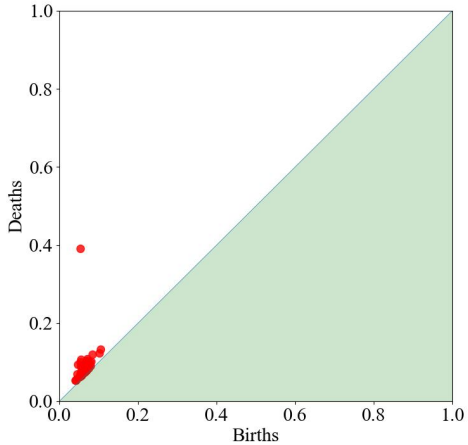
**Step 3 : Persistence Diagrams** The Sub-samples are put through Vietoris-rips filtration. Persistent diagrams for each $X_i$ are generated. $D = \{D_1, D_2, ..., D_B\}$, each $D_i$ consist of $\Omega$ representing the off-diagonal points and $\partial\Omega$ representing the diagonal.




(a) PD for Sub-Sample 1




(b) PD for Sub-Sample 2




(c) PD for Sub-Sample 3

**Step 4 : Fréchet Mean** The Persistent diagram which minimizes the Fréchet function,

$$Fr_\rho(D_j) = \frac{1}{B}\sum_{i=1}^{B} W_p^p(D, D_i)$$

The Fréchet mean $D_f$, is not unique, as in the Fréchet function is concave in space of Persistent diagrams.

$$D_f = \arg\min_{D_j \in D} Fr_\rho(D_j)$$

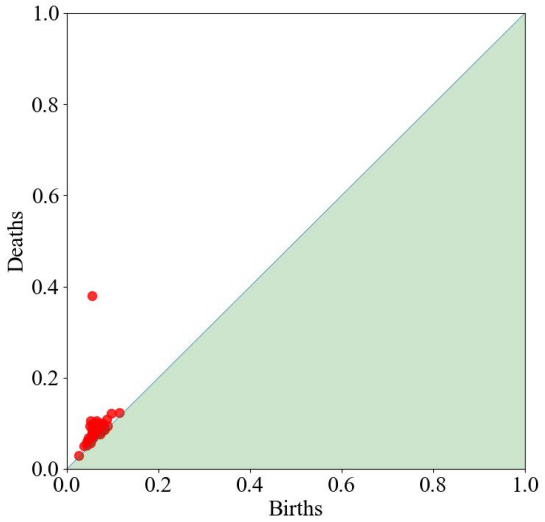



Figure: Fréchet Mean

**Mean Persistence Measure** The Empirical mean of the persistence diagrams is given by

$$\mu = \frac{1}{B}\left(\sum_{i=1}^{B} \mu_i\right)$$

where each, $\mu_i$ represents a persistent diagram's off-diagonal elements $\mu_i = \{(x_1, n_1), ..., (x_k, n_k)\}$ where $x_i$ represents the coordinates and $n_i$ the multiplicity of the point.

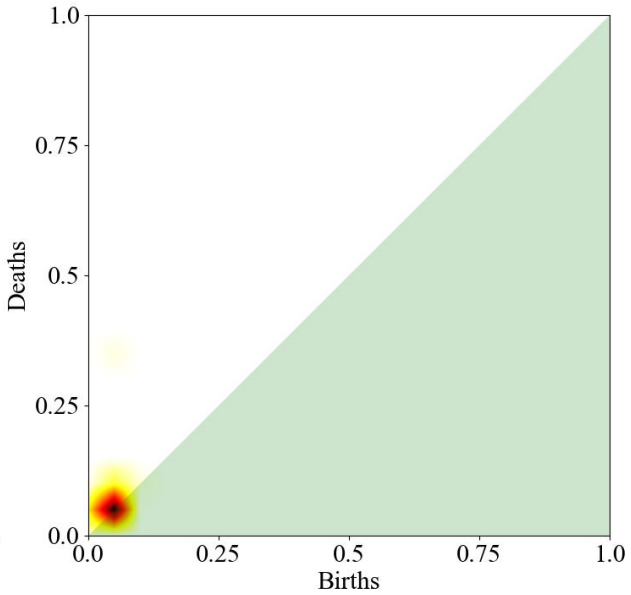



Figure: Mean Persistence

# Conclusion
## Future Work

**Weighted Wasserstein Barycenter** The weighted Wasserstein barycenter method is a technique for merging multiple persistence diagrams into a single diagram. It is based on the concept of Wasserstein distance, which measures the dissimilarity between probability distributions.
Given a set of $n$ persistence diagrams $D_1, D_2, \ldots, D_n$, the Wasserstein barycenter method computes a weighted average of the diagrams using the Wasserstein distance as the weighting function. More specifically, the method solves the following optimization problem:

$$\min_{D} \sum_{i=1}^{n} w_i W_2(D, D_i)^2$$

where $D$ is the merged persistence diagram, $w_i$ are the weights assigned to each input diagram, and $W_2(D, D_i)$ is the Wasserstein distance between $D$ and $D_i$.

After obtaining the merged persistence diagram using the Wasserstein barycenter method, we can further refine it using post-processing techniques, such as clustering, or outlier removal, to obtain a more accurate approximation of the persistence diagram of the large point cloud.

Overall, the Wasserstein barycenter method provides an efficient and effective way to merge multiple persistence diagrams into a single diagram, while retaining the topological information of the original data.

**Challenges:**

Although this seems viable option , a few things are still to be seen.

- A method to compute weights, one of the options is $w_j = e^{\left(-\gamma W_p(B_{i-1}, D_j)^p\right)}$
- Getting a persistence diagram out of this "weighted sum"
- Proving mathematically why and how this algorithm converges better.

# THANK YOU