Persistent Homology

MATH 476: Mathematical Data Science

Semi-rigorous introduction to persistent homology.

Metric Space

Metric Space

A metric space is a pair (M, d) of a set M and a distance function $d : M \times M \to \mathbb{R}$ satisfying the following properties:

1. $d(x, y) \geq 0$ for all $x, y \in M$
2. $d(x, y) = 0$ if and only if $x = y$
3. $d(x, y) = d(y, x)$ for all $x, y \in M$
4. $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in M$

Examples:

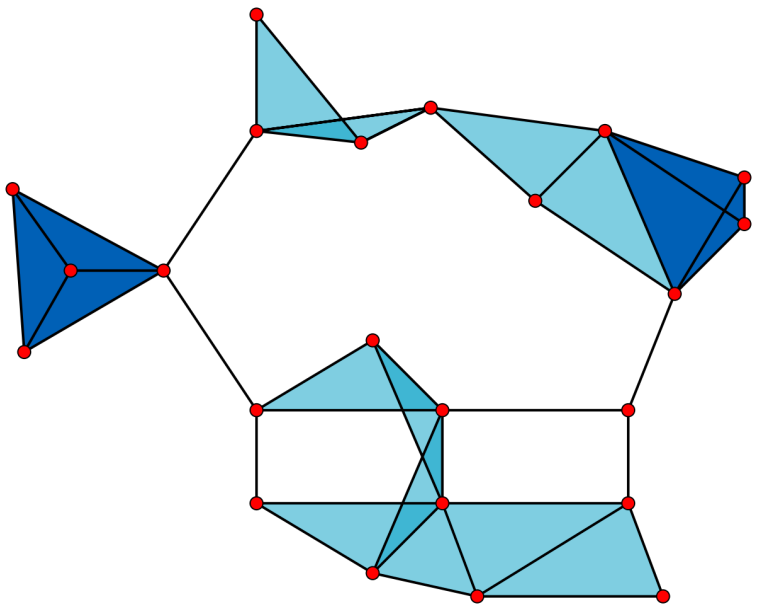
1. The set of real numbers $\mathbb{R}$ with the metric $d(x, y) = |x - y|$.
2. The set $\mathbb{R}^2$ with the metric $d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 - (y_1 - y_2)^2}$.

Čech and Rips Complex

Vietoris-Rips Complex

Let (P, d) be a metric space where P is a point set. Given a real number $r > 0$, the Vietoris-Rips or Rips complex is the simplicial complex $\mathcal{R}^r(P)$ such that a simplex σ is in $\mathcal{R}^r(P)$ if and only if $d(p, q) \leq r$ for every pair of vertices of σ.

Example: A rips complex on a set of 23 points



Čech and Rips Complex

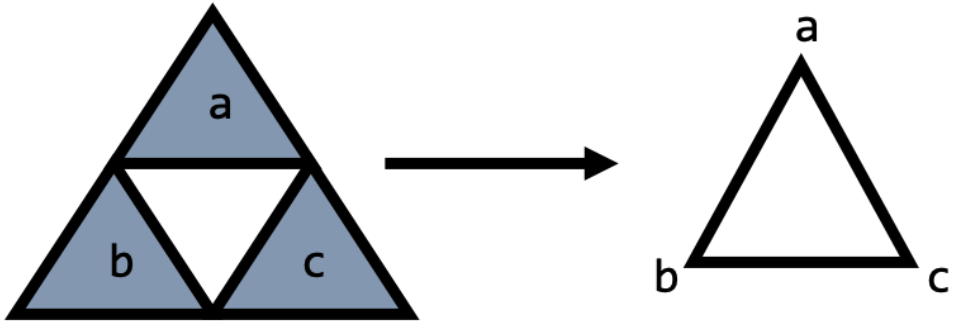
Nerve

Let M be a topological space and $\mathcal{M}$ be a set of subsets of M. The nerve of $\mathcal{M}$ is the simplicial complex $\mathcal{K}$ defined on the set $\mathcal{M}$ where a simplex $\{c_1, \ldots, c_k\} \subseteq \mathcal{M}$ is in $\mathcal{K}$ if

$$\cap_{i=1}^{k} c_i \neq \emptyset$$

In other words, the nerve is the collection of sets of facets that intersect at at least one point.

Example: A simplicial complex and its nerve



Čech and Rips Complex

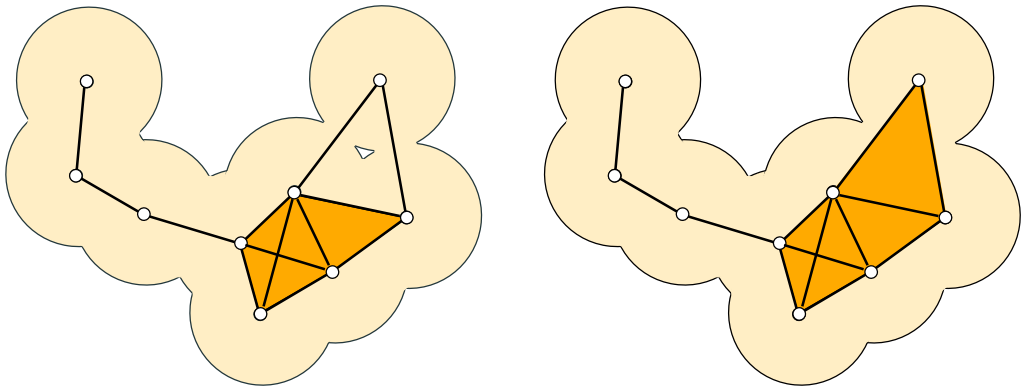
Čech Complex (formal definition)

Let (M, d) be a metric space with a topology induced by its metric and P be a subset of M. Given a real $r > 0$, the Čech complex $\mathcal{C}^r(P)$ is the nerve of the set $\{B(p, r/2) \mid p \in P\}$ where

$$B(p, r/2) = \{x \in M | d(p, x) < r/2\}$$

is the metric open ball of radius $r/2$ centering p.

Example: A Čech and Rips Complex on the same point cloud



Čech and Rips Complex

Čech Complex (constructive/algorithmic definition)

Given a point set P in some metric space and a number $r > 0$, the simplicies of the Čech complex $\mathcal{C}^r(P)$ can formed as follows:

- For each subset $\sigma \subseteq P$ of points, form a $(r/2)$-ball around each point in σ, and include σ as a simplex if there is a common point contained in all of the balls in σ.

It is easy to see that the object created with the above method is a simplicial complex, in other words, it is closed under the operation of taking subset. If σ is a simplex in $\mathcal{C}^r(P)$ and $\sigma' \subseteq \sigma$, then σ' will be also be a simplex in $\mathcal{C}^r(P)$.

Čech and Rips Complex

- How do we know that this simplicial complex resembles the topological space we used to construct it? In other words, does C_ε resemble the structure of P?
- We know that the union of these epsilon-balls forms some topological space, denote it $X(\varepsilon)$, that is close in structure to P.
- But how do we know if the Čech complex C_ε has same topoligcal structure as $X(\varepsilon)$?

To answer that question we need the following theorem.

Nerve Theorem

The homotopy types of $X(\varepsilon)$ and C_ε are the same.

We can roughly conclude that the Čech complex is a good representation of the data.

Čech and Rips Complex

We are primarily interested in the Rips complex. Knowing that the Čech complex is a good representation of the data, can we relate it to the Rips complex?

Proposition

Let P be a subset of a metric space (M, d). Then,

$$\mathcal{C}^r(P) \subseteq \mathcal{R}^r(P) \subseteq \mathcal{C}^{2r}(P)$$

This proposition tells us that if $\mathcal{C}^r(P)$ and $\mathcal{C}^{2r}(P)$ are good approximations of the underlying data, then so is the Rips complex. And the nerve theorem tells us that the Čech complexes are good representations of the data. We are finally ready to do persistent homology.

Proof of Proposition

It is clear that if $x \in \cap_{i=1}^{k} B(p_i, r/2)$, then the distances $d(p_i, p_j)$ for every pair $(i, j), 1 \leq i, j \leq k$, can be at most r. It follows that for every simplex $\{p_1, \ldots, p_k\} \in \mathcal{C}^r(P)$ is also in $\mathcal{R}^r(P)$. Thus, $\mathcal{C}^r(P) \subseteq \mathcal{R}^r(P)$.

Consider a simplex $\{p_1, \ldots, p_k\} \in \mathcal{R}^r(P)$. By definition of the Rips complex $d(p_i, p_1) \leq r$ for every $p_i, i = 1, \ldots, k$, we have $\cap_{i=1}^{k} B(p_i, r) \supset p_1 \neq \emptyset$. Then, by definition, $\{p_1, \ldots, p_k\}$ is also a simplex in $\mathcal{C}^{2r}(P)$. Therefore, $\mathcal{R}^r(P) \subseteq \mathcal{C}^{2r}(P)$.

Sequence of Rips Complexes

Persitent Homology associates to a point set S in a metric space (S, d) a sequence of Rips complexes. Then, we will find the homology groups of each complex in the sequence and analyze them. The general construction is as follows:

- One of these simplicial complexes in this sequence is formed by first viewing all the points as the 0-simplexes in the simplicial complex.
- Then we choose a number $\varepsilon > 0$, and connect any two 0-simplexes with a 1-simplex if the two points are less than distance ε apart.
- Then we connect 2-simplexes onto this space wherever there is a triple of 0-simplexes, all of which are less than distance ε apart from one another.
- The higher dimensional simplexes are added similarly.

Sequence of Rips Complexes

Note that:

- Converting a point cloud into a sequence of simplicial complexes requires a choice of parameter ε.
- For ε sufficiently small, the complex is a discrete set
- For ε sufficiently large, the complex is a single high-dimensional simplex.
- Is there an optimal choice for ε which best captures the topology of the data set?

To answer this question, consider the sampling of points on a planar annulus on the next page.

Sequence of Rips Complexes

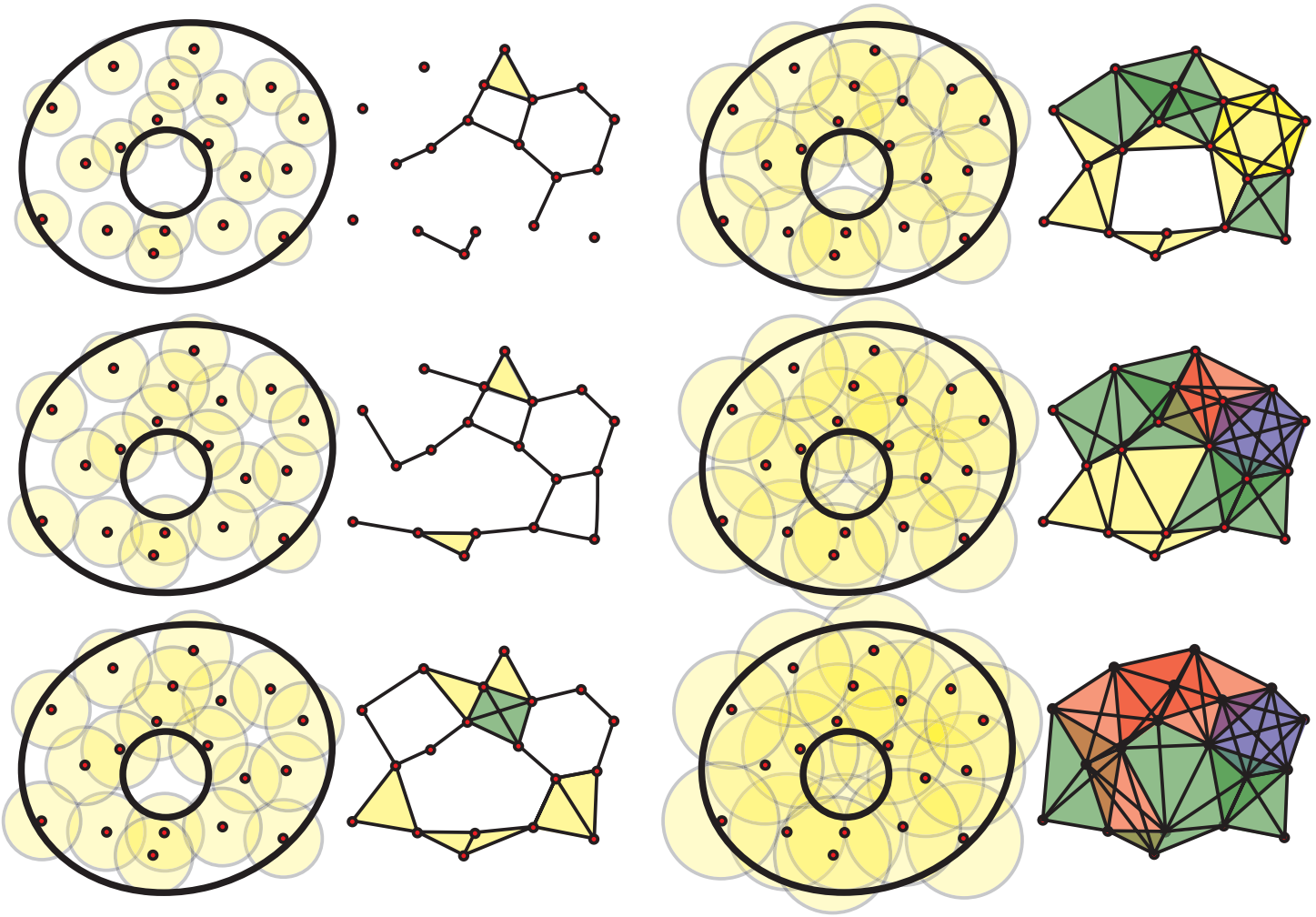


Figure: A sequence of Rips complexes for a point cloud data set representing an annulus. Upon increasing ε, holes appear and disappear.

Sequence of Rips Complexes

From the figure, we can deduce that:

- An ideal choice of ε may not even exist:
  - By the time ε is increased so as to remove small holes from within the annulus, the large hole distinguishing the annulus from the disk is filled in.
- It is insufficient to know the number of components and different dimensional holes (Betti numbers) of a single simplicial complex created from clouds of data with a particular ε.

In fact, it is a mistake to ask which value of ε is optimal. That is not the goal of persistent homology. What we need is a tool to declare which holes are essential and which can be ignored (as noise). To answer that question, we will develop the concept of persistence.

Filtrations

Filtration

A filtration of a simplicial complex, K, is a nested sequence of *subcomplexes* starting at the empty set and ending with the full simplicial complex, i.e.,

$$\emptyset \subset K_0 \subset K_1 \subset \cdots \subset K_m = K.$$

Going back to the Rips complex $\mathcal{R}^{\varepsilon}(S)$, we consider ε to be a free parameter. If we vary ε, we get different Rips complexes. In many data analysis situations, the value of ε that best describes the data is unknown or does not exist, so why not look at all of them? Observe if we increase ε continuously, then we get a family of nested Rips complexes: the *Rips filtration*. Now we will work through an example.

Filtrations

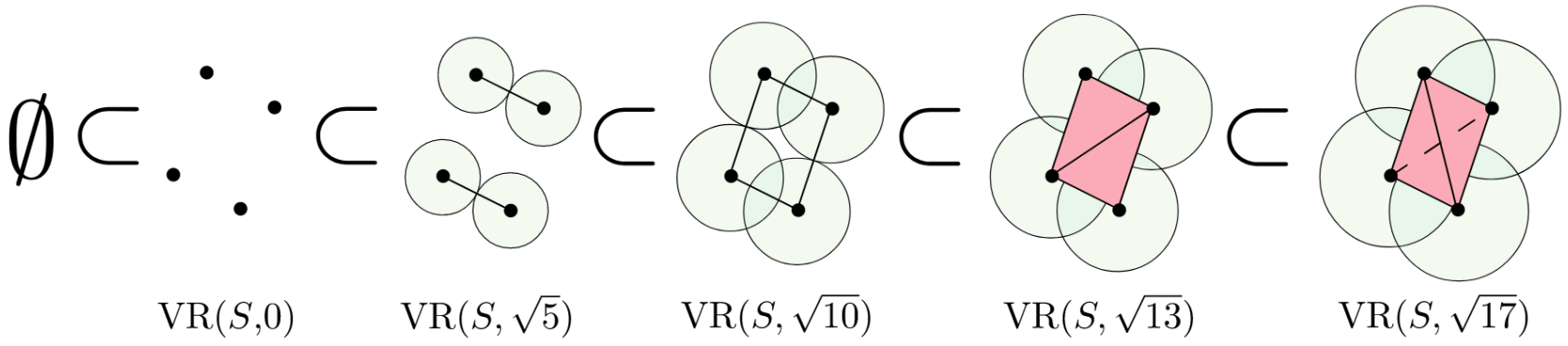


Figure: Visual representation of a Rips filtration for S

Example: Consider the point cloud
$S = \{(0,0),(1,3),(2,-1),(3,2)\} \subset \mathbb{R}^2$. We want to compute a Rips filtration on S for all $\varepsilon \geq 0$.

Filtrations

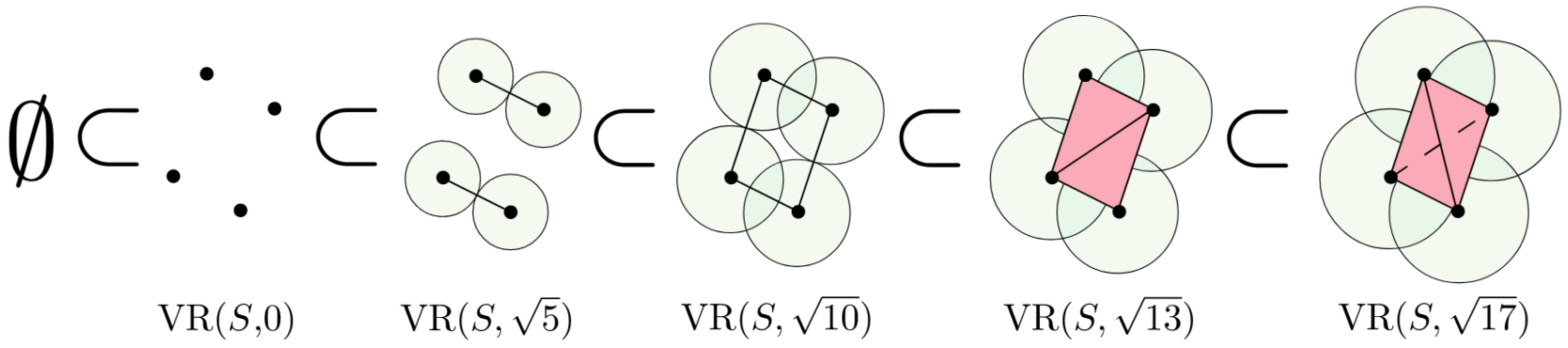


Figure: Visual representation of a Rips filtration for S

- When $\varepsilon < \sqrt{5}$, none of the balls of radius $\varepsilon/2$ intersect and so $\mathcal{R}^{\varepsilon}(S)$ is four points.
- When $\varepsilon = \sqrt{5}$, the balls of radius $\sqrt{5}/2$ centered at $(0,0)$ and $(2,-1)$ intersect which means we add a 1-simplex between $(0,0)$ and $(2,-1)$. Similarly, we add a 1-simplex between $(1,3)$ and $(3,2)$.

Filtrations

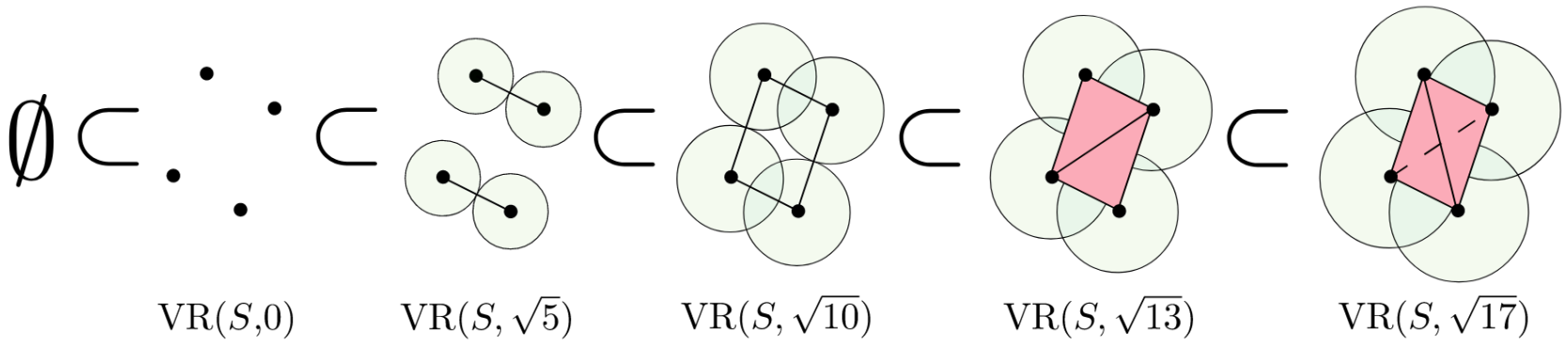


Figure: Visual representation of a Rips filtration for S

- When $\varepsilon \in (\sqrt{5}, \sqrt{10})$, no additional balls of radius $\varepsilon/2$ intersect which means $\mathcal{R}^{\varepsilon}(S) = \mathcal{R}^{\sqrt{5}}(S)$.
- When $\varepsilon = \sqrt{10}$, we add a two more 1-simplices between $(0, 0), (1, 3)$, and $(2, -1), (3, 2)$.

Filtrations

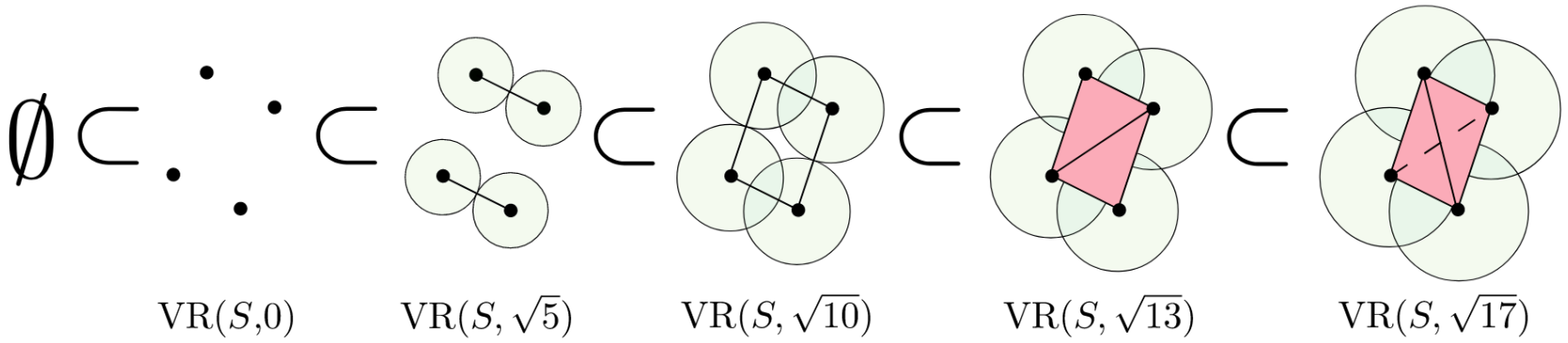


Figure: Visual representation of a Rips filtration for S

- When $\varepsilon \in (\sqrt{10}, \sqrt{13})$, $\mathcal{R}^{\varepsilon}(S) = \mathcal{R}^{\sqrt{10}}(S)$.
- When $\varepsilon = \sqrt{13}$, we add two 2-simplices between $(0,0), (1,3), (2,-1)$ and $(1,3), (2,-1), (3,2)$.
- When $\varepsilon \in (\sqrt{13}, \sqrt{17})$, $\mathcal{R}^{\varepsilon}(S) = \mathcal{R}^{\sqrt{13}}(S)$.

Filtrations

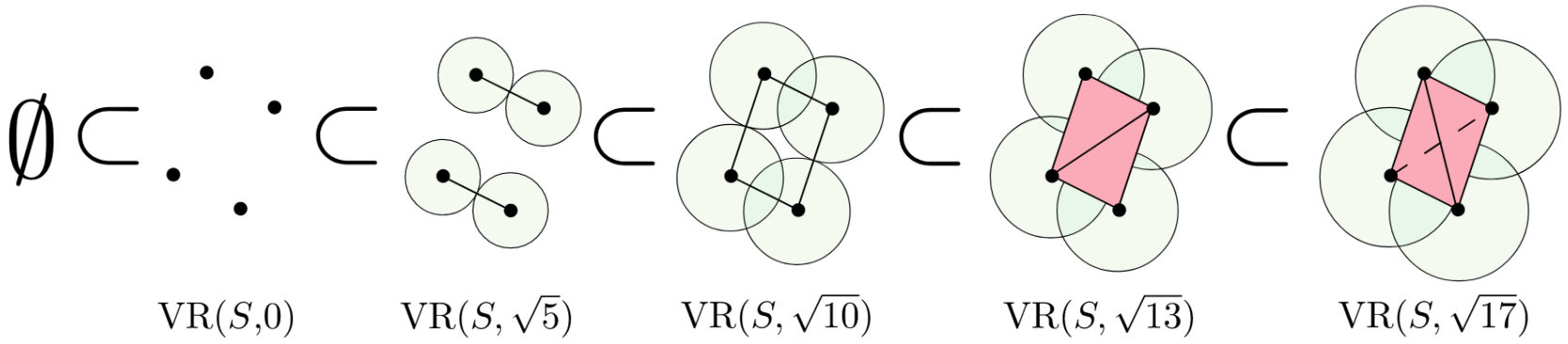


Figure: Visual representation of a Rips filtration for S

- When $\varepsilon = \sqrt{17}$, we add a 3-simplex.
- When $\varepsilon > \sqrt{17}$, $\mathcal{R}^{\varepsilon}(S) = \mathcal{R}^{\sqrt{17}}(S)$.

Persistence

Given a filtration, our goal now is to analyze it to identify persistence intervals in the complex filtered by ε.

Birth

For a filtered complex K and subcomplexes K_{i-1}, K_i, a topological feature $x \in H_p(K_i)$ is born at i if $x \notin H_p(K_{i-1})$.

Death

For a filtered complex K and subcomplexes K_{i-1}, K_i, a topological feature $x \in H_p(K_{i-1})$ dies at i if $x \notin H_p(K_i)$. A feature will also die if the feature merges with a feature born earlier in the filtration.

We can now use these concepts to define the persistence interval.

Persistence

Persistence interval

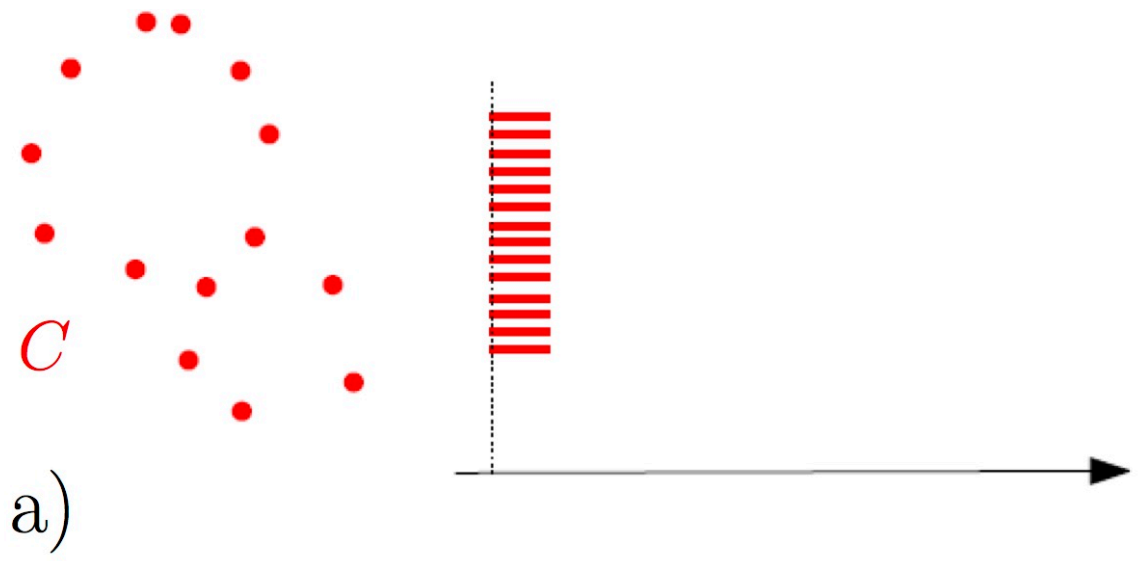
For a given topological feature x with birth point i and death point j, the persistence interval for the feature is given by $[i, j)$. If $j = \infty$, then the component does not die during the filtration (persists forever).

We can visualize this multiset of lifepsans or (birth, death) tuples as a persistence barcode.

Persistence barcode

A persistence barcode is a graphical representation of persistent intervals as a collection of horizontal line segments in a plane whose horizontal axis corresponds to the parameter ε and whose vertical axis represents an (arbitrary) ordering of topological feature (homology generators).

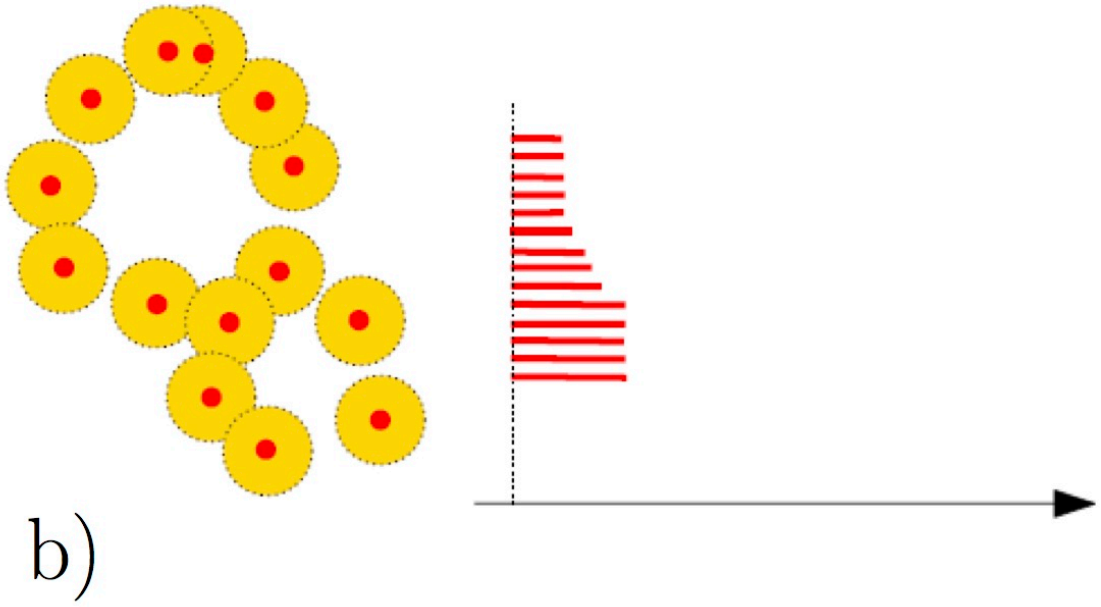
Persistence



Example: We will now walk throught the construction of a persistence barcode.

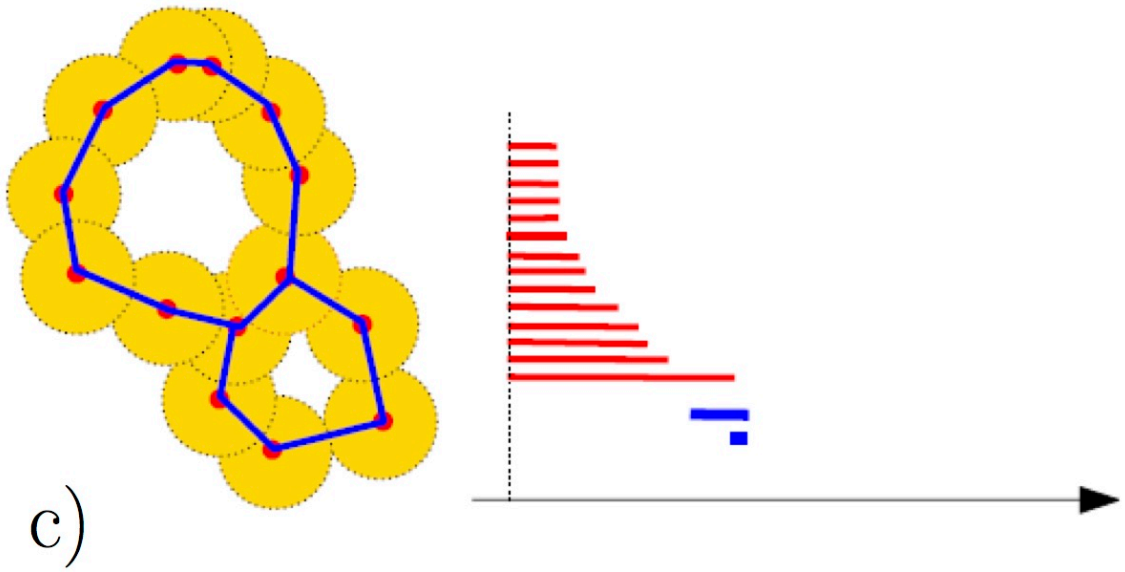
For the radius $r = 0$, the union of balls is reduced to the initial finite set of point, each of them corresponding to a 0-dimensional feature, i.e. a connected component; an interval is created for the birth for each of these features at $r = 0$.

Persistence



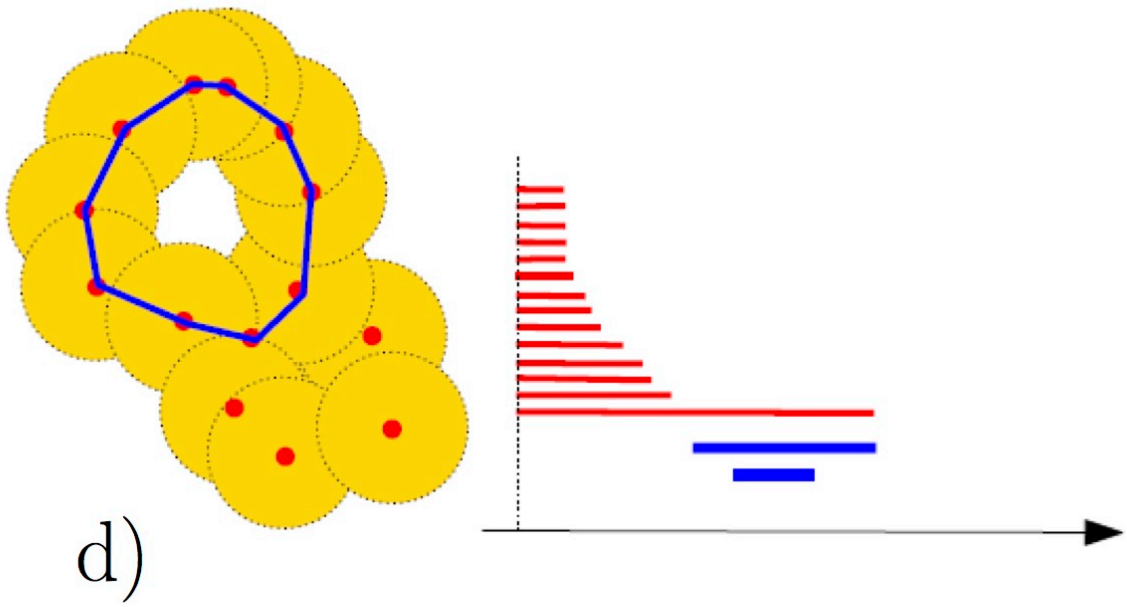
Some of the balls started to overlap resulting in the death of some connected components that get merged together.

Persistence



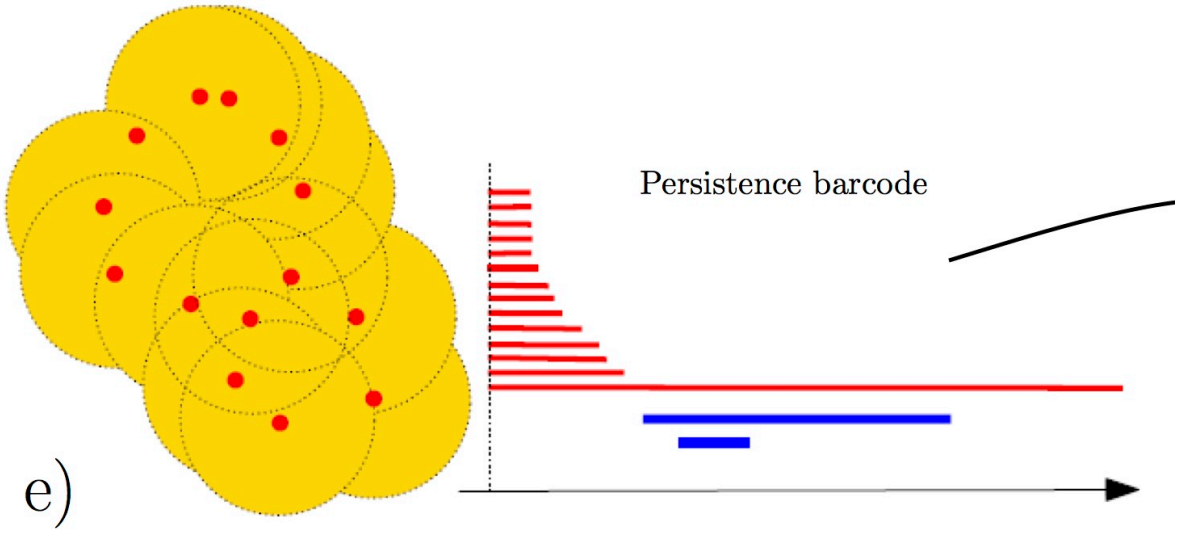
New components have merged giving rise to a single connected component and, so, all the intervals associated to a 0-dimensional feature have been ended, except the one corresponding to the remaining component; two new 1-dimensional features have appeared resulting in two new intervals (in blue) starting at their birth scale.

Persistence



One of the two 1-dimensional cycles has been filled, resulting in its death in the filtration and the end of the corresponding blue interval.

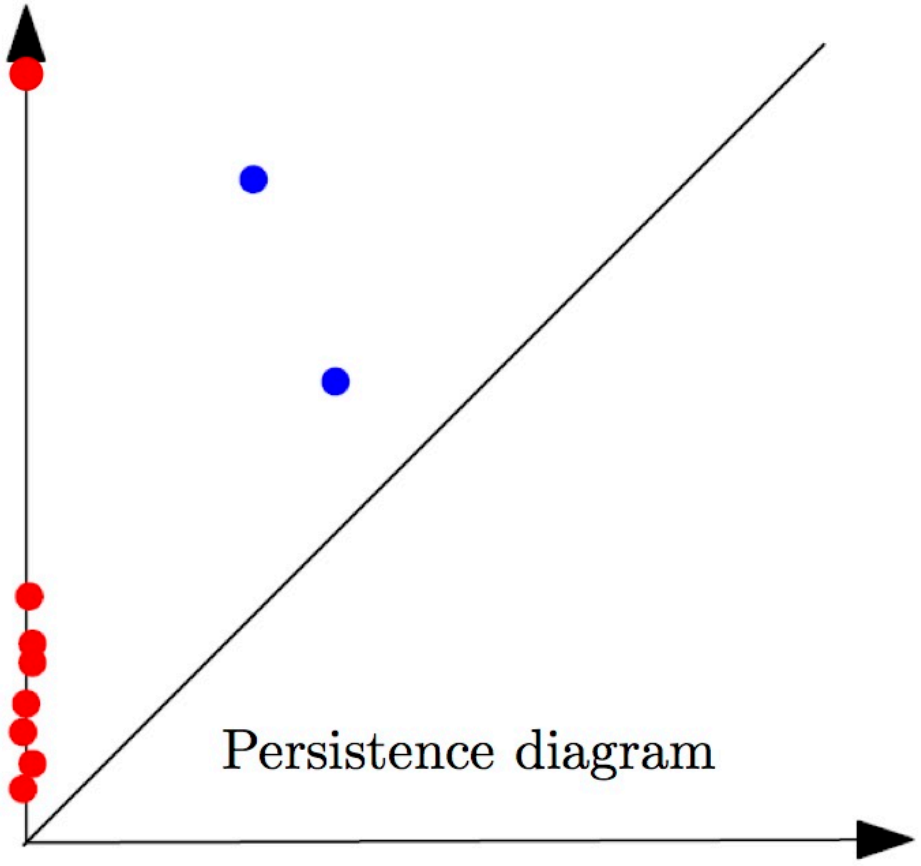
Persistence



All the 1-dimensional features have died, it only remains the long (and never dying) red interval.

Persistence

Another way of capturing this information is to use a persistence diagram. For example this is the persistence diagram for the above example.



Persistence

Notes:

- A barcode is best thought of as the persistence analogue of a Betti number.
- Recall from simplicial homology that the kth Betti number of a complex, $\beta_k = \text{rank } H_k$, acts as a coarse numerical measure of H_k.
- As with β_k, the barcode for H_k does not give any information about the finer structure of the homology, but merely a continuously parameterized rank.
- The value in a barcode representation is the ability to qualitatively filter out topological noise and capture significant features.

Persistence

Notes: (continued)

- It has been shown that barcodes are stable in the presence of noise added to certain filtrations. For example, in the filtration example, one sees (from a very coarse sampling) that the point cloud likely represents a connected object with one or two significant 'holes' as measured by H_1 and no significant higher homology.
- Due to the practical value of barcodes, Ripser, at the moment the fastest and one of the most popular TDA software packages, specializes in calculating those.