ETICS 2021

Research School on Uncertainty in Scientific Computing

Introduction to Topological Data Analysis

Bertrand Michel

Laboratoire de Mathémariques Jean Leray

UMR 6629 - Nantes

CENTRALE NANTES

Summary

1. Introduction
2. A brief look at topology
3. Simplicial Complexes and Homology
4. Homology inference
5. Persistent homology
6. Topological Data Analysis and Statistics
7. Topological Data Analysis and Machine Learning
8. Mapper

1 - Introduction

Topological Data Analysis and Topological Inference

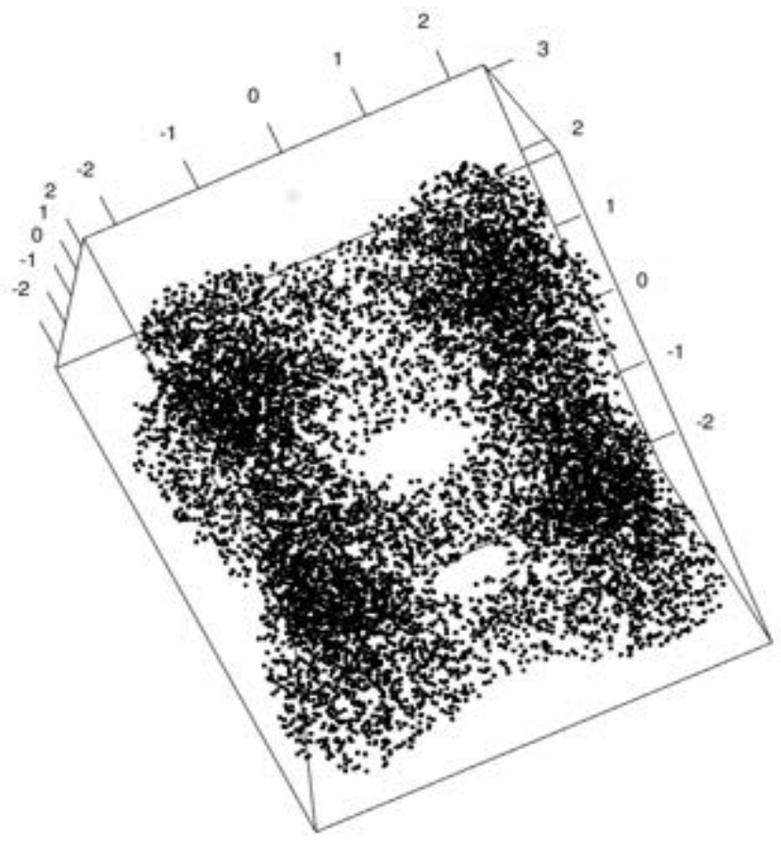
- **Geometric inference** and **algebraic topology tools**, **computational topology** has recently witnessed important developments with regards to data analysis, giving birth to the field of **topological data analysis** (TDA).

- The aim of TDA is to infer relevant, qualitative and quantitative **topological structures** (clusters, holes ...) directly from the data.

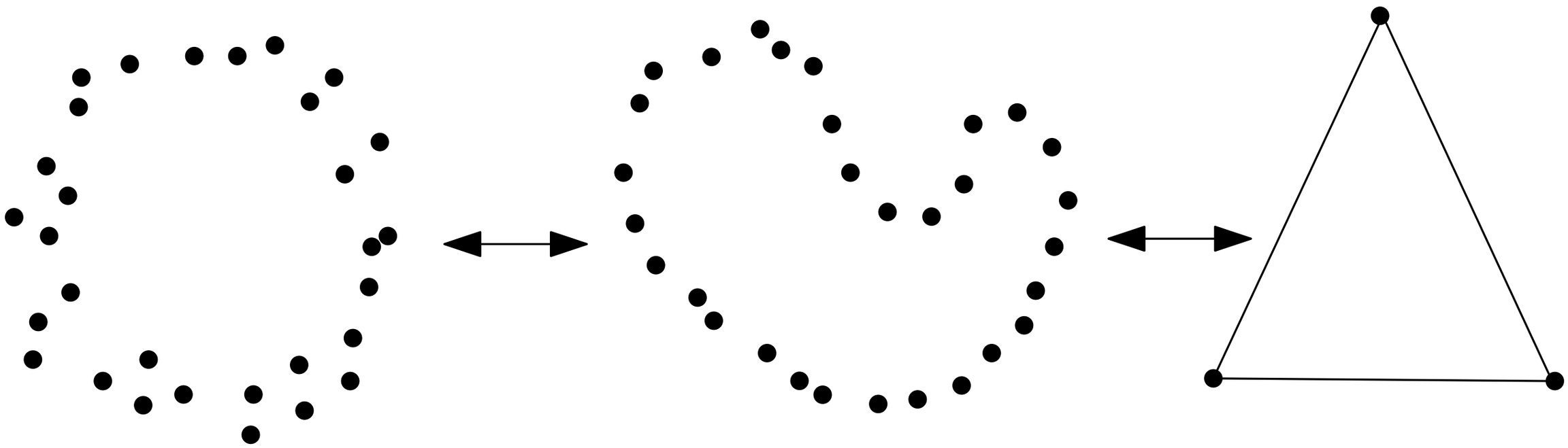
- The two popular methods in TDA : **Mapper algorithm** [Singh et al., 2007] and **persistent homology** [Edelsbrunner et al., 2002].

- TDA methods relies on **Topological Inference** methods / results.

- **Topological inference** methods aim to infer topological properties of an unknown topological space $\mathbb{X}$, typically from a point cloud $\mathbb{X}_n$ "close" to $\mathbb{X}$.



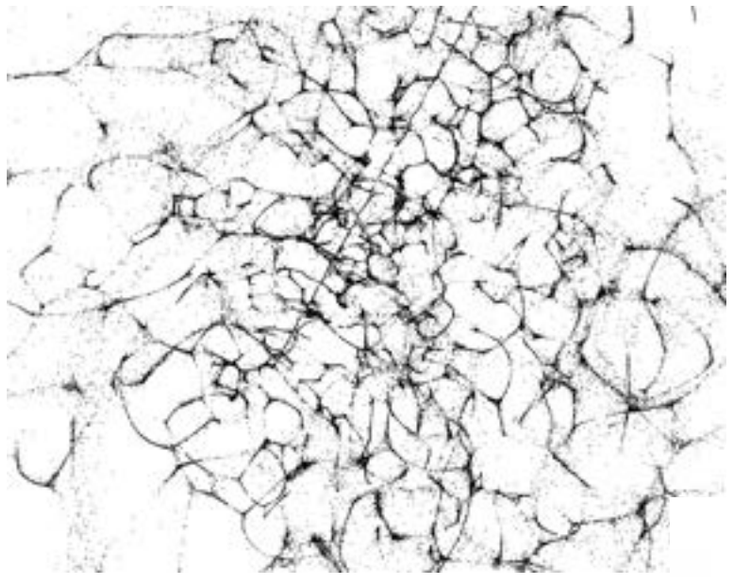
Why is topology interesting for data analysis?



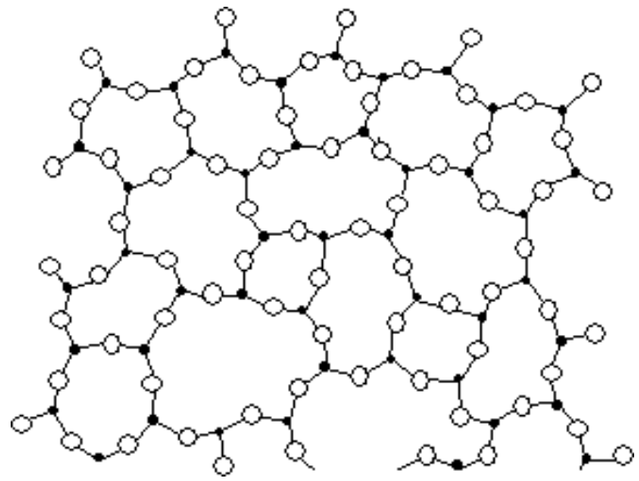
- **Coordinate invariance:** topological features/invariants do not rely on any coordinate system. $\Rightarrow$ no need to have data with coordinate or to embed data in spaces with coordinates... But the metric (distance/similarity between data points) is important.
- **Deformation invariance:** topological features are invariant under homeomorphism.
- **Compressed representation:** Topology offer a set of tools to summarize and represent the data in compact ways while preserving its global topological structure.
- **Informative:** can improve inference and prediction in ML problems

Application fields of TDA methods

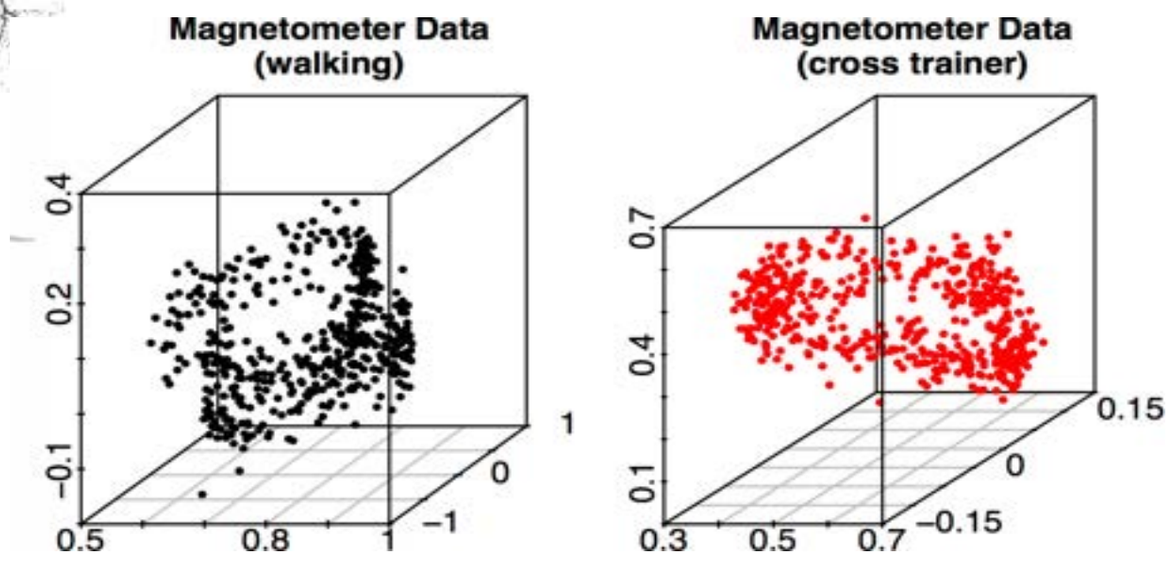
[distribution of galaxies]



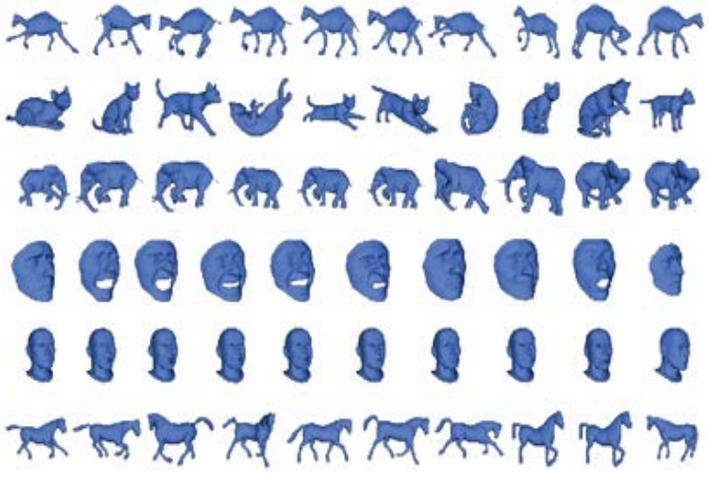
[nano materials]

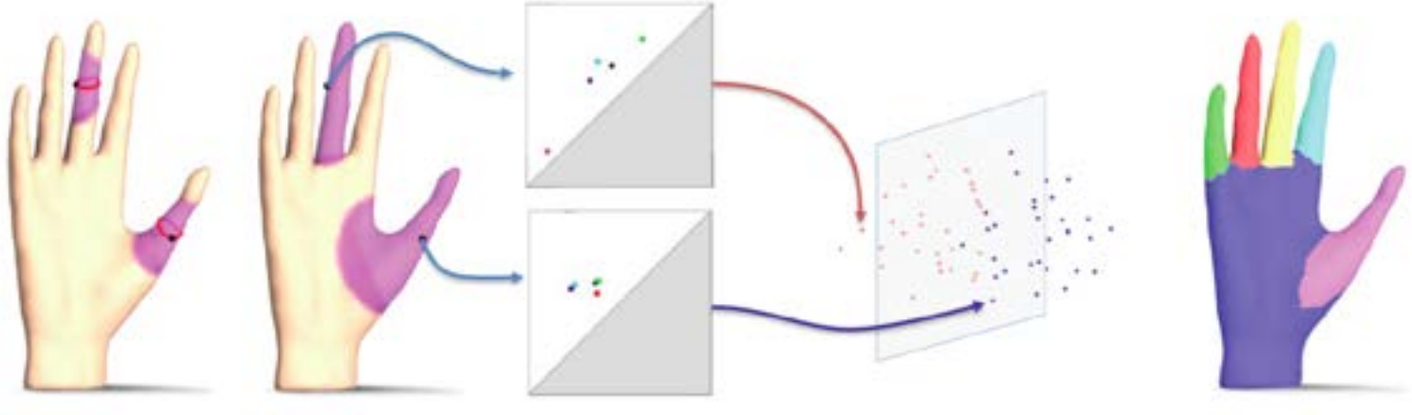


[Magnetometer Data]



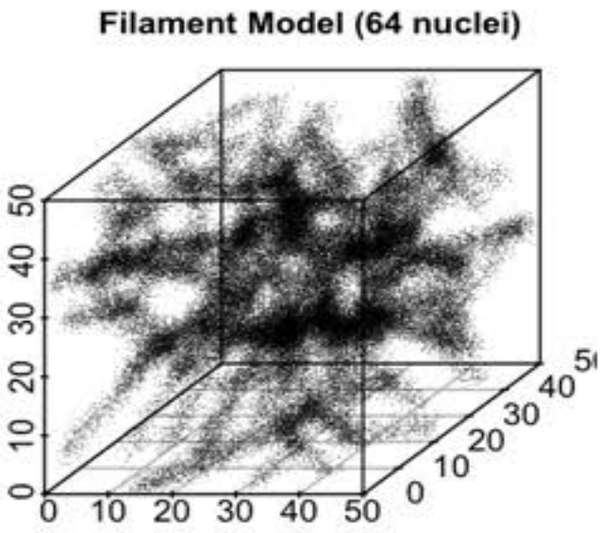
[3D shape database]

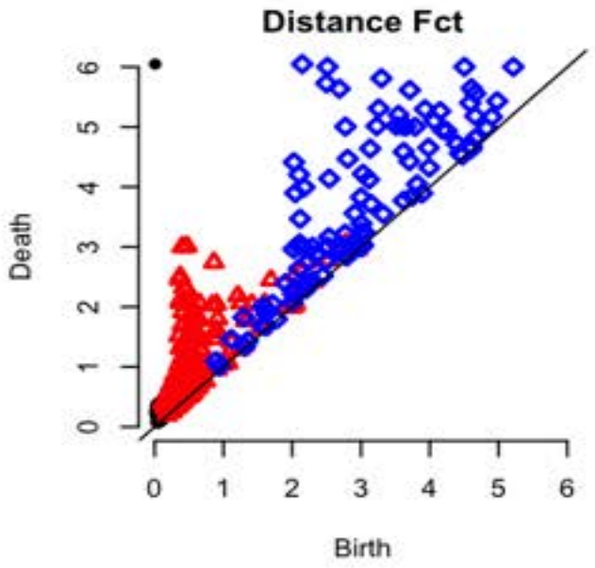


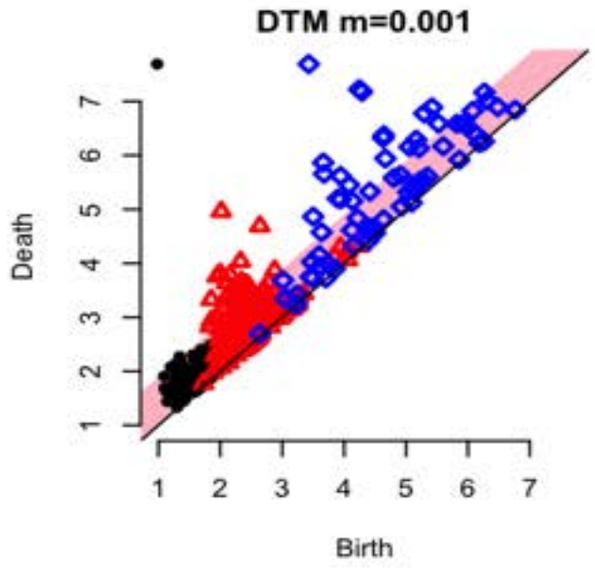


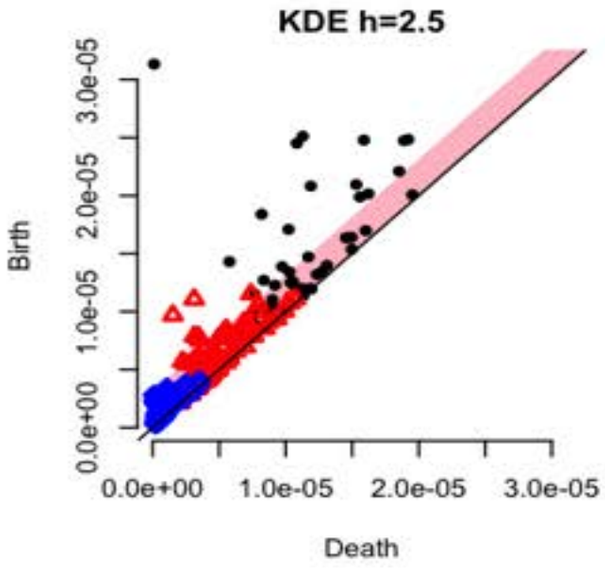
Topological Data Analysis (TDA)

- For **exploratory analysis**, visualization





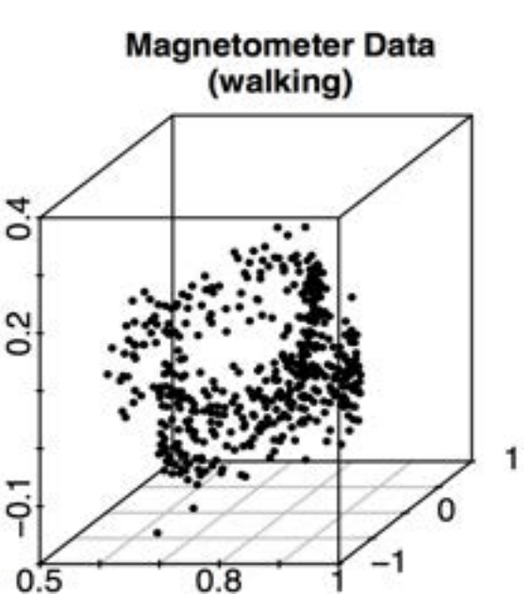


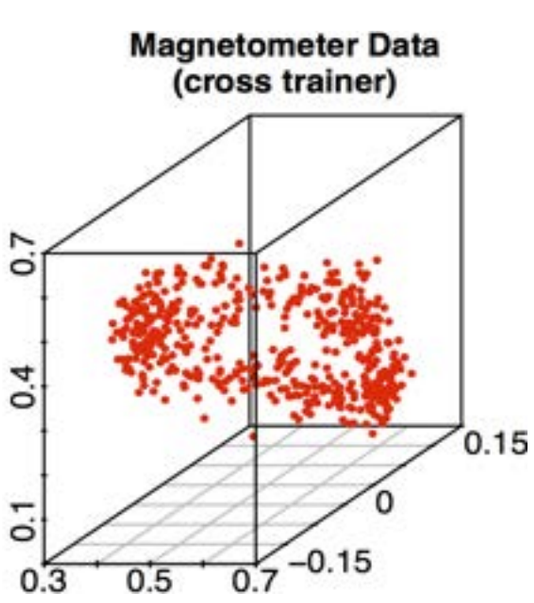


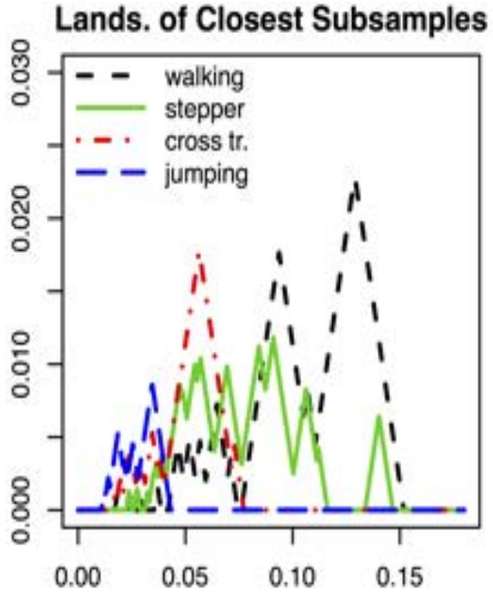
Topological Data Analysis (TDA)

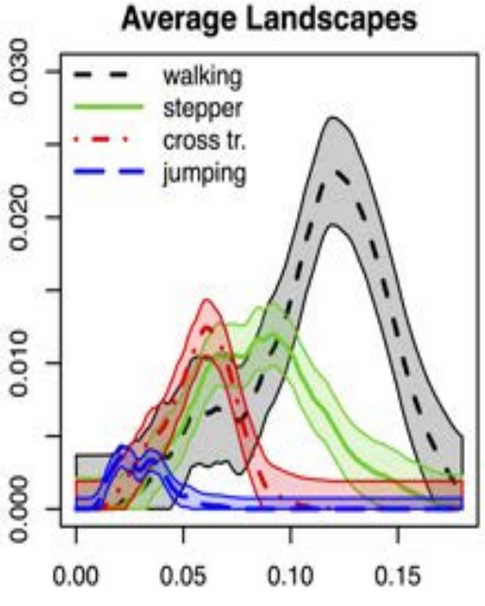
- For **exploratory analysis**, visualization
- For **feature extraction** and statistical learning

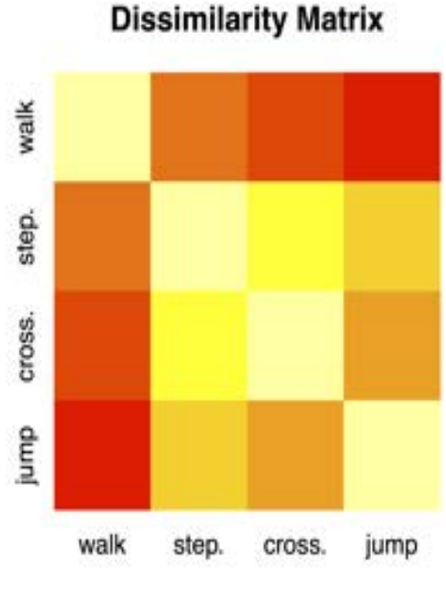
raw data → • Topological descriptors • Geometric descriptors • Other signatures → • supervised methods • unsupervised methods





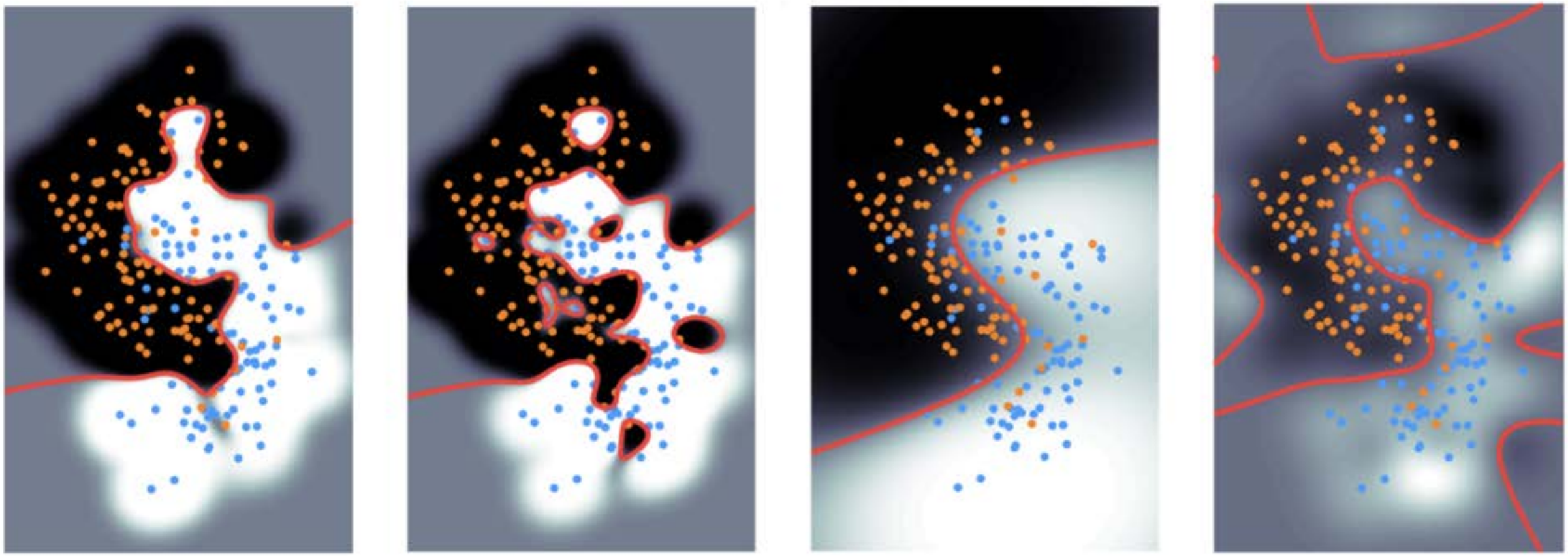






Topological Data Analysis (TDA)

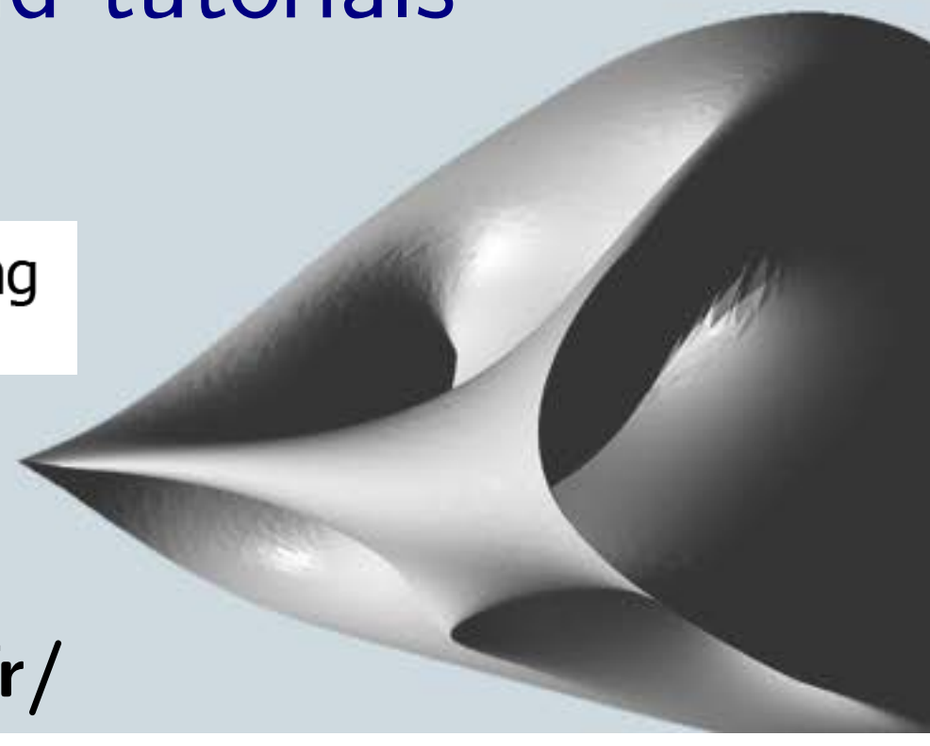
- For **exploratory analysis**, visualization
- For **feature extraction** and statistical learning
- For **model decision** (topology is used as a measure of complexity)



[Chen etal. - AISTAT 2019]

GUDHI library : Github and tutorials

गुढी GUDHI Geometry Understanding in Higher Dimensions



http://gudhi.gforge.inria.fr/

GUDHI :

- a C++/Python open source software library for TDA,
- a developers team, an editorial board, open to external contributions,
- provides state-of-the-art TDA data structures and algorithms : design of filtrations, computation of pre-defined filtrations, persistence diagrams,...
- part of GUDHI is interfaced to R through the TDA package.

2 - A brief look at topology

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces* and more specifically *topological spaces*.

A topology on a space specifies how the points of this space are "connected" by listing out what points constitute a neighborhood, the so-called an open set.

A brief look at topology

Roughly speaking, the goal of topology is to *classify spaces* and more specifically *topological spaces*.

A topology on a space specifies how the points of this space are "connected" by listing out what points constitute a neighborhood, the so-called an open set.

Def: A *topological space* is a set X equipped with a *topology*, i.e., a family $\mathcal{O}$ of subsets of X, called the *open sets* of X, such that:
(i) the empty set $\emptyset$ and X are elements of $\mathcal{O}$,
(ii) any union of elements of $\mathcal{O}$ is an element of $\mathcal{O}$,
(iii) any finite intersection of elements of $\mathcal{O}$ is an element of $\mathcal{O}$.

Examples ...

A brief look at topology

A very common family of topological spaces is comprised of the *metric spaces*.

Def: A metric (or distance) on X is a map $d : X \times X \to [0, +\infty)$ such that:
(i) for any $x, y \in X$, $d(x, y) = d(y, x)$,
(ii) for any $x, y \in X$, $d(x, y) = 0$ if and only if $x = y$,
(iii) for any $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$.
The set X together with d is a metric space.

The smallest topology containing all the open balls $B(x, r) = \{y \in X : d(x, y) < r\}$ is called the metric topology on X induced by d.

Ex: the standard topology in an Euclidean space is the one induced by the metric defined by the norm: $d(x, y) = \|x - y\|$.

A brief look at topology

- Equivalence of two topological spaces is determined by how the points that comprise them are connected.
  Ex : the surface of a cube can be deformed into a sphere without cutting or gluing it because they are connected the same way (they have the same topology.

- This notion of topological equivalence can be formalized via functions between topological spaces: the preservation of connectivity is achieved by preserving the open sets.

- A function from one space to another that preserves the open sets is called a continuous function.

- Continuity is a vehicle to define topological equivalence, because a continuous function can send many points to a single point in the target space, or send no points to a given point in the target space

Def: a map $f : X \to Y$ is *continuous* if and only if the pre-image $f^{-1}(O_Y) = \{x \in X : f(x) \in O_Y\}$ of any open set $O_Y \subseteq Y$ is an open set of X.

A brief look at topology

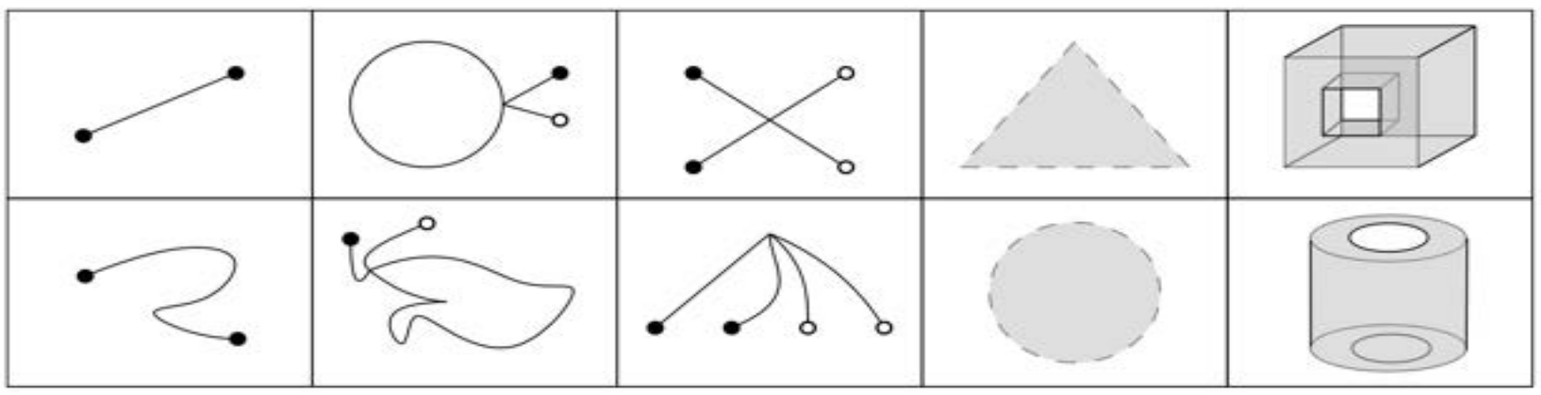
Def: Let X and Y be topological spaces. An homeomorphism is a bijective and continuous function $f : X \to Y$ whose inverse is continuous too.

- Two topological spaces are homeomorphic if there exists a homeomorphism between them.
- Homeomorphism induces an equivalence relation among topological spaces (topologically equivalent).

A brief look at topology

Def: Let X and Y be topological spaces. An homeomorphism is a bijective and continuous function $f : X \to Y$ whose inverse is continuous too.

- Two topological spaces are homeomorphic if there exists a homeomorphism between them.
- Homeomorphism induces an equivalence relation among topological spaces (topologically equivalent).



Computational Topology for Data Analysis, Tamal Krishna Dey and Yusu Wang, to be Published by Cambridge University Press, 2021

Each point set in this figure is homeomorphic to the point set above or below it, but not to any of the others.
Open circles indicate points missing from the point set, as do the dashed edges in the point sets second from the right.

A brief look at topology

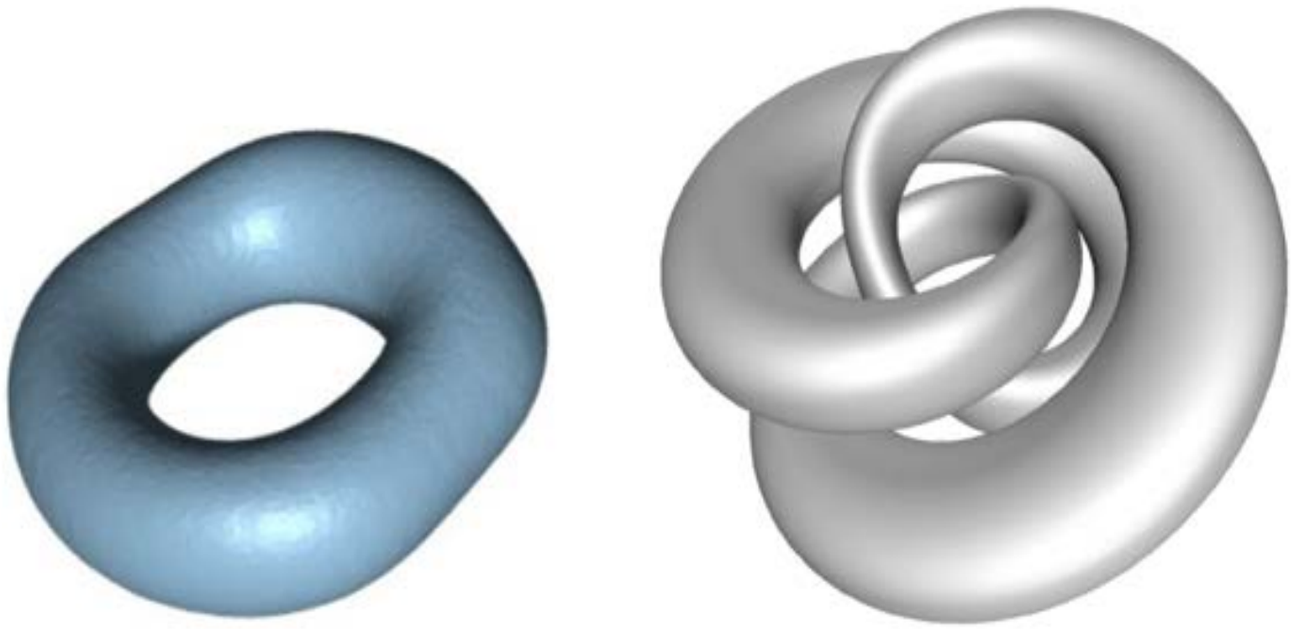
Two subspaces are said to be isotopic if one can be continuously deformed to the other while keeping the deforming subspace homeomorphic to its original form all the time.

Def: X and Y are isotopic if $\exists$ a continuous map (isotopy) $F : X \times [0,1] \to Y$ s.t. $F(.,0) = \mathrm{id}_X$, $F(X,1) = Y$ and $\forall t \in [0,1]$, $F(.,t)$ is an homeomorphism.

A brief look at topology

Two subspaces are said to be isotopic if one can be continuously deformed to the other while keeping the deforming subspace homeomorphic to its original form all the time.

Def: X and Y are isotopic if $\exists$ a continuous map (isotopy) $F : X \times [0,1] \to Y$ s.t. $F(.,0) = \mathrm{id}_X$, $F(X,1) = Y$ and $\forall t \in [0,1]$, $F(.,t)$ is an homeomorphism.



Computational Topology for Data Analysis, Tamal Krishna Dey and Yusu Wang, to be Published by Cambridge University Press, 2021

- Turus and knotted torus: the two tori are homeomorphic but not isotopic
- Any attempt to continuously deform one to the other while keeping it homeomorphic to the original would force the torus to be "self-intersecting"

A brief look at topology

- Homotopy equivalence is a weaker notion of similarity among topological spaces than homeomorphism.

- It relates spaces that can be continuously deformed to one another but the transformation may not preserve homeomorphism.
  Ex.: a ball can shrink to a point, which is not homeomorphic to it (no bijection)

Def: • Two maps $f_0 : X \to Y$ and $f_1 : X \to Y$ are *homotopic* if $\exists$ a continuous map $F : [0,1] \times X \to Y$ s.t. $\forall x \in X,\ F(0,x) = f_0(x)$ and $F_1(1,x) = f_1(x)$.
• X and Y are *homotopy equivalent* if $\exists$ continuous maps $f : X \to Y$ and $g : Y \to X$ s.t. $g \circ f$ is homotopic to id_X and $f \circ g$ is homotopic to id_Y.

Examples

A brief look at topology

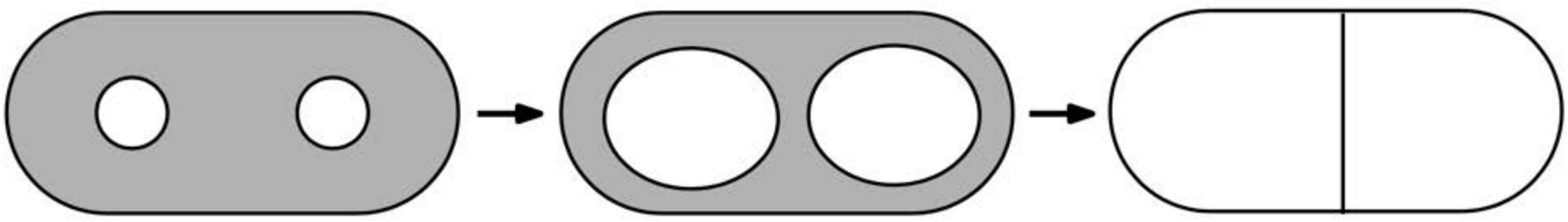
Def: If $Y \subseteq X$ and if there exists a continuous map $F : [0,1] \times X \to X$ s.t.:
(i) $\forall x \in X,\ F(0,x) = x$
(ii) $\forall x \in X,\ F(1,x) \in Y$
(iii) $\forall y \in Y,\ \forall t \in [0,1],\ F(t,y) = y$
then F is a *deformation retract* of X onto Y and in particular X and Y are **homotopy equivalent**.

Ex : any point on a line segment is a deformation retract of the line segment and is homotopy equivalent to it.

A brief look at topology

Def: If $Y \subseteq X$ and if there exists a continuous map $F : [0,1] \times X \to X$ s.t.:
(i) $\forall x \in X,\ F(0,x) = x$
(ii) $\forall x \in X,\ F(1,x) \in Y$
(iii) $\forall y \in Y,\ \forall t \in [0,1],\ F(t,y) = y$
then F is a *deformation retract* of X onto Y and in particular X and Y are **homotopy equivalent**.

Ex : any point on a line segment is a deformation retract of the line segment and is homotopy equivalent to it.



Computational Topology for Data Analysis, 2021

All three of the topological spaces are homotopy equivalent, because they are all deformation retracts of the left most space.

A brief look at topology

Pb 1: How to encode topological spaces for computational purposes?

A brief look at topology

Pb 1: How to encode topological spaces for computational purposes?

Pb 2: Looking for homotopy equivalences/homeomorphisms/isotopies is extremely difficult. Are there mathematical quantities that are invariant to homotopy equivalences **and** easy to compute?

3 - Simplicial Complexes and Homology

A topological space fit for computation

Pb 1: How to encode topological spaces for computational purposes?

Offsets

Point clouds in themselves do not carry any non trivial topological or geometric structure.

For a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$ (or in a metric space),
the α-offset of $\mathbb{X}_n$ is defined by

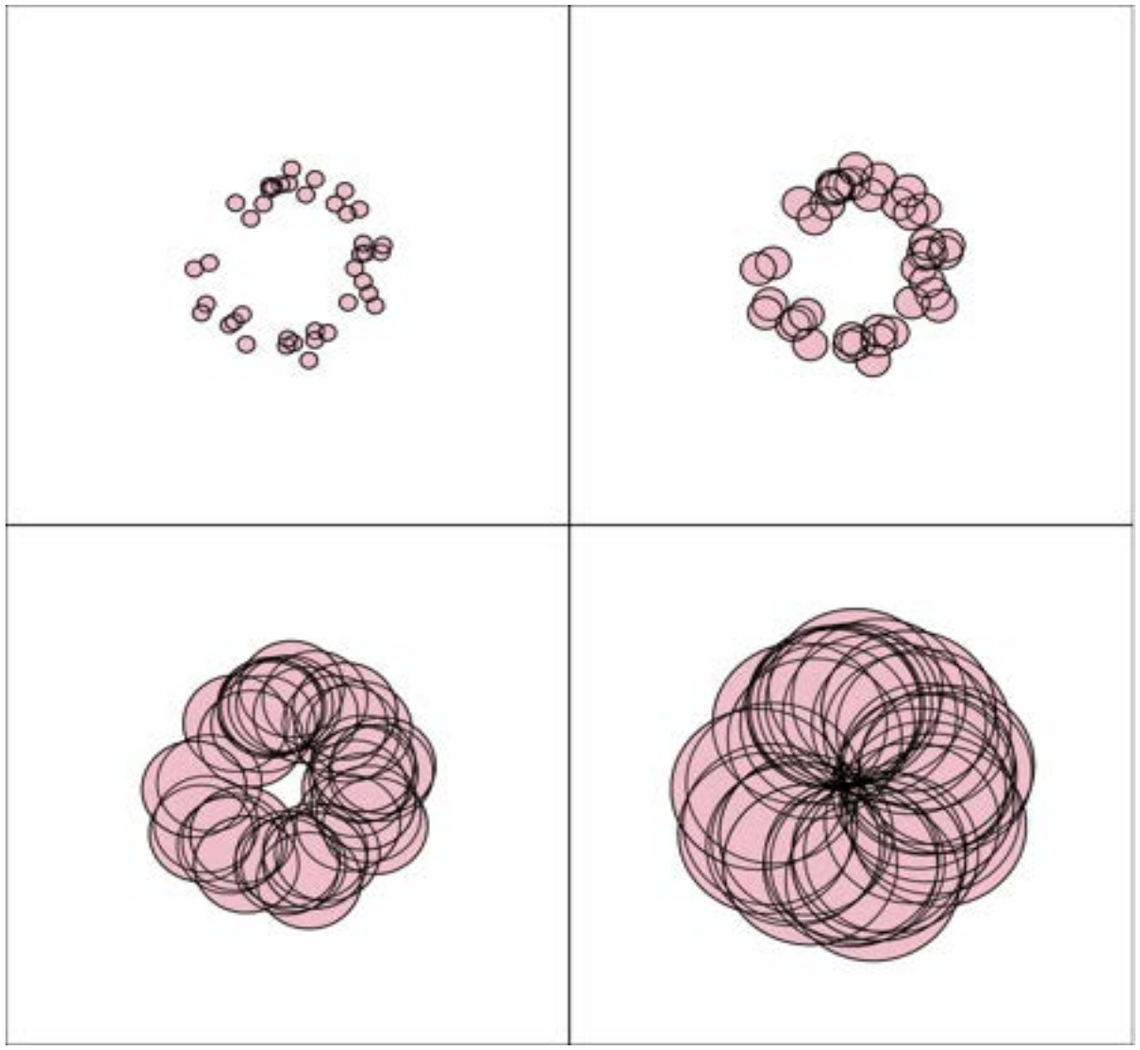
$$\mathbb{X}_n^\alpha = \bigcup_{x \in \mathbb{X}_n} B(x, \alpha).$$

More generally, for any compact set $\mathbb{X}$,

$$\mathbb{X}^\alpha := \bigcup_{x \in \mathbb{X}} B(x, \alpha) = d_{\mathbb{X}}^{-1}([0, \alpha])$$

where the distance function $d_{\mathbb{X}}$ to $\mathbb{X}$ is

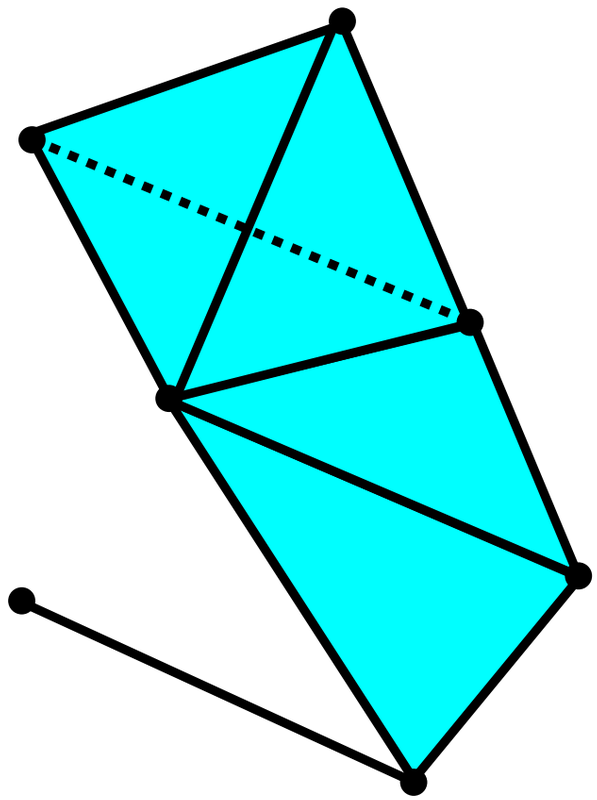
$$d_{\mathbb{X}}(y) = \inf_{x \in \mathbb{X}} \|x - y\| \qquad \text{(in } \mathbb{R}^d\text{)}$$



General idea: deduce from $(\mathbb{X}_n^\alpha)_{r>0}$ some topological and geometric information of an underlying object.

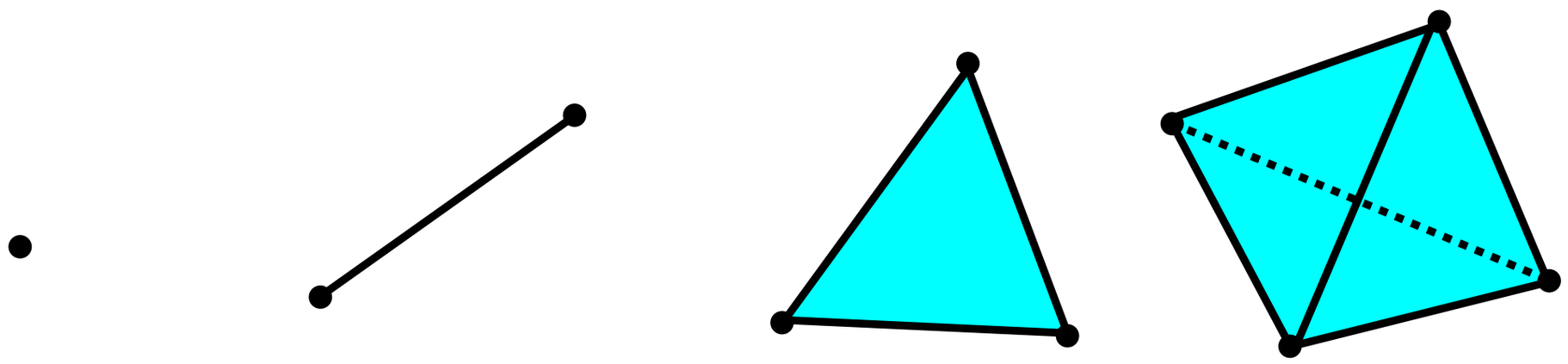
Simplicial Complexes

Non-discrete sets such as offsets, and also continuous mathematical shapes like curves, surfaces cannot easily be encoded as finite discrete structures.



We consider spaces made of small convex bricks, namely the *simplicial complexes* made of *simplices*.

Simplicial complexes



0-simplex: vertex

1-simplex: edge

2-simplex: triangle

3-simplex: tetrahedron

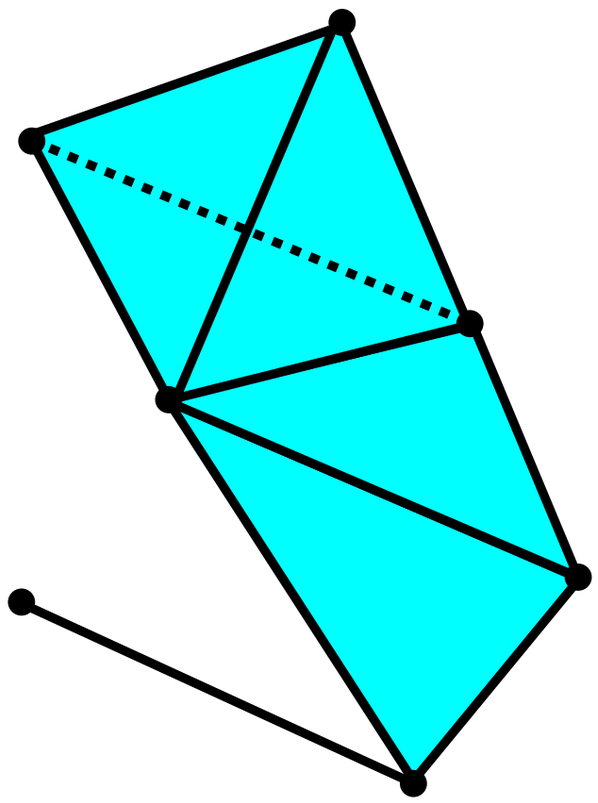
etc...

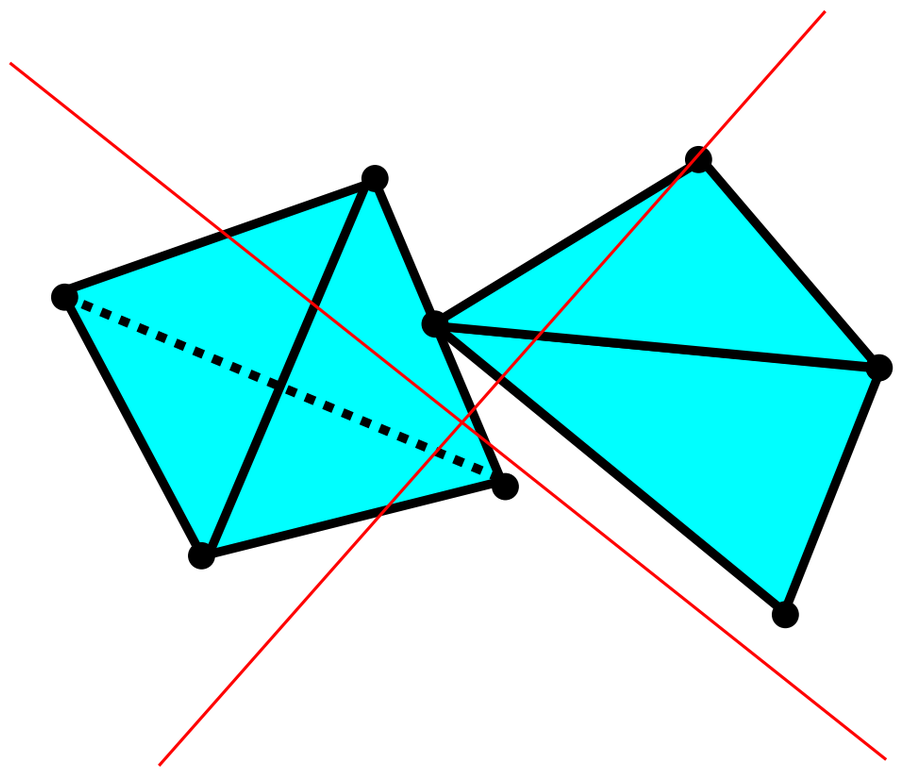
Given a set $P = \{p_0, \ldots, p_k\} \subset \mathbb{R}^d$ of $k+1$ affinely independent points, the k-dimensional simplex σ, or k-simplex for short, spanned by P is the set of convex combinations

$$\sum_{i=0}^{k} \lambda_i \, p_i, \quad \text{with} \quad \sum_{i=0}^{k} \lambda_i = 1 \quad \text{and} \quad \lambda_i \geq 0.$$

The points $p_0, \ldots, p_k$ are called the vertices of σ.

Simplicial complexes





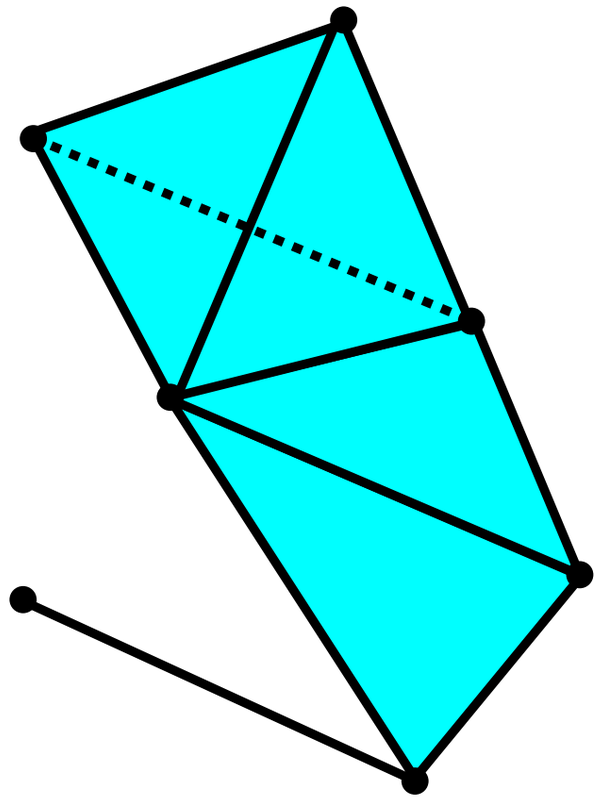
A (finite) simplicial complex K in $\mathbb{R}^d$ is a (finite) collection of simplices such that:

1. any face of a simplex of K is a simplex of K,
2. the intersection of any two simplices of K is either empty or a common face of both.

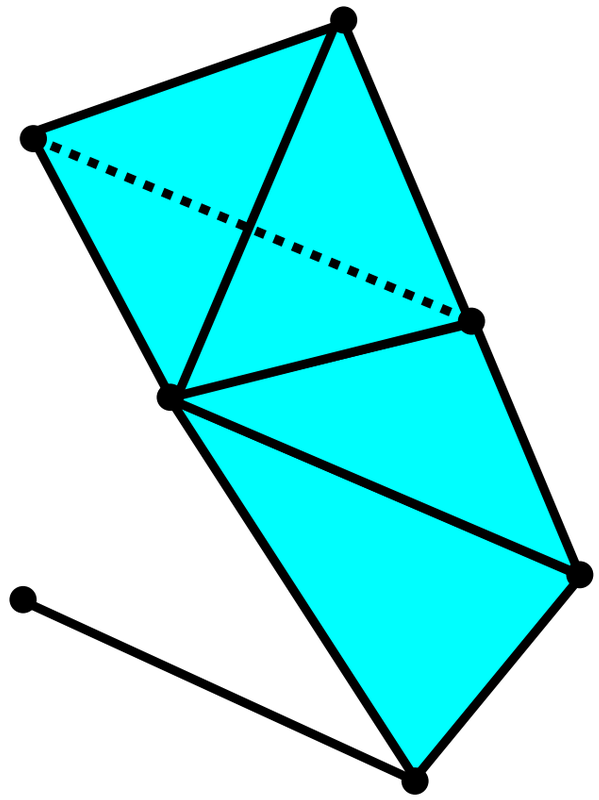
The underlying space of K, denoted by $|K| \subset \mathbb{R}^d$ is the union of the simplices of K.

Abstract simplicial complexes

- Given a set V, an *abstract simplicial complex* with vertex set V is a set $\tilde{K}$ of finite subsets of V such that
  - the elements of V belongs to $\tilde{K}$ and
  - for any $\sigma \in \tilde{K}$ any subset of σ belongs to $\tilde{K}$.
- The elements of $\tilde{K}$ are called the faces or the simplices of $\tilde{K}$.



Abstract simplicial complexes

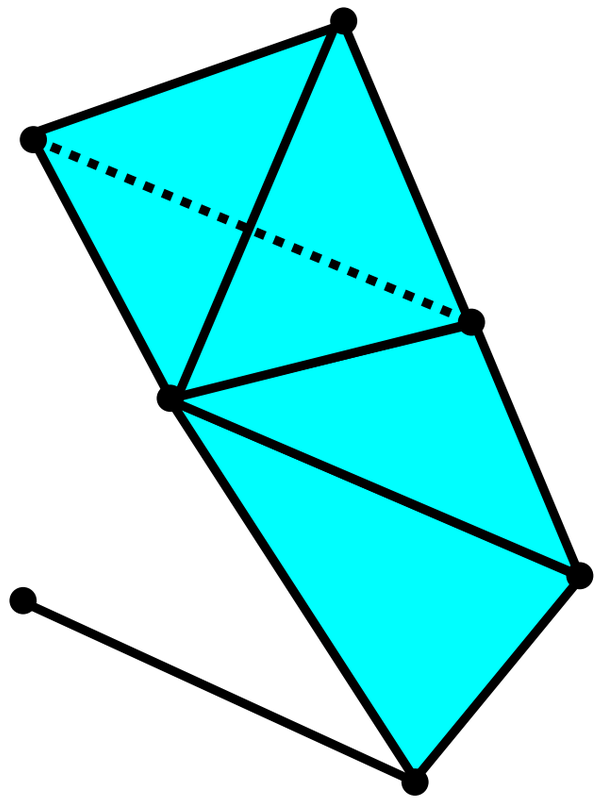


- Given a set V, an *abstract simplicial complex* with vertex set V is a set $\tilde{K}$ of finite subsets of V such that
  - the elements of V belongs to $\tilde{K}$ and
  - for any $\sigma \in \tilde{K}$ any subset of σ belongs to $\tilde{K}$.
- The elements of $\tilde{K}$ are called the faces or the simplices of $\tilde{K}$.

- The combinatorial description of any geometric simplicial K obviously gives rise to an abstract simplicial complex $\tilde{K}$.
- Conversely, one can always associate to an abstract simplicial complex $\tilde{K}$, a topological space $|\tilde{K}|$ such that if K is a geometric complex whose combinatorial description is the same as $\tilde{K}$, then the underlying space of K is homeomorphic to $|\tilde{K}|$. Such a K is called a *geometric realization* of $\tilde{K}$.

Abstract simplicial complexes

- Given a set V, an *abstract simplicial complex* with vertex set V is a set $\tilde{K}$ of finite subsets of V such that
  - the elements of V belongs to $\tilde{K}$ and
  - for any $\sigma \in \tilde{K}$ any subset of σ belongs to $\tilde{K}$.
- The elements of $\tilde{K}$ are called the faces or the simplices of $\tilde{K}$.



IMPORTANT

Simplicial complexes can be seen at the same time as geometric/topological spaces (good for top./geom. inference) and as combinatorial objects (abstract simplicial complexes, good for computations).

An invariant fit for computation

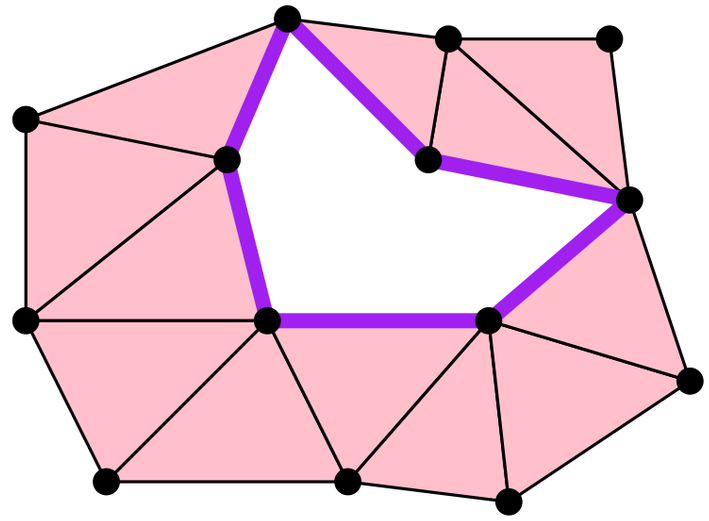
Pb 2: Looking for homotopy equivalences/homeomorphisms/isotopies is extremely difficult. Are there mathematical quantities that are invariant to homotopy equivalences **and** easy to compute?

A: The *holes*, encoded in the *homology groups* H_k, $k \in \mathbb{N}$

Simplicial Homology

How to characterize a hole in a simplicial complex?

A hole (in 1D) is a path whose first and end points are the same, a loop.

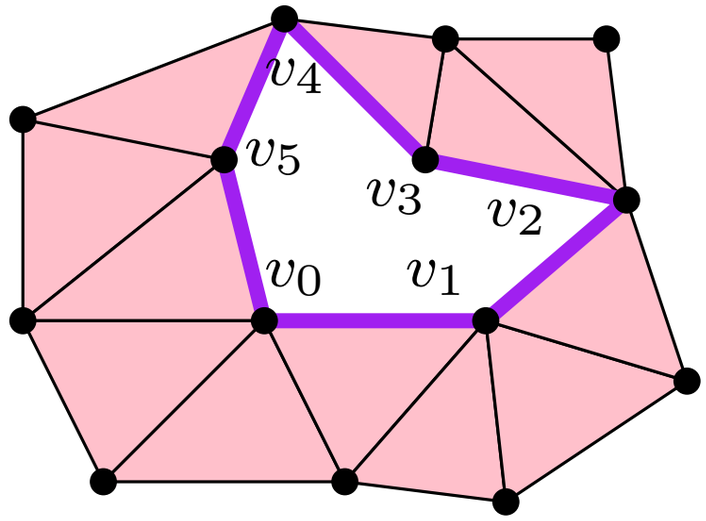


Simplicial Homology

How to characterize a hole in a simplicial complex?

A hole (in 1D) is a path whose first and end points are the same, a loop.

The sequence of 1-dimensional simplices $[v_0, v_1]$, $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$, $[v_4, v_5]$, $[v_5, v_0]$ is a hole

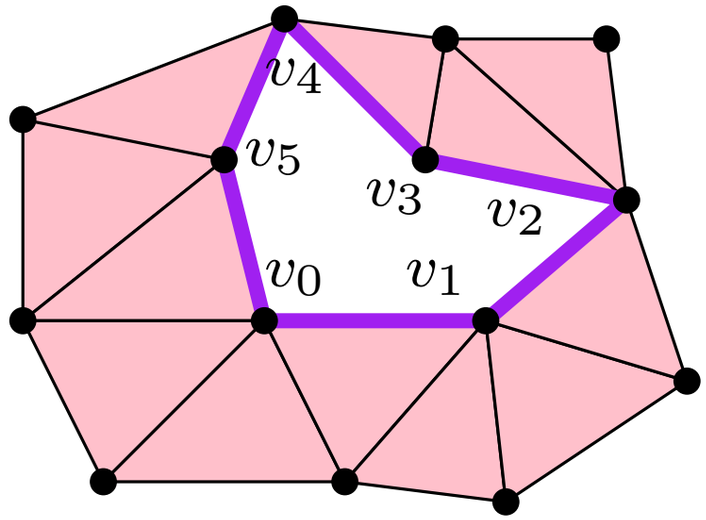


Simplicial Homology

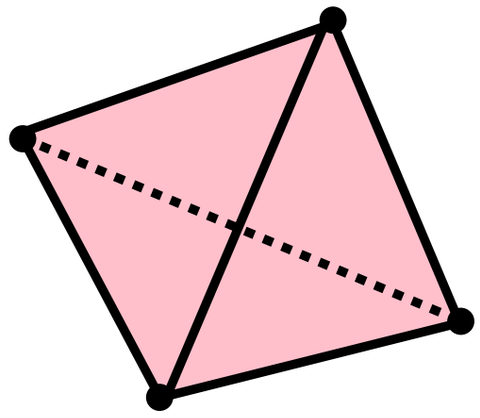
How to characterize a hole in a simplicial complex?

A hole (in 1D) is a path whose first and end points are the same, a loop.

The sequence of 1-dimensional simplices $[v_0, v_1]$, $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$, $[v_4, v_5]$, $[v_5, v_0]$ is a hole



But what about higher dimensional holes (like the inside of a tetrahedron)?

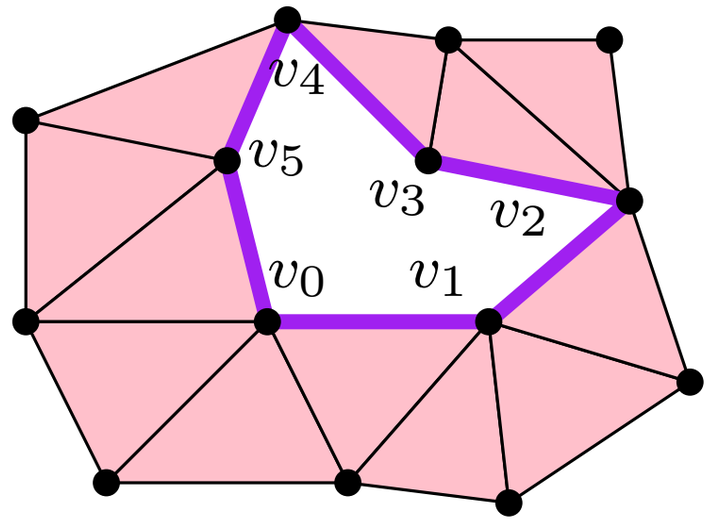


Simplicial Homology

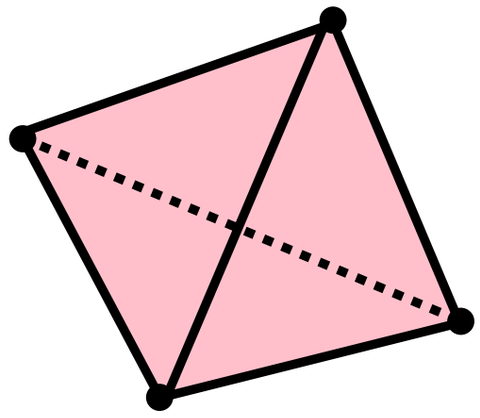
How to characterize a hole in a simplicial complex?

A hole (in 1D) is a path whose first and end points are the same, a loop.

The sequence of 1-dimensional simplices $[v_0, v_1]$, $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$, $[v_4, v_5]$, $[v_5, v_0]$ is a hole



But what about higher dimensional holes (like the inside of a tetrahedron)?



A hole in dimension d is a simplicial complex in which each $(d-1)$-simplex appears an even number of times.

Simplicial Homology

Def: Let K be a simplicial complex. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K.

Example : $C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0]$.

Simplicial Homology

Def: Let K be a simplicial complex. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K.

Example : $C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0]$.

- Geometrically a k-chain can be seen as a finite collection of k-simplices,
- The sum of two k-chains as the symmetric difference of the two corresponding collection (coefficient in $\mathbb{Z}_2$)

Symmetric difference of two sets A and B is the set $A \Delta B = (A \setminus B) \cup (B \setminus A)$.

Simplicial Homology

Def: Let K be a simplicial complex. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K.

Example : $C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0]$.

Def: The *boundary* of a k-simplex is the chain made of its $(k-1)$-simplices.

Simplicial Homology

Def: Let K be a simplicial complex. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K.

Example : $C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0]$.

Def: The *boundary* of a k-simplex is the chain made of its $(k-1)$-simplices.

$$\partial_k [v_1, \ldots, v_{k+1}] = \sum_{i=1}^{k+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{k+1}]$$

Simplicial Homology

Def: Let K be a simplicial complex. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K.

Example : $C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0]$.

Def: The *boundary* of a k-simplex is the chain made of its $(k-1)$-simplices.

$$\partial_k [v_1, \ldots, v_{k+1}] = \sum_{i=1}^{k+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{k+1}]$$

$$\partial_1 C = \partial_1 [v_0, v_1] + \partial_1 [v_1, v_2] + \partial_1 [v_2, v_3] + \partial_1 [v_3, v_4] + \partial_1 [v_4, v_5] + \partial_1 [v_5, v_0]$$

Simplicial Homology

Def: Let K be a simplicial complex. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K.

Example : $C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0]$.

Def: The *boundary* of a k-simplex is the chain made of its $(k-1)$-simplices.

$$\partial_k [v_1, \ldots, v_{k+1}] = \sum_{i=1}^{k+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{k+1}]$$

$$\partial_1 C = \partial_1 [v_0, v_1] + \partial_1 [v_1, v_2] + \partial_1 [v_2, v_3] + \partial_1 [v_3, v_4] + \partial_1 [v_4, v_5] + \partial_1 [v_5, v_0]$$

$$= [v_0] + [v_1] + [v_1] + [v_2] + [v_2] + [v_3] + [v_3] + [v_4] + [v_4] + [v_5] + [v_5] + [v_0]$$

Simplicial Homology

Def: Let K be a simplicial complex. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K.

Example : $C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0]$.

Def: The *boundary* of a k-simplex is the chain made of its $(k-1)$-simplices.

$$\partial_k [v_1, \ldots, v_{k+1}] = \sum_{i=1}^{k+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{k+1}]$$

$$\partial_1 C = \partial_1 [v_0, v_1] + \partial_1 [v_1, v_2] + \partial_1 [v_2, v_3] + \partial_1 [v_3, v_4] + \partial_1 [v_4, v_5] + \partial_1 [v_5, v_0]$$

$$= [v_0] + [v_1] + [v_1] + [v_2] + [v_2] + [v_3] + [v_3] + [v_4] + [v_4] + [v_5] + [v_5] + [v_0]$$

$$= [v_0] + [v_0] = 0.$$

Def: The kernel $Z_k(K) = \{c \in C_k(K) : \partial_k = 0\}$ of ∂_k is called the *space of k-cycles of K* and the image $B_k(K) = \{c \in C_k(K) : \exists c' \in C_{k+1}(K), \partial_{k+1}(c') = c\}$ of ∂_{k+1} is called the *space of k-boundaries* of K.

Simplicial Homology

Def: Let K be a simplicial complex. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K.

Example : $C = [v_0, v_1] + [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_5] + [v_5, v_0]$.

Def: The *boundary* of a k-simplex is the chain made of its $(k-1)$-simplices.

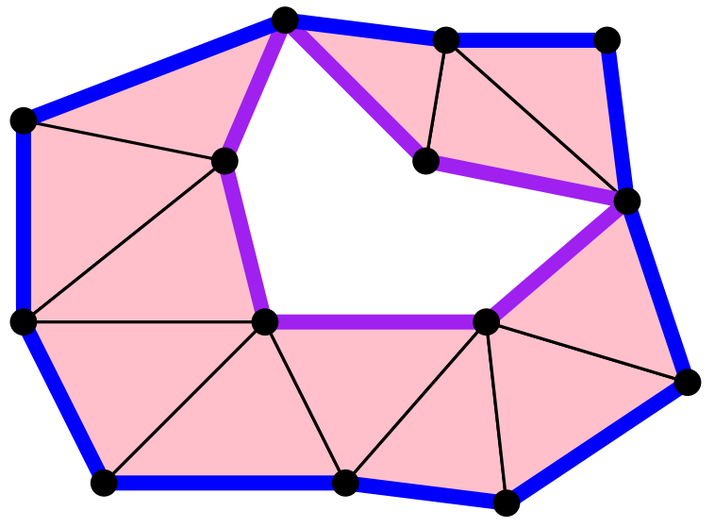
$$\partial_k [v_1, \ldots, v_{k+1}] = \sum_{i=1}^{k+1} [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{k+1}]$$

$$\partial_1 C = \partial_1 [v_0, v_1] + \partial_1 [v_1, v_2] + \partial_1 [v_2, v_3] + \partial_1 [v_3, v_4] + \partial_1 [v_4, v_5] + \partial_1 [v_5, v_0]$$

$$= [v_0] + [v_1] + [v_1] + [v_2] + [v_2] + [v_3] + [v_3] + [v_4] + [v_4] + [v_5] + [v_5] + [v_0]$$

$$= [v_0] + [v_0] = 0.$$

Def: The kernel $Z_k(K) = \{c \in C_k(K) : \partial_k = 0\}$ of ∂_k is called the *space of k-cycles of K* and the image $B_k(K) = \{c \in C_k(K) : \exists c' \in C_{k+1}(K), \partial_{k+1}(c') = c\}$ of ∂_{k+1} is called the *space of k-boundaries* of K.



Two cycles

Simplicial Homology

Lemma: Any k-boundary is a k-cycle: $\partial_k \circ \partial_{k+1} = 0$.

Simplicial Homology

Lemma: Any k-boundary is a k-cycle: $\partial_k \circ \partial_{k+1} = 0$.

Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$' (boundary):

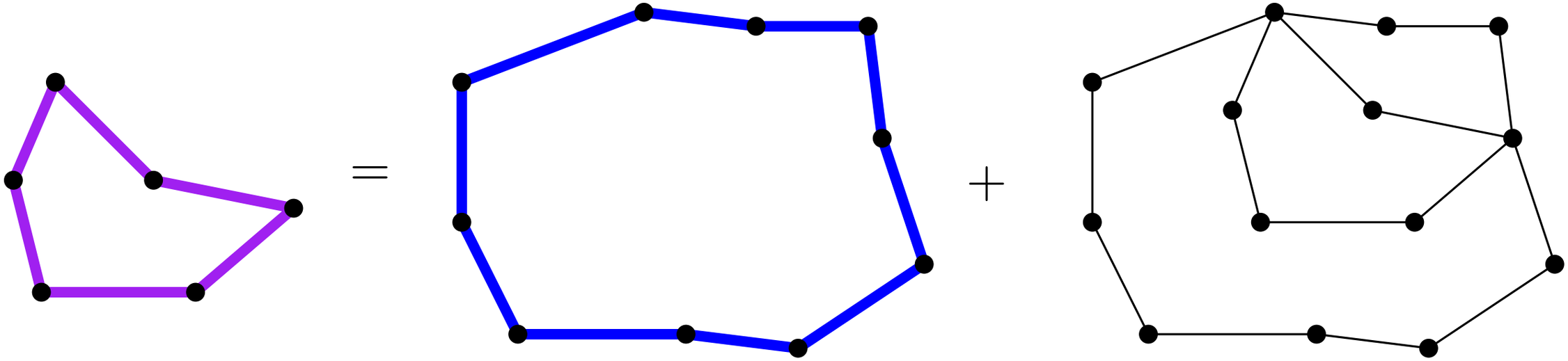
$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$

Simplicial Homology

Lemma: Any k-boundary is a k-cycle: $\partial_k \circ \partial_{k+1} = 0$.

Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$' (boundary):

$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$

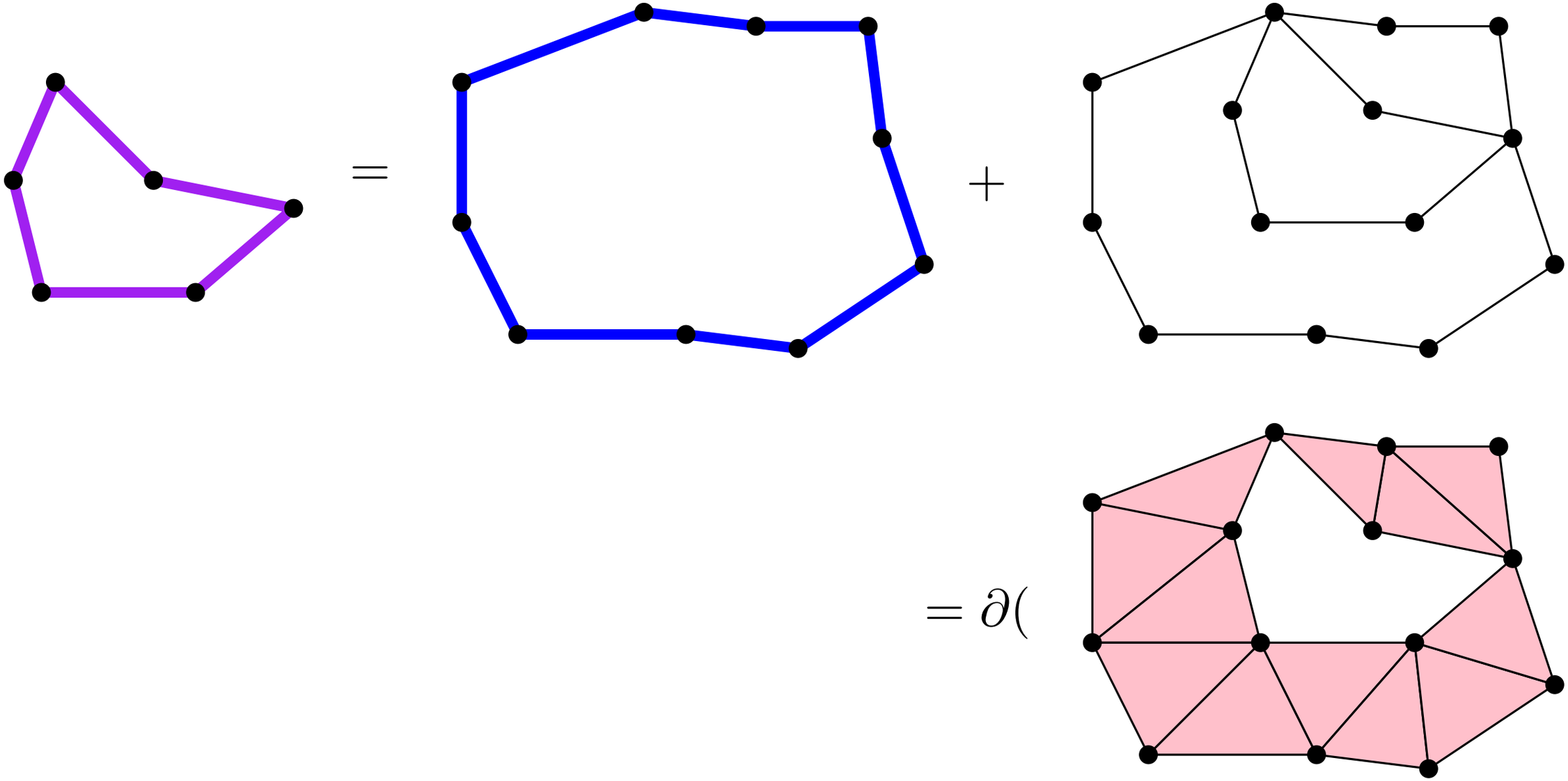


Simplicial Homology

Lemma: Any k-boundary is a k-cycle: $\partial_k \circ \partial_{k+1} = 0$.

Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$' (boundary):

$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$

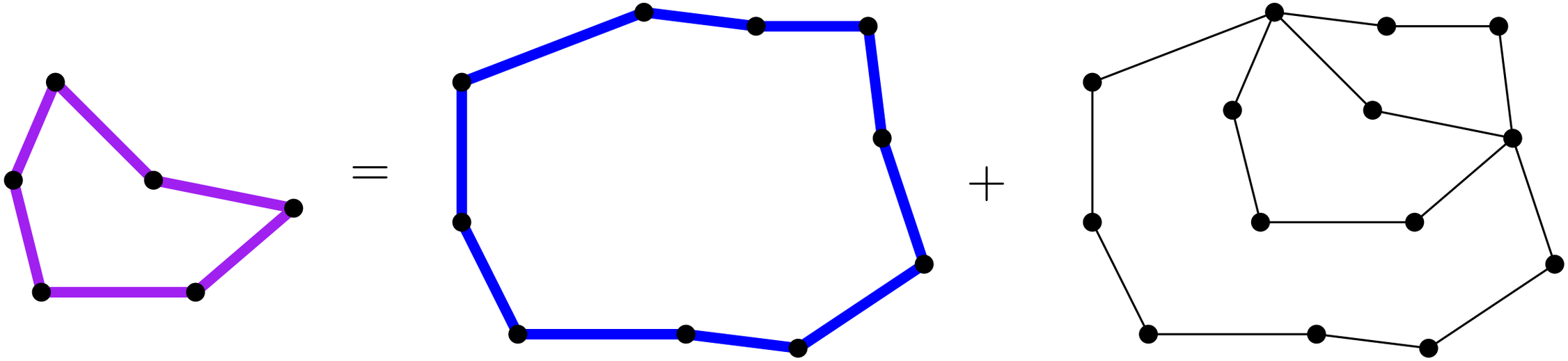


Simplicial Homology

Lemma: Any k-boundary is a k-cycle: $\partial_k \circ \partial_{k+1} = 0$.

Def: Two cycles are the same (homologous) if 'their difference is in $\mathrm{im}(\partial)$' (boundary):

$$C \sim C' \iff C + C' \in \mathrm{im}(\partial)$$



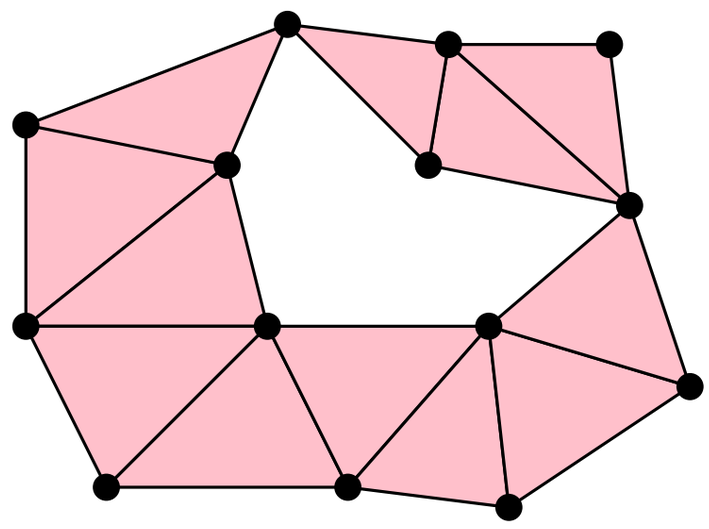
group of k-cycles

$$H_k = Z_k / B_k$$

group of 'cycles minus boundaries'

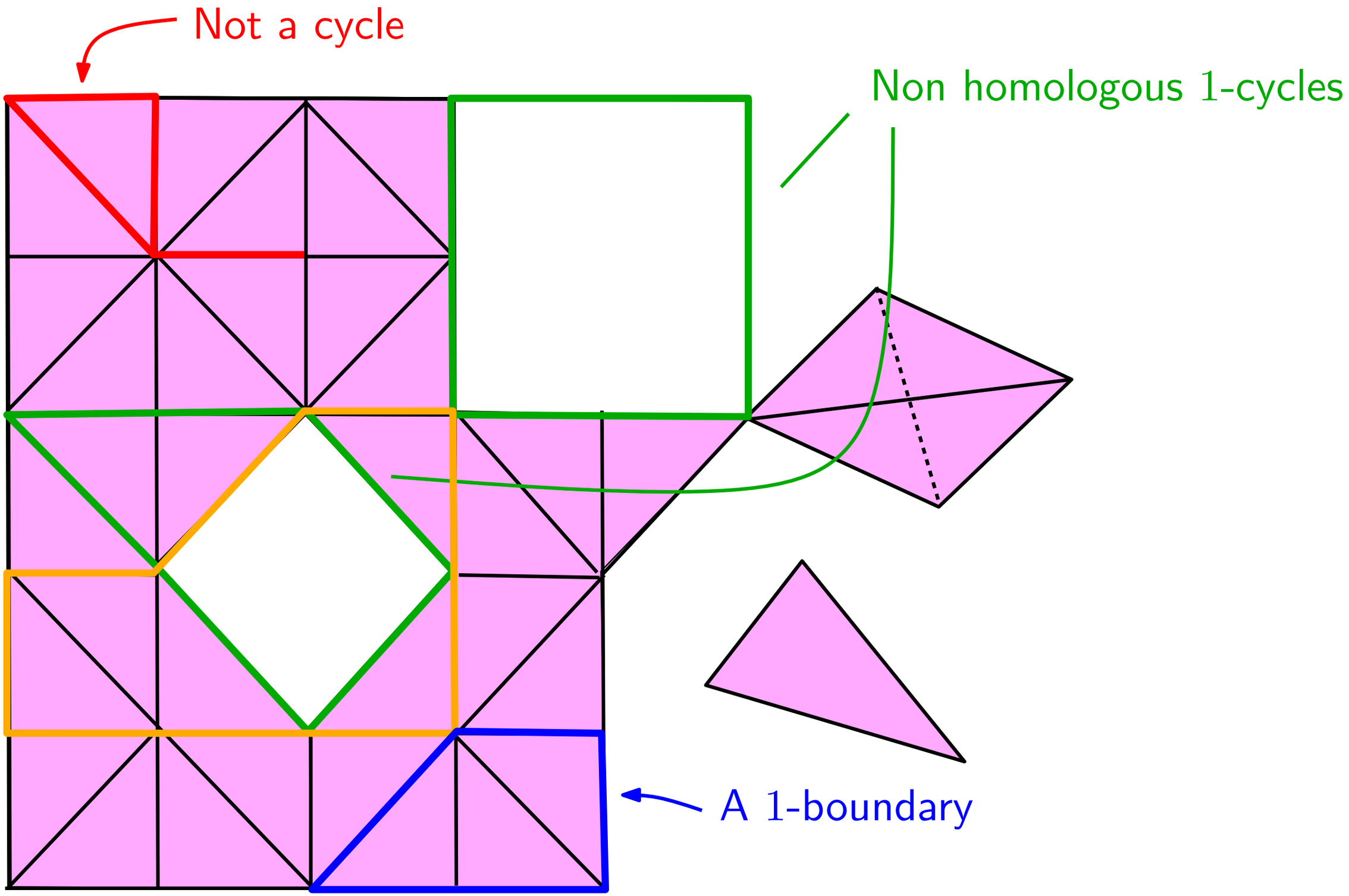
$\mathrm{im}(\partial_{k+1})$

$$= \partial($$



$$)$$

Simplicial Homology



Simplicial Homology

H_k is a group (vector space) in which each element is an equivalence class of cycles associated to the same hole.

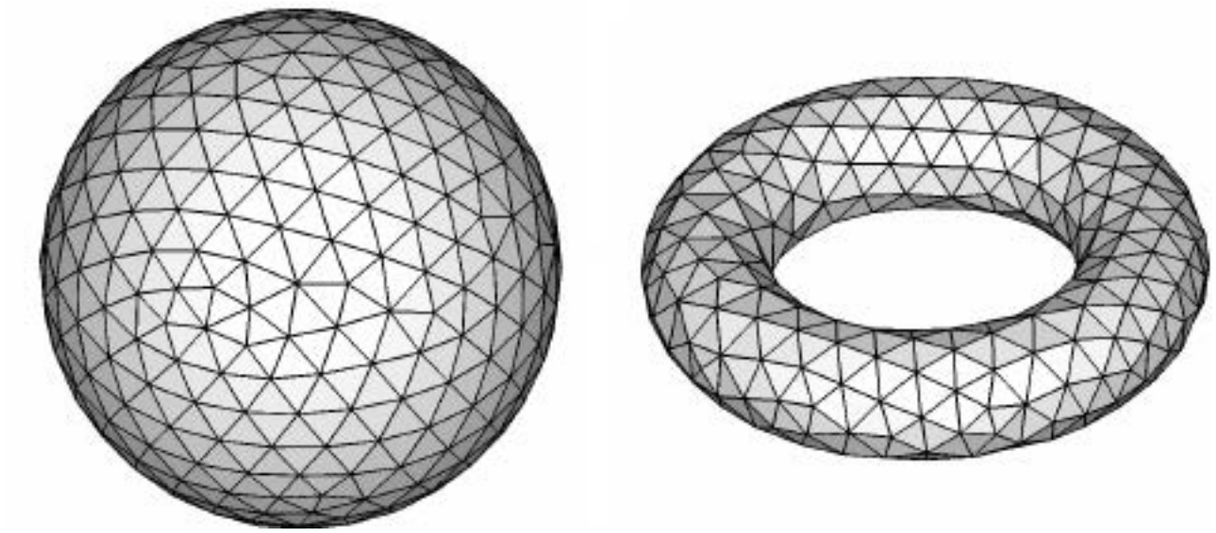
Def: The dimension of H_k is called the *Betti number* β_k.

Simplicial Homology

H_k is a group (vector space) in which each element is an equivalence class of cycles associated to the same hole.

Def: The dimension of H_k is called the *Betti number* β_k.

Q: What are the Betti numbers of:

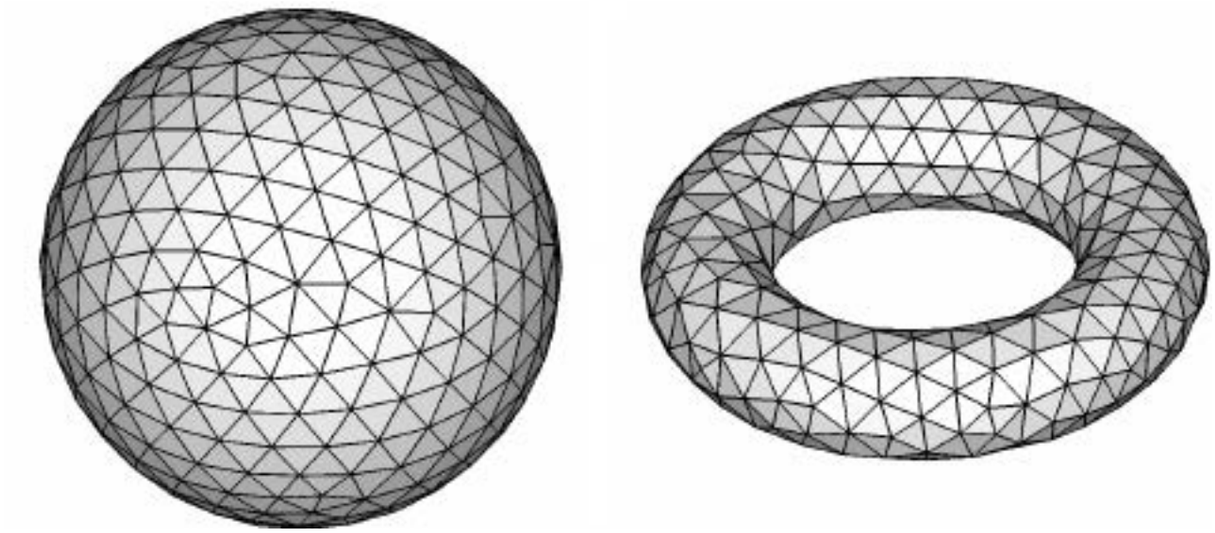


Simplicial Homology

H_k is a group (vector space) in which each element is an equivalence class of cycles associated to the same hole.

Def: The dimension of H_k is called the *Betti number* β_k.

Q: What are the Betti numbers of:



- Homology groups and Betti numbers quantify topological features in a space.
- It does not capture all topological aspects of a space but

$$H_k(X) \not\sim H_k(Y) \Longrightarrow X \not\sim Y$$

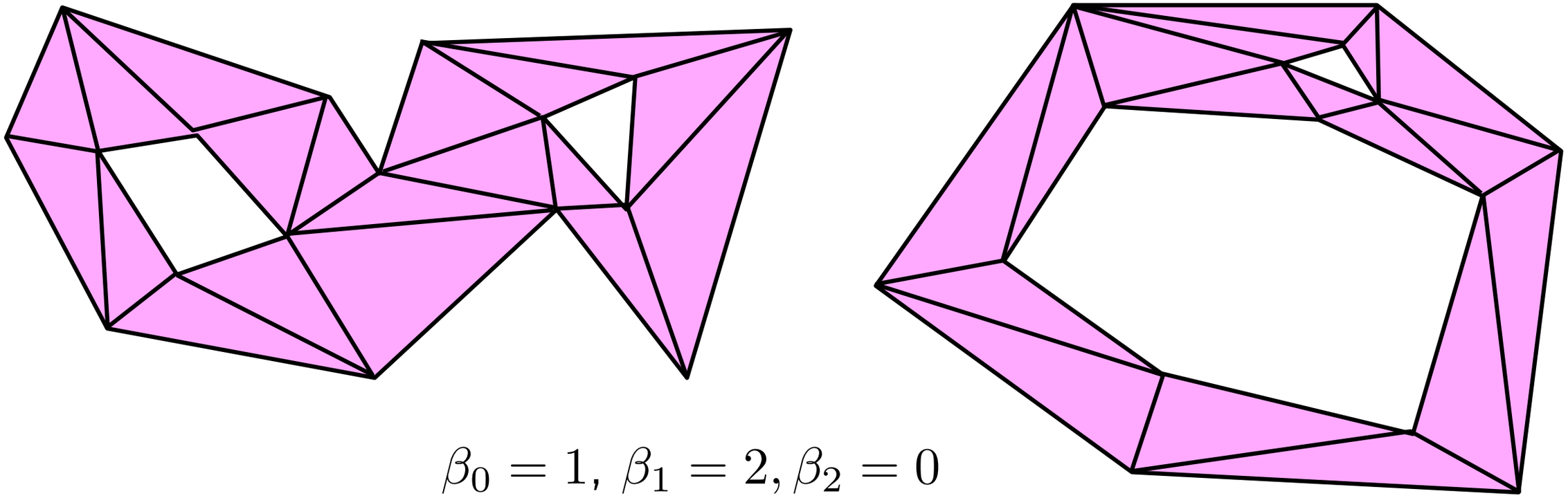
- Homology groups are computationally tractable

Simplicial Homology

By understanding how the topology of the 'filtration' of a simplicial complex (see further) evolves each time we add a simplex, we can propose a simple algorithm for Betti numbers computation:

Input: A filtration $\mathbb{F}$ of a d-dimensional simplicial complex K containing m simplices.
$\beta_0 \leftarrow 0;\ \beta_1 \leftarrow 0; \cdots\ \beta_d \leftarrow 0$
for $i = 1$ to m **do**
  $k = \dim \sigma^i - 1$
  if σ^i is contained in a $(k+1)$-cycle in K^i **then**
    $\beta_{k+1} \leftarrow \beta_{k+1} + 1$
  else
    $\beta_k \leftarrow \beta_k - 1$
Output: The Betti numbers $\beta_0, \beta_1, \cdots, \beta_d$ of K.

Topological invariance and singular homology

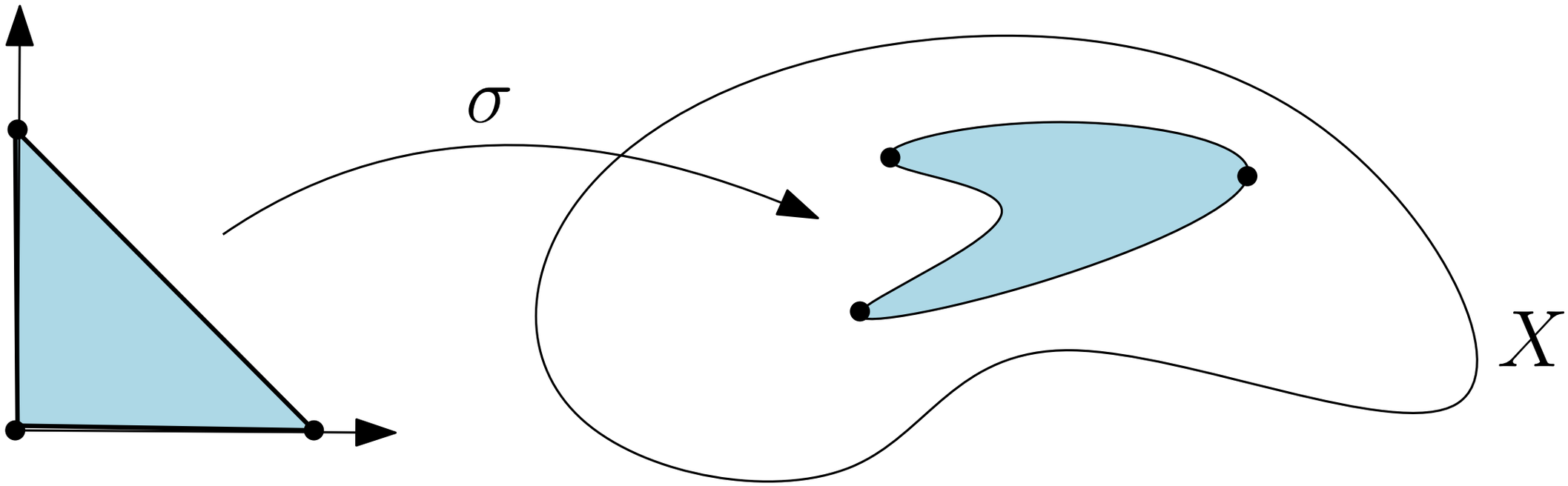


$\beta_0 = 1,\ \beta_1 = 2, \beta_2 = 0$

Theorem: If K and K' are two simplicial complexes with homeomorphic supports then their homology groups are isomorphic and their Betti numbers are equal.

- This is a classical result in algebraic topology but the proof is not obvious.
- Rely on the notion of singular homology $\rightarrow$ defined for any topological space.

Topological invariance and singular homology



Let Δ_k be the standard simplex in $\mathbb{R}^k$. A singular k-simplex in a topological space X is a continuous map $\sigma : \Delta_k \to X$.

The same construction as for simplicial homology can be done with singular complexes $\to$ Singular homology

Important properties:

- Singular homology is defined for any topological space X.
- If X is homotopy equivalent to the support of a simplicial complex, then the singular and simplicial homology coincide

4 - Homology inference

Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

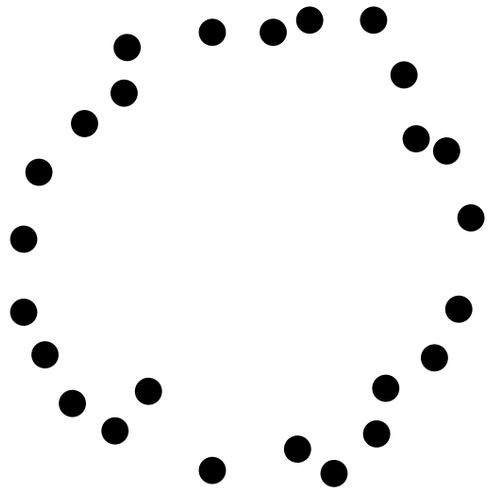
- Define 'covers' from the data (for instance by grouping data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of these 'covers'.

Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

- Define 'covers' from the data (for instance by grouping data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of these 'covers'.

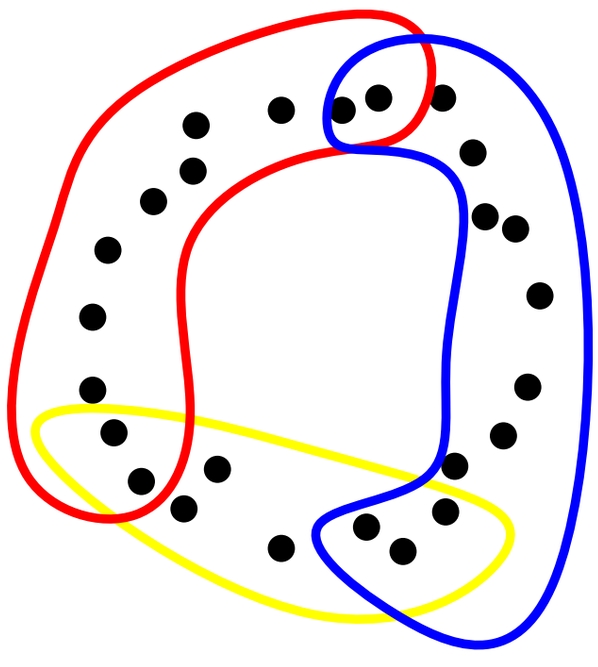


Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

- Define 'covers' from the data (for instance by grouping data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of these 'covers'.

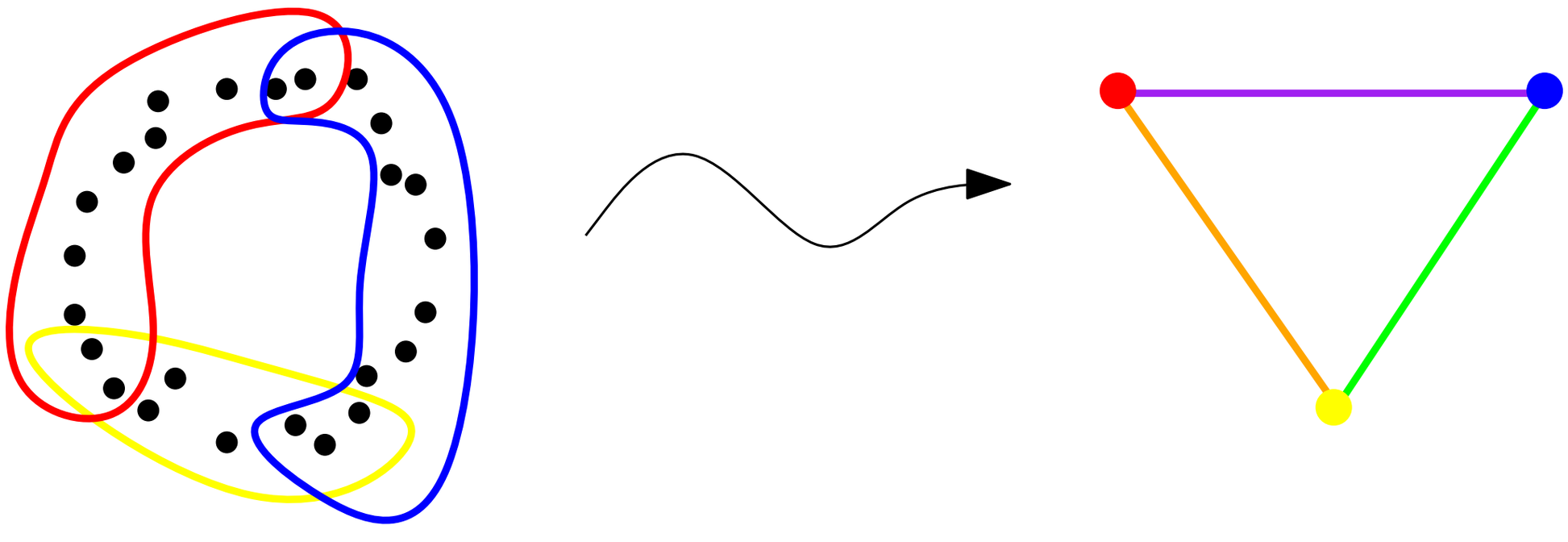


Topological exploratory data analysis

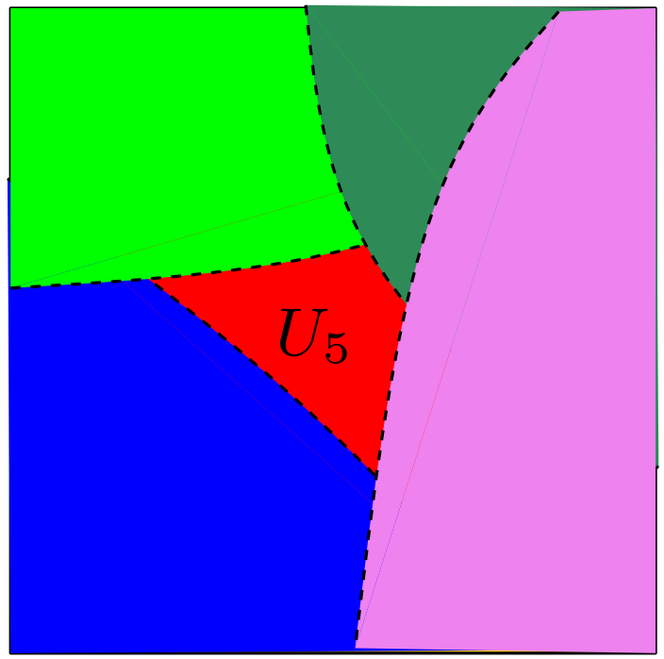
Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

- Define 'covers' from the data (for instance by grouping data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of these 'covers'.

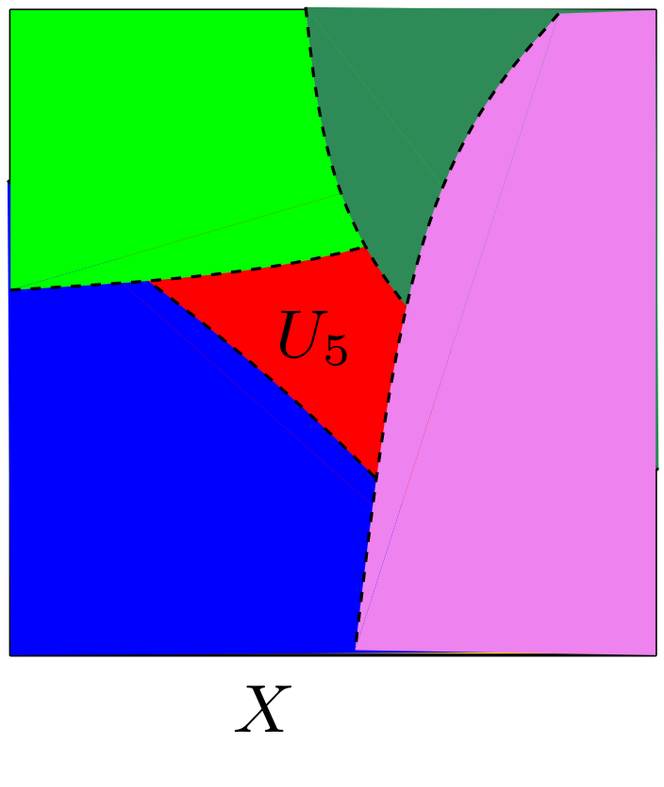


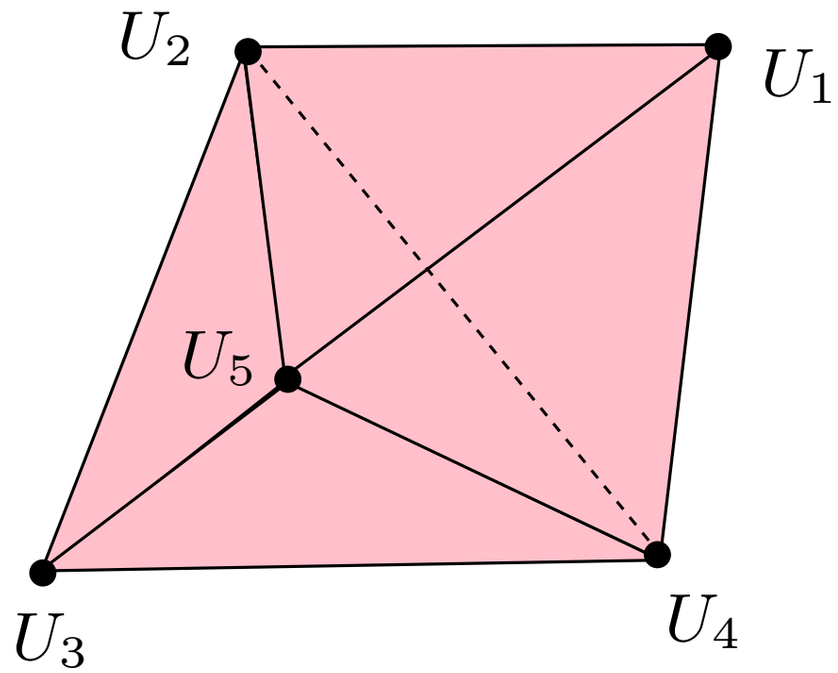
Topological exploratory data analysis



Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

Topological exploratory data analysis



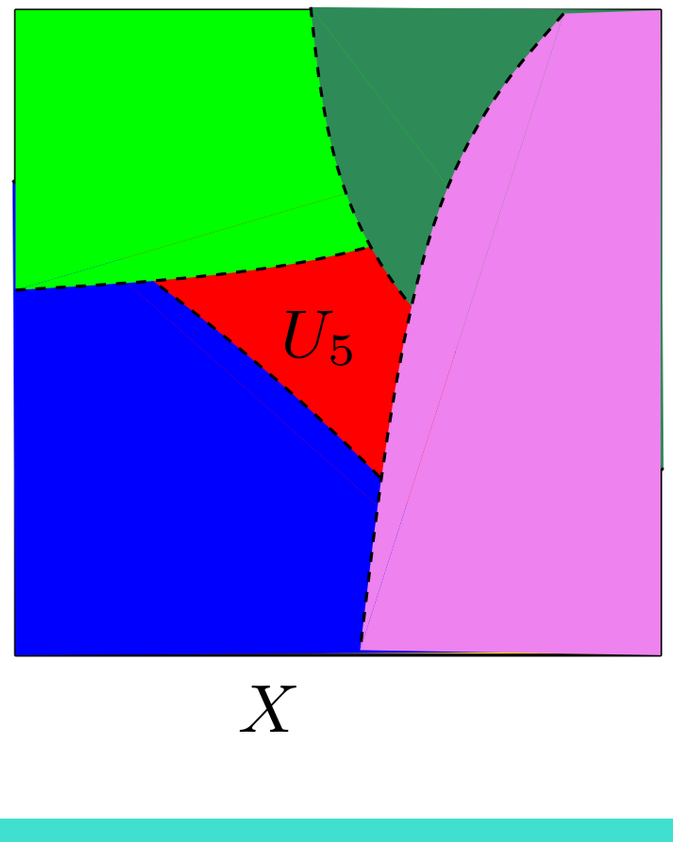


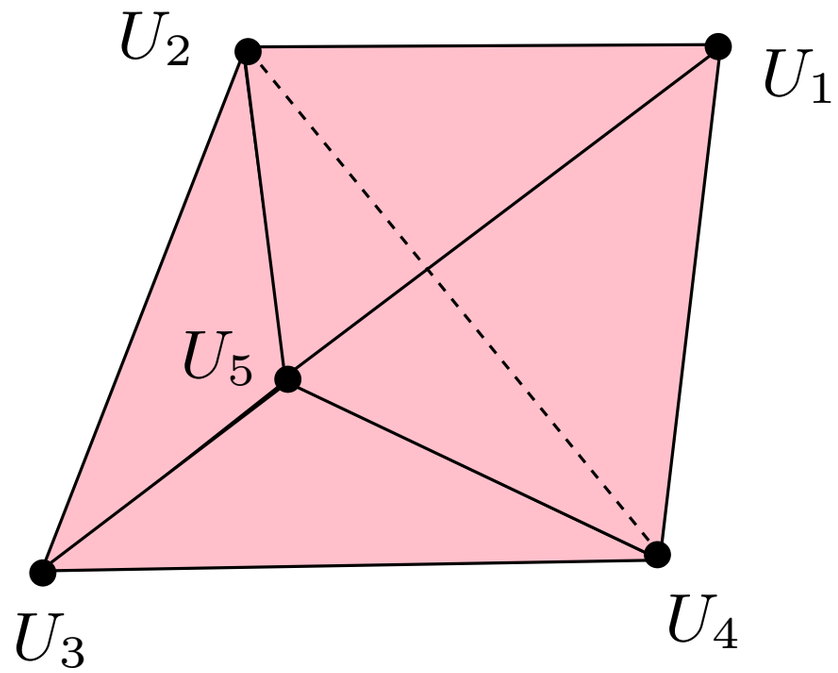
Def: An open cover of a topological space X is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where I is a set, such that $X \subseteq \cup_{i \in I} U_i$.

Def: Given a cover of a topological space X, $\mathcal{U} = (U_i)_{i \in I}$, its nerve is the abstract simplicial complex $C(\mathcal{U})$ whose vertex set is $\mathcal{U}$ and s.t.

$$\sigma = [U_{i_0}, U_{i_1}, \ldots, U_{i_k}] \in C(\mathcal{U}) \quad \text{if and only if} \quad \cap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

Topological exploratory data analysis





The Nerve Theorem: Let $\mathcal{U} = (U_i)_{i \in I}$ be a finite open cover of a subset X of $\mathbb{R}^d$ such that any intersection of the U_i's is either empty or contractible. Then X and $C(\mathcal{U})$ are homotopy equivalent. In particular, their homology groups are isomorphic.

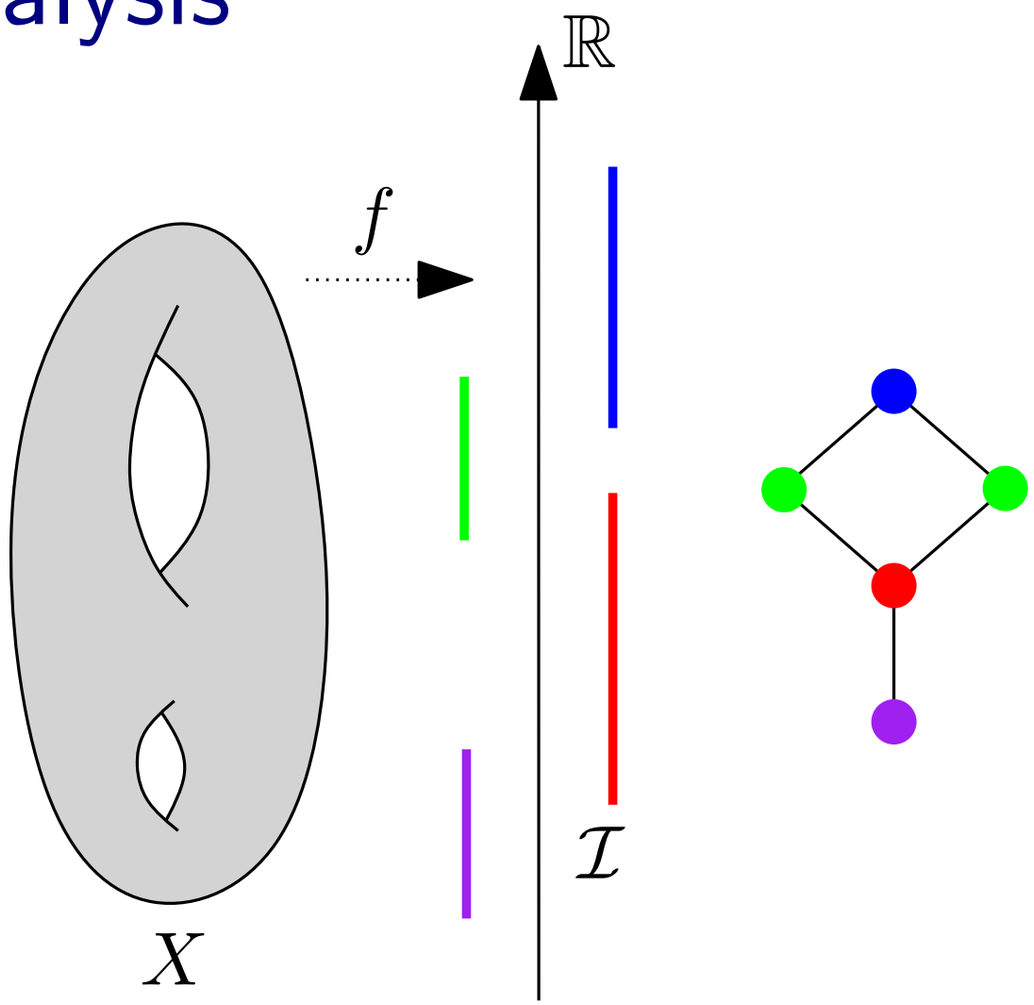
Simpler version of the Nerve Theorem: replace contractible by convex (balls for instance).

Topological exploratory data analysis

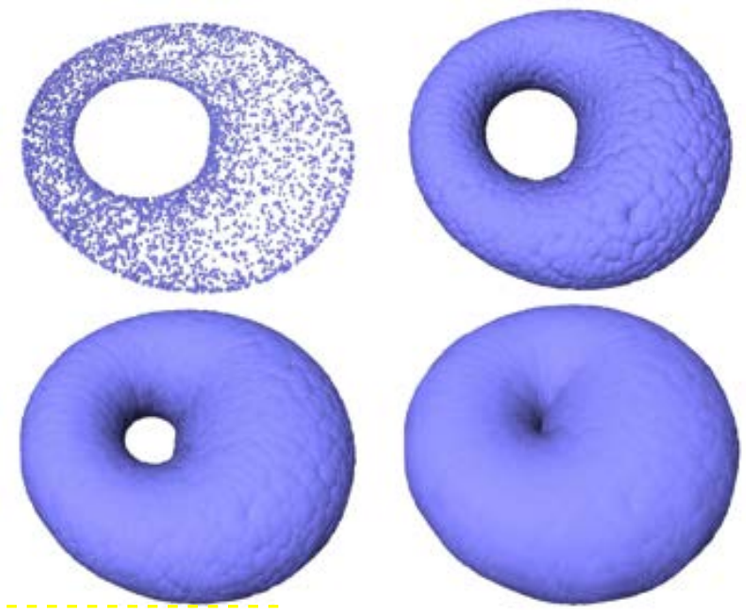
Q: How to build meaningful covers?

Two directions:

1. Using a function (lens) defined on the data:
$\rightarrow$ the Mapper algorithm
$\rightarrow$ exploratory data analysis



2. Covering data by balls:
$\rightarrow$ distance functions frameworks, persistence-based signatures,...
$\rightarrow$ geometric inference, provide a framework to establish various theoretical results in TDA.

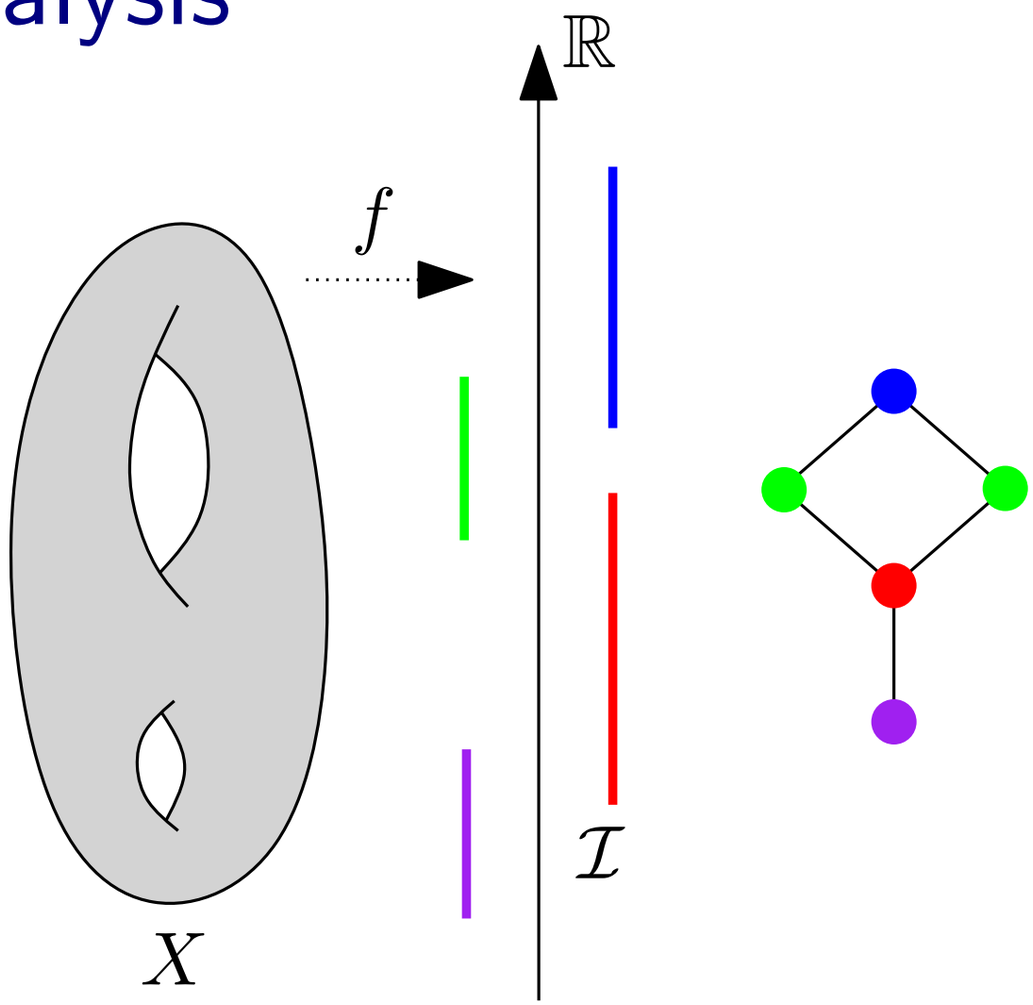


Topological exploratory data analysis

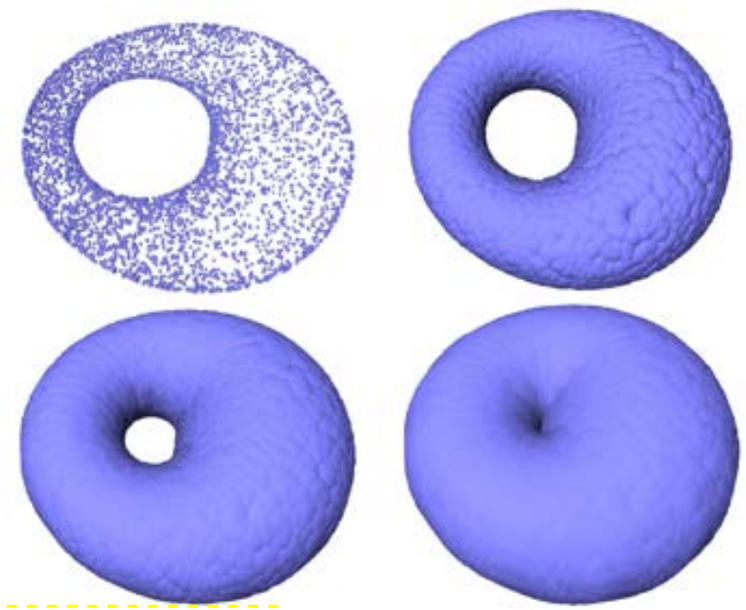
Q: How to build meaningful covers?

Two directions:

1. Using a function (lens) defined on the data:
$\rightarrow$ the Mapper algorithm (see further)
$\rightarrow$ exploratory data analysis



2. Covering data by balls:
$\rightarrow$ distance functions frameworks, persistence-based signatures,...
$\rightarrow$ geometric inference, provide a framework to establish various theoretical results in TDA.



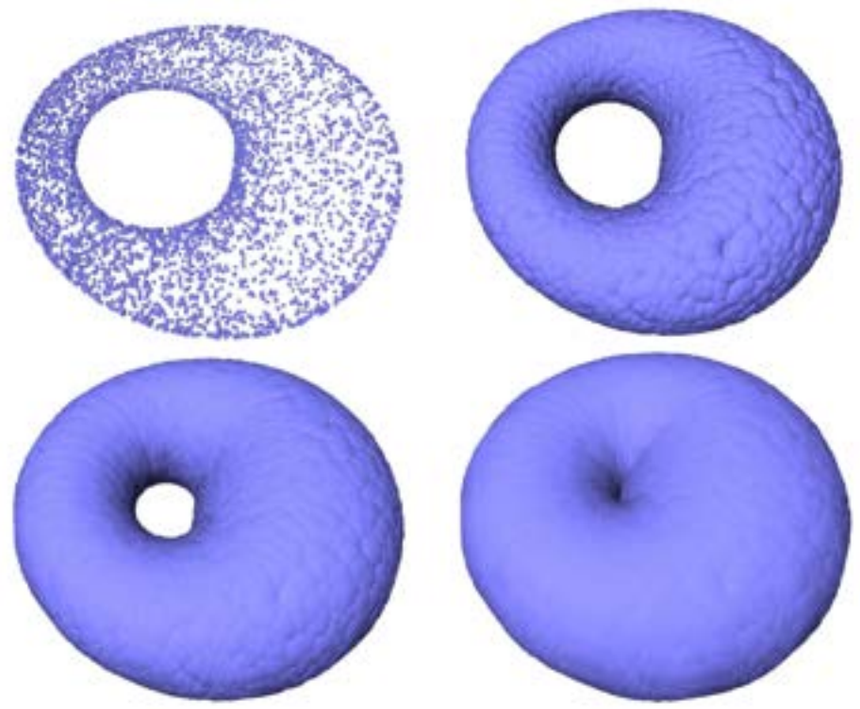
Union of balls and distance functions

Data set : a point cloud P embedded in $\mathbb{R}^d$, sampled around a compact set M.

General idea:

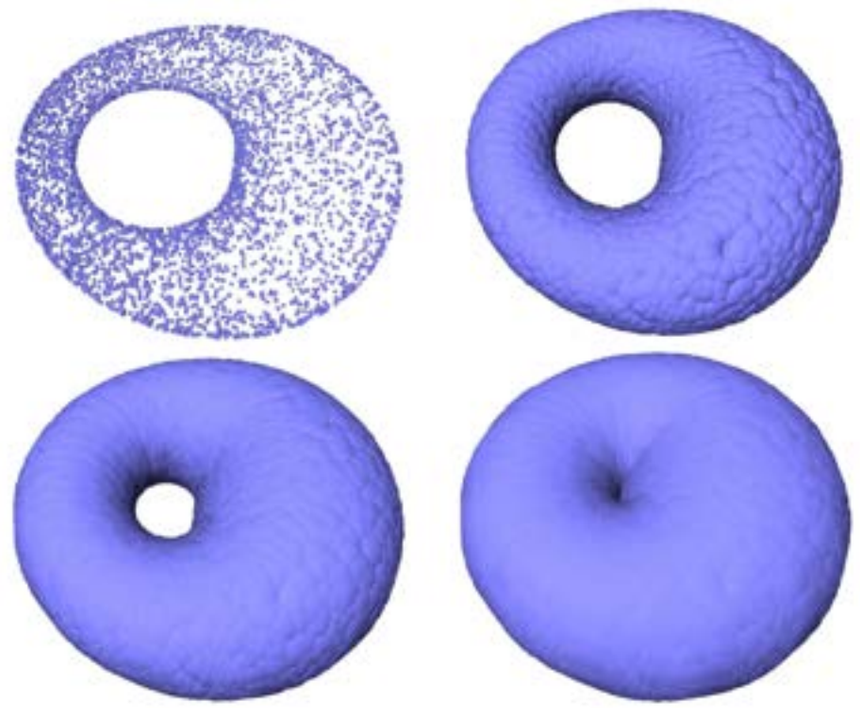
1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about M from (the nerve of) the union of balls centered on P.



Union of balls and distance functions

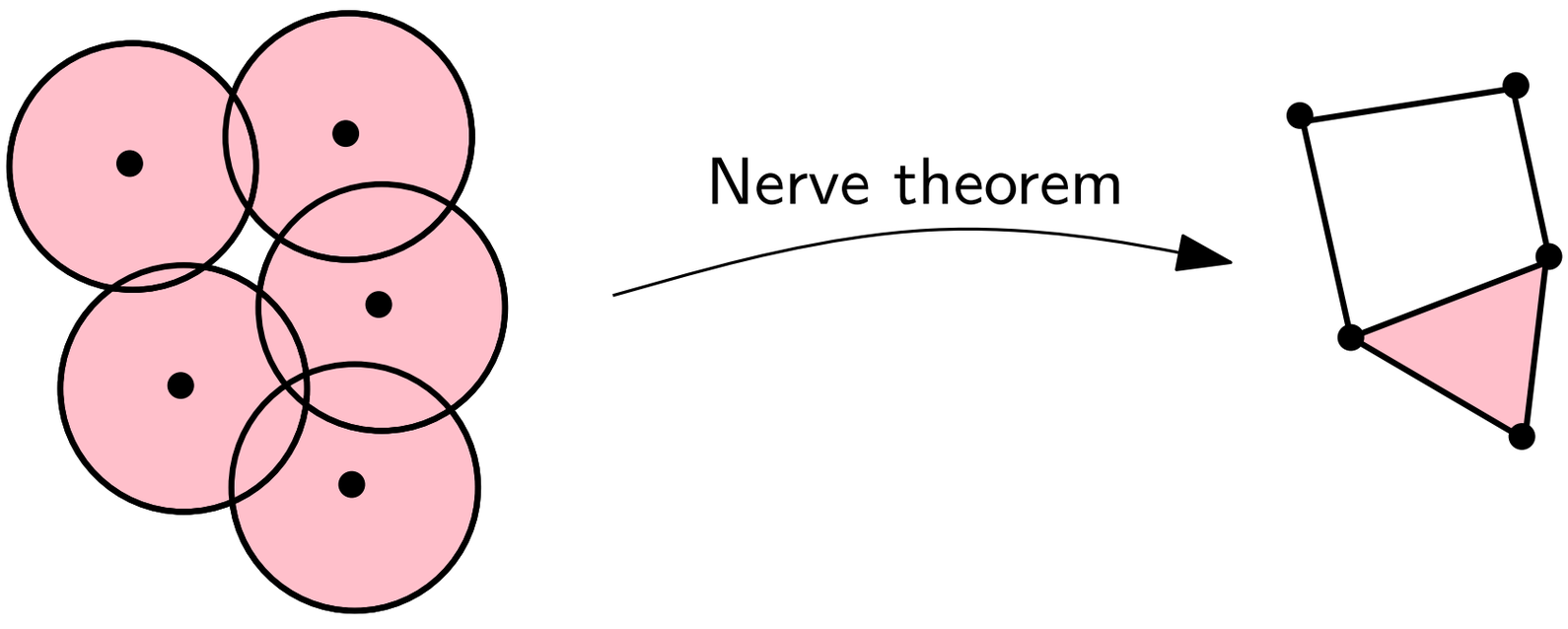
Data set : a point cloud P embedded in $\mathbb{R}^d$, sampled around a compact set M.



General idea:

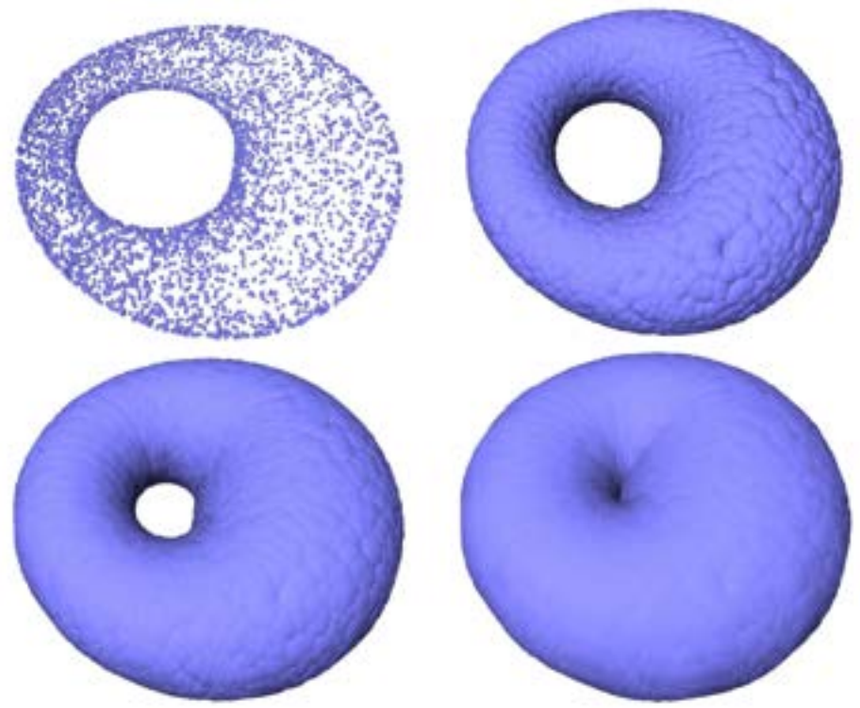
1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about M from (the nerve of) the union of balls centered on P.



Union of balls and distance functions

Data set : a point cloud P embedded in $\mathbb{R}^d$, sampled around a compact set M.



General idea:

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about M from (the nerve of) the union of balls centered on P.

Sublevel set of the distance function $d_P : \mathbb{R}^d \to \mathbb{R}_+$ is defined by

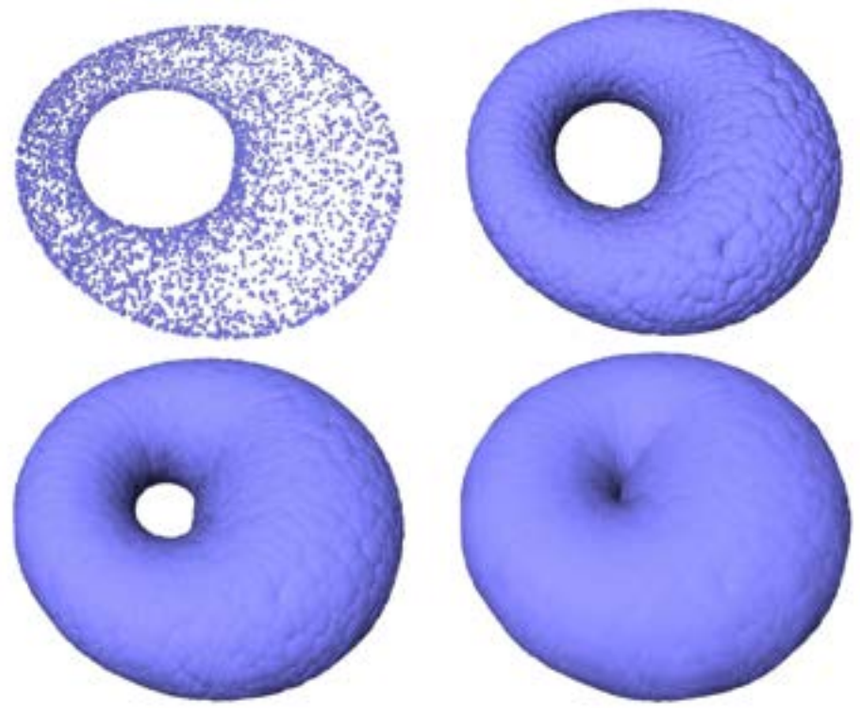
$$d_P(x) = \inf_{p \in P} \|x - p\|$$

Compare the topology of the offsets

$$M^r = d_M^{-1}([0, r]) \text{ and } P^r = d_P^{-1}([0, r])$$

Union of balls and distance functions

Data set : a point cloud P embedded in $\mathbb{R}^d$, sampled around a compact set M.



General idea:

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about M from (the nerve of) the union of balls centered on P.

Sublevel set of the distance function $d_P : \mathbb{R}^d \to \mathbb{R}_+$ is defined by

$$d_P(x) = \inf_{p \in P} \|x - p\|$$

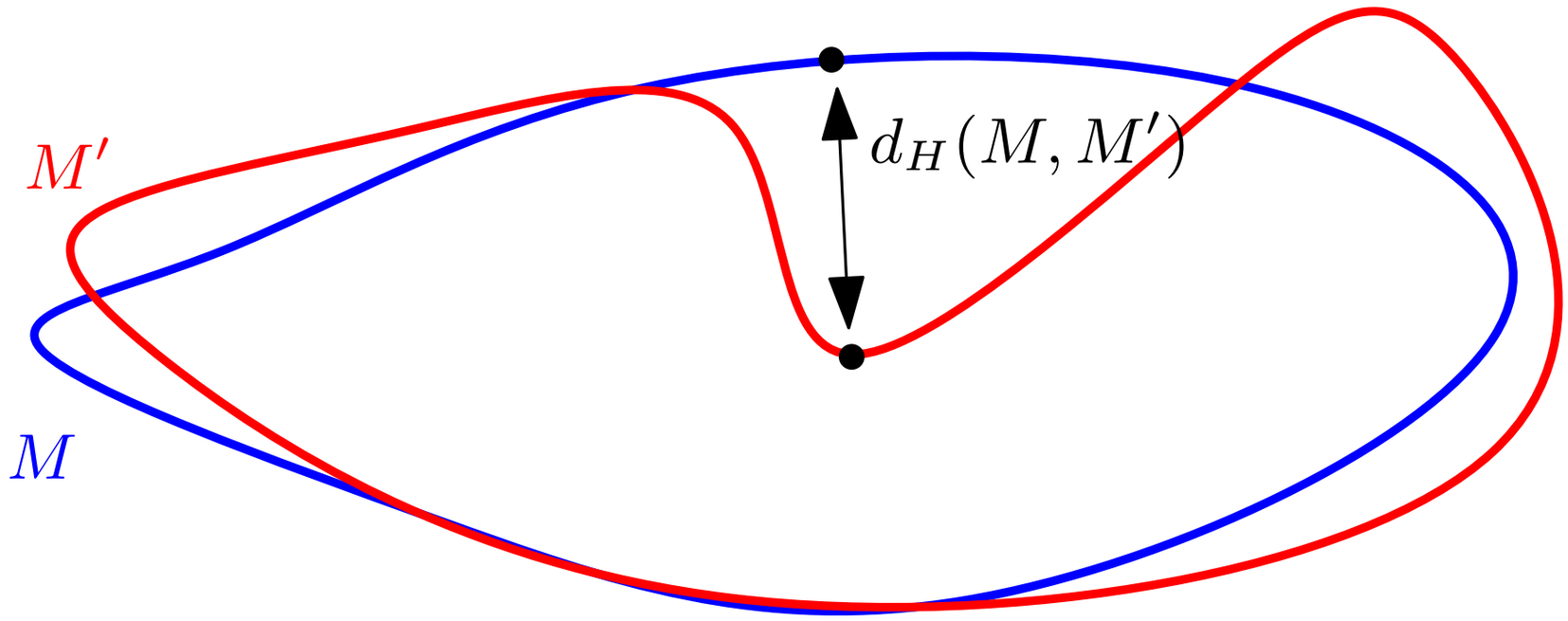
Regularity conditions?

Sampling conditions?

Compare the topology of the offsets

$$M^r = d_M^{-1}([0, r]) \text{ and } P^r = d_P^{-1}([0, r])$$

The Hausdorff distance



The distance function to a compact $M \subset \mathbb{R}^d$, $d_M : \mathbb{R}^d \to \mathbb{R}_+$ is defined by:

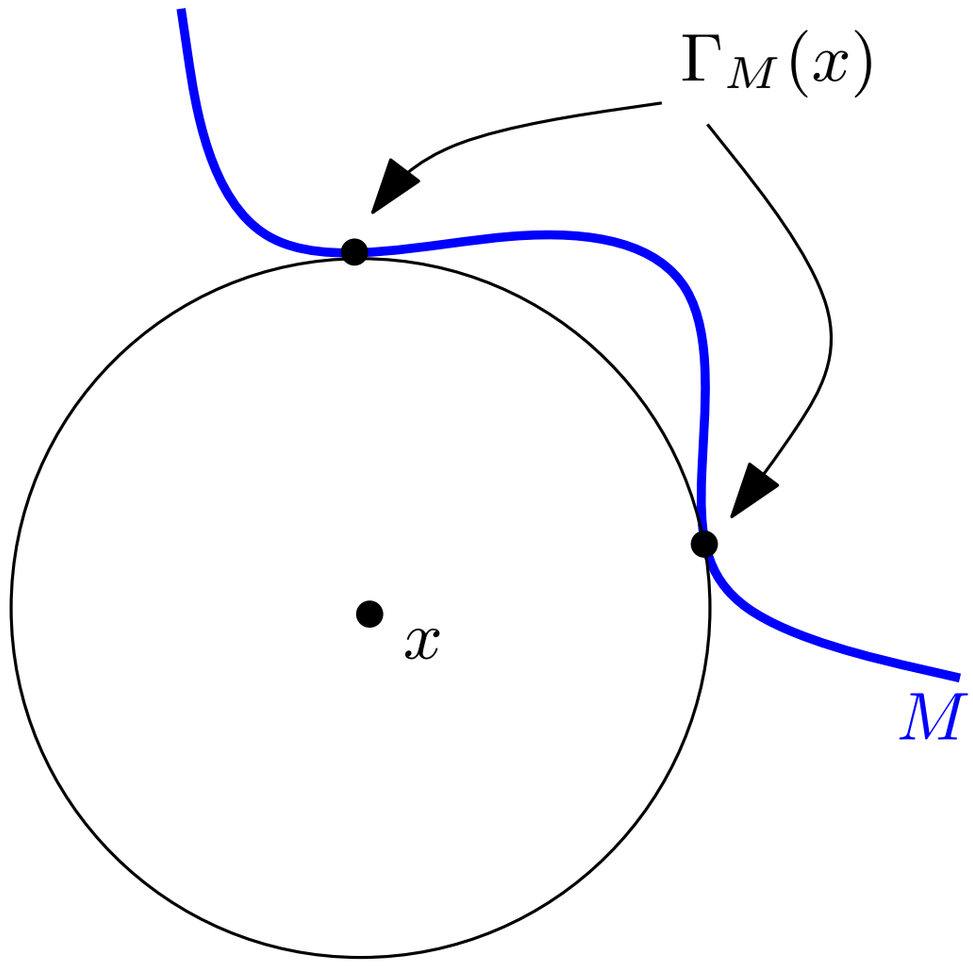
$$d_M(x) = \inf_{p \in M} \|x - p\|$$

The Hausdorff distance between two compact sets $M, M' \subset \mathbb{R}^d$ is:

$$d_H(M, M') = \sup_{x \in \mathbb{R}^d} |d_M(x) - d_{M'}(x)|$$

Medial axis

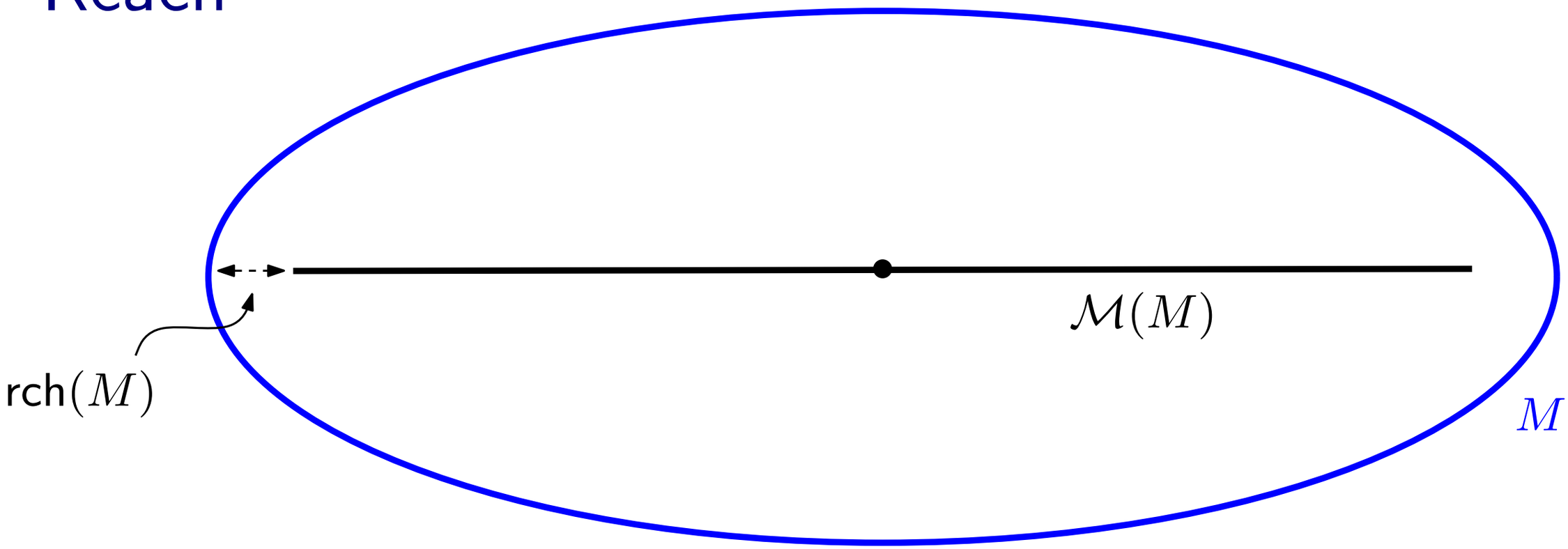
$$\Gamma_M(x) = \{y \in M : d_M(x) = \|x - y\|\}$$



Def: The medial axis of M:

$$\mathcal{M}(M) = \{x \in \mathbb{R}^d : |\Gamma_M(x)| \geq 2\}$$

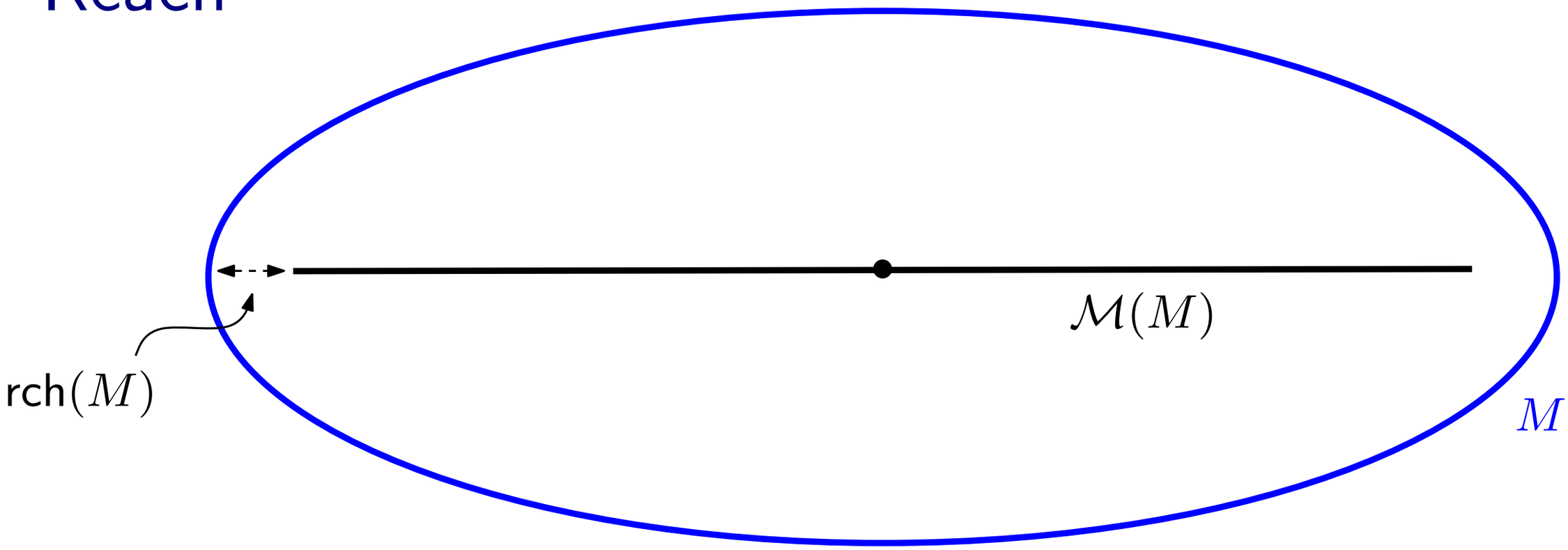
Reach



Def: The reach of M, $\mathrm{rch}(M)$ is the smallest distance from $\mathcal{M}(M)$ to M:

$$\mathrm{rch}(M) = \inf_{y \in \mathcal{M}(M)} d_M(y)$$

Reach

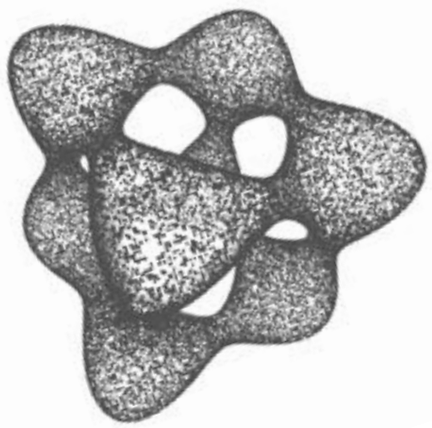


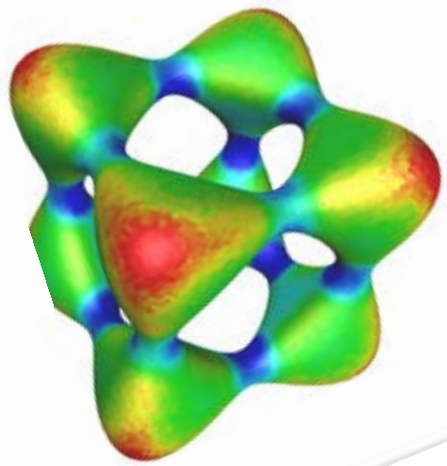
Def: The reach of M, $\mathrm{rch}(M)$ is the smallest distance from $\mathcal{M}(M)$ to M:

$$\mathrm{rch}(M) = \inf_{y \in \mathcal{M}(M)} d_M(y)$$

- The projection on M is well defined outside $\mathcal{M}(M)$
- $\mathrm{rch}(M) =$ largest $\rho > 0$ s.t. the projection is well defined on the off-set M^ρ.
- The reach controls both the local curvature and the auto-intersection.

Geometric inference

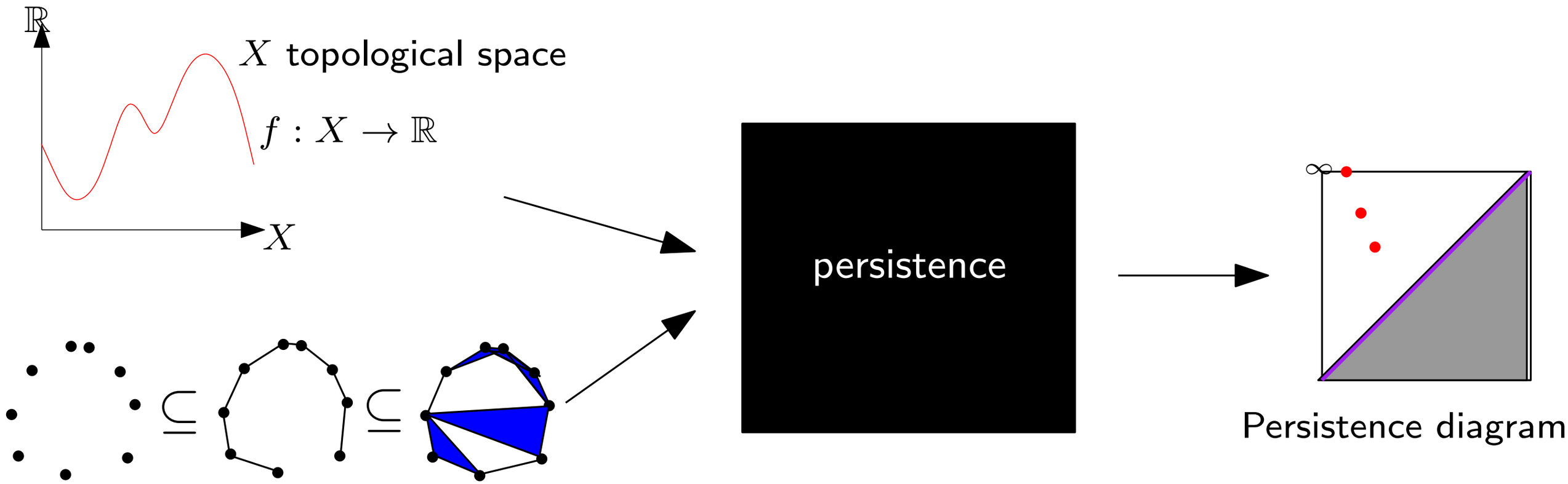




Homology inference [Niyogi et al., 2008 and 2011, Balakrishnan et al., 2012]: The Betti number (actually the homotopy type) of Riemannian manifolds with positive reach can be recovered with high probability from offsets of a sample on (or close to) the manifold.

5 - Persistent homology

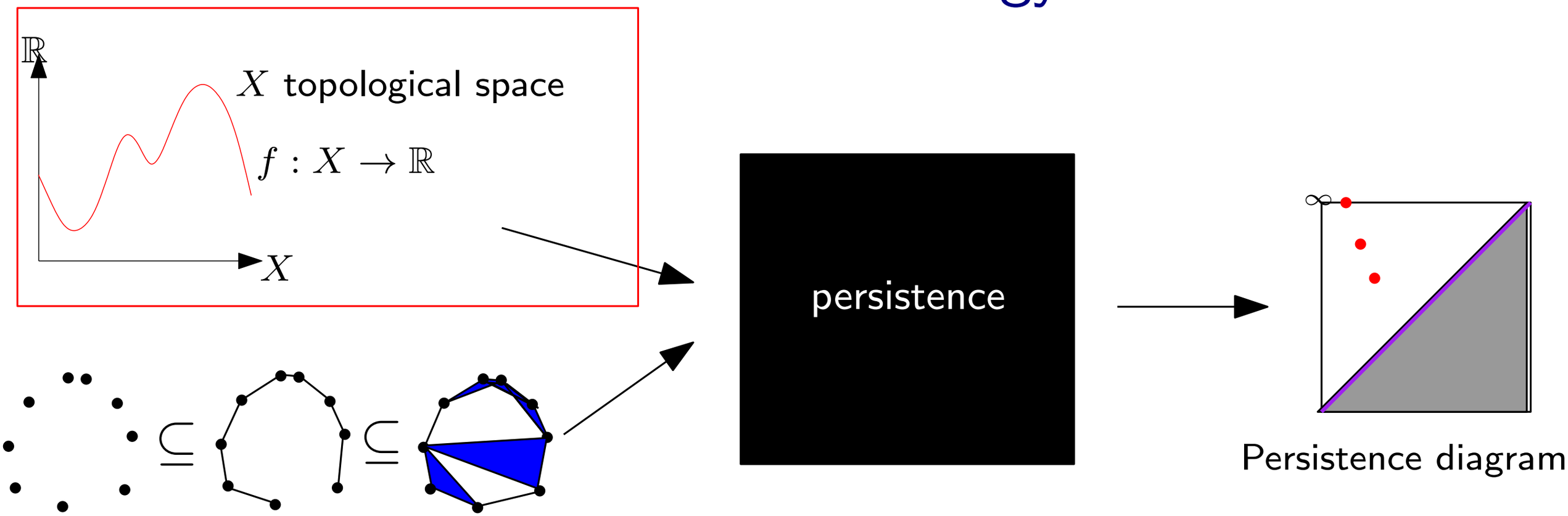
Persistent homology



Filtrations

- A general mathematical framework to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Formalized by H. Edelsbrunner (2002) et al and G. Carlsson et al (2005) - wide development during the last decade. Ideas tracing back to M. Morse (1940)!
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed (e.g. Gudhi library!).
- Stability properties

Persistent homology

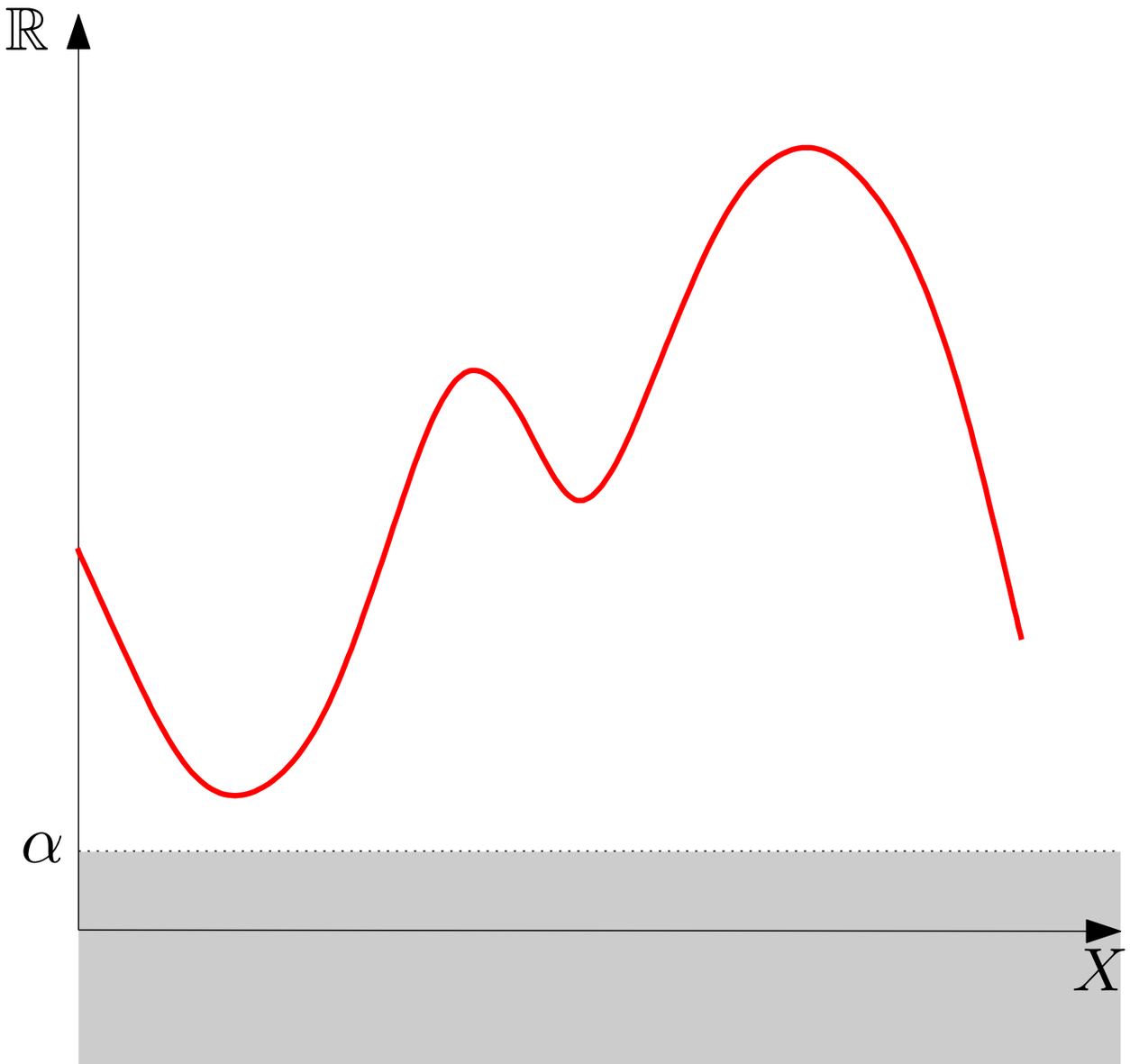


Filtrations

- A general mathematical framework to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Formalized by H. Edelsbrunner (2002) et al and G. Carlsson et al (2005) - wide development during the last decade. Ideas tracing back to M. Morse (1940)!
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed (e.g. Gudhi library!).
- Stability properties

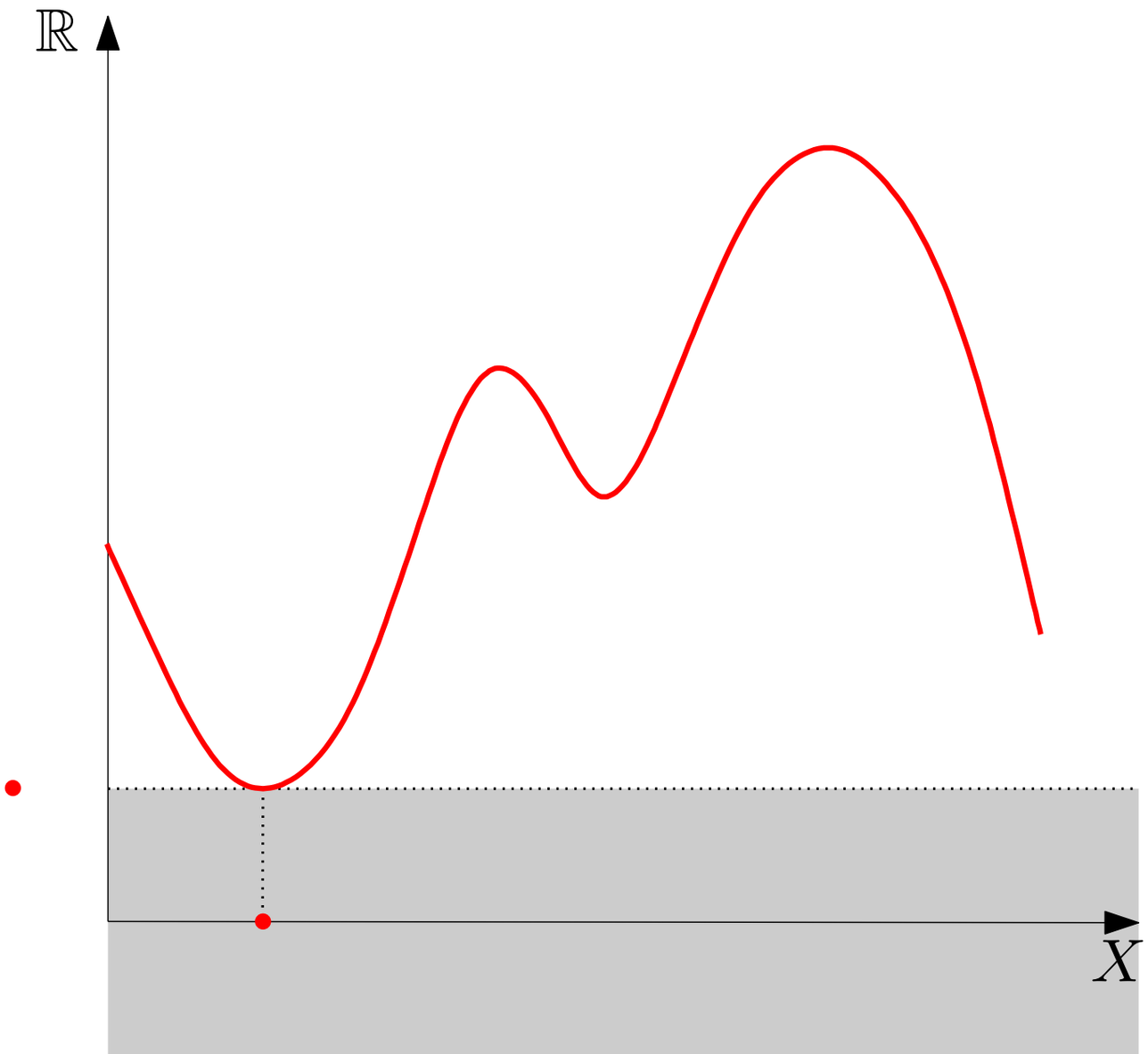
Persistent homology for functions

- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.



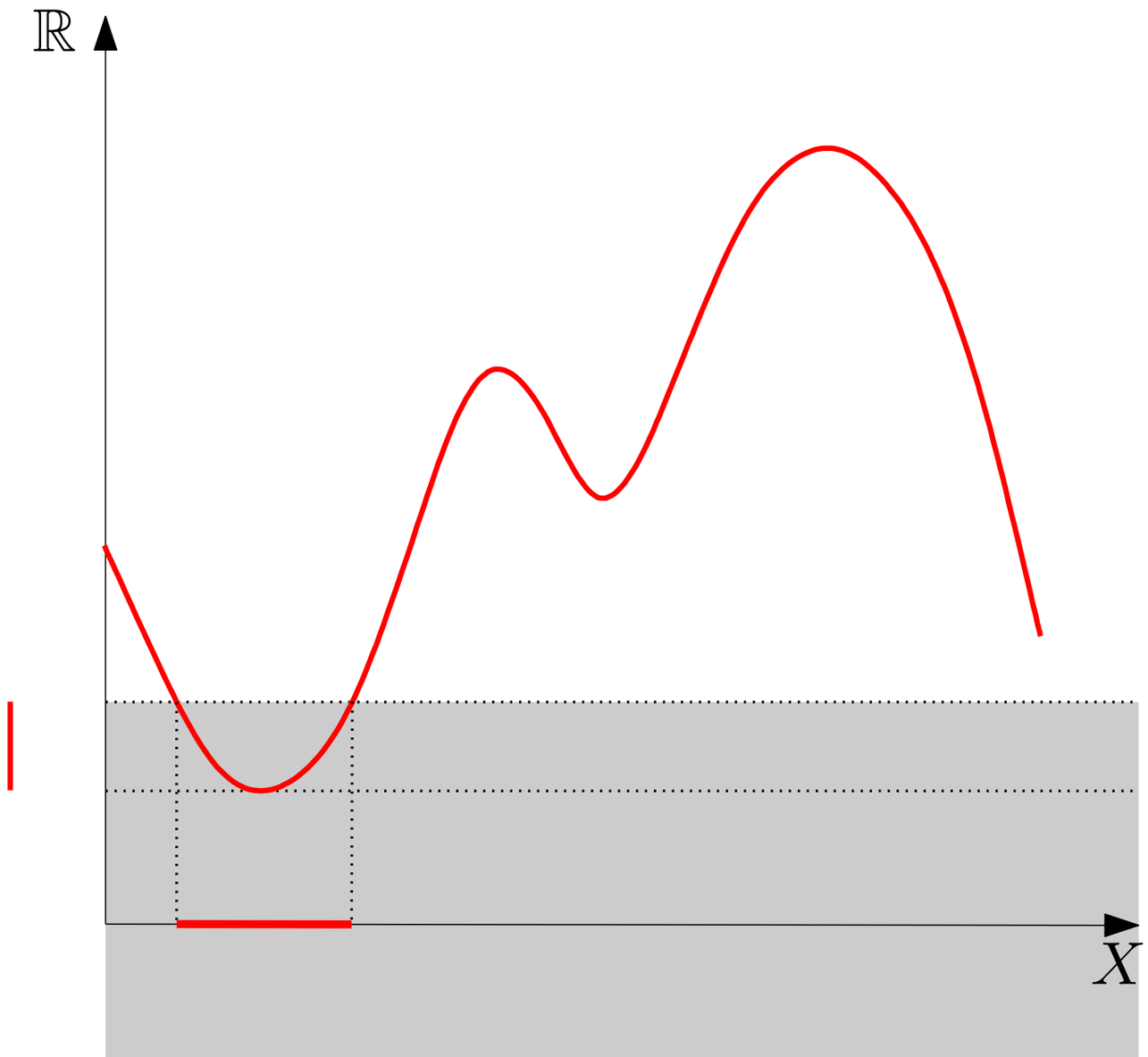
Persistent homology for functions

- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.



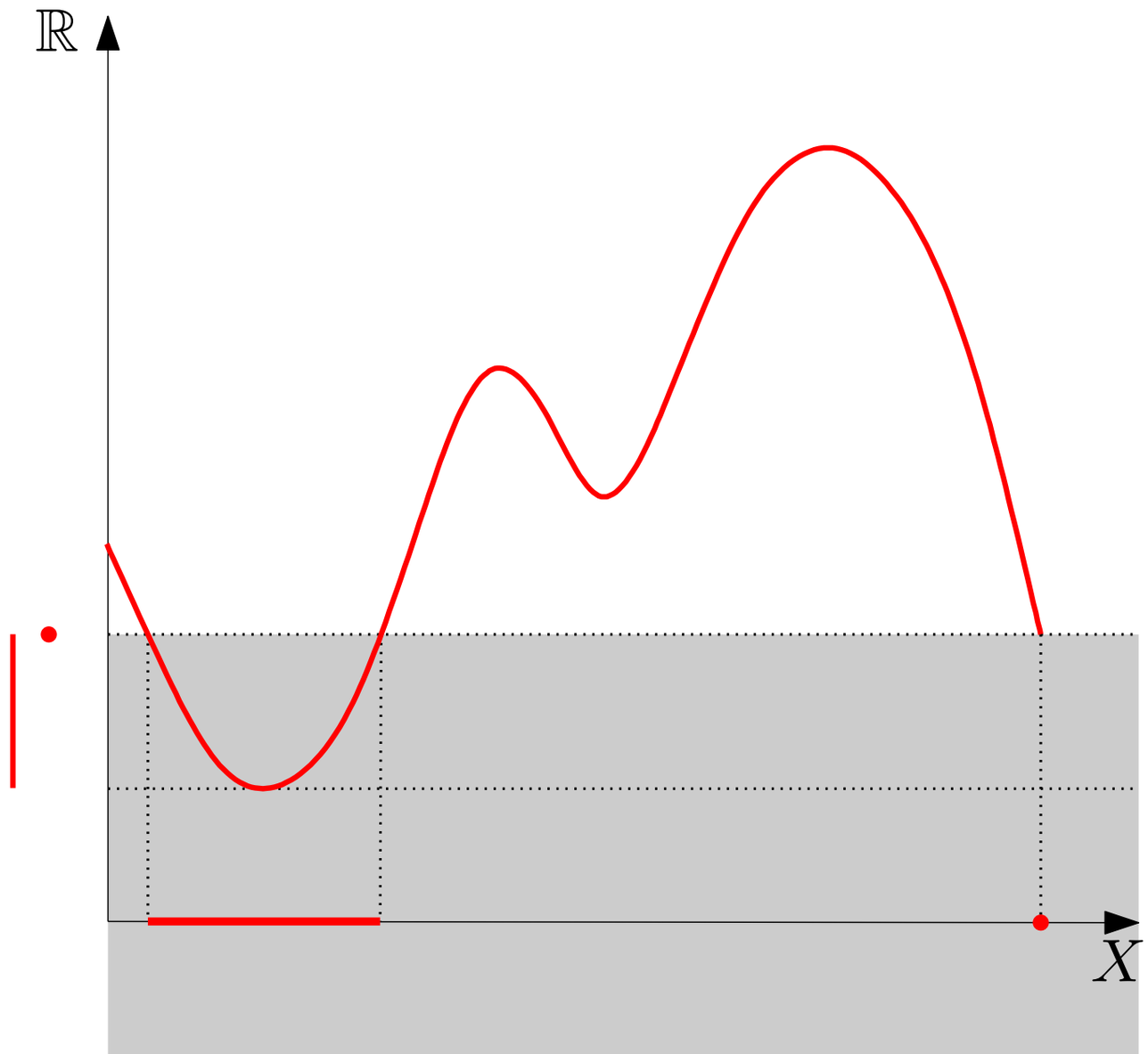
Persistent homology for functions

- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.



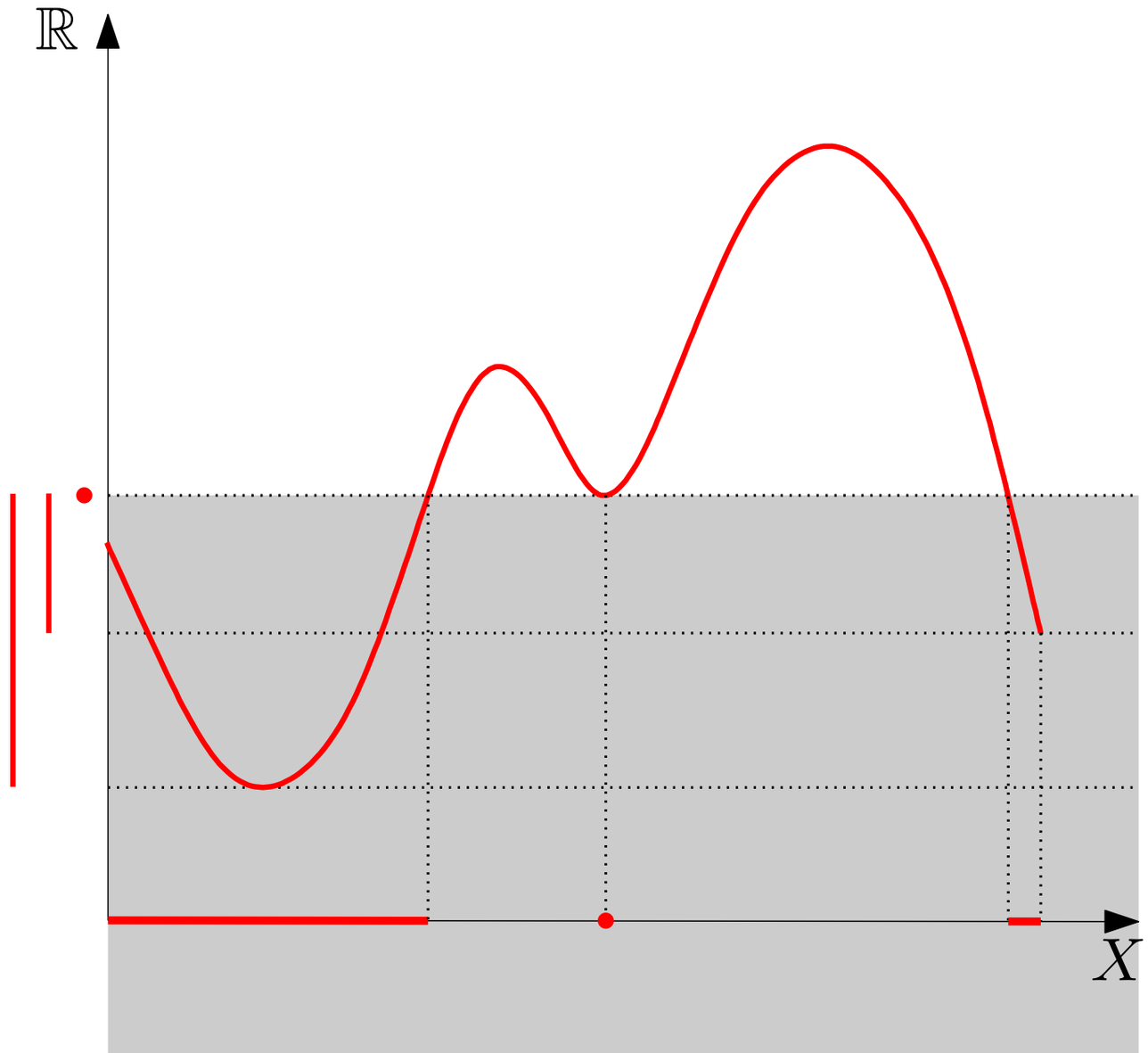
Persistent homology for functions

- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.



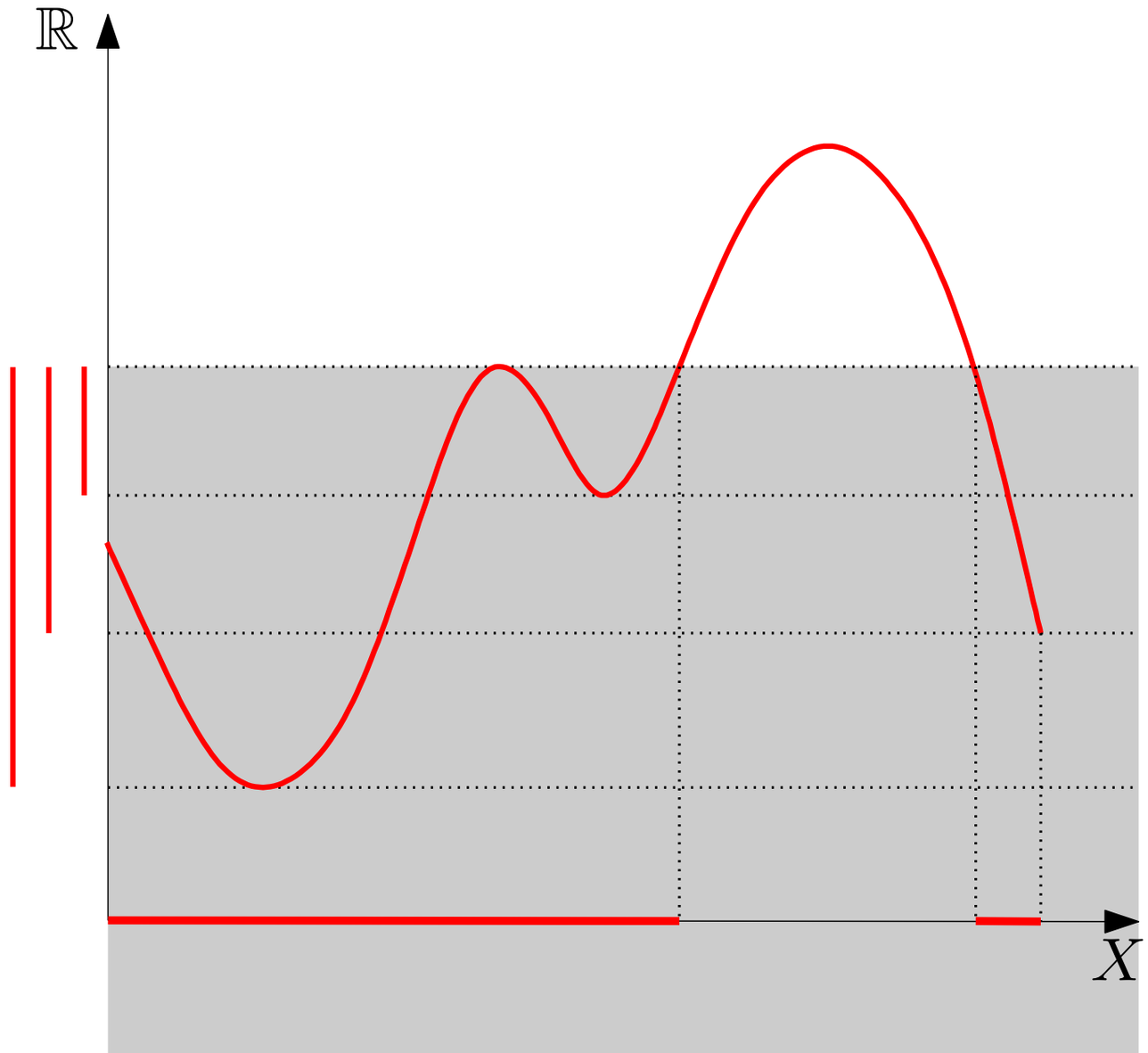
Persistent homology for functions

- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.



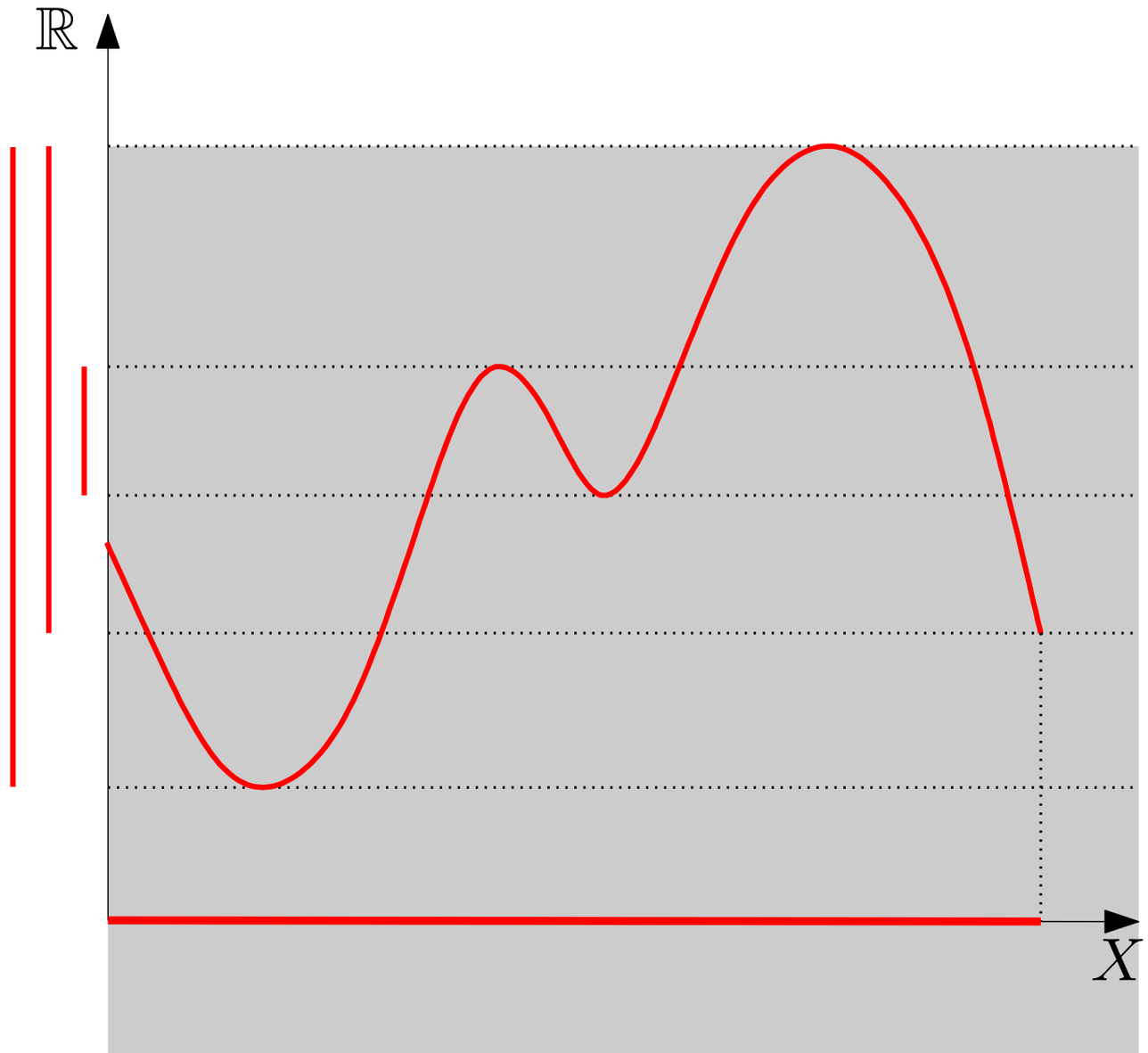
Persistent homology for functions

- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.



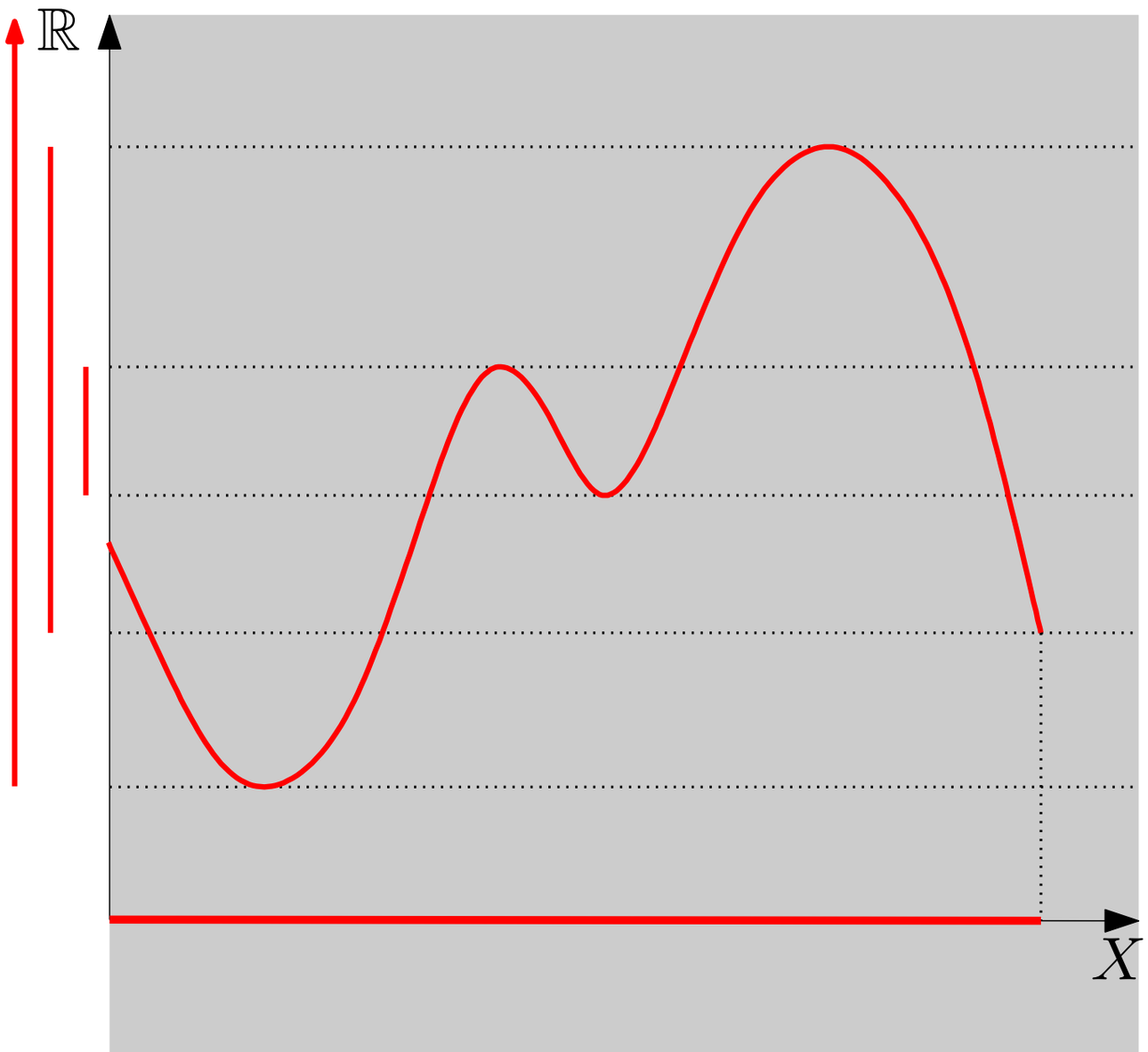
Persistent homology for functions

- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.



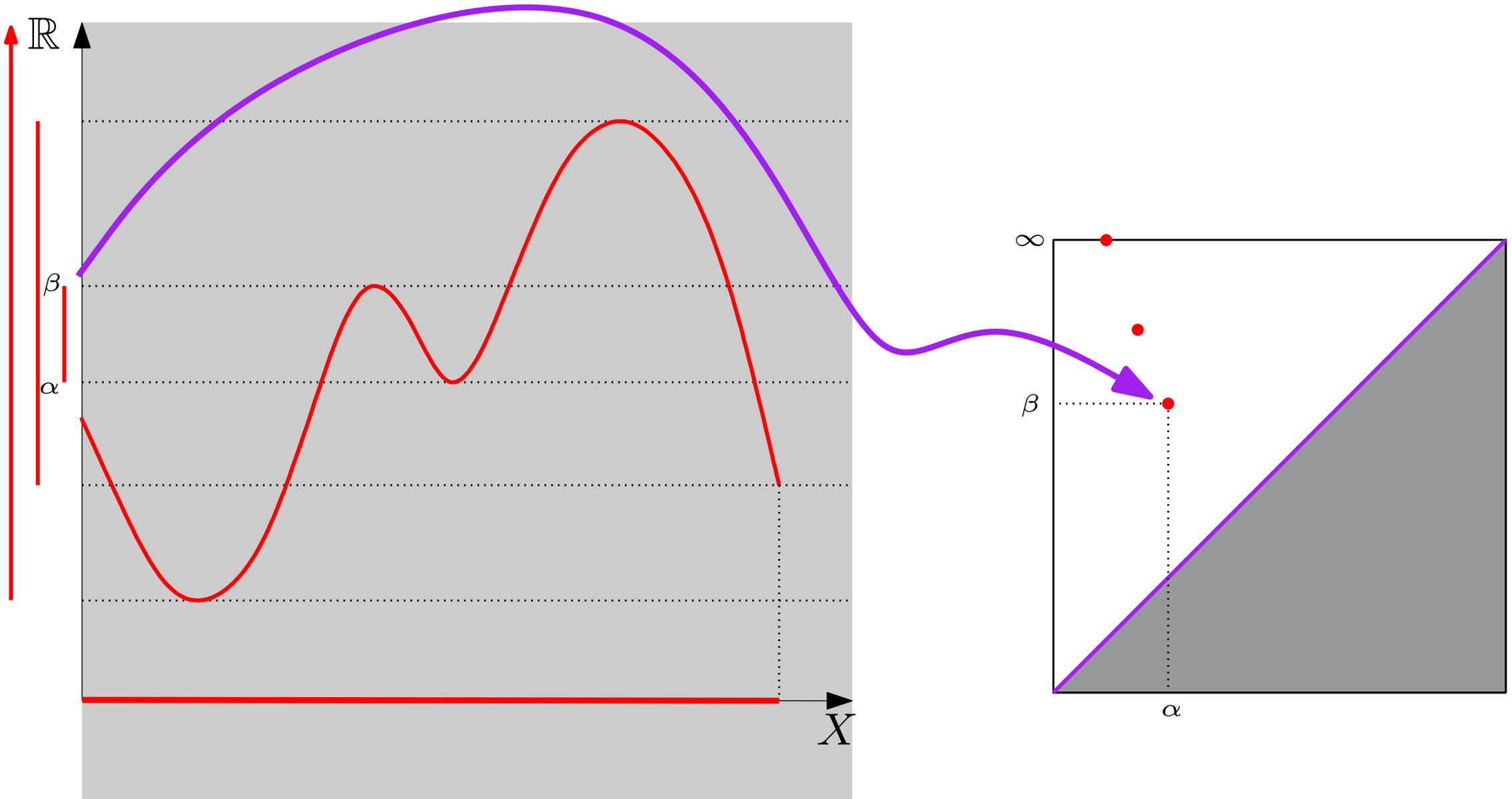
Persistent homology for functions

- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.
- Finite set of intervals (barcode) encodes births/deaths of topological features.



Persistent homology for functions

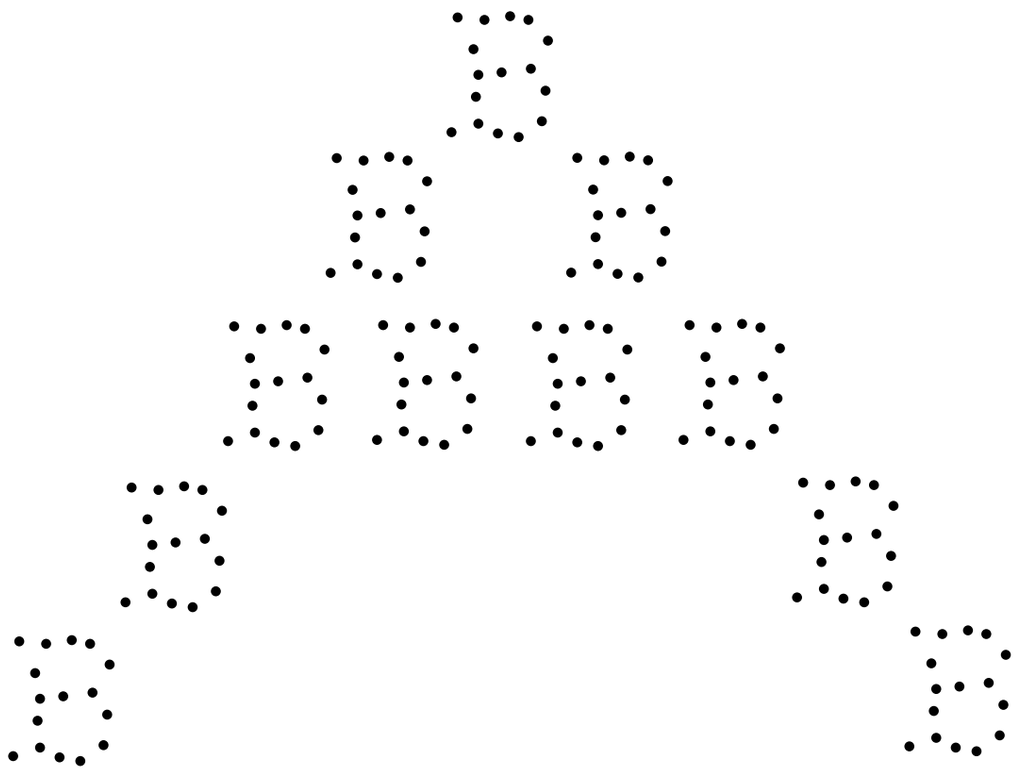
- Nested family (filtration) of sublevel-sets $f^{-1}((-\infty, \alpha])$ for $\alpha = -\infty$ to $+\infty$.
- Track evolution of topology throughout the family.
- Finite set of intervals (barcode) encodes births/deaths of topological features.



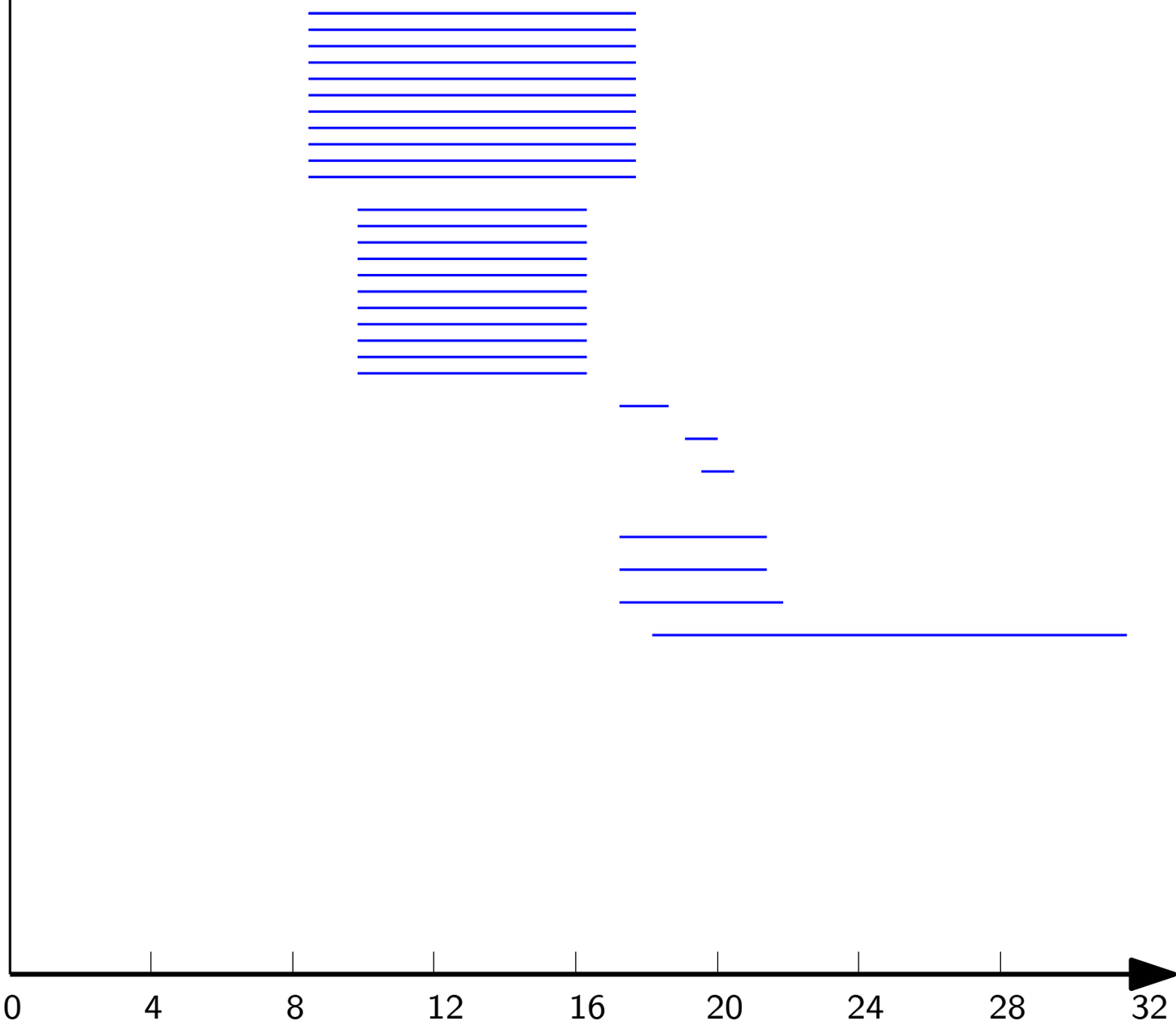
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



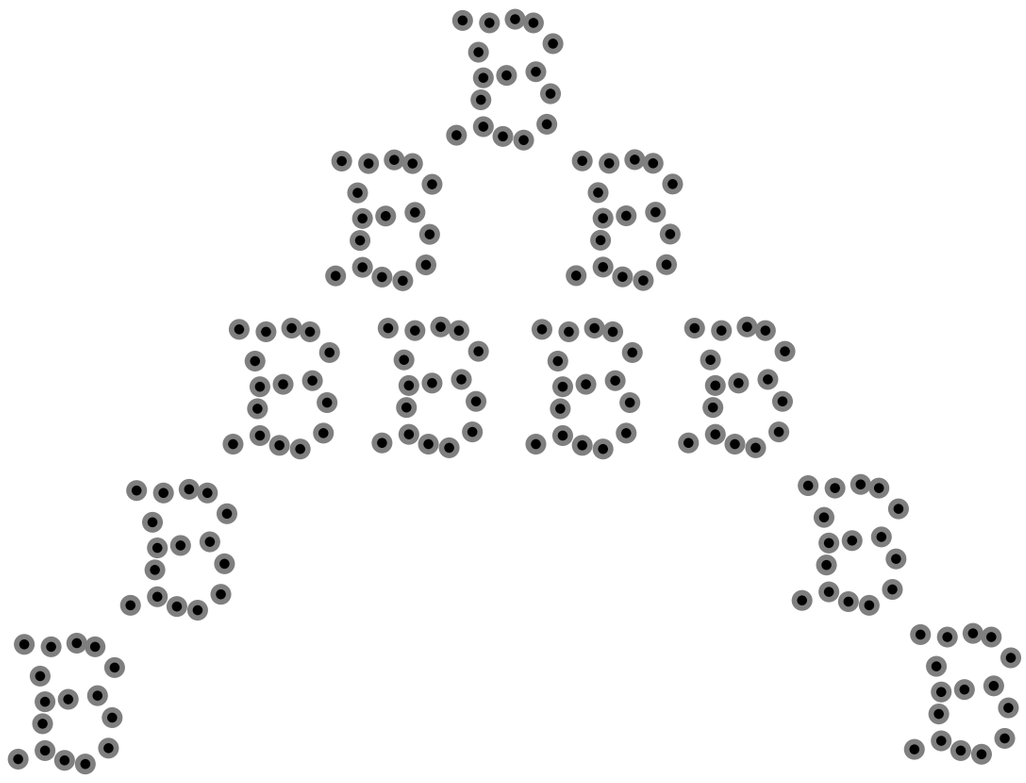
barcode for holes (1-d homology)



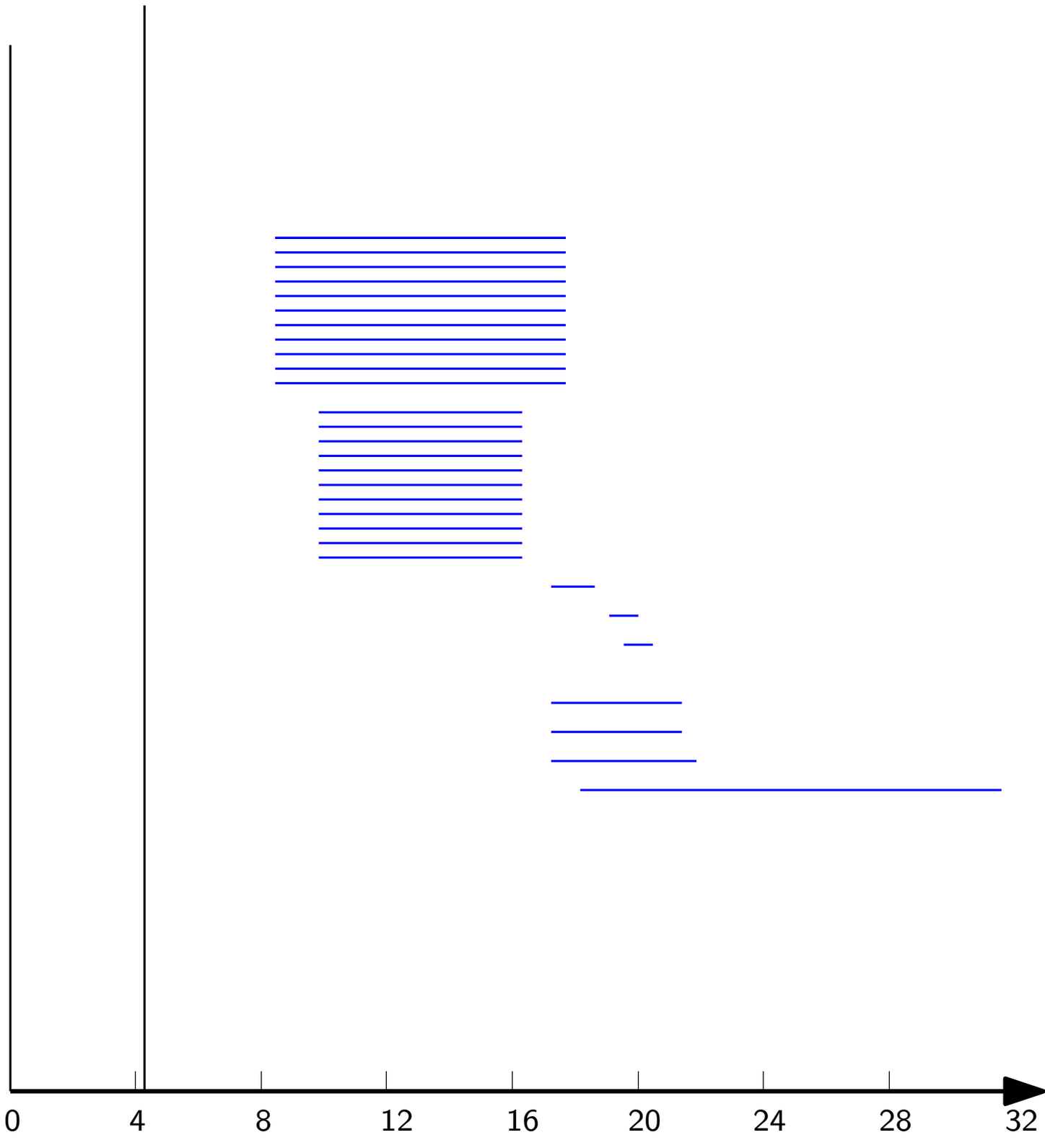
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



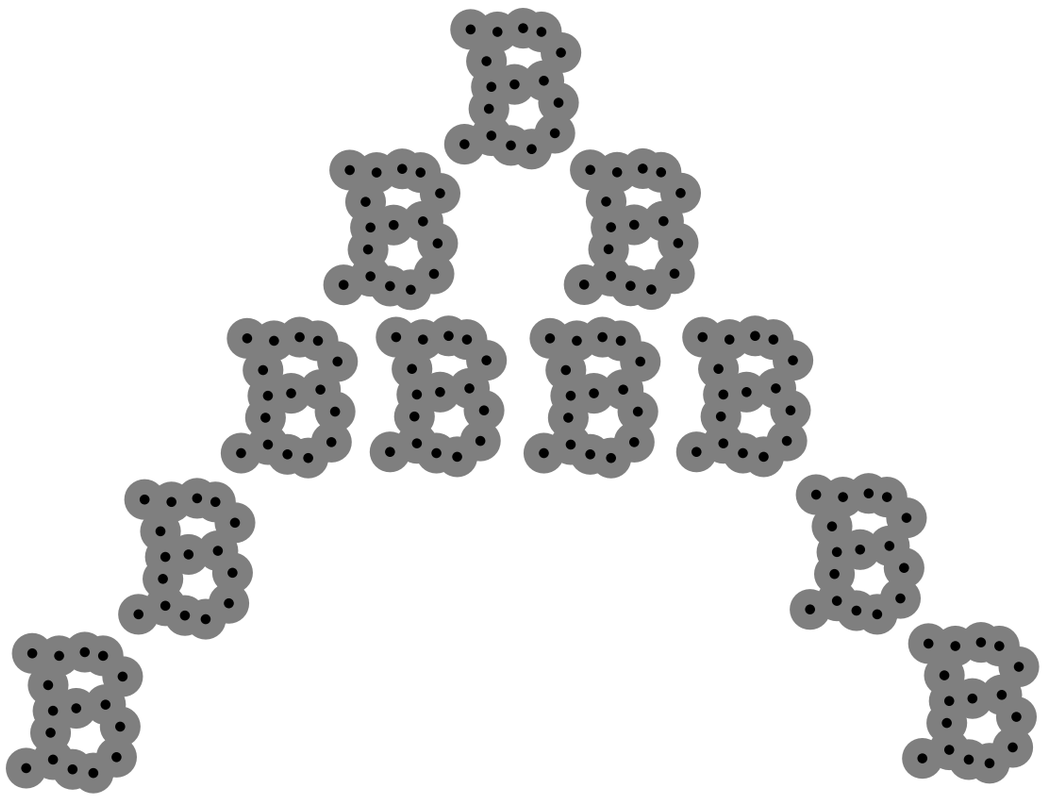
barcode for holes (1-d homology)



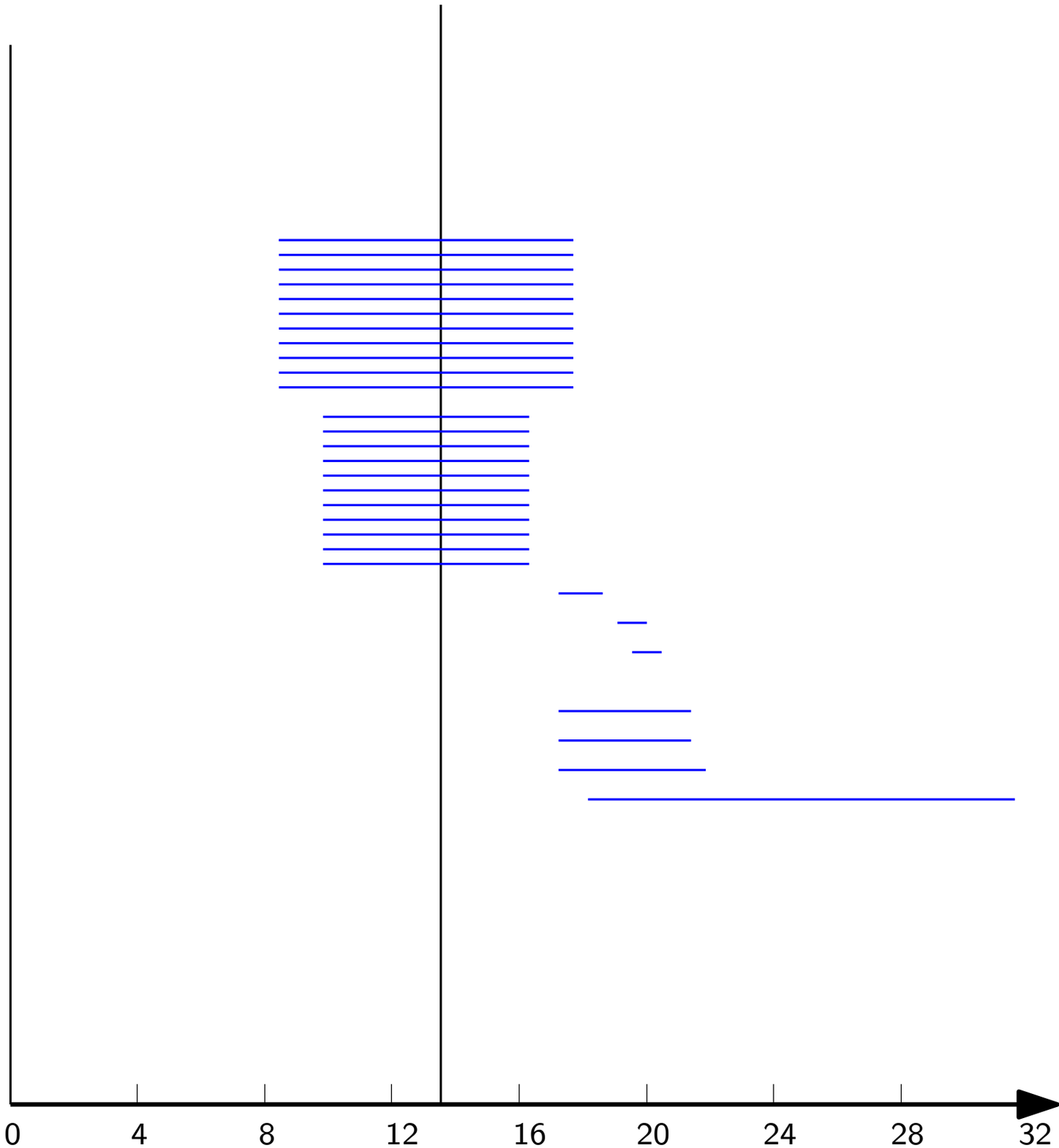
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



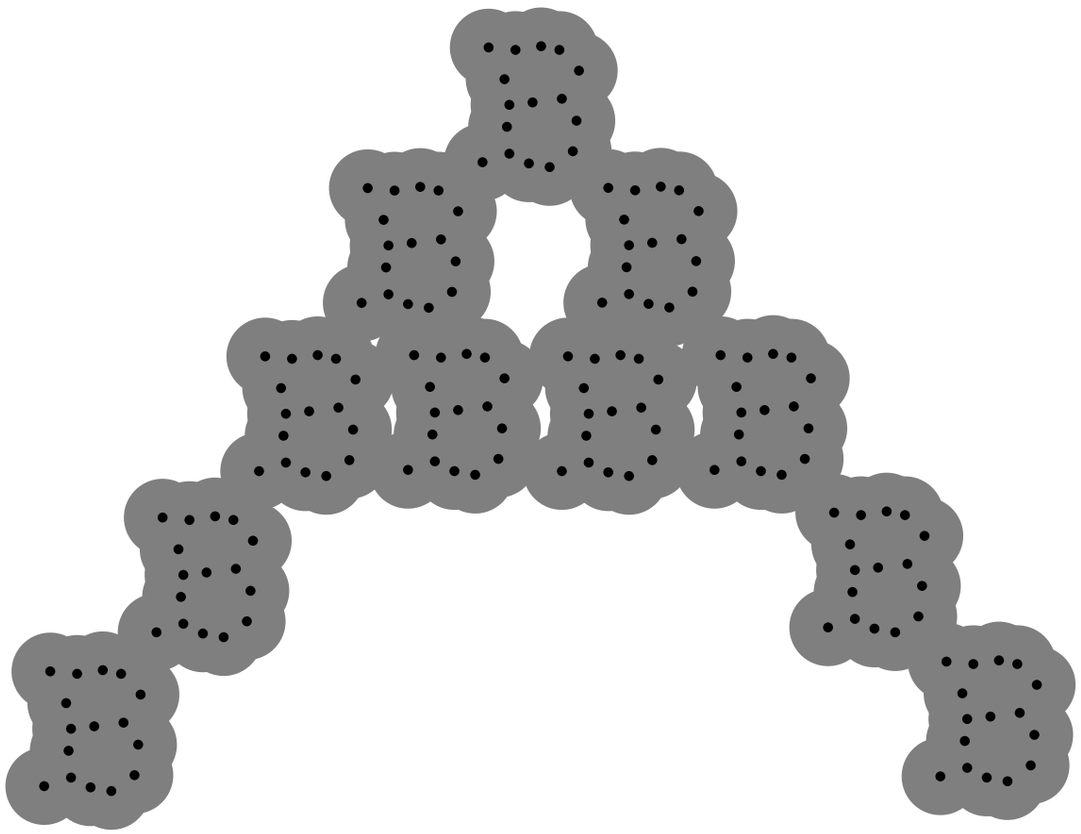
barcode for holes (1-d homology)



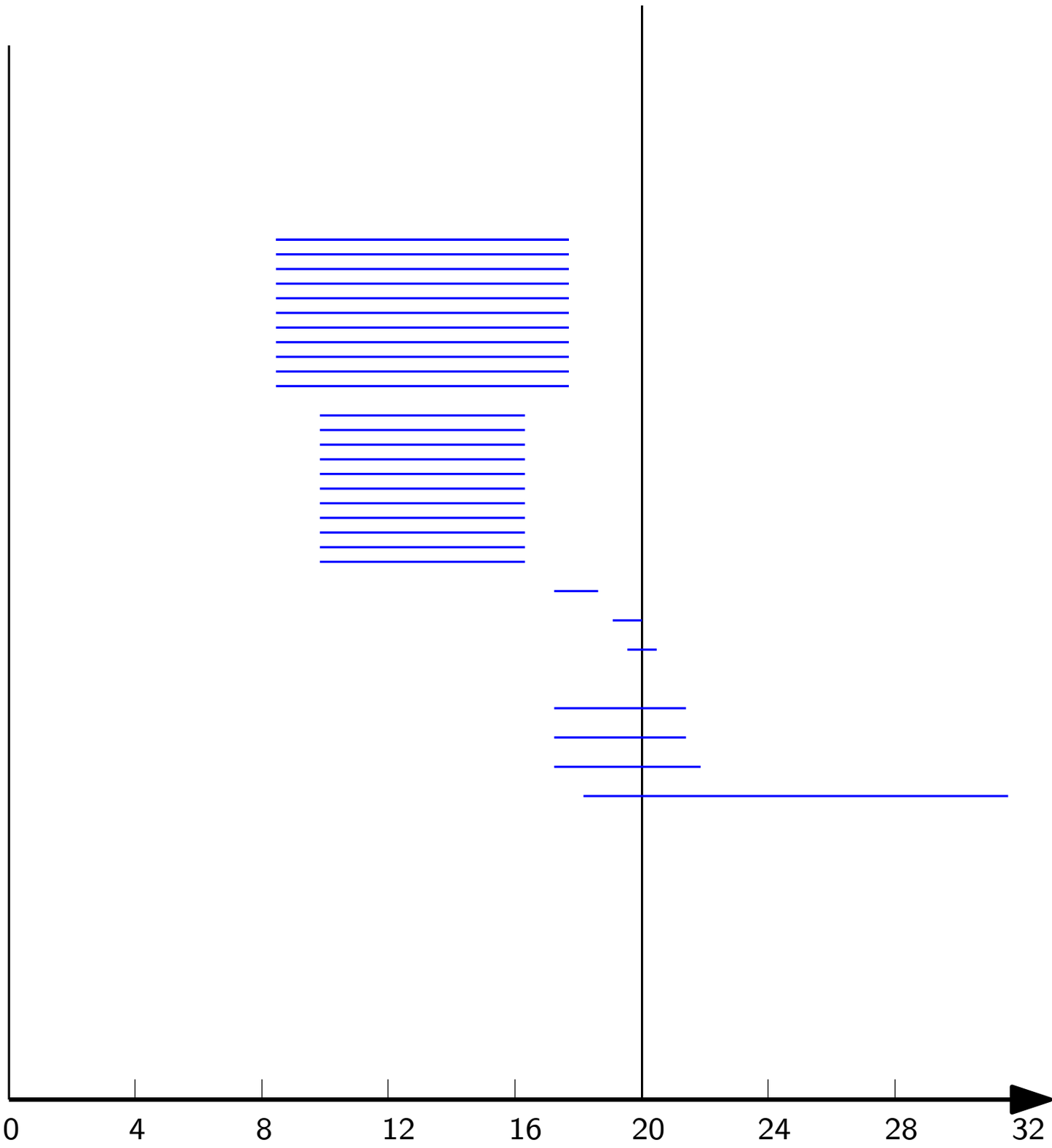
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



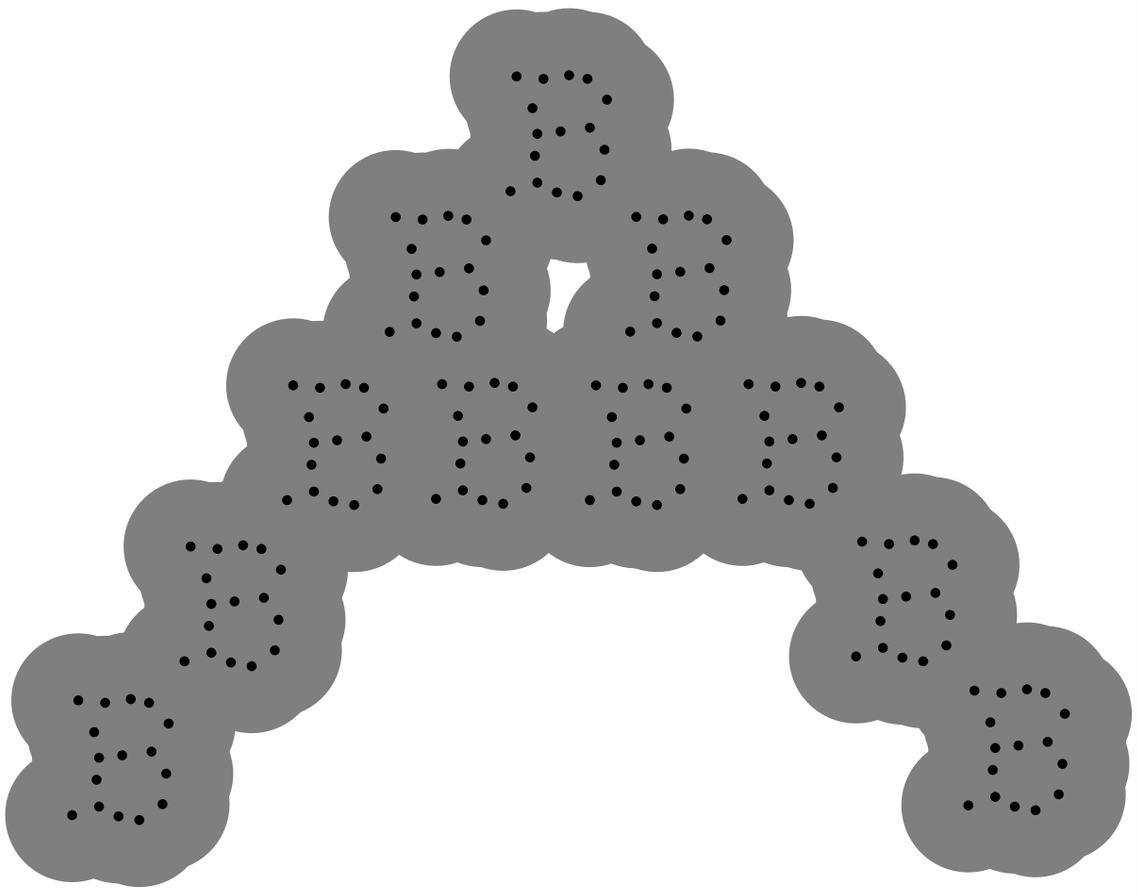
barcode for holes (1-d homology)



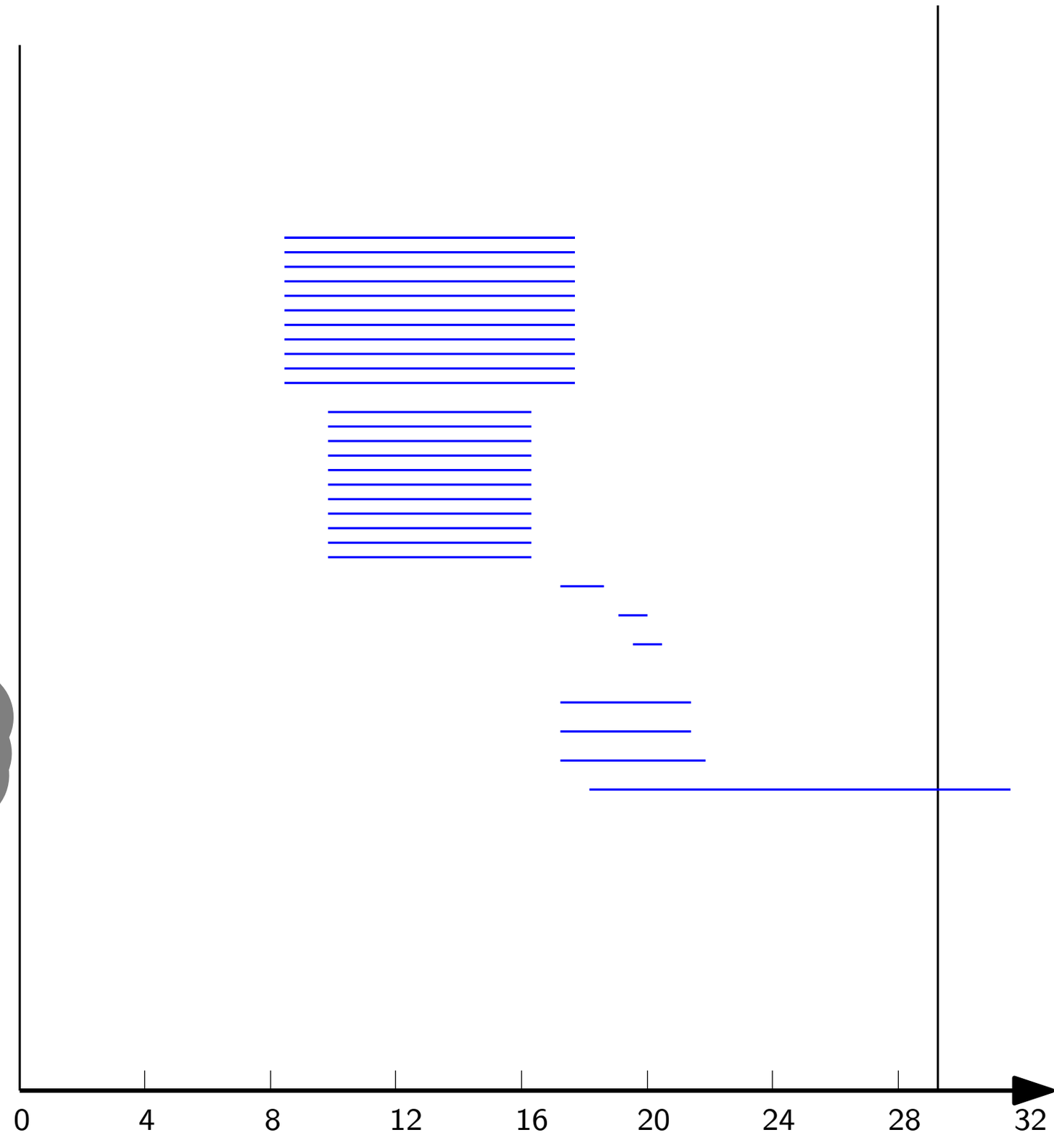
Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

$$x \to \min_{p \in P} \|x - p\|_2$$



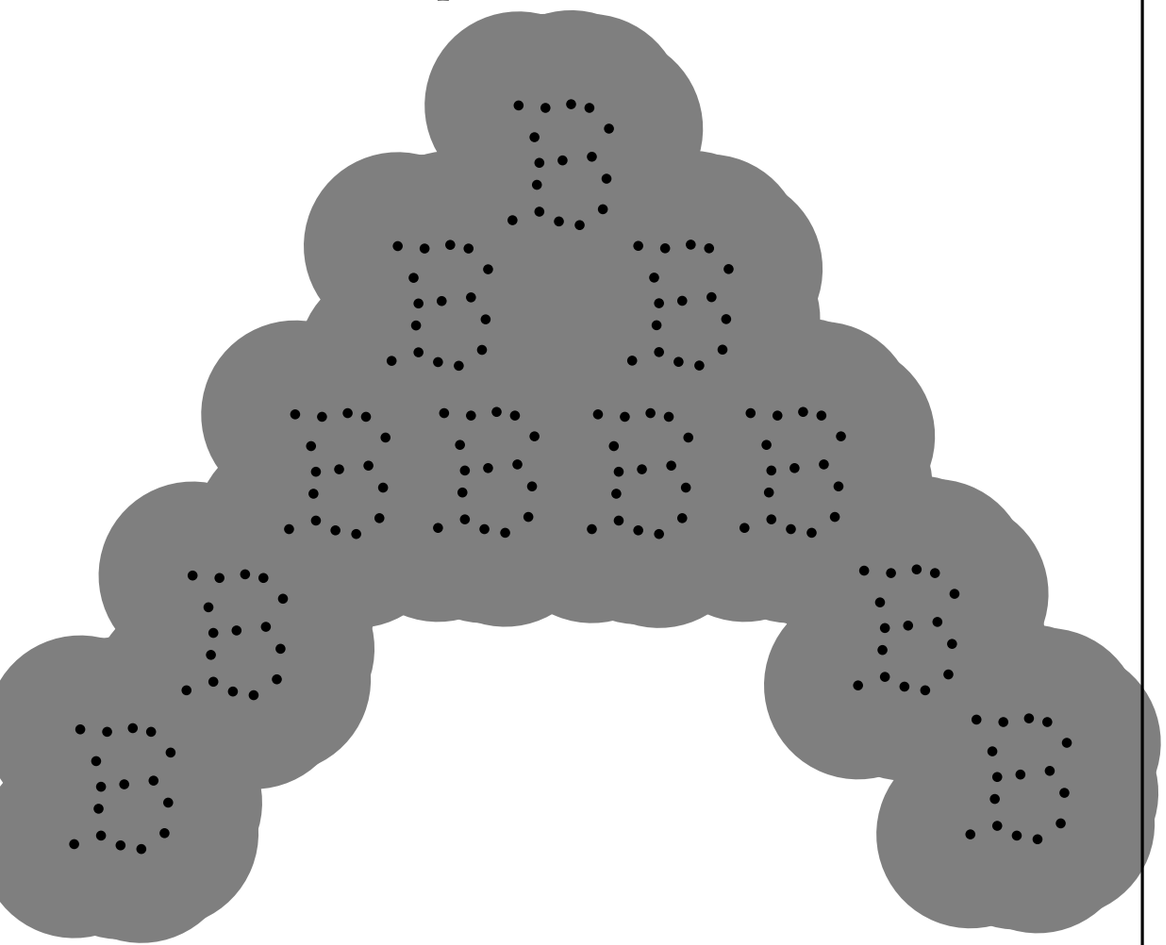
barcode for holes (1-d homology)

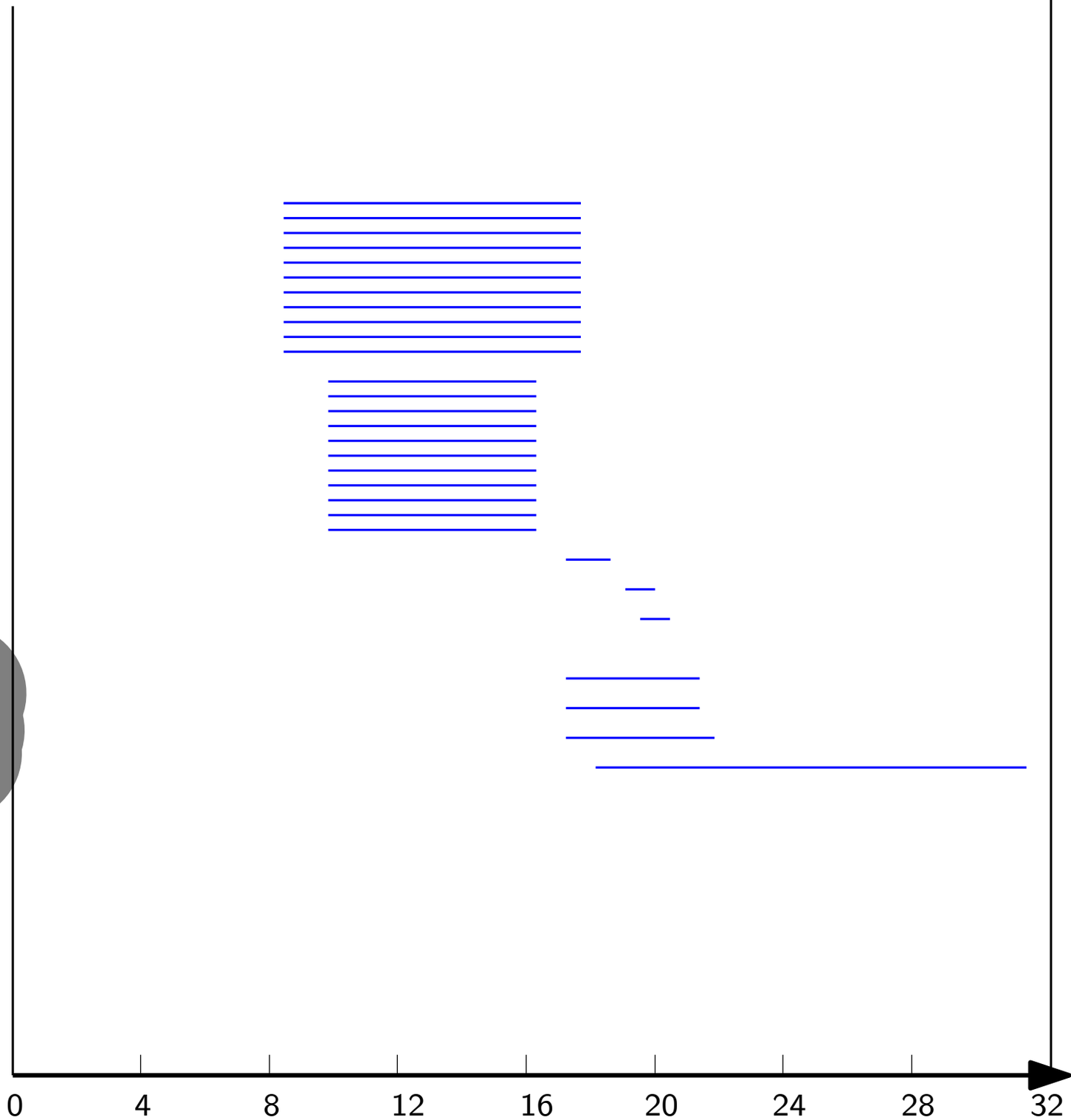


Persistent homology for functions

$$f_P : \mathbb{R}^2 \to \mathbb{R}$$

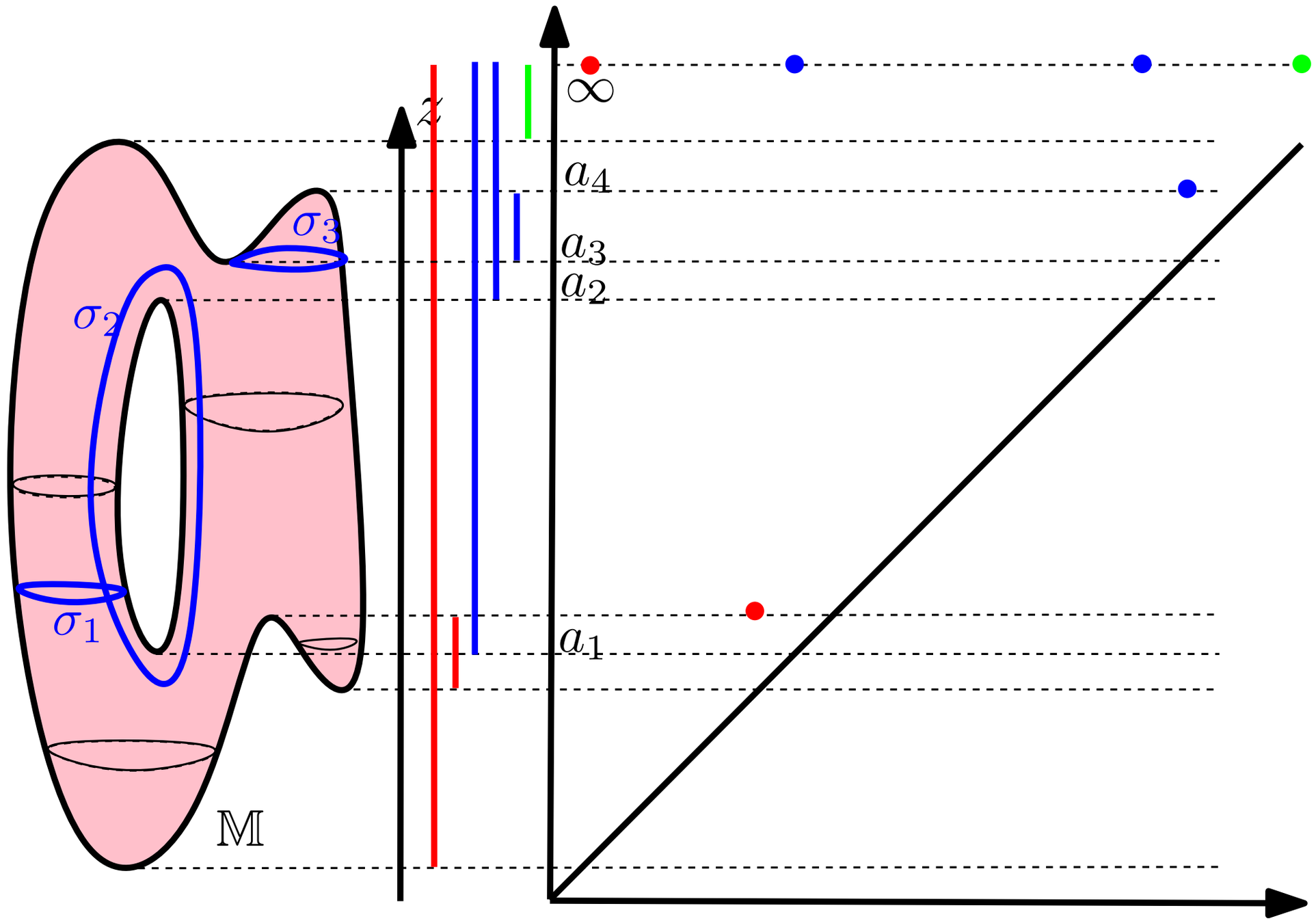
$$x \to \min_{p \in P} \|x - p\|_2$$





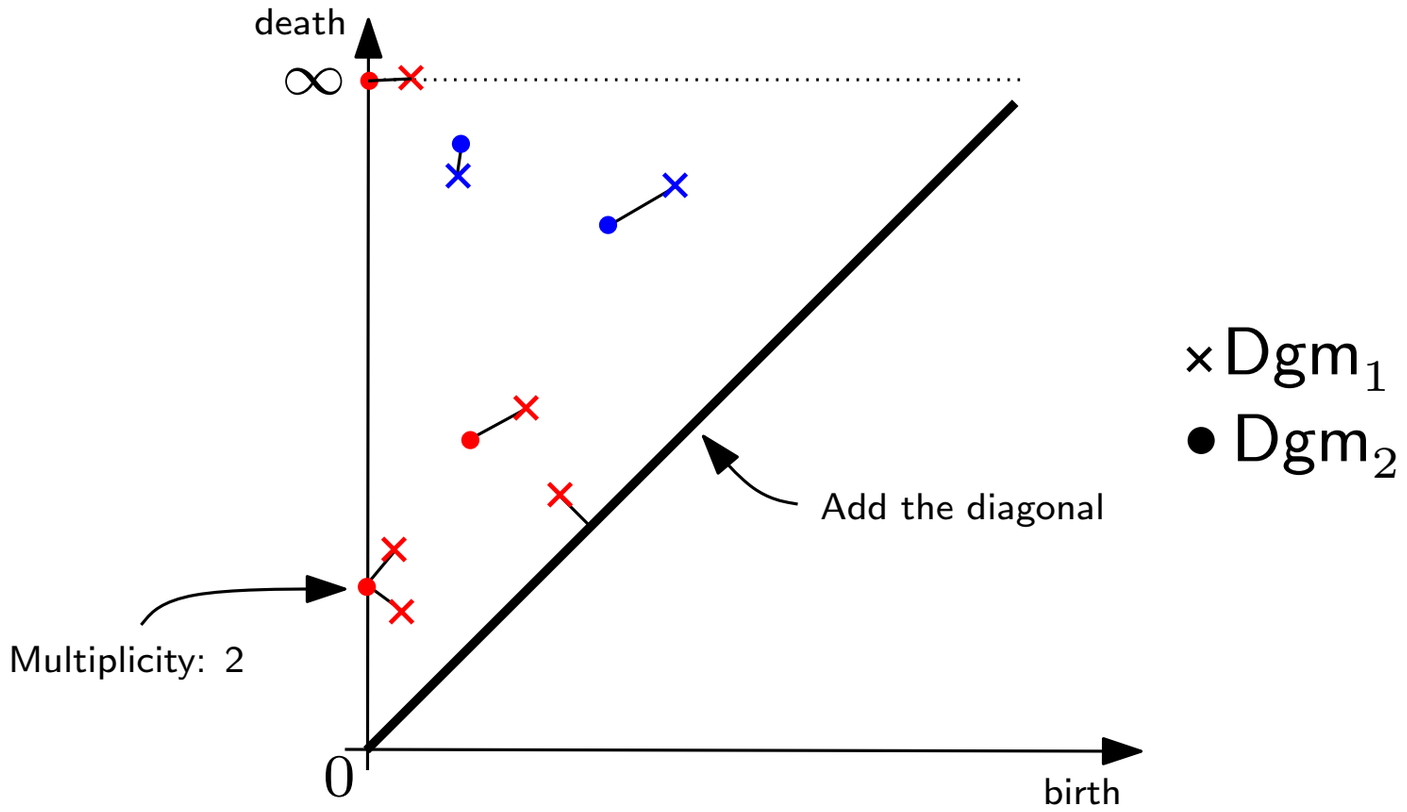
barcode for holes (1-d homology)

Persistent homology for functions



Tracking and encoding the evolution of the 0-dimensional homology, 1-dimensional homology and 2-dimensional homology of the sublevel sets.

Comparing persistence diagrams



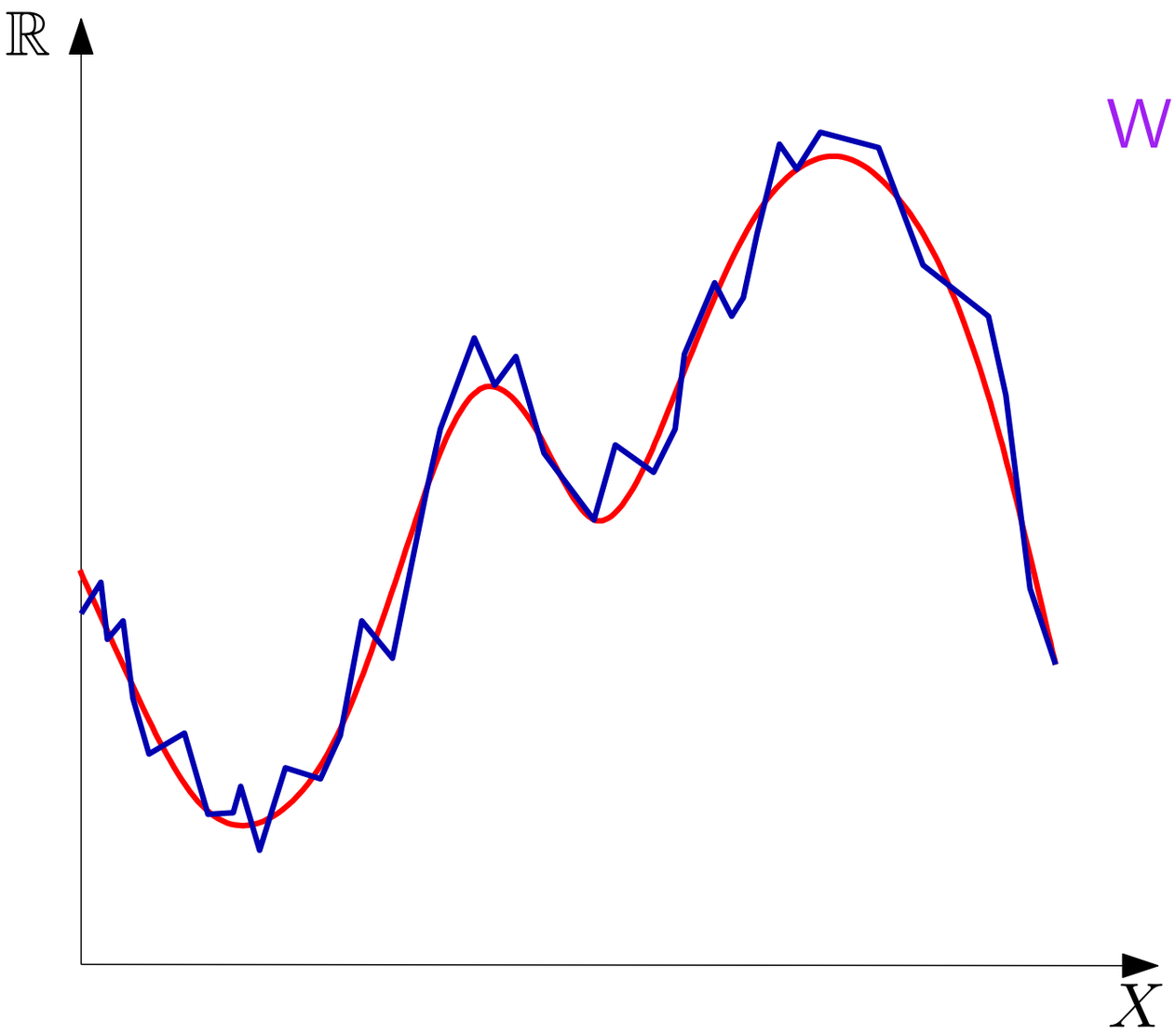
The bottleneck distance between two diagrams Dgm_1 and Dgm_2 is

$$\mathrm{d}_\infty(\mathsf{Dgm}_1, \mathsf{Dgm}_2) = \inf_{\gamma \in \Gamma} \sup_{p \in \mathsf{Dgm}_1} \|p - \gamma(p)\|_\infty$$

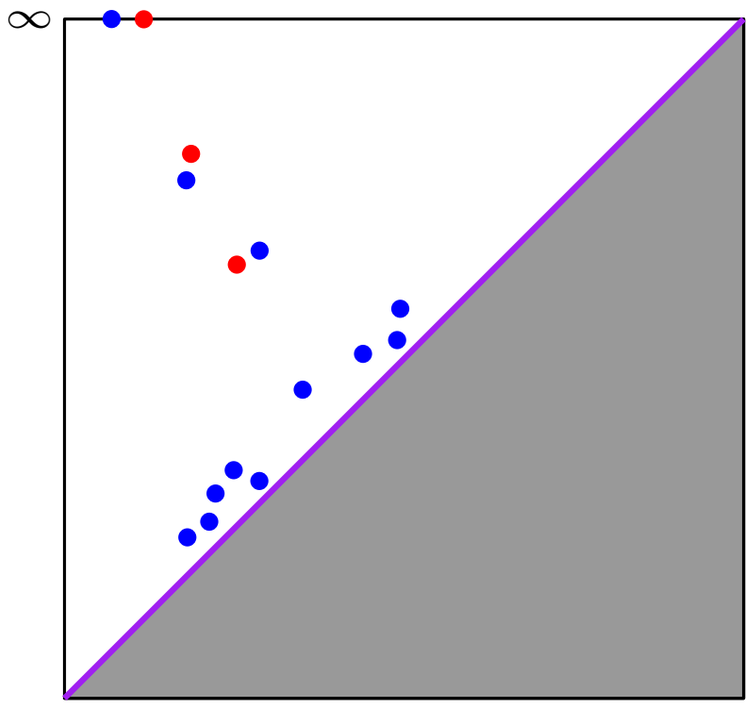
where Γ is the set of all the bijections between Dgm_1 and Dgm_2 and $\|p-q\|_\infty = \max(|x_p - x_q|, |y_p - y_q|)$.

$\rightarrow$ Persistence diagrams provide easy to compare topological signatures.

Stability properties



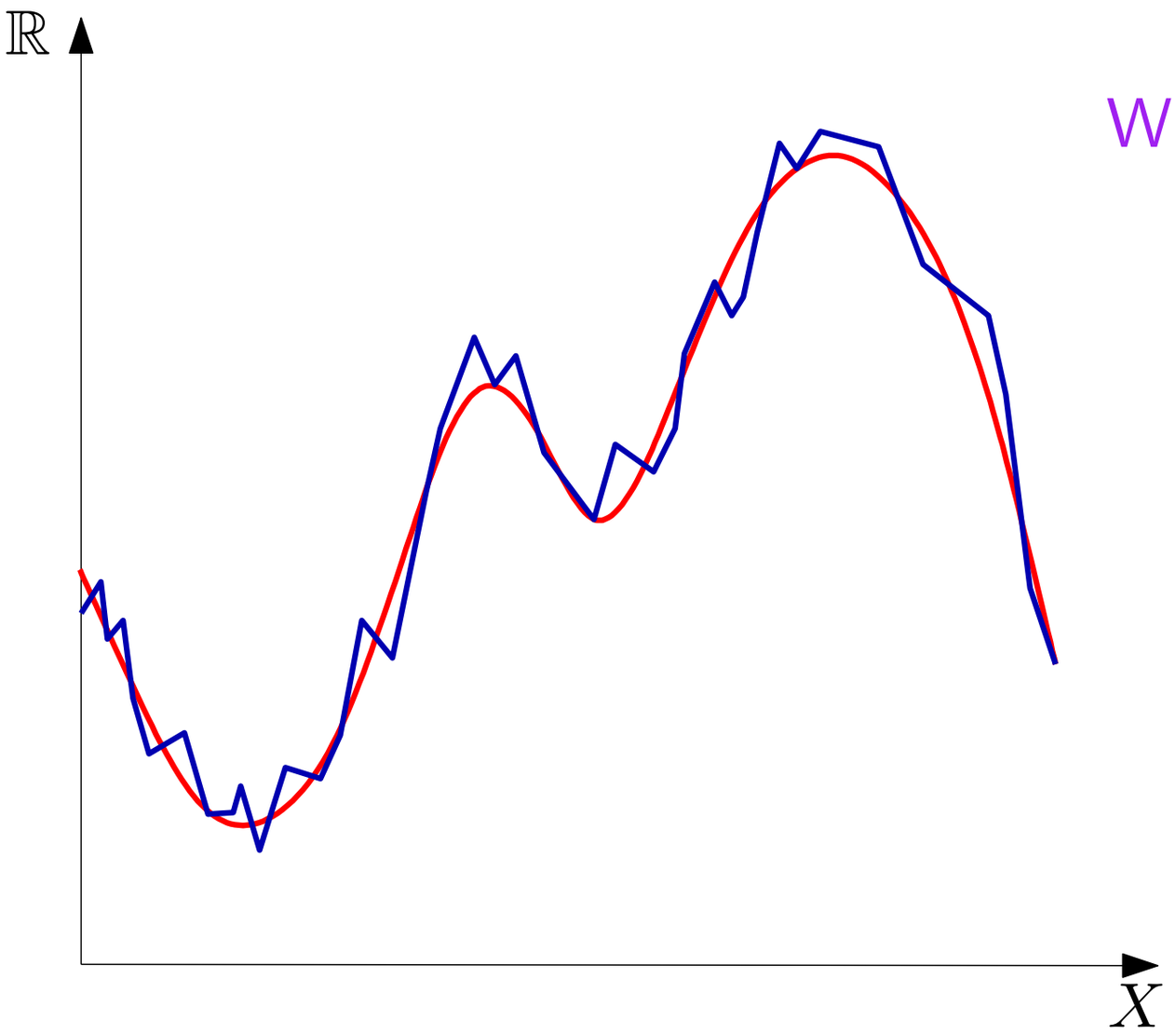
What if f is slightly perturbed?



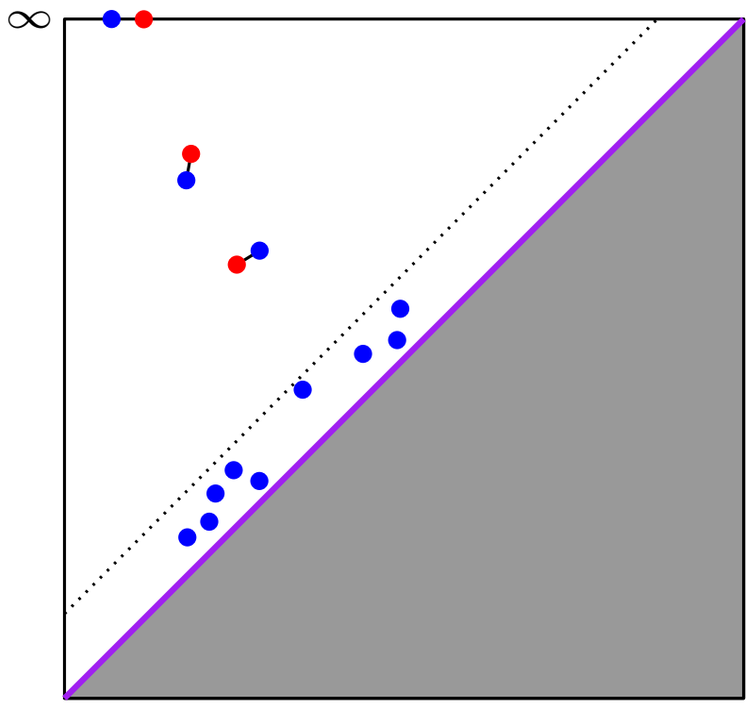
Stability properties

Theorem (Stability):
For any *tame* functions $f, g : \mathbb{X} \to \mathbb{R}$, $\mathrm{d_\infty}^\infty(\mathrm{Dgm}f, \mathrm{Dgm}g) \leq \|f - g\|_\infty$.

[Cohen-Steiner, Edelsbrunner, Harer 05], [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG 09], [C., de Silva, Glisse, Oudot 12]



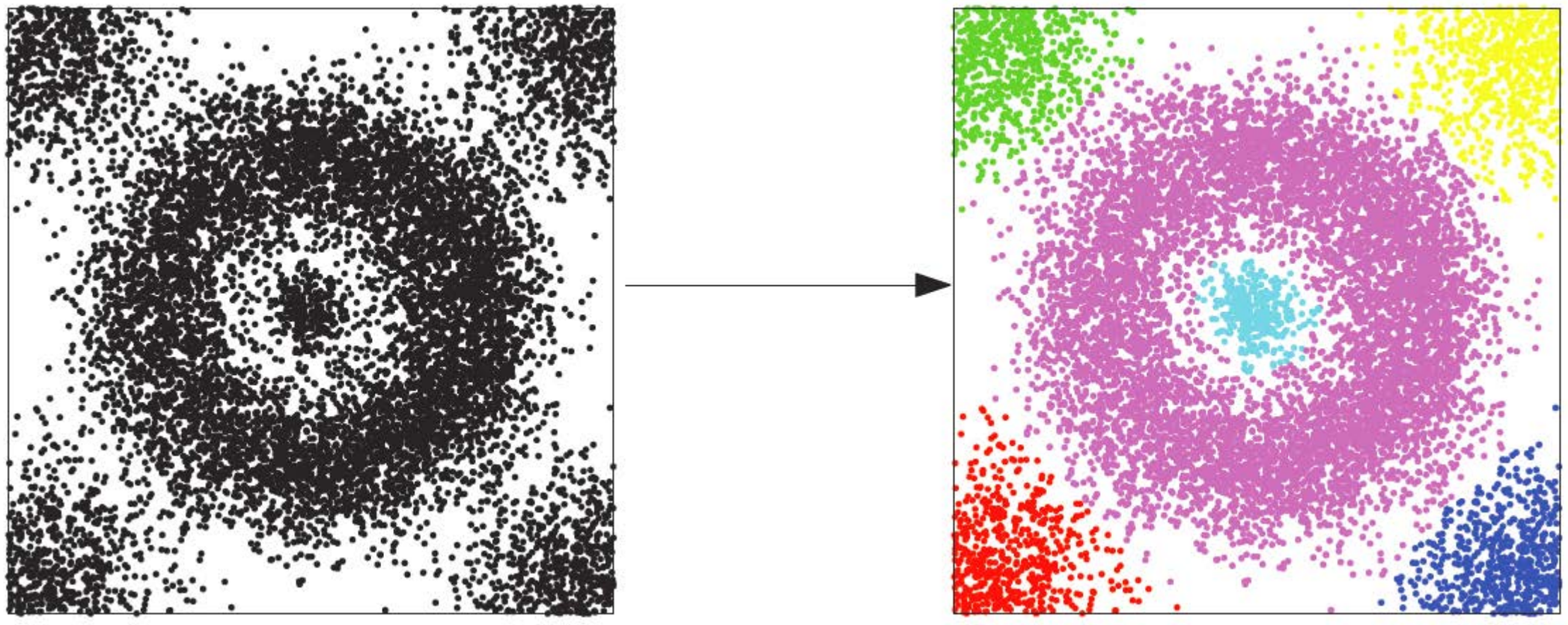
What if f is slightly perturbed?



Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

[C.,Guibas,Oudot,Skraba - J. ACM 2013]



Input:

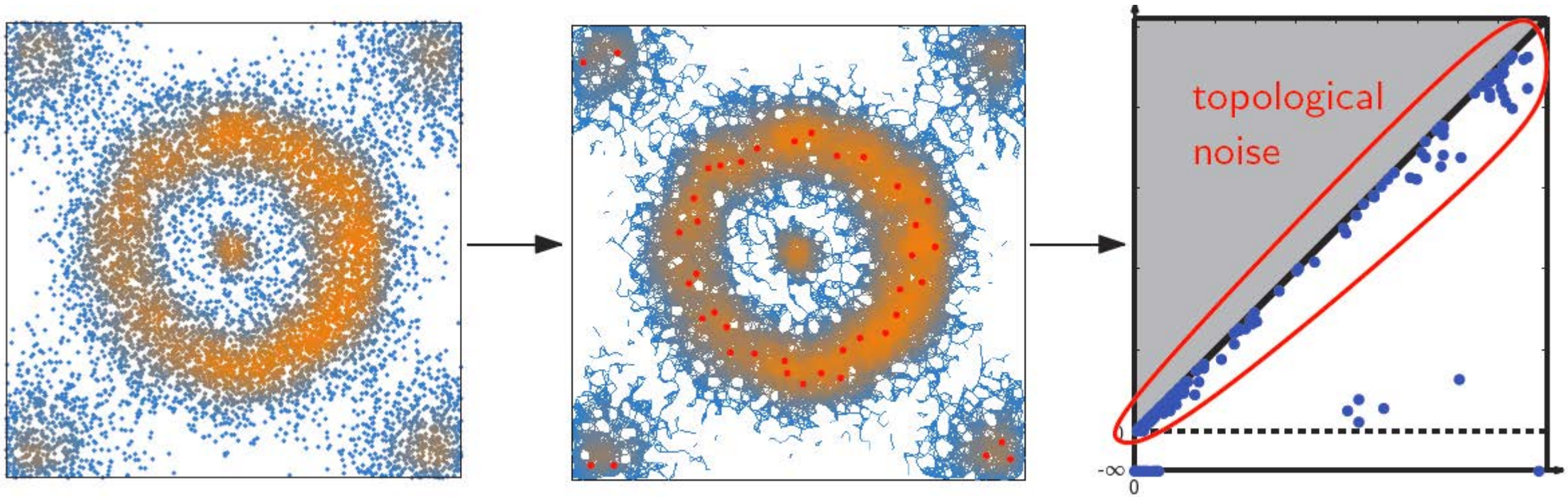
1. A finite set X of observations (point cloud with coordinates or pairwise distance matrix),
2. A real valued function f defined on the observations (e.g. density estimate).

Goal: Partition the data according to the basins of attraction of the peaks of f

Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

[C.,Guibas,Oudot,Skraba - J. ACM 2013]

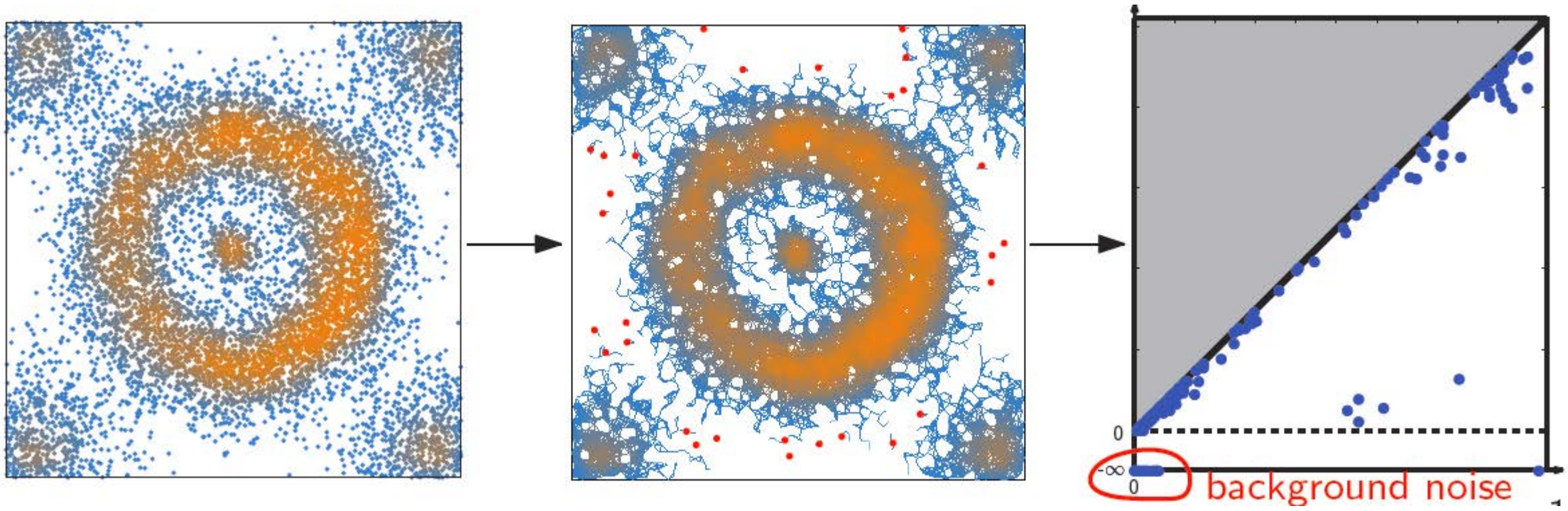


1. Build a neighborhing graph G on top of X.
2. Compute the (0-dim) persistence of f to identify prominent peaks

Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

[C.,Guibas,Oudot,Skraba - J. ACM 2013]

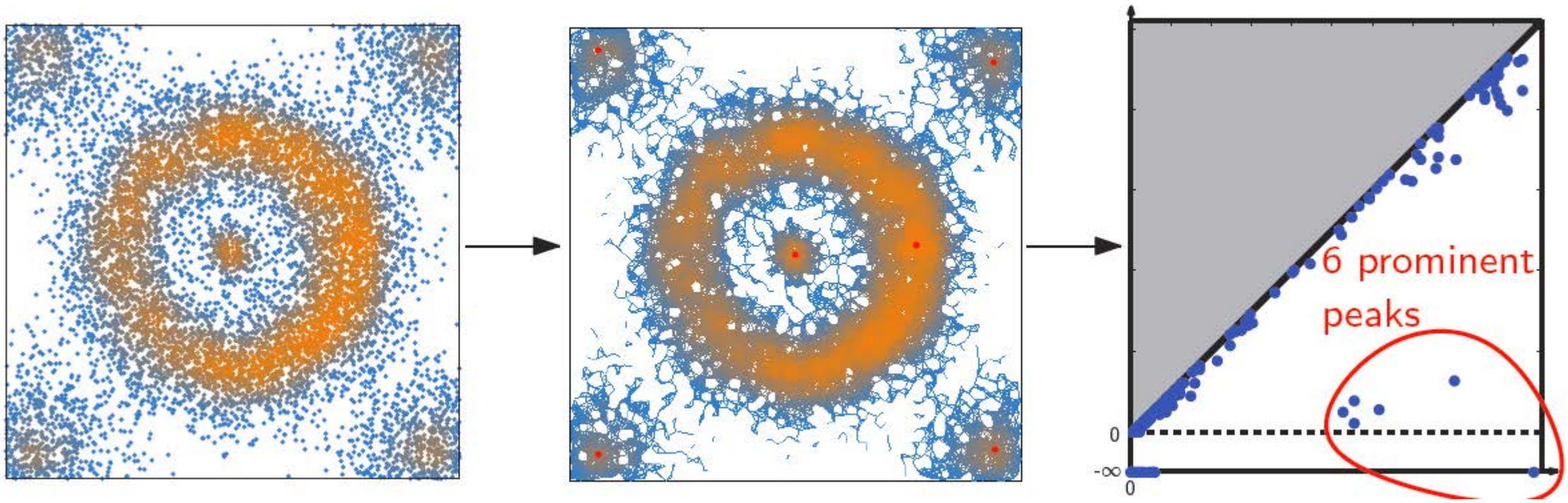


1. Build a neighborhing graph G on top of X.
2. Compute the (0-dim) persistence of f to identify prominent peaks

Persistence-based clustering

Combine a mode seeking approach with (0-dim) persistence computation.

[C.,Guibas,Oudot,Skraba - J. ACM 2013]

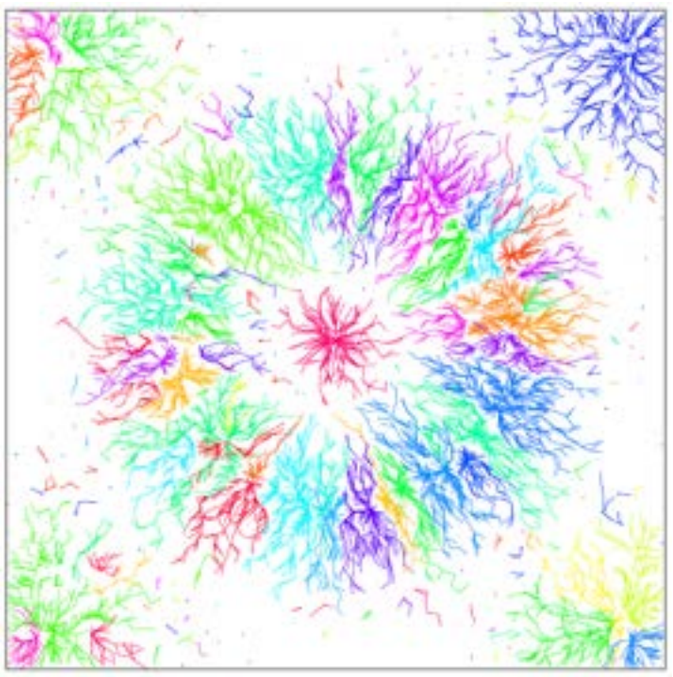


1. Build a neighborhing graph G on top of X.
2. Compute the (0-dim) persistence of f to identify prominent peaks

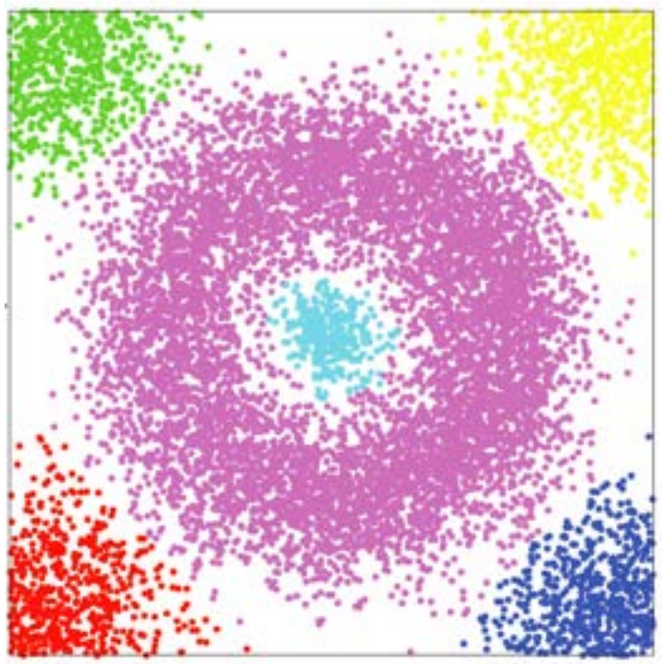
Persistence-based clustering

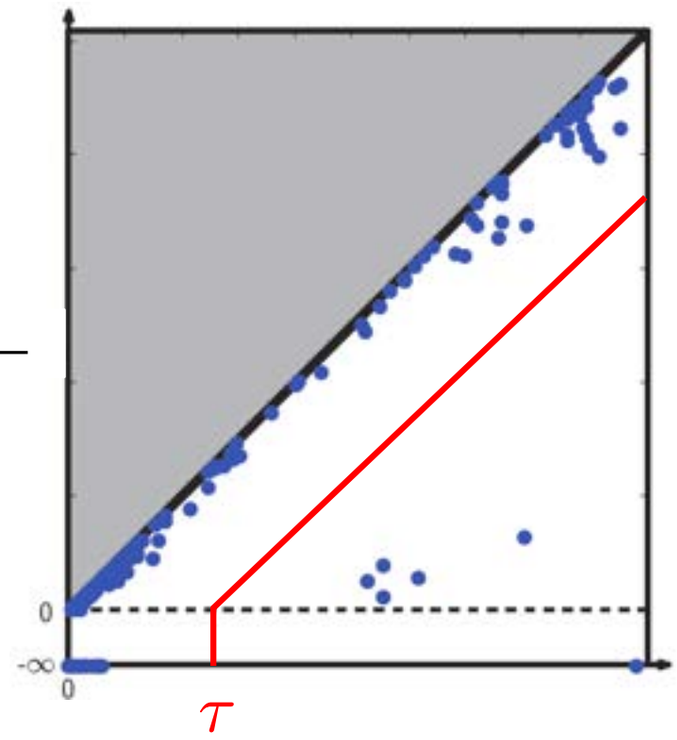
Combine a mode seeking approach with (0-dim) persistence computation.

[C.,Guibas,Oudot,Skraba - J. ACM 2013]



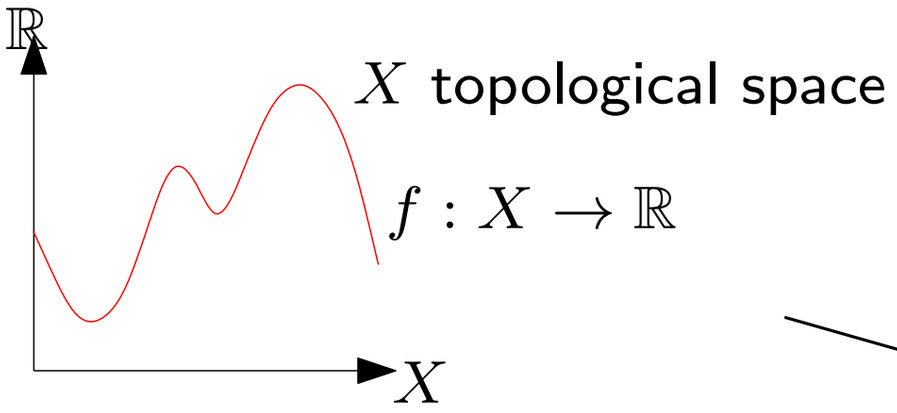
$\tau = 0$

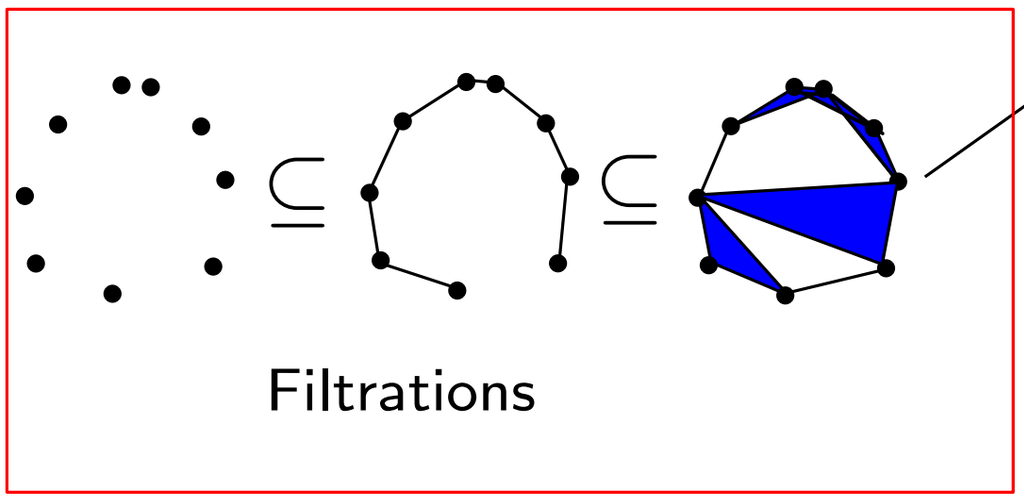




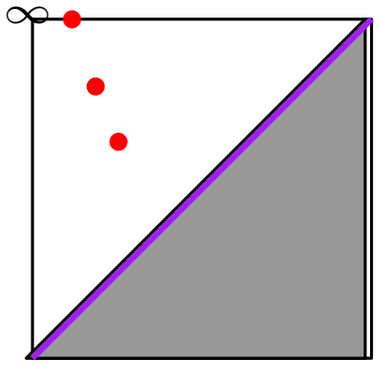
1. Build a neighborhing graph G on top of X.
2. Compute the (0-dim) persistence of f to identify prominent peaks
3. Chose a threshold $\tau > 0$ and use the persistence algorithm to merge components with prominence less than τ.

Persistent homology



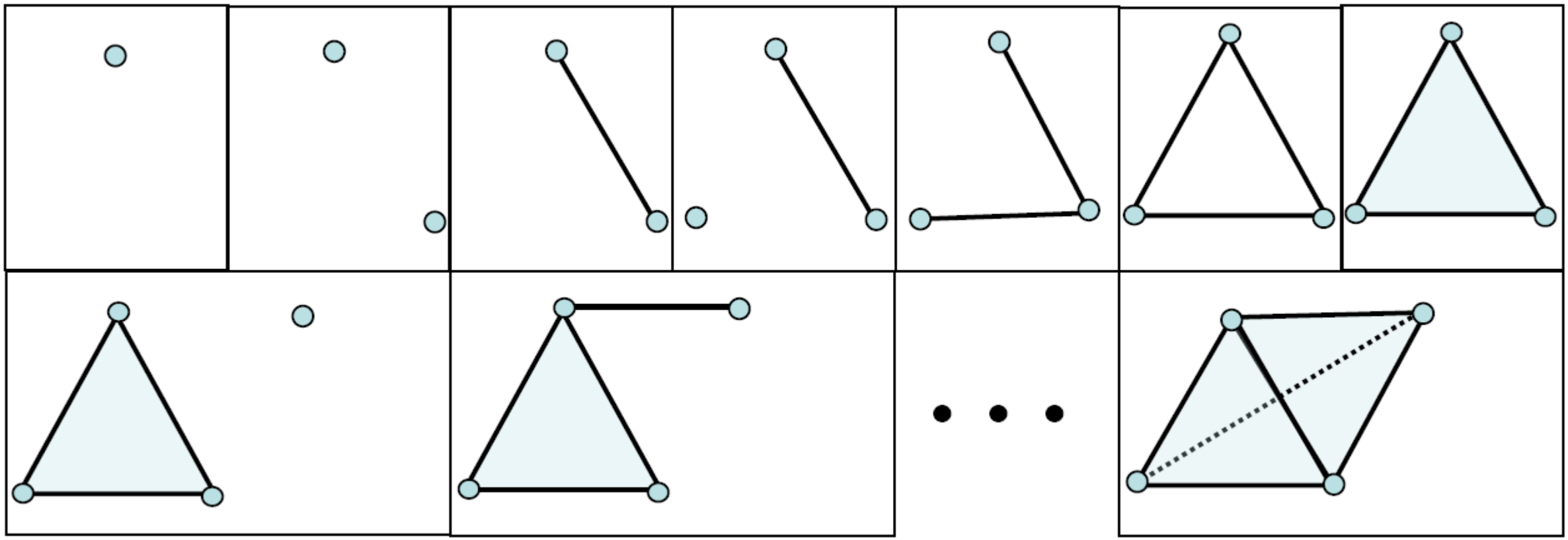


persistence



Persistence diagram

Filtrations of simplicial complexes

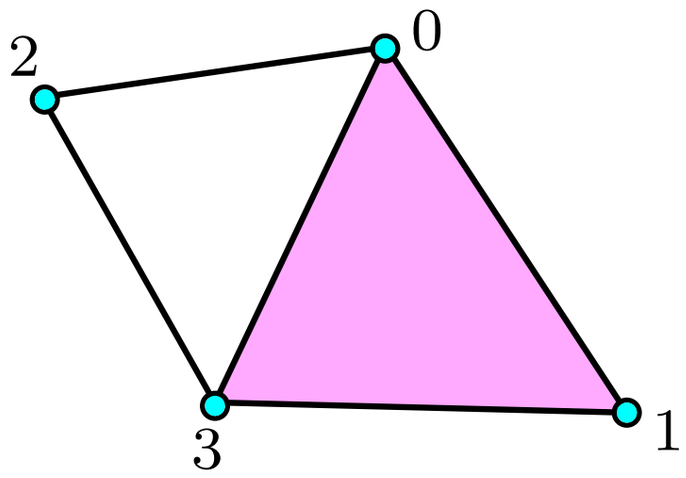


A filtration of a (finite) simplicial complex K is a sequence of subcomplexes such that

i) $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$,

ii) $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

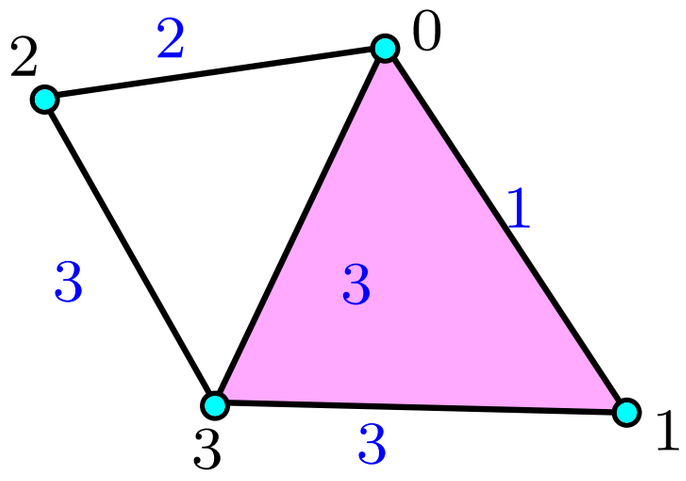
Example: filtration associated to a function



- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

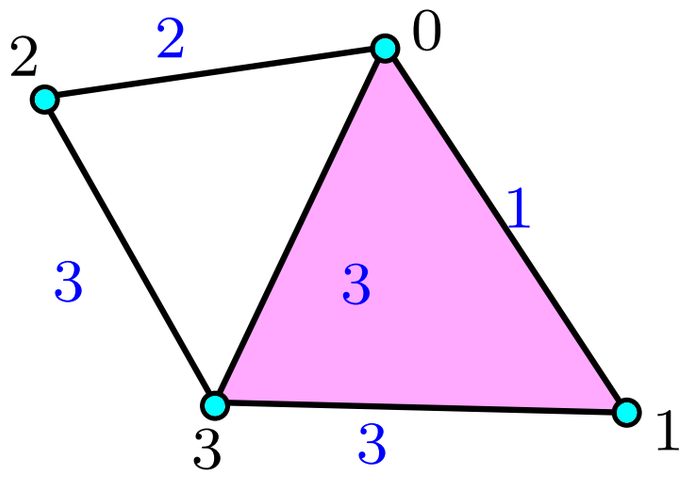
Example: filtration associated to a function



- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

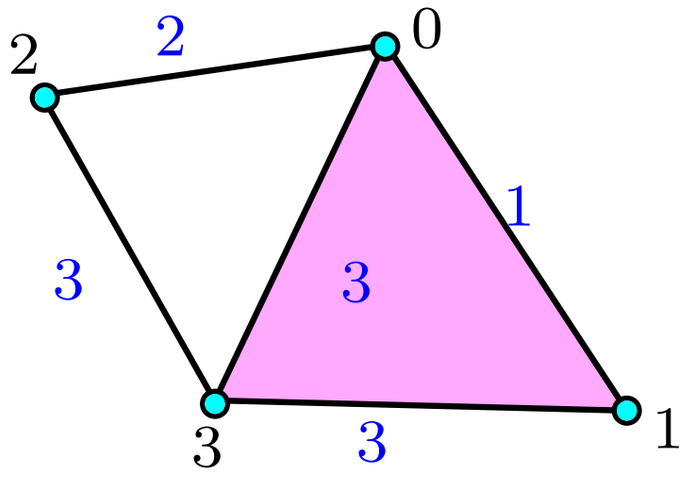
Example: filtration associated to a function



- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

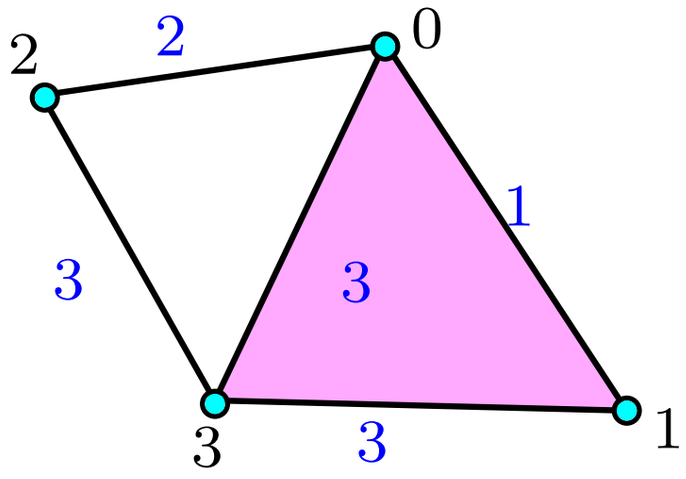
Example: filtration associated to a function

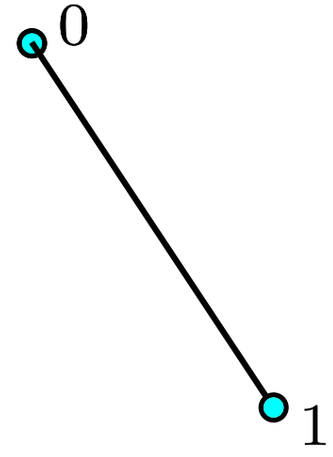


- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

Example: filtration associated to a function

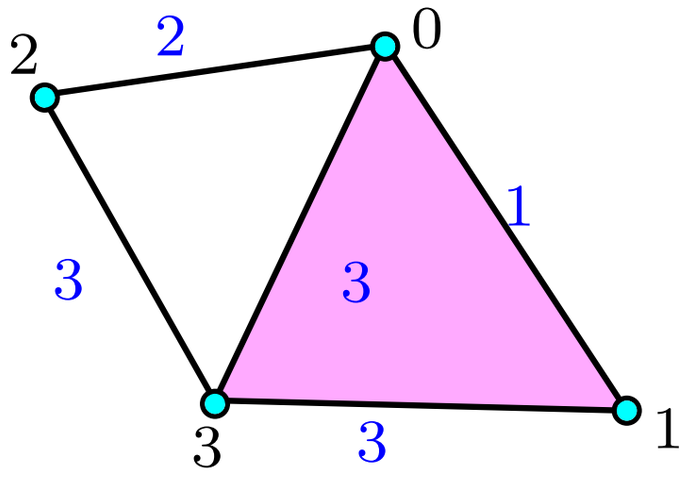


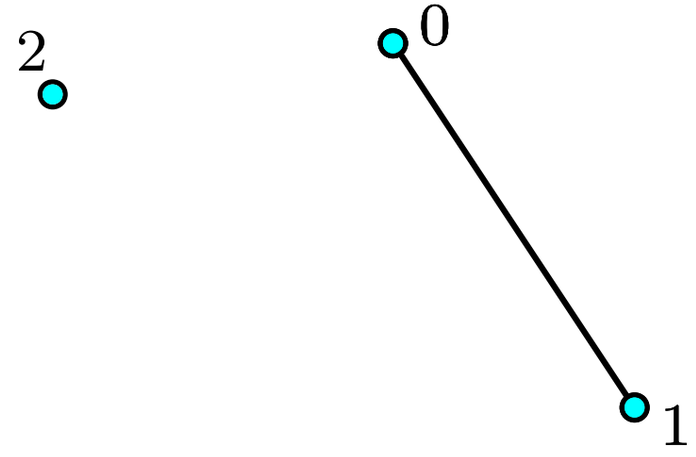


- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

Example: filtration associated to a function

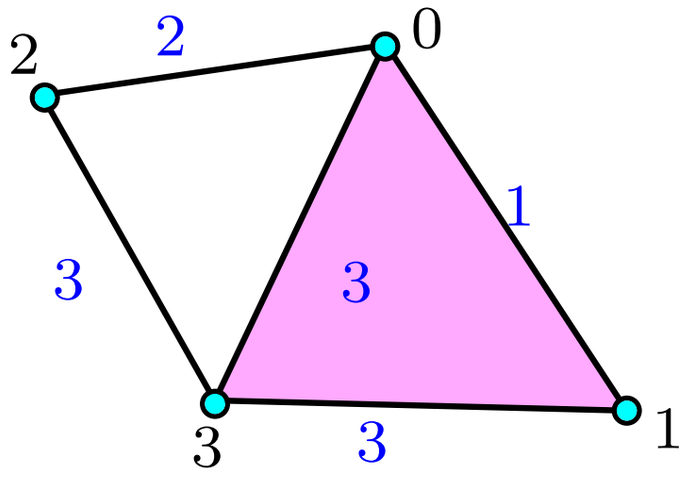


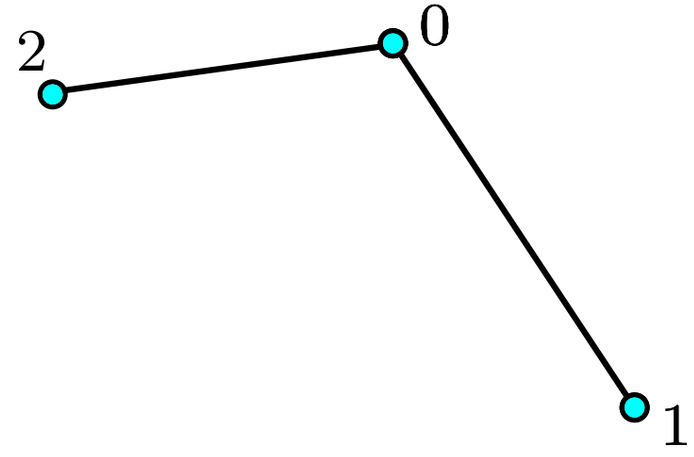


- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

Example: filtration associated to a function

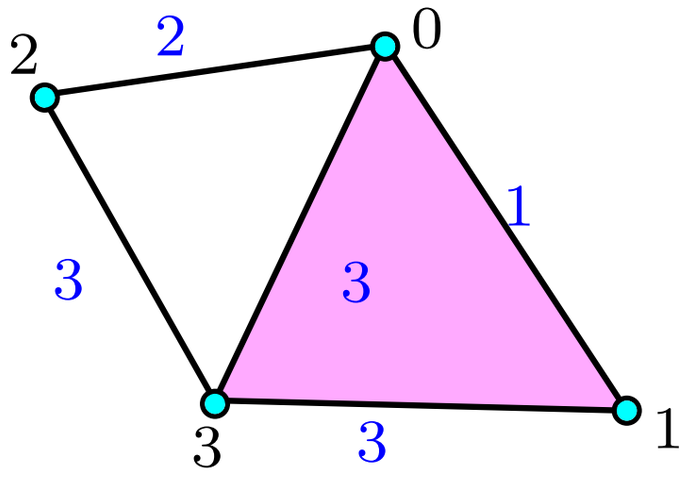


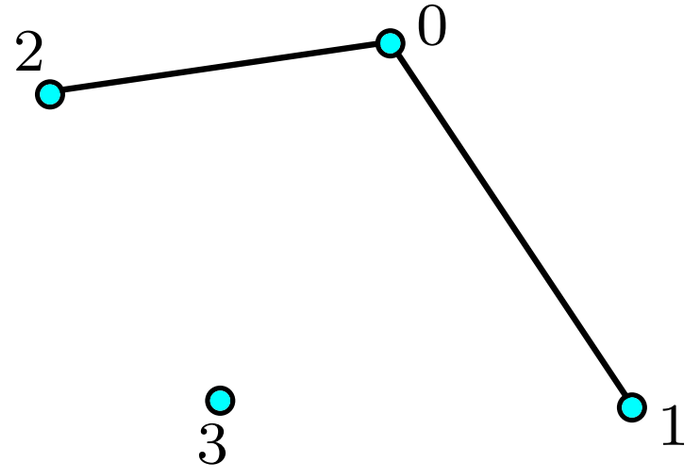


- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

Example: filtration associated to a function

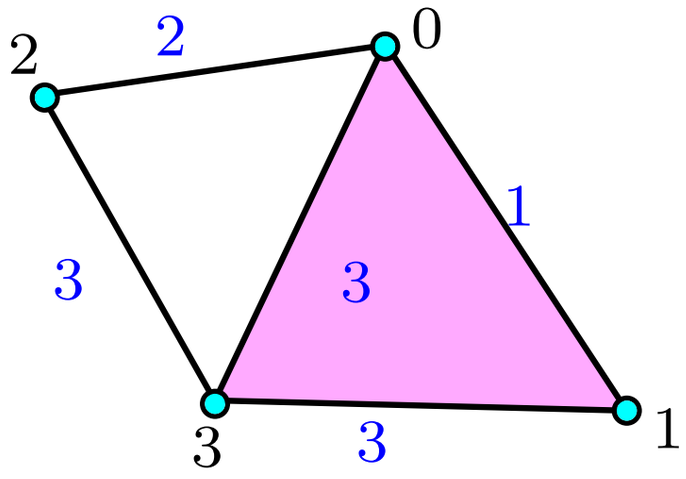


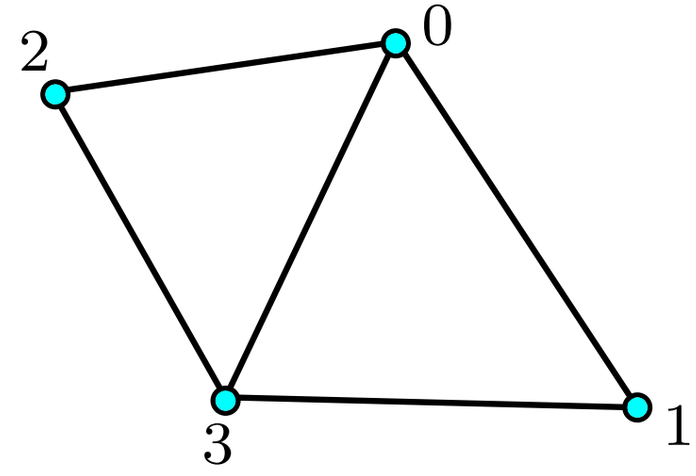


- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

Example: filtration associated to a function

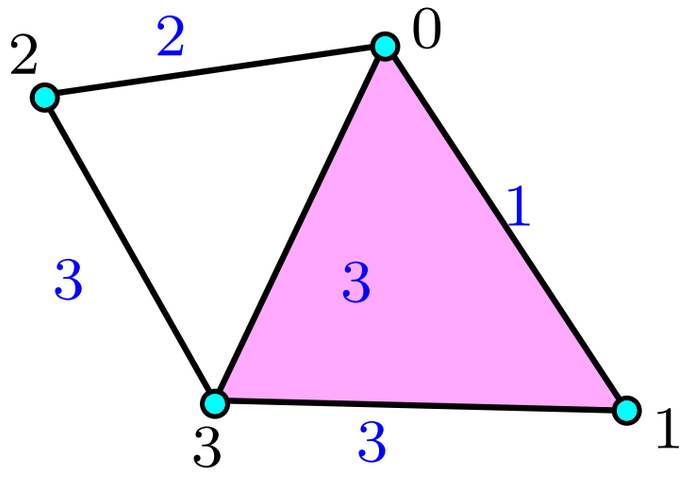


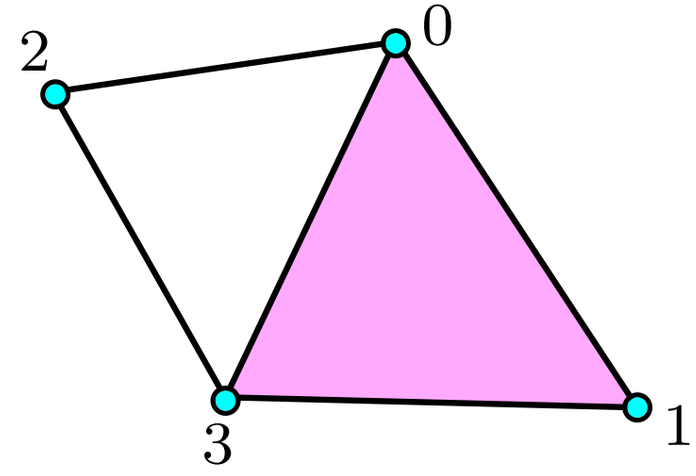


- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

Example: filtration associated to a function

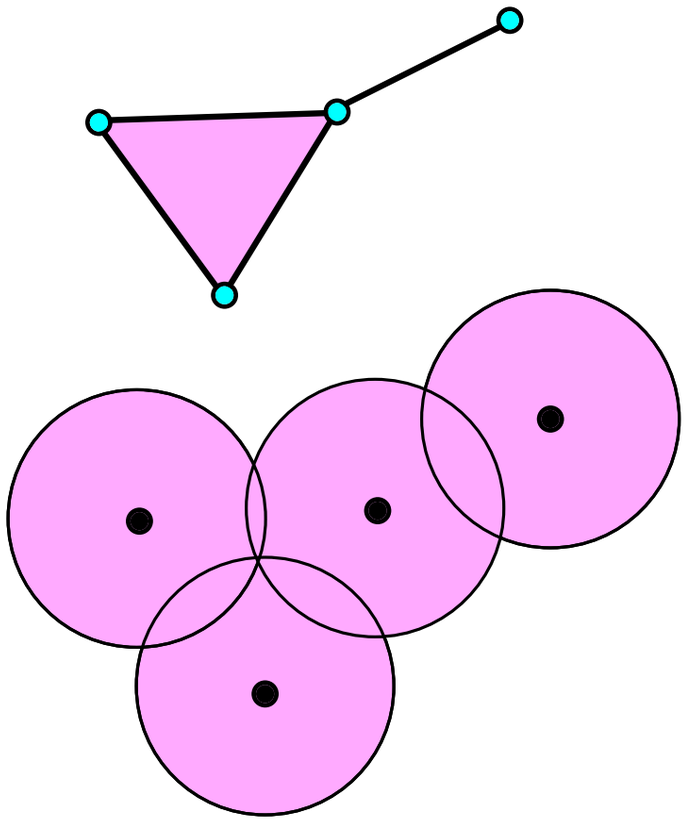


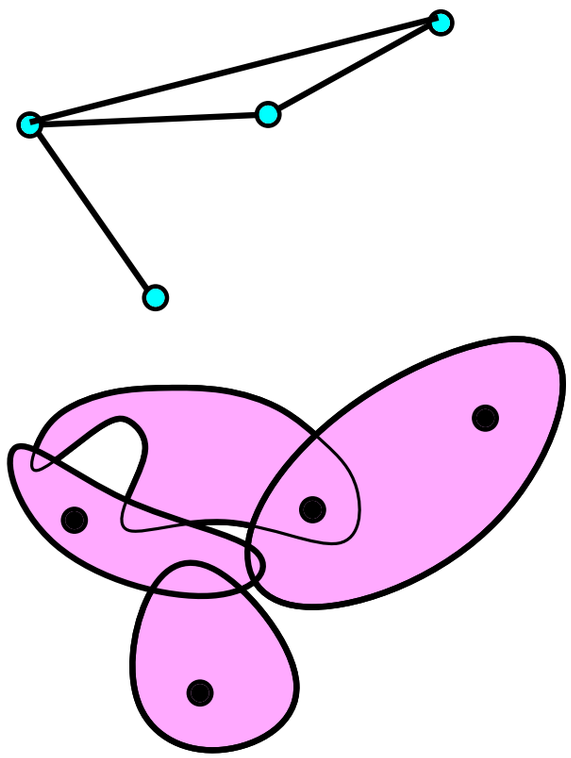


- f a real valued function defined on the vertices of K
- For $\sigma = [v_0, \cdots, v_k] \in K$, $f(\sigma) = \max_{i=0,\cdots,k} f(v_i)$
- The simplices of K are ordered according increasing f values (and dimension in case of equal values on different simplices).

$\Rightarrow$ The sublevel sets filtration.

Example: The Cěch complex

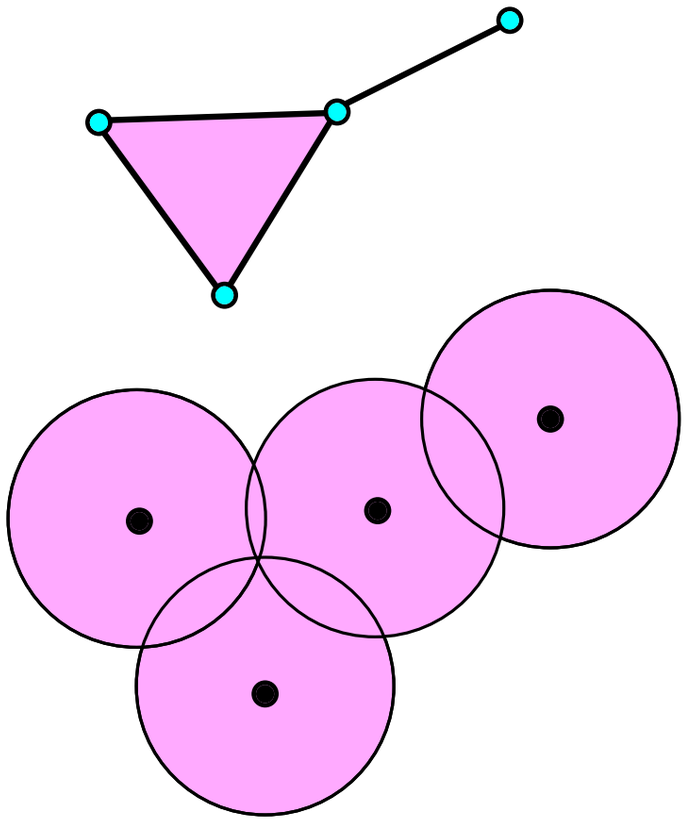


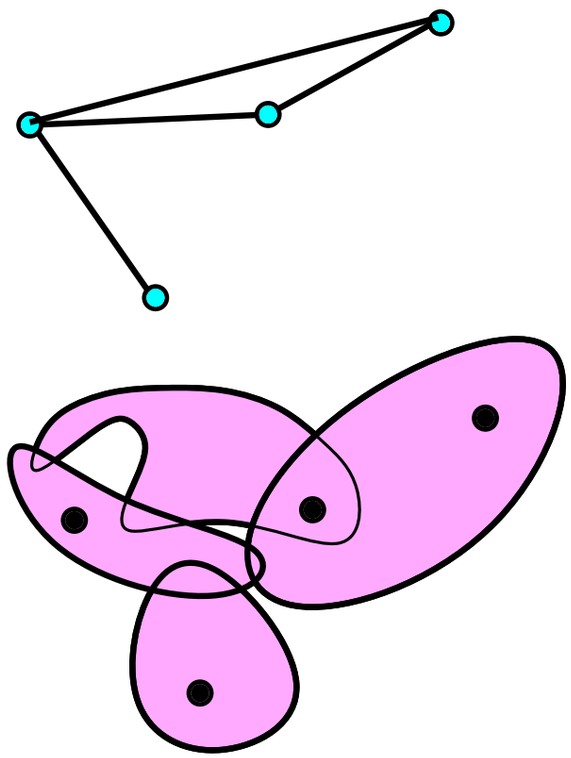


- Let $\mathcal{U} = (U_i)_{i \in I}$ be a covering of a topological space X by open sets: $X = \cup_{i \in I} U_i$.

- The Cěch complex $C(\mathcal{U})$ associated to the covering $\mathcal{U}$ is the simplicial complex defined by:
  - the vertex set of $C(\mathcal{U})$ is the set of the open sets U_i
  - $[U_{i_0}, \cdots, U_{i_k}]$ is a k-simplex in $C(\mathcal{U})$ iff $\cap_{j=0}^{k} U_{i_j} \neq \emptyset$.

Example: The Čech complex

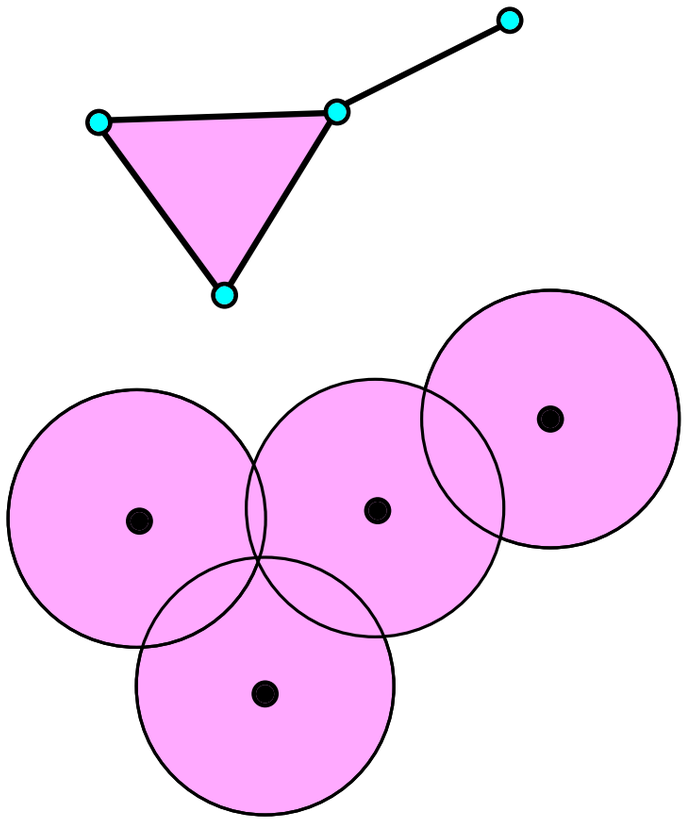


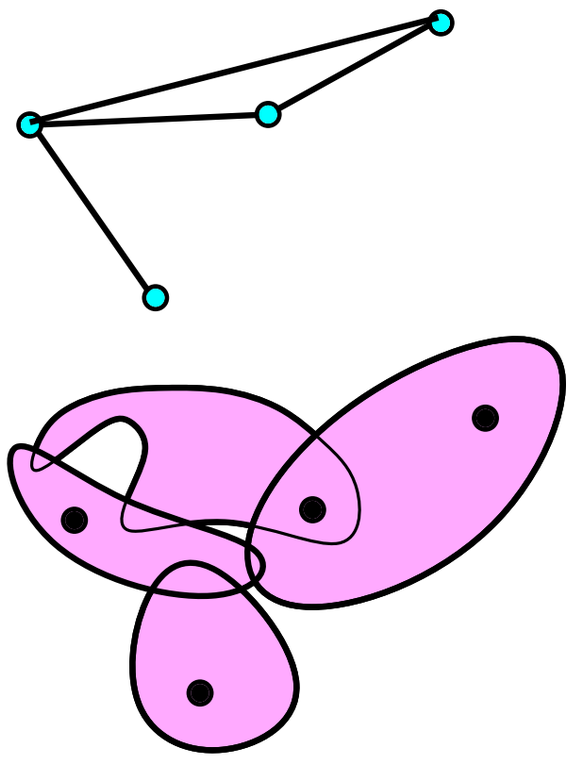


Nerve theorem: If all the intersections between opens in $\mathcal{U}$ are either empty or contractible then $C(\mathcal{U})$ and $X = \cup_{i \in I} U_i$ are homotopy equivalent.

$\Rightarrow$ The combinatorics of the covering (a simplicial complex) carries the topology of the space.

Example: The Cěch complex



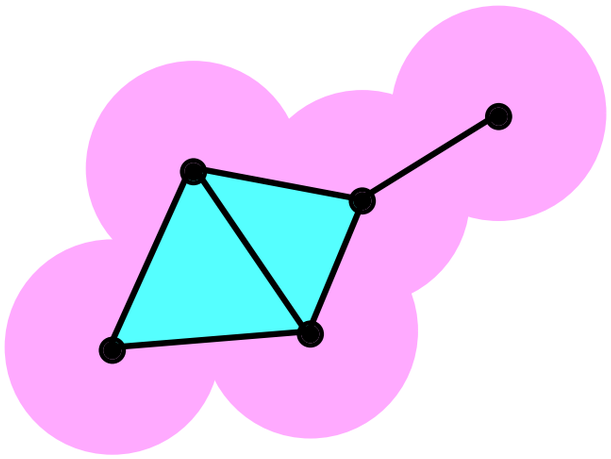


Nerve theorem: If all the intersections between opens in $\mathcal{U}$ are either empty or contractible then $C(\mathcal{U})$ and $X = \cup_{i \in I} U_i$ are homotopy equivalent.

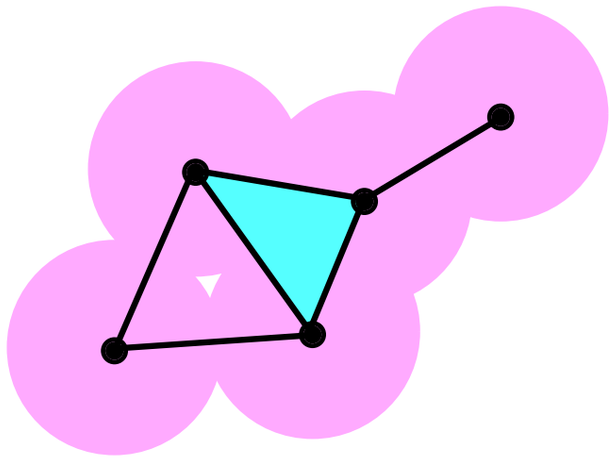
$\Rightarrow$ The combinatorics of the covering (a simplicial complex) carries the topology of the space.

Warning: even when the open sets are euclidean balls, the computation of the Cěch complex is a very difficult task!

Rips and Čech filtrations



Rips vs Čech



Let $L = \{p_0, \cdots p_n\}$ be a (finite) point cloud (in a metric space).

The Rips complex $\mathcal{R}^\alpha(L)$: for $p_0, \cdots p_k \in L$,

$$\sigma = [p_0 p_1 \cdots p_k] \in \mathcal{R}^\alpha(L) \text{ iff } \forall i, j \in \{0, \cdots k\},\ d(p_i, p_j) \leq \alpha$$

- Easy to compute and fully determined by its 1-skeleton
- Rips-Čech interleaving: for any $\alpha > 0$,

$$\mathcal{C}^{\frac{\alpha}{2}}(L) \subseteq \mathcal{R}^\alpha(L) \subseteq \mathcal{C}^\alpha(L) \subseteq \mathcal{R}^{2\alpha}(L) \subseteq \cdots$$

Persistent homology of filtered complexes

Let $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$ be a filtration of a simplicial complex K s. t. $K^{i+1} = K^i \cup \sigma^{i+1}$ where σ^{i+1} is a simplex of K.

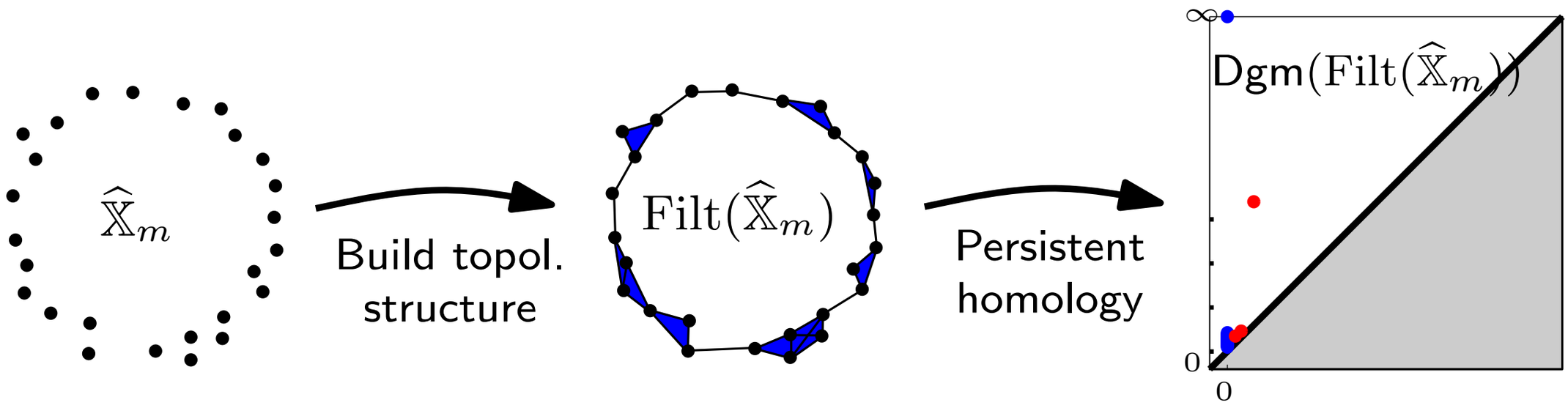
Algorithm to compute the Betti numbers $\beta_0, \beta_1, \cdots, \beta_d$ of K:

$\beta_0 = \beta_1 = \cdots = \beta_d = 0$;
for $i = 1$ to m
  $k = \dim \sigma^i - 1$;
  if σ^i is contained in a $(k+1)$-cycle in K^i
    then $\beta_{k+1} = \beta_{k+1} + 1$;
    else $\beta_k = \beta_k - 1$;
  end if;
end for;
output $(\beta_0, \beta_1, \cdots, \beta_d)$;

Adapt the algorithm to keep track of an homology basis and pairs positive simplices (birth of a new homological class) to negative simplices (death of an existing homology class).

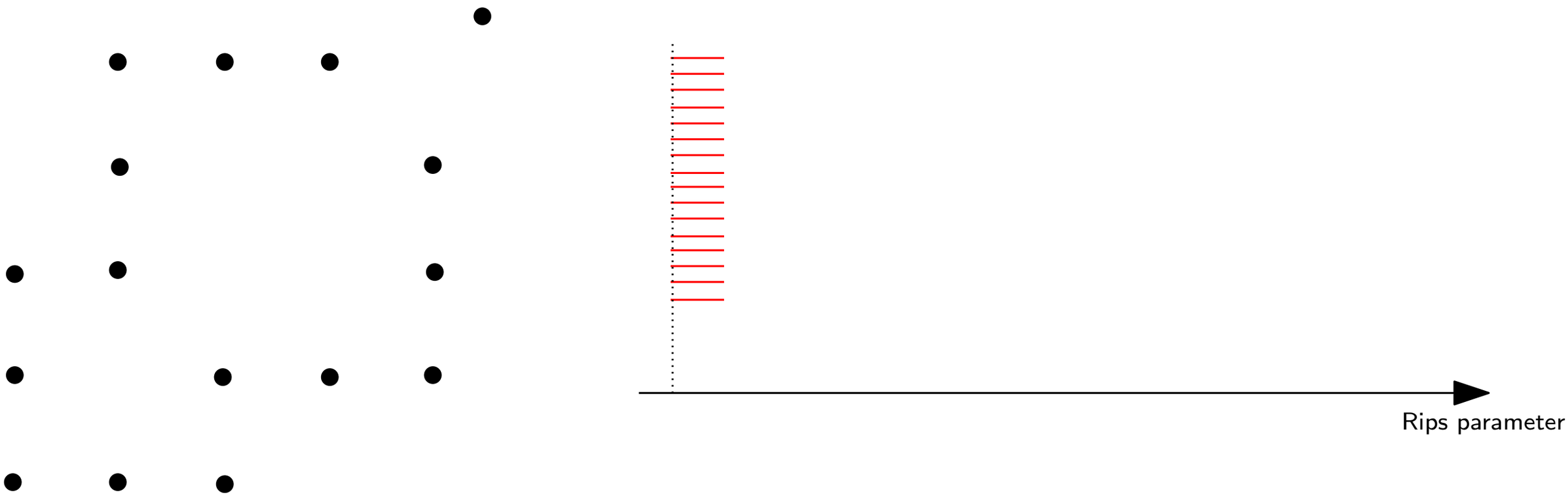
Notation: $H_k^i = H_k(K^i)$

Persistent homology for (point cloud) data



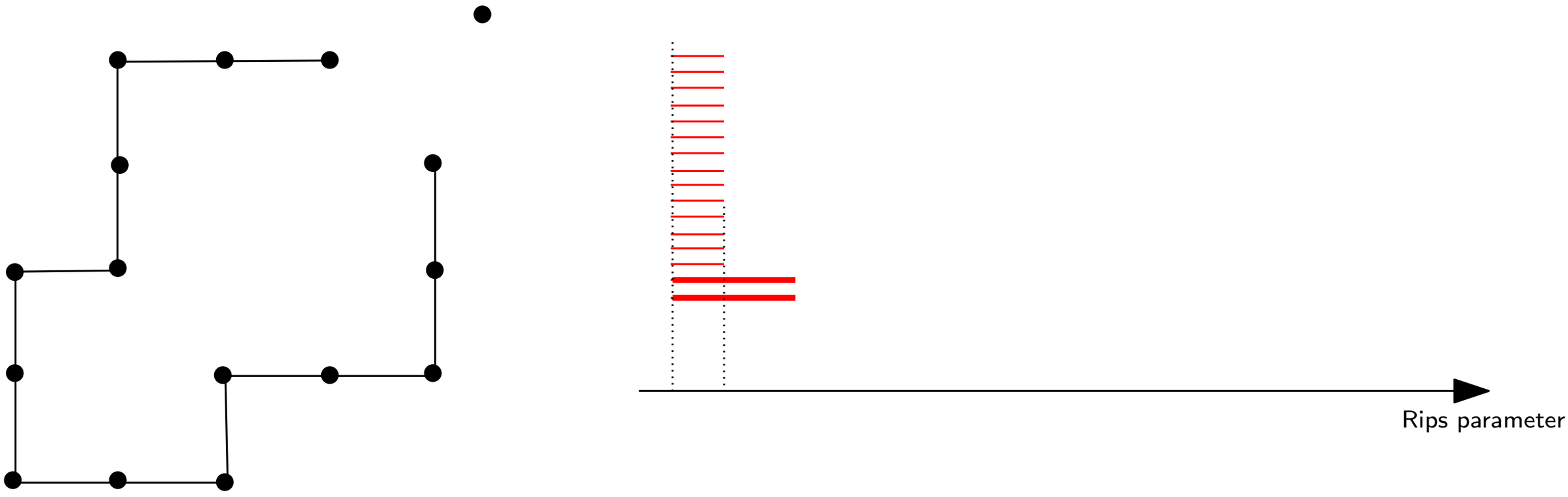
- Build a geometric filtered simplicial complex on top of $\widehat{\mathbb{X}}_m \rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures

Persistent homology of filtered complexes



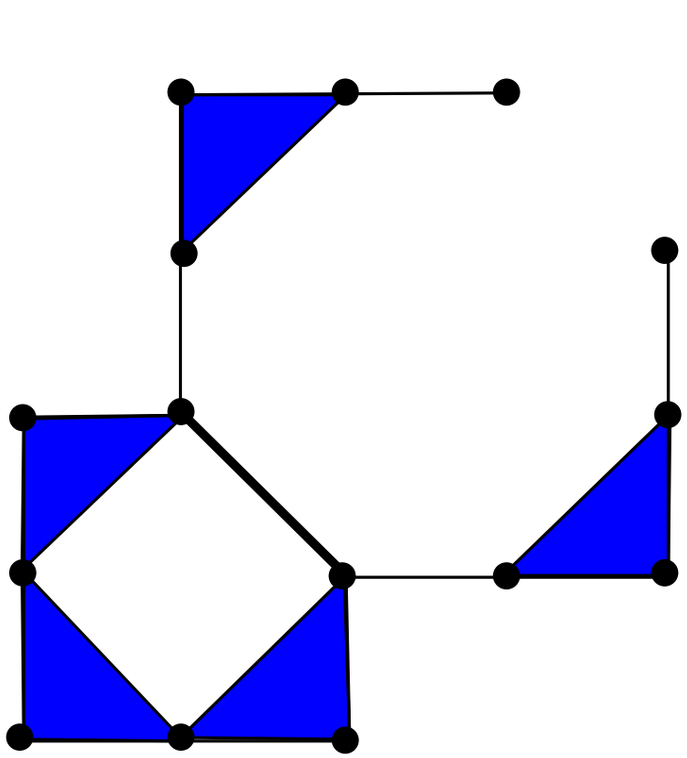
- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

Persistent homology of filtered complexes



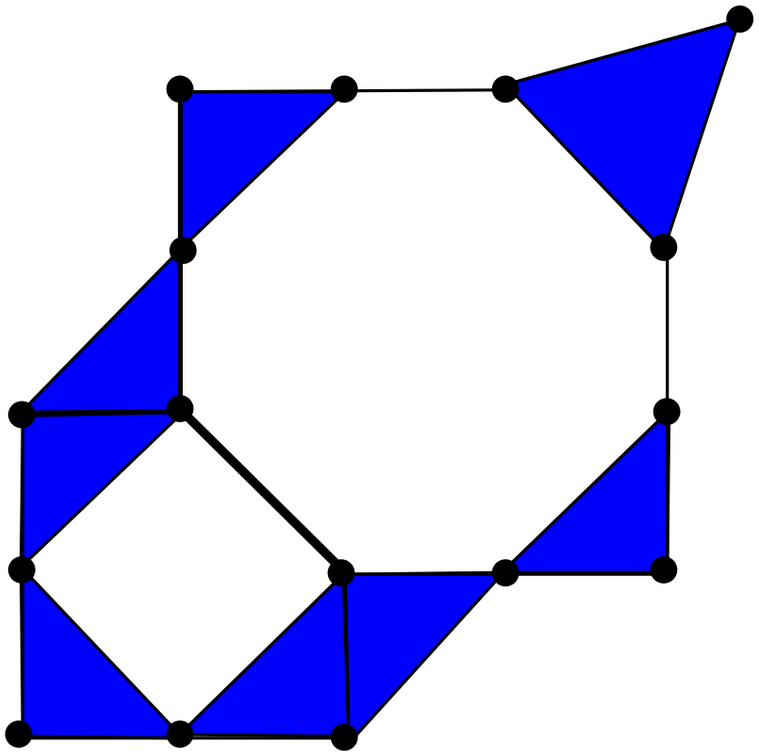
- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

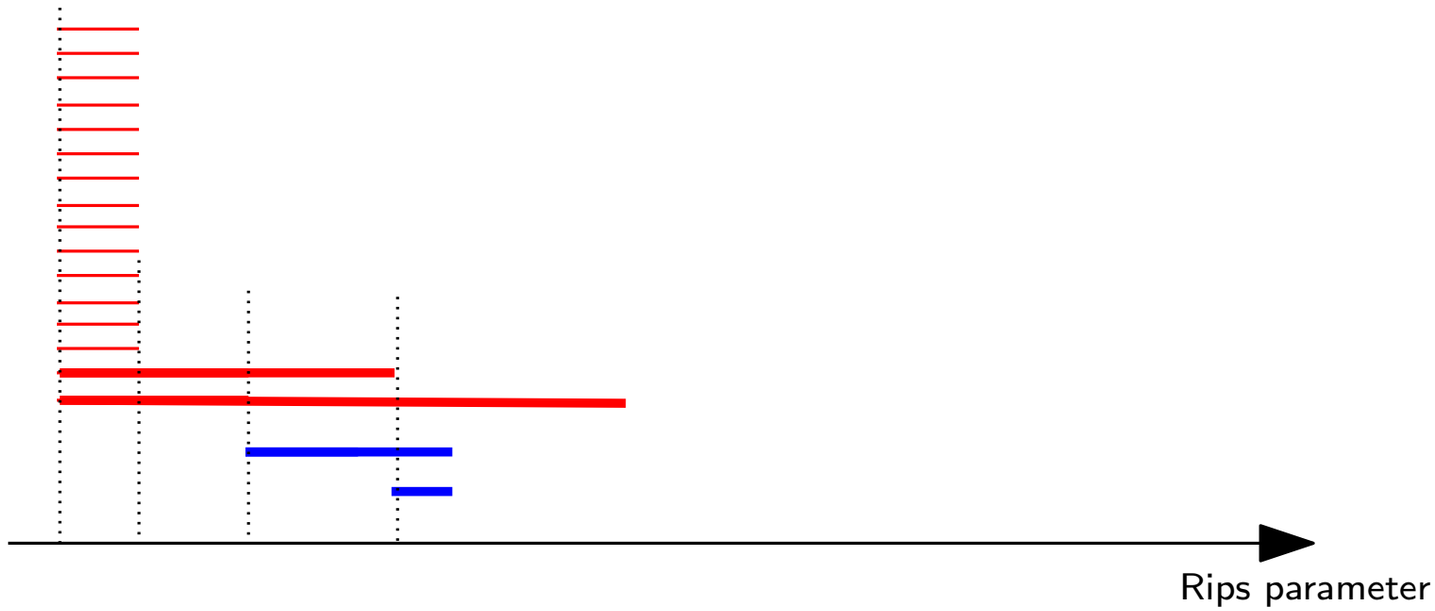
Persistent homology of filtered complexes



- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

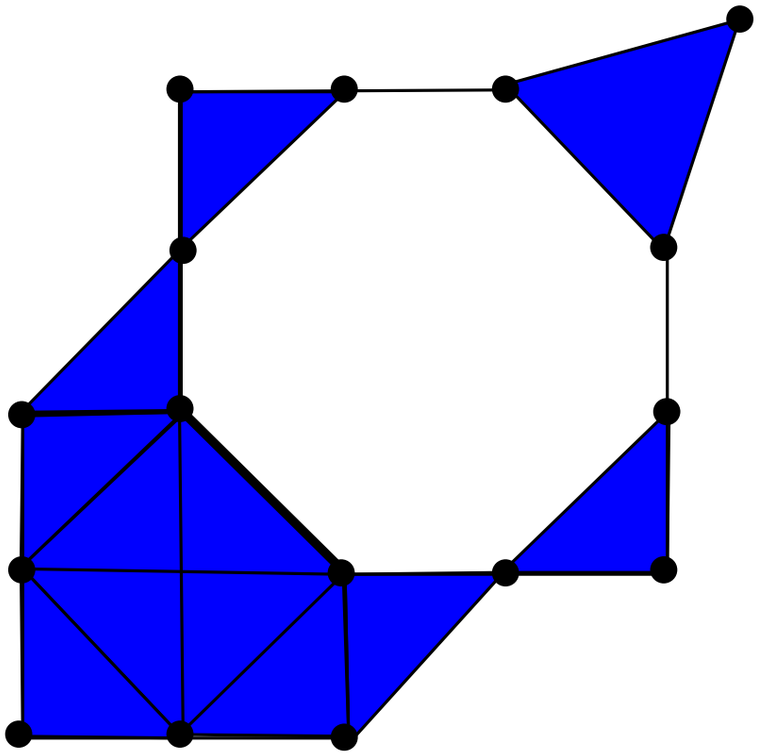
Persistent homology of filtered complexes

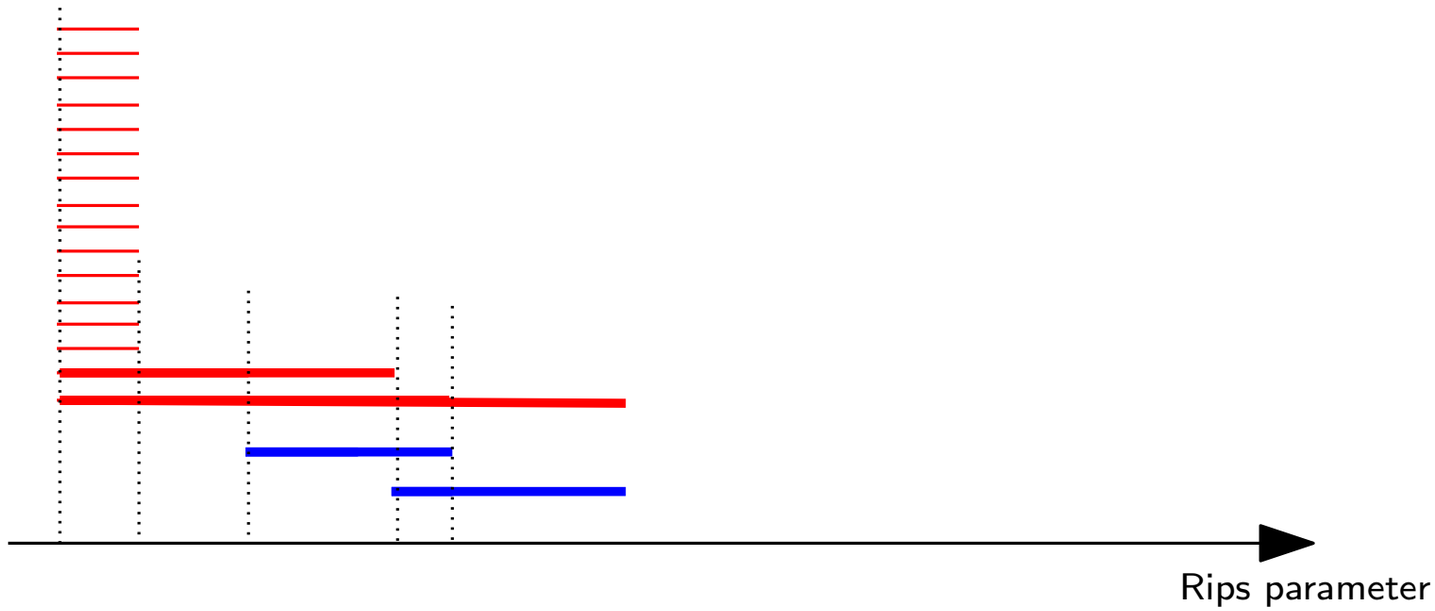




- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

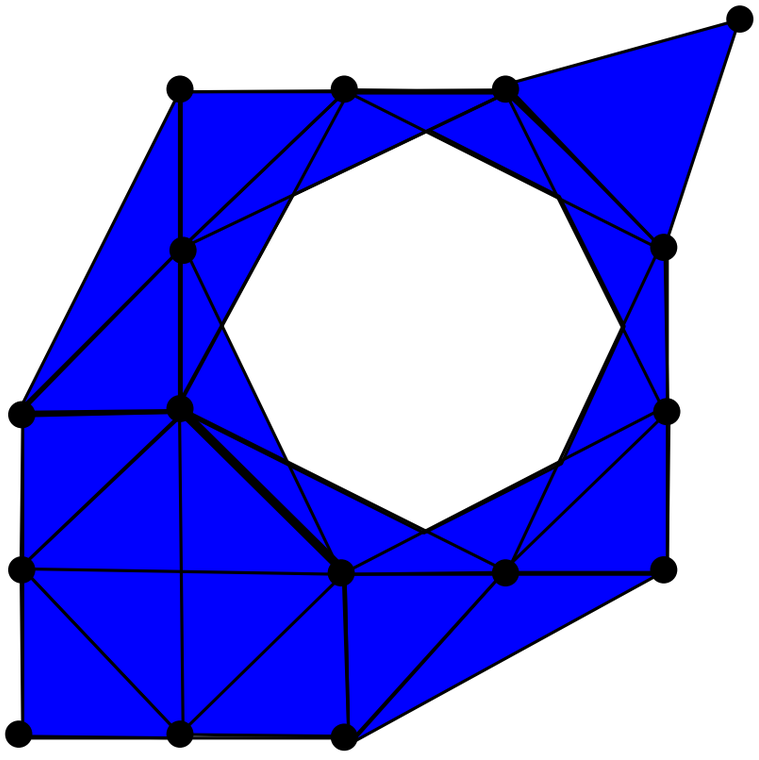
Persistent homology of filtered complexes

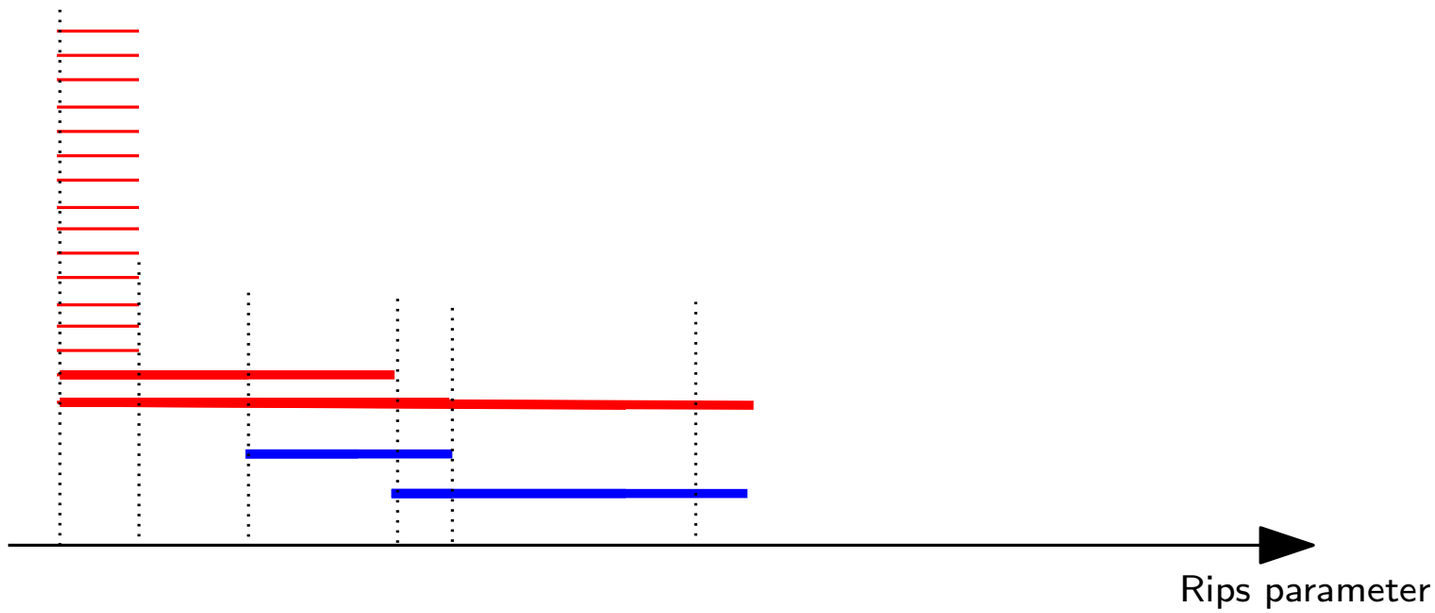




- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

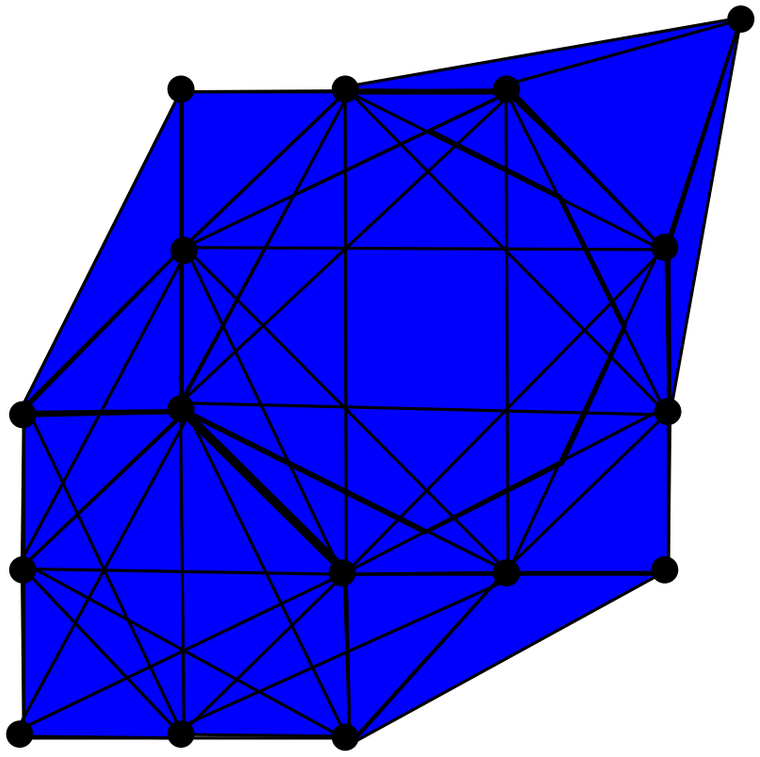
Persistent homology of filtered complexes

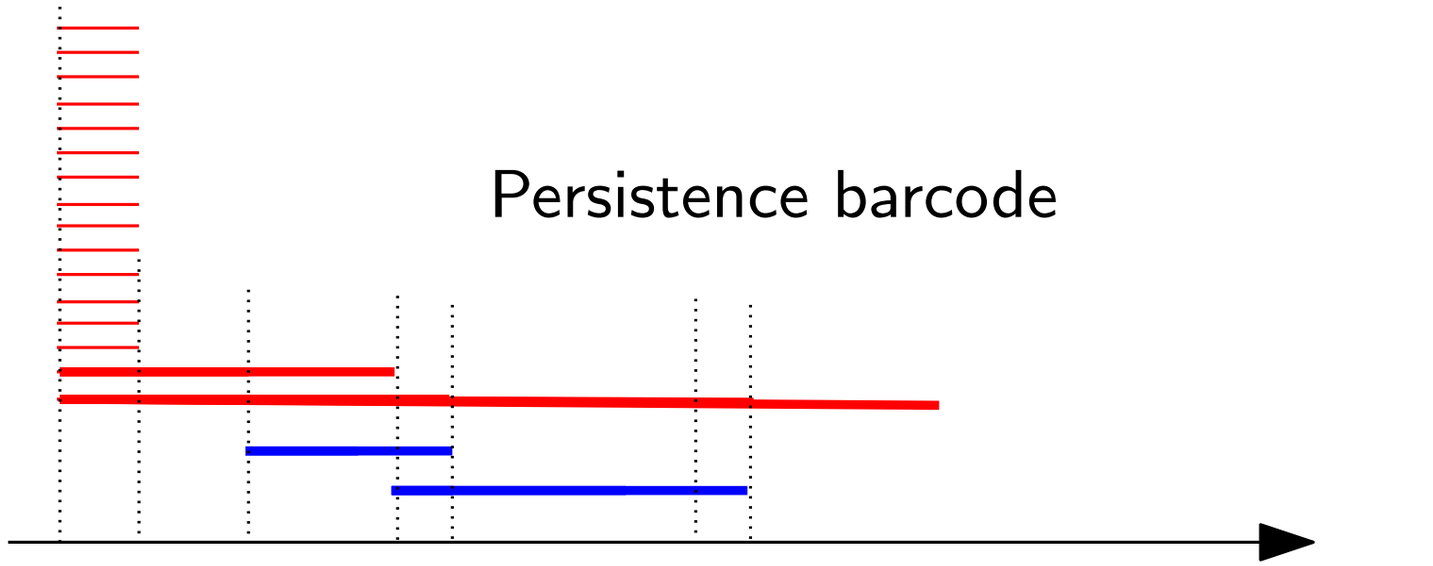


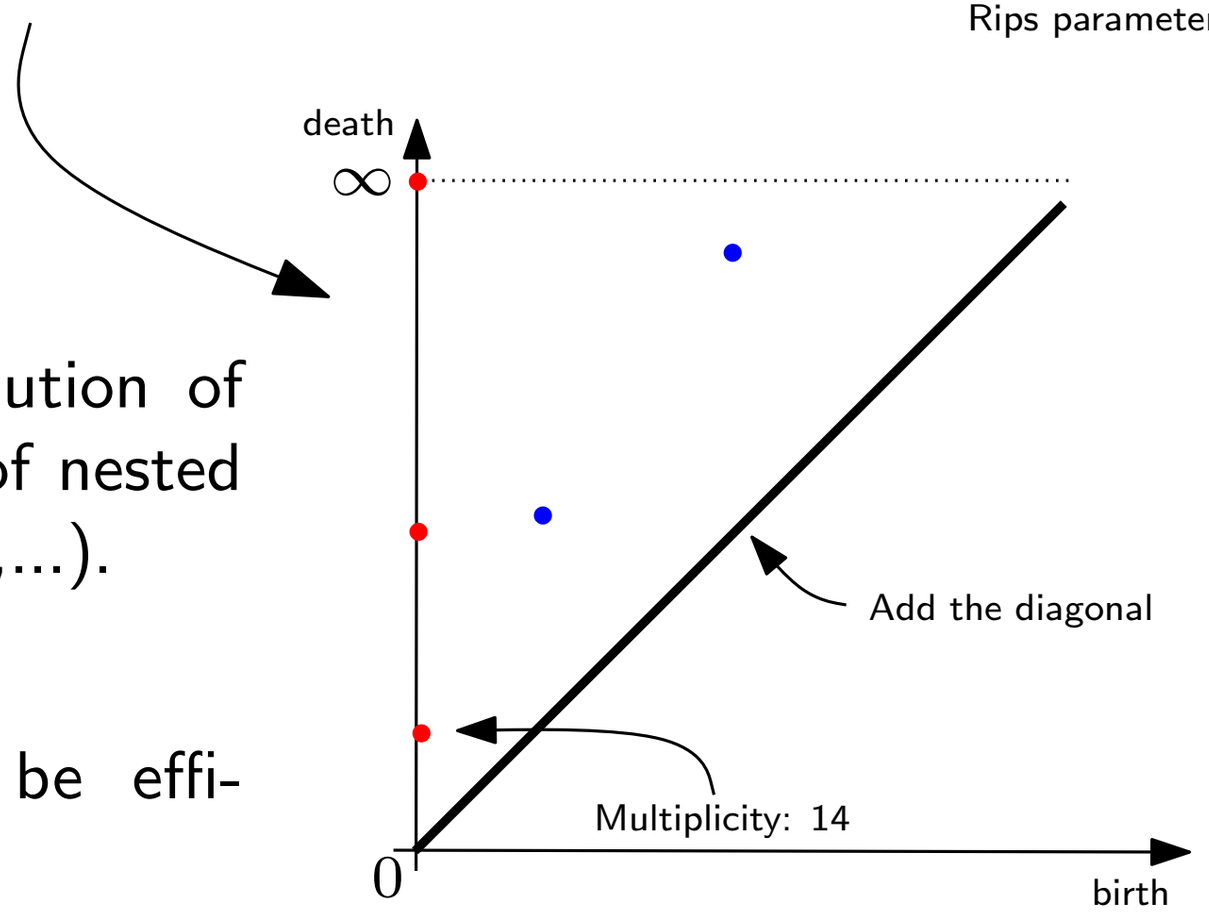


- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

Persistent homology of filtered complexes







Persistence diagram

- An efficient way to encode the evolution of the topology (homology) of families of nested spaces (filtered complex, sublevel sets,...).
- Multiscale topological information.
- Barcodes/persistence diagrams can be efficiently computed.
- Stability properties

Stability properties

"Stability theorem": Close spaces/data sets have close persistence diagrams!

[C., de Silva, Oudot - Geom. Dedicata 2013].

If $\mathbb{X}$ and $\mathbb{Y}$ are pre-compact metric spaces, then

$$\mathrm{d}_\infty(\mathrm{Dgm}(\mathrm{Rips}(\mathbb{X})), \mathrm{Dgm}(\mathrm{Rips}(\mathbb{Y}))) \leq d_{GH}(\mathbb{X}, \mathbb{Y}).$$

Bottleneck distance

Gromov-Hausdorff distance

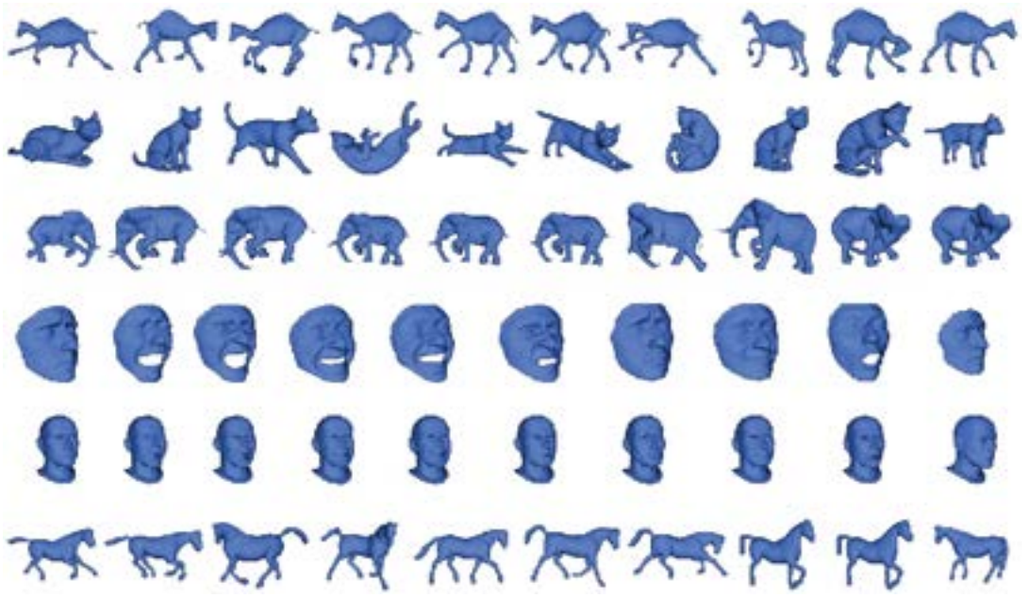
$$d_{GH}(\mathbb{X}, \mathbb{Y}) := \inf_{\mathbb{Z}, \gamma_1, \gamma_2} d_H(\gamma_1(\mathbb{X}), \gamma_2(\mathbb{Y}))$$

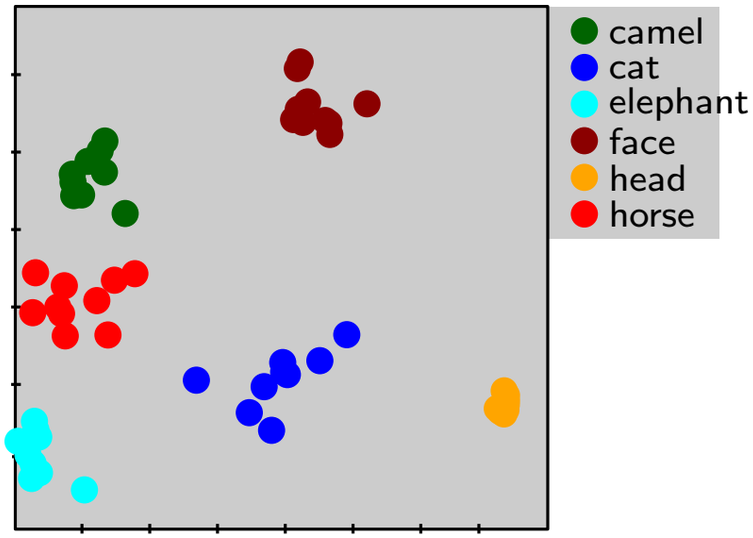
$\mathbb{Z}$ metric space, $\gamma_1 : \mathbb{X} \to \mathbb{Z}$ and $\gamma_2 : \mathbb{Y} \to \mathbb{Z}$ isometric embeddings.

Rem: This result also holds for other families of filtrations (particular case of a more general thm).

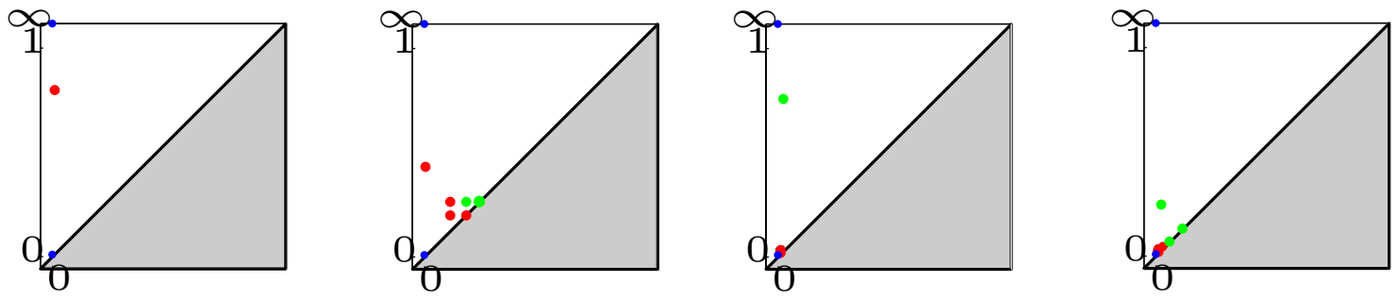
Application: non rigid shape classification

[C., Cohen-Steiner, Guibas, Mémoli, Oudot - SGP '09]



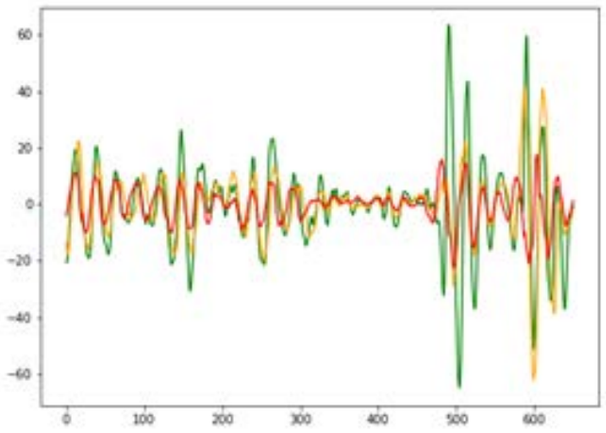


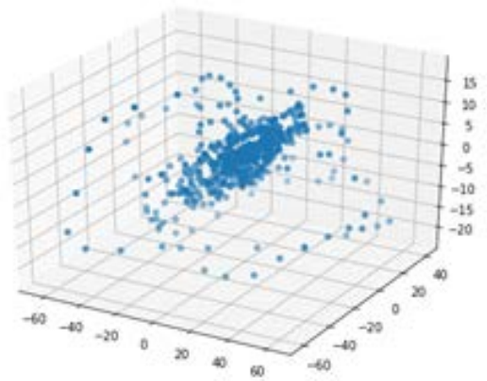
MDS using bottleneck distance.

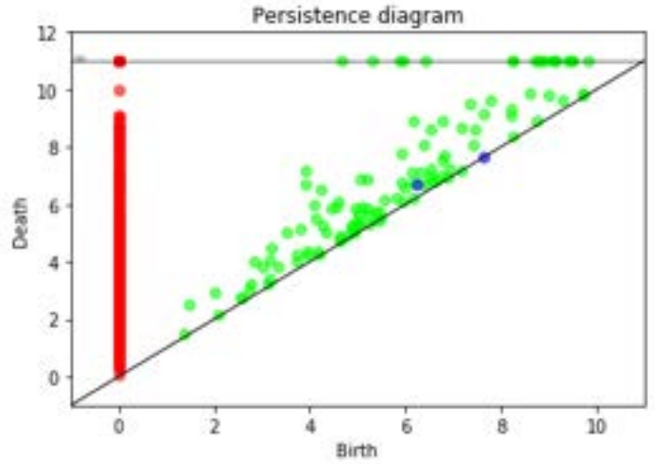


- Non rigid shapes in a same class are almost isometric, but computing Gromov-Hausdorff distance between shapes is extremely expensive.
- Compare diagrams of sampled shapes instead of shapes themselves.

The problem of representation of persistence





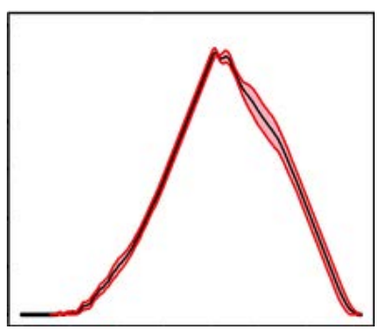


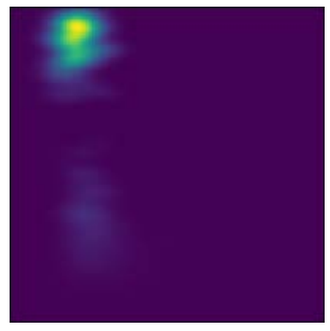
Persistence diagrams are not well-suited for classical ML algorithms (the space of PD is highly non linear)

Not always clear which part of the diagrams carries the relevant information.

An active research area, but still at a very early stage (in particular regarding math. aspects).

Machine Learning / AI





Representations of persistence

A zoo of representations of persistence

(non exhaustive list)

- Collections of 1D functions

  → landscapes [Bubenik 2012]

  → Betti curves [Umeda 2017]

- discrete measures: (interesting statistical properties)

  → persistence images [Adams et al 2017]

  → convolution with Gaussian kernel [Reininghaus et al. 2015] [Chepushtanova et al. 2015] [Kusano Fukumisu Hiraoka 2016-17] [Le Yamada 2018]

  → sliced on lines [Carrière Oudot Cuturi 2017]

  → quantization [C. Ike Royer Umeda 2019]

- finite metric spaces [Carrière Oudot Ovsjanikov 2015]
- polynomial roots or evaluations [Di Fabio Ferri 2015] [Kališnik 2016]
- Neural networks [Carrière, C. Ike, Lacombe, Royer, Umeda 2019]

6 - Topological Data Analysis and Statistics

Statistics, Learning and TDA

A **statistical approach to TDA** means that :

- we consider data as generated from an unknown distribution
- the inferred topological features by TDA methods are seen as estimators of topological quantities describing an underlying object.

Statistics, Learning and TDA

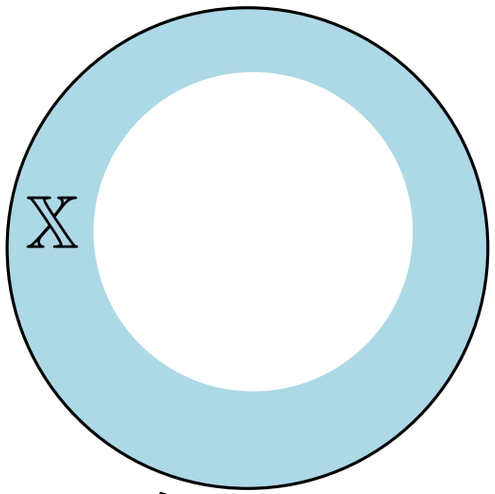
Contributions about statistical TDA (**non-exhaustive** list):

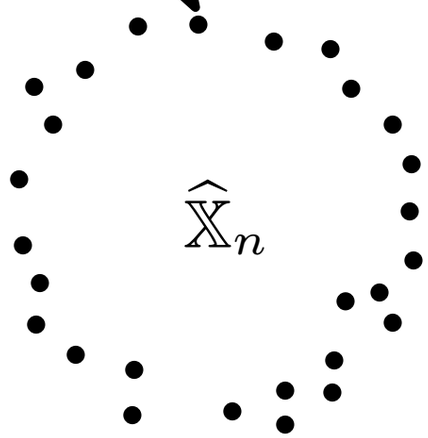
- Consistency / convergence of TDA methods: [Chazal15 JMLR], [Bobrowski 17 Bernouilli]
- Confidence regions for TDA [Fasy 14 AoS] [Chazal 15 JOCG]
- Central tendency for persistent homology [Turner 14 DCG] [Fasy15 Nips]
- Asymptotic normality of persistent Betti numbers [Krebs 2019]
- Robust methods for TDA [Chazal 17, EJS Chazal 17 JMLR]
- Bayesian Statistics for TDA [Maroulas et al. 2019]
- Representations of persistence in Euclidean spaces [Bubenik15 JMLR] [Adams15]
- Develop kernels for topological descriptors [Reininghaus 15 IEEE] [Carriere 17 ICML]
- Statistical analysis of Mapper [Carriere 18, Brown 19]
- ...

Persistent homology Inference on metric spaces

[Chazal et al. 2015]

$(\mathbb{M}, \rho)$ metric space
$\mathbb{X}$ compact set in $\mathbb{M}$.

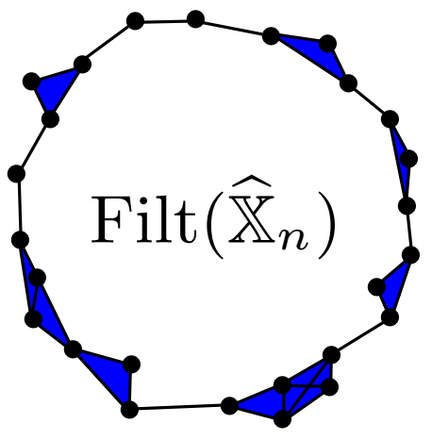




n points sampled in $\mathbb{X}$ according to μ

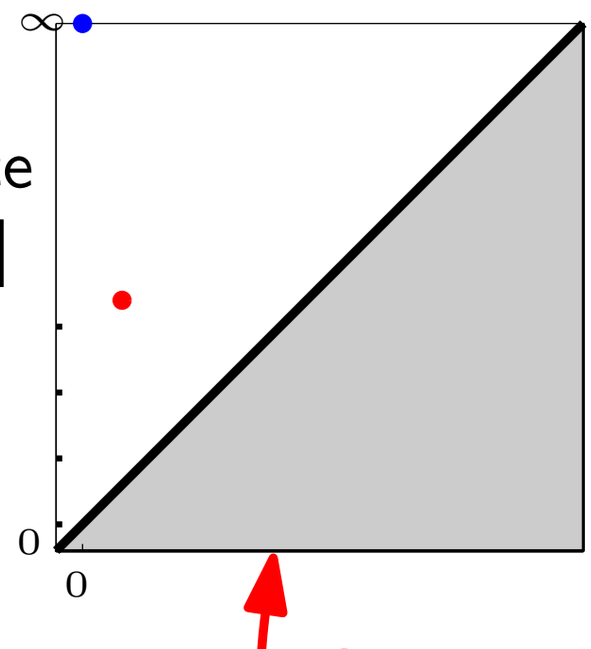
well-defined for any compact metric space [Chazal et al., 2012]

$\mathrm{Filt}(\mathbb{X})$

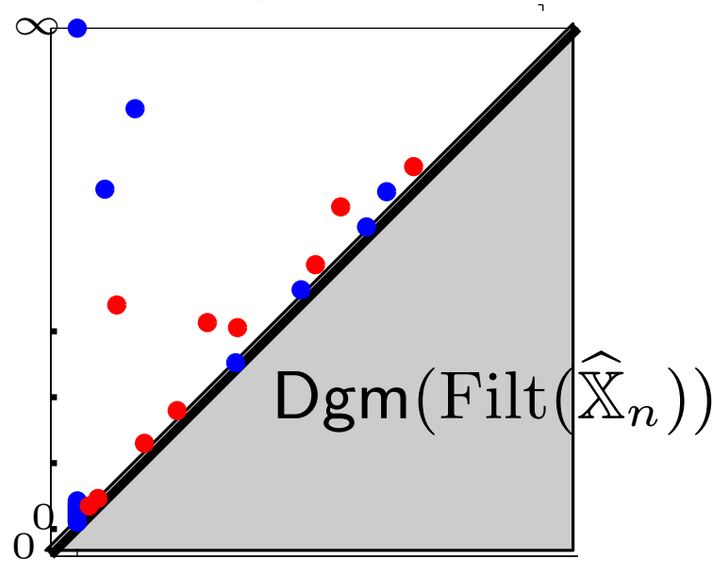


Estimator of $\mathrm{Dgm}(\mathrm{Filt}(K))$

$\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}))$



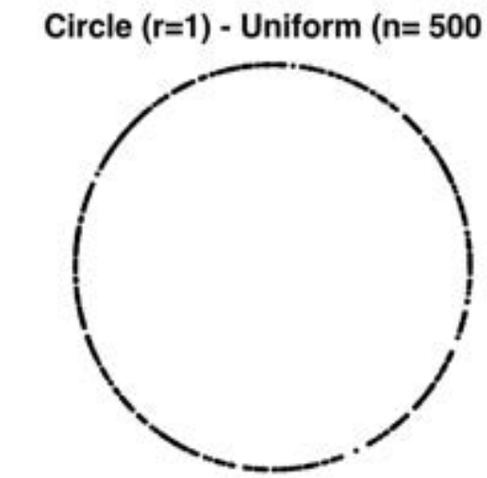
Convergence ???

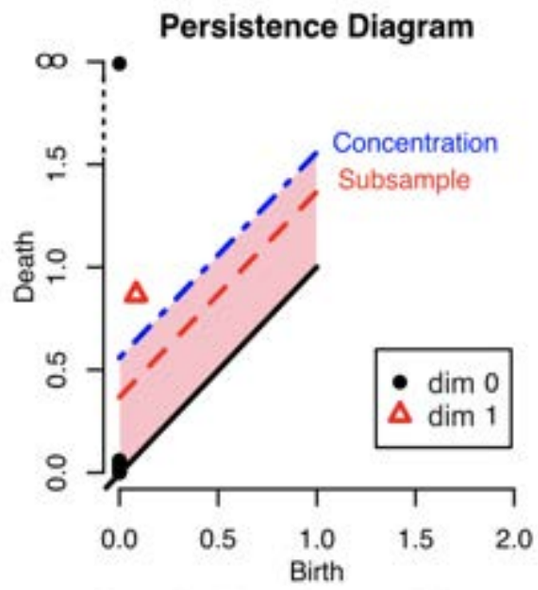


Persistent Homology Inference for upper level sets

Point cloud : $X_1, \dots X_n \sim f$ i.i.d. in $\mathbb{R}^p$.

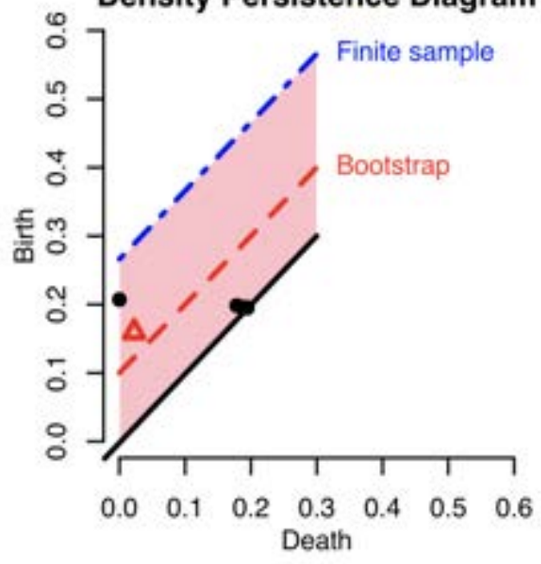
[Fasy et al. 2014] and [Bobrowski et al. 2017] : estimation of the persistence diagram for the upper level sets of densities using kernel estimators.



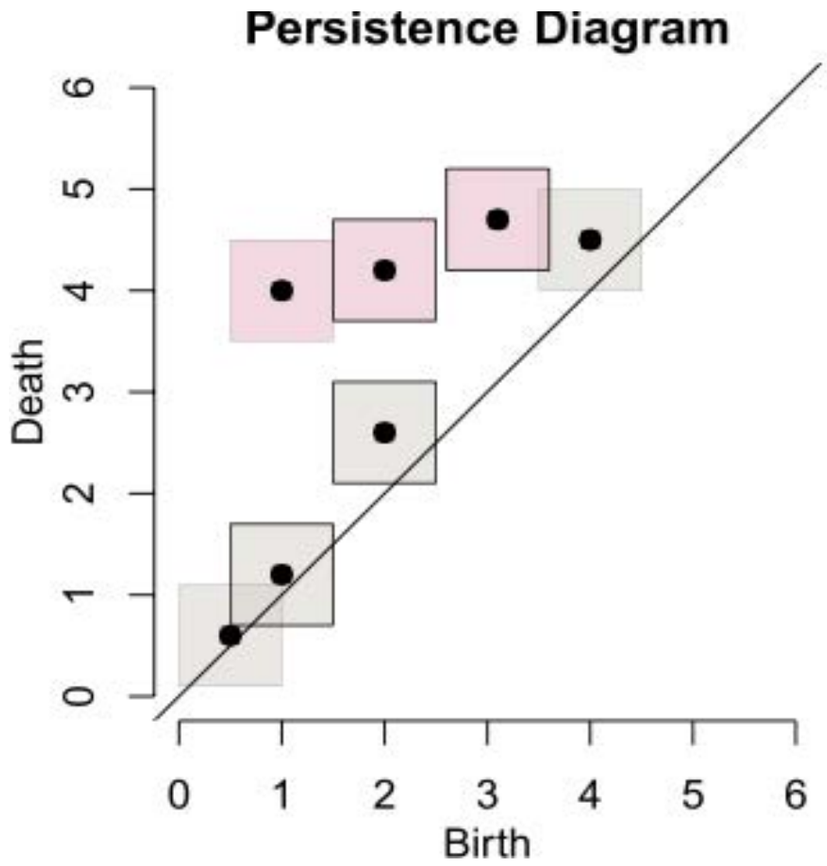


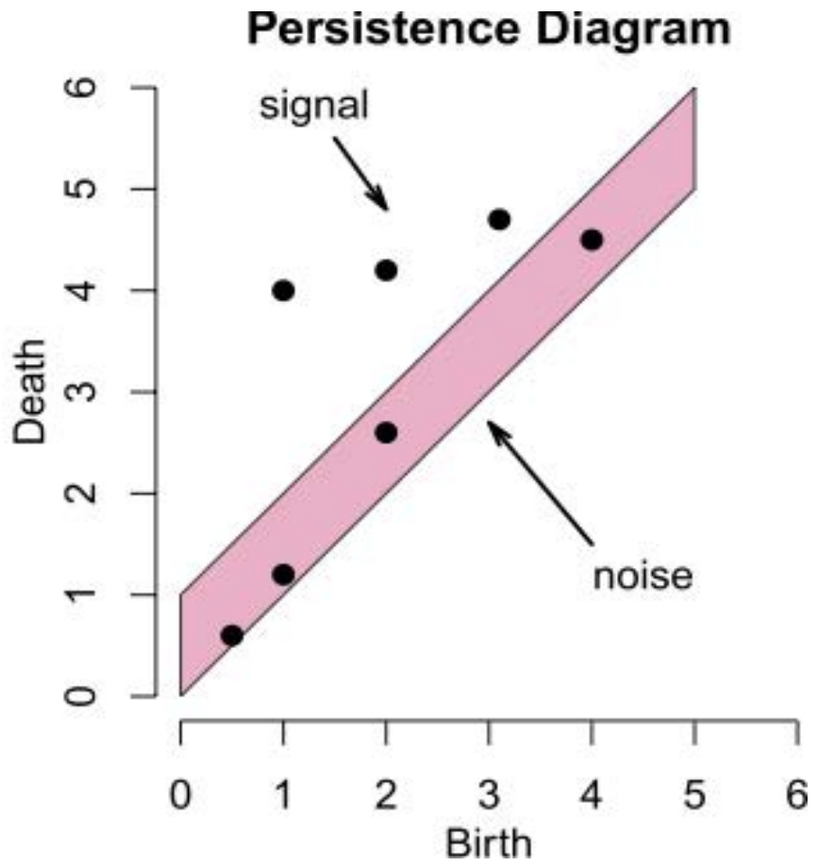
Kernel Density Estimator (h= 0.3)

Density Persistence Diagram



Confidence sets for persistence diagrams

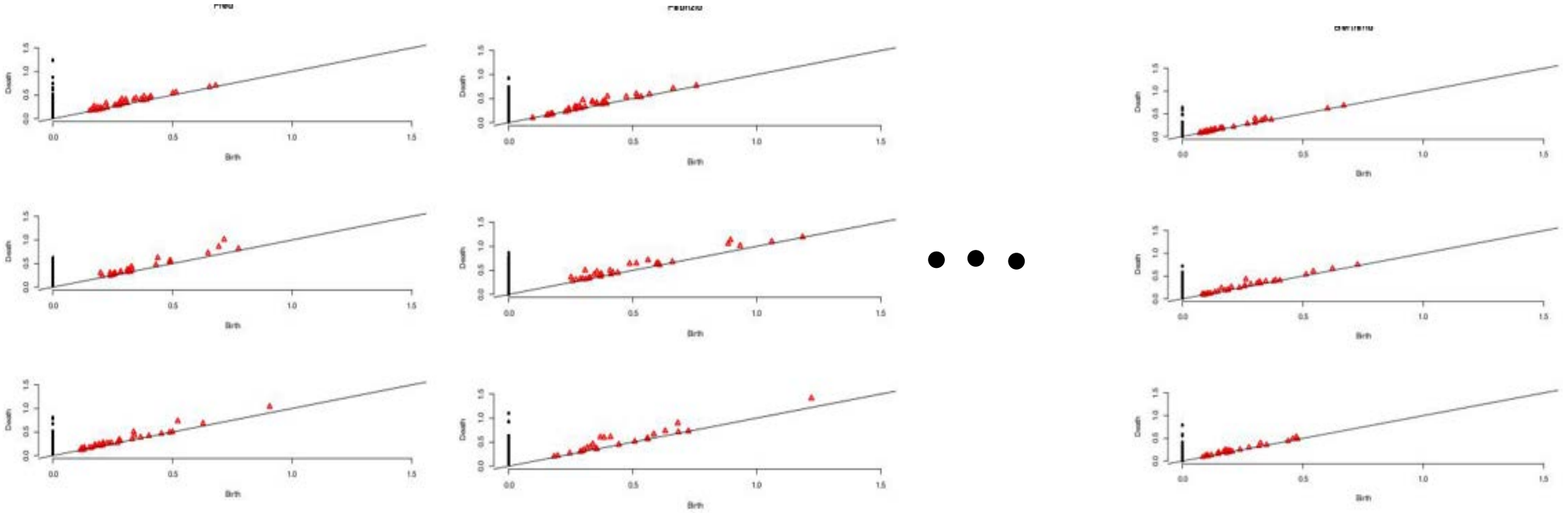




- Confidence sets [Fasy et al. 2014] :
  - For a compact manifold
  - For the upper level sets of a density function

- Various bootstrap and subsampling strategies.
  Bottleneck bootstrap : Chazal etal. 2017.

Central tendency for persistent homology ?



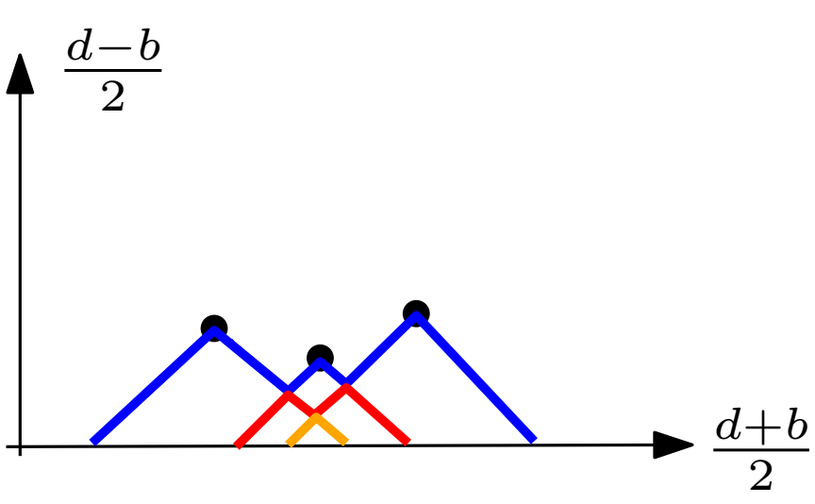
- Frechet mean [Turner etal. 2014] : difficult to compute and no unicity.

- Use an alternative descriptor of persistence using :
  - functions
  - discrete measures.

Functional Representations of Persistence Homology

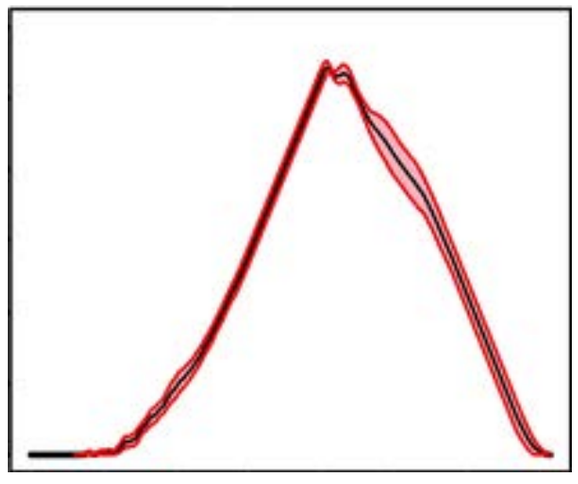
Berry etal. 2020

A functional summary of a persistence diagram is a map between the space of diagrams and a functional space $\mathcal{F}$.



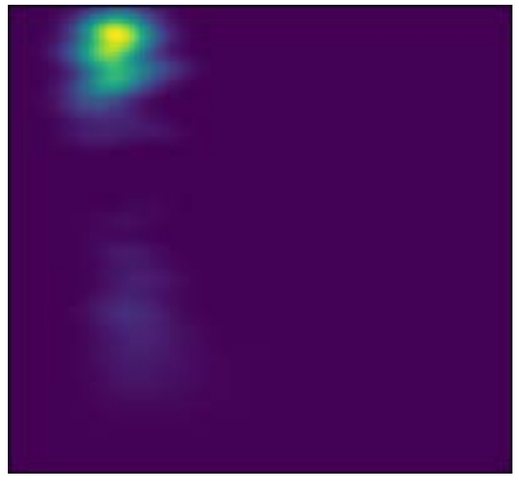
persistence landscape

[Bubenik 2015]



Persistent silhouette

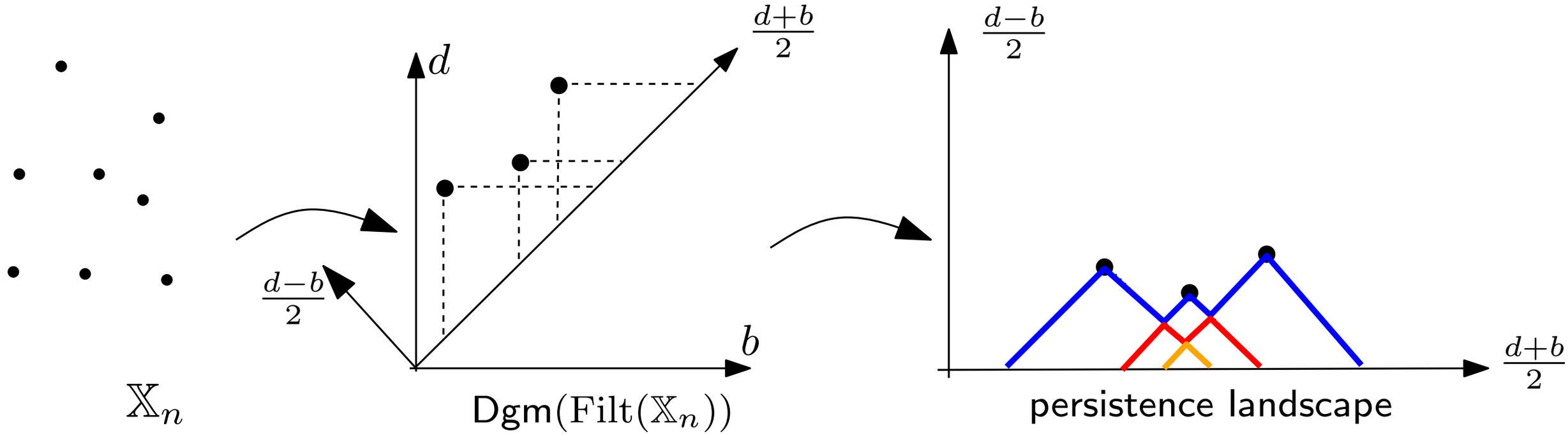
[Chazal & al, 2013]



Persistent surface

[Adams & al, 2016]

Persistence landscapes [Bubnik JMLR 2015]



$$\mathsf{Dgm} = \left\{ \left(\tfrac{d_i+b_i}{2}, \tfrac{d_i+b_i}{2}\right),\ i \in I \right\}$$

For $p = \left(\frac{b+d}{2}, \frac{d-b}{2}\right) \in \mathsf{Dgm}$,

$$\Lambda_p(t) \ = \ \begin{cases} t-b & t \in [b, \frac{b+d}{2}] \\ d-t & t \in (\frac{b+d}{2}, d] \\ 0 & \text{otherwise.} \end{cases}$$

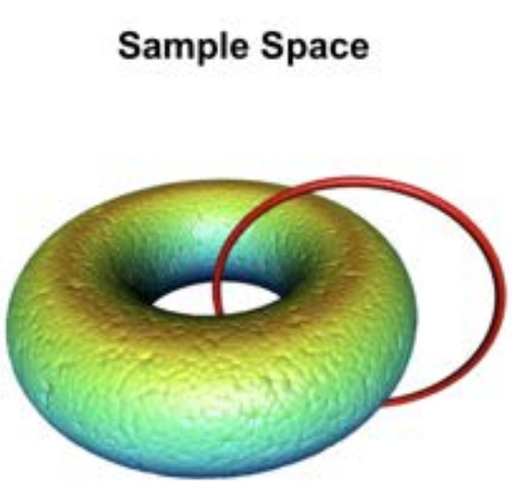
Persistence landscape λ of Dgm:

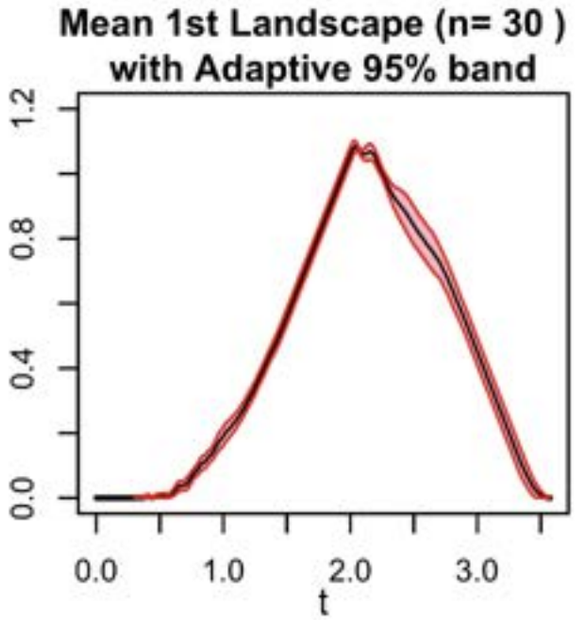
$$\lambda(k,t) = \operatorname*{kmax}_{p \in D} \Lambda_p(t), \quad t \in \mathbb{R},\ k \in \mathbb{N},$$

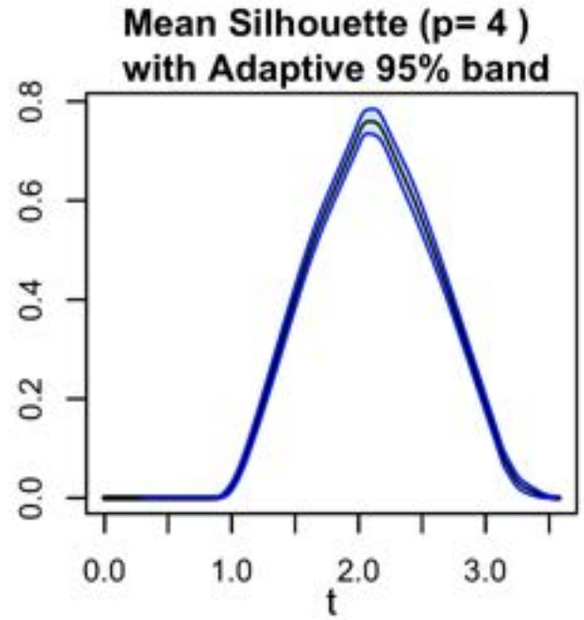
where kmax is k-th largest value in the set.

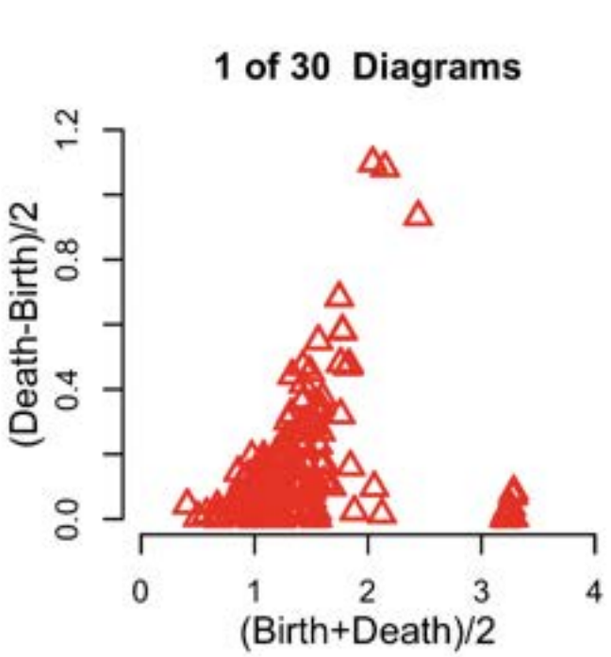
Stability: For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, $|\lambda(k,t) - \lambda'(k,t)| \leq \mathrm{d}_\infty(\mathsf{Dgm}, \mathsf{Dgm}')$.

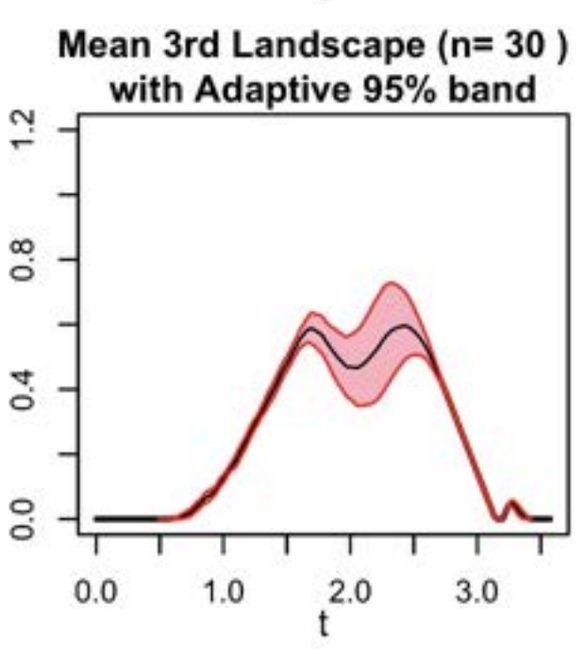
Asymptotic normality for fonctional representations of persistence

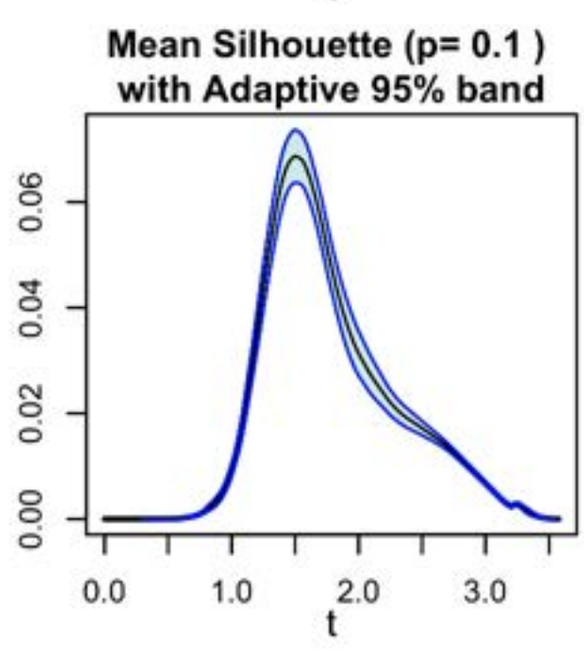








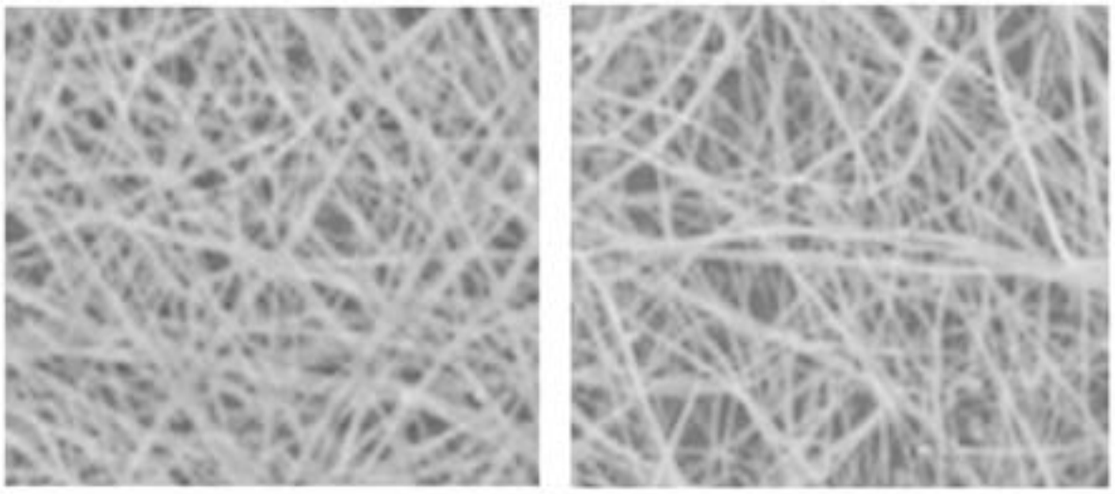




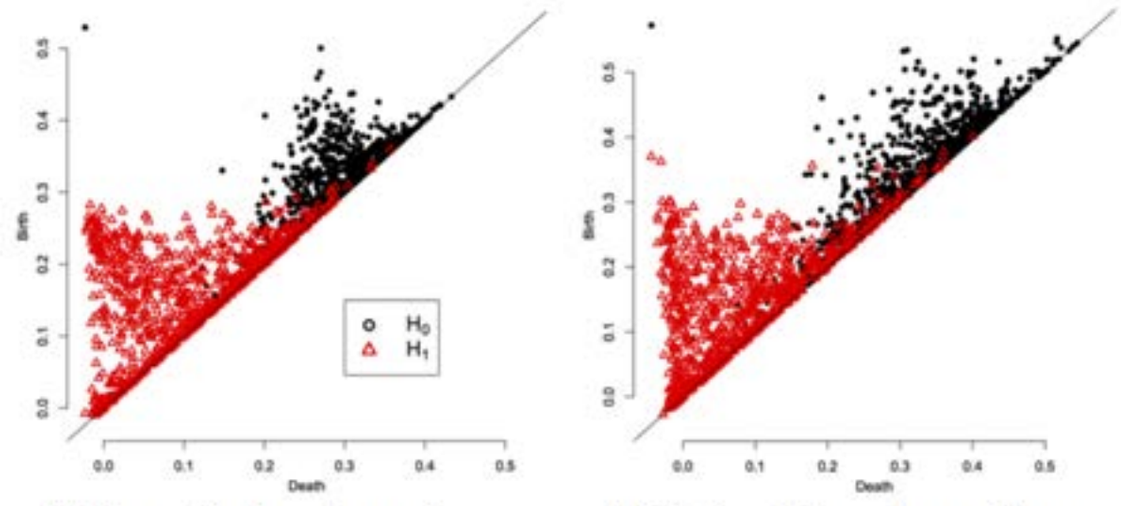
[Chazal etal. 2014, Berry etal. 2020]

Hypothesis tests for Persistence Homology

- Asymptotic normality and bootstrap allows us to propose hypothesis tests (one sample, two sample tests ...)
- Permutation tests



(a) Modeled human fibrin (b) Modeled monkey fibrin



(a) Human fibrin persistence diagram (b) Monkey fibrin persistence diagram

Fig. 9. Persistence diagrams for the modeled human (left) and monkey (right) fibrin networks displayed in Fig. 8 computed using an upper-level set filtration on the modeled images.

[Pretorius et al. (2009) and Berry et al. (2020)]

Subsampling methods for pers. homology [Chazal etal. ICML15]

- Let $X = \{X_1, \cdots, X_m\}$ sampled from μ.
- λ_X: corresponding persistence landscape.
- Ψ_μ^m: the measure induced by $\mu^{\otimes m}$ on the space of persistence landscapes.
- We consider the point-wise expectations of the (random) persistence landscape under this measure:

$$\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)], t \in [0, T]$$

- For $S_1^m, \ldots, S_\ell^m$ some independent samples of size m from $\mu^{\otimes m}$, the empirical counterpart of $\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)]$ is

$$\overline{\lambda_\ell^m}(t) = \frac{1}{\ell} \sum_{i=1}^{\ell} \lambda_{S_i^m}(t), \qquad \text{for all } t \in [0, T],$$

Subsampling methods for pers. homology [Chazal etal. ICML15]

Definition: The p-th Wasserstein distance between two measures μ, ν defined on $(\mathbb{M}, \rho)$ is

$$W_{\rho,p}(\mu,\nu) = \left(\inf_{\Pi} \int_{\mathbb{M}\times\mathbb{M}} [\rho(x,y)]^p d\Pi(x,y) \right)^{\frac{1}{p}},$$

where the infimum is taken over all measures on $\mathbb{M} \times \mathbb{M}$ with marginals μ and ν.

Stability of the average landscape:

Theorem: Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where μ and ν are two probability measures on $\mathbb{M}$. For any $p \geq 1$ we have

$$\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2\, m^{\frac{1}{p}}\, W_{\rho,p}(\mu,\nu).$$

Asymptotic normality for persistent Betti numbers

- Pointwise asymptotic normality of persistent Betti numbers for stationary Poisson processes and binomial processes with constant intensity function: Yogeshwaran et al. (2017) and Hiraoka et al. (2018).
- Krebs and Polonik (2019): stabilizing property of persistent Betti numbers and generalization on the asymptotic normality to the multivariate case and to a broader class of underlying Poisson and binomial processes.

7 - Topological Data Analysis and Machine Learning

1- Kernel Methods

Kernels as pairwise comparisons

- Kernels are usually introduced for studying complex objects on a space $\mathcal{X}$ which is not necessary endowed with a metric.
- For many settings, we know how to construct a **comparison function** K on $\mathcal{X}^2$ (e.g. images, words, texts, trees, graphs ...)
- Examples:
  - $K(x, x') = \exp(-cd(x, x'))$ where $c > 0$ and d is a pseudo distance on $\mathcal{X}$. When d is the norm on $\mathbb{R}^p$, K is the **Gaussian kernel**.
  - Kernels on string data with n-grams and suffix trees : compare the strings by means of the substrings they contain.
  - An example of kernel between graphs is the **random walk kernel**: performs random walks on two graphs simultaneously and counts the number of paths that were produced by both walks.
  - Motif kernels on genetic sequences.

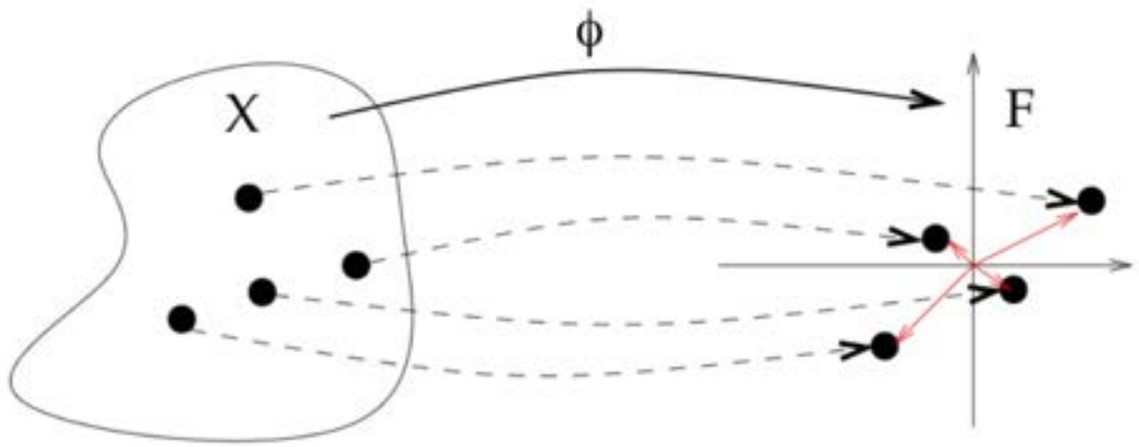
Kernels

Definition A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a positive definite (p.d.) kernel if

- K is symmetric,
- for any $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}^N$, the similarity matrix $[K(x_i, x_j)]$ is definite positive: for any $a = (a_1, \ldots, a_N)' \in \mathbb{R}^N$, we have

$$a'[K(x_i, x_j)]a = \sum_{i,j} a_i a_j K(x_i, x_j) \geq 0.$$

Feature map and kernels



Theorem [Aronszajn, 1950] K is a p.d. kernel on $\mathcal{X}$ **if and only if** there exists an Hilbert space $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ and a mapping $\Phi : \mathcal{X} \to \mathcal{F}$ such that

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}}.$$

$\mathcal{F}$ is call a reproducing kernel Hilbert space (RKHS). In short:

- $\forall x \in \mathcal{X}$, $K_x := K(\cdot, x) \in \mathcal{F}$ $\quad$ $x \to K_x$ is a possible mapping
- $\forall f \in \mathcal{F}$, $\langle f, K(\cdot, x) \rangle_{\mathcal{F}} = f(x)$ $\quad$ Reproducing property

Motivations for Kernel Methods on RKHS

- **Kernel trick:** Any algorithm defined on finite-dimensional vectors that can be expressed only in terms of pairwise inner products can be applied to (potentially) infinite-dimensional vectors in the feature space of a p.d. kernel by replacing each inner product evaluation by a kernel evaluation.

- **Representer Theorems:** Statistical learning problems can often be written as an optimization problem of the form

$$\min_{f \in \mathcal{F}} c(f(x_1), \ldots, f(x_n)) + \lambda \|f\|_{\mathcal{F}} \tag{1}$$

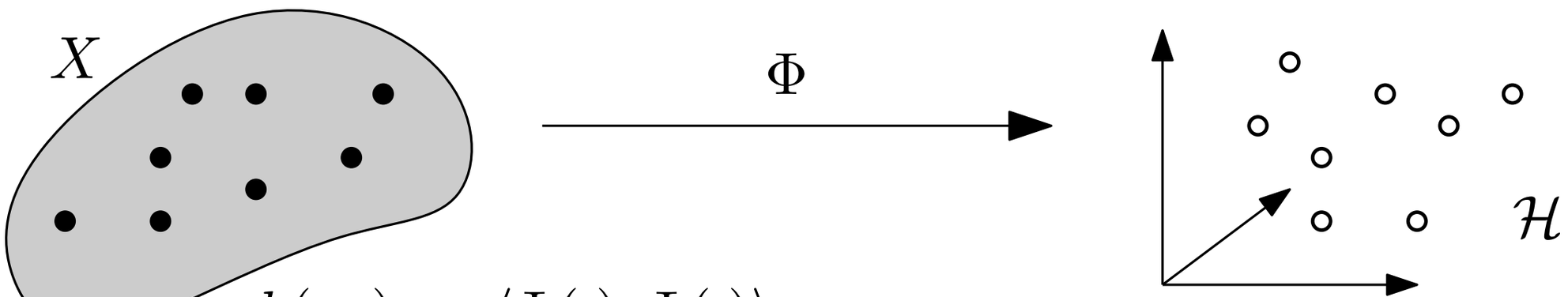
where c measures the fit of f to a given problem and Ω is strictly increasing. The so-called **Representer Theorems** show that any solution of (1) on the RKHS associated to K admits a representation of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i).$$

Kernels for persistence diagrams

Two main approaches:

- define explicit feature map $\Phi : X \to \mathcal{H}$ (vectorization)



$$k(\cdot,\cdot) := \langle \Phi(\cdot), \Phi(\cdot) \rangle_{\mathcal{H}}$$

Kernels for persistence diagrams

Two main approaches:

- define explicit feature map $\Phi : X \to \mathcal{H}$ (vectorization)

- define kernel from metric via radial basis function

Thm:

If $d : X \times X \to \mathbb{R}_+$ symmetric is *conditionally negative semidefinite*, i.e.:

$$\forall n \in \mathbb{N},\ \forall x_1, \ldots, x_n \in X,\ \sum_{i=1}^{n} \alpha_i = 0 \Longrightarrow \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, d(x_i, x_j) \leq 0,$$

then $k(x, y) = \exp\left(-\frac{d(x,y)}{2\sigma^2}\right)$ is positive definite for all $\sigma > 0$.

Persistence weighted Gaussian kernel (PWGK)

Kusano et al., 2016; 2017

- Ω : the subspace of $\mathbb{R}^2$ on or above the diagonal.
- ω a strictly positive function on Ω. Ex: $\omega(b,d) = \arctan(C|d-b|^p)$
- Gaussian kernel: $k_g(x,y) = \exp(\|x-y\|^2/2r^2)$.
- Weighted Gaussian kernel $k_g^\omega(x,y) = \omega(x)\omega(y)k_g(x,y)$ is pos. def.
- The following map is a valid feature map to the RKHS $\mathcal{H}_{k_g,\omega}$ associated to k_g^ω:

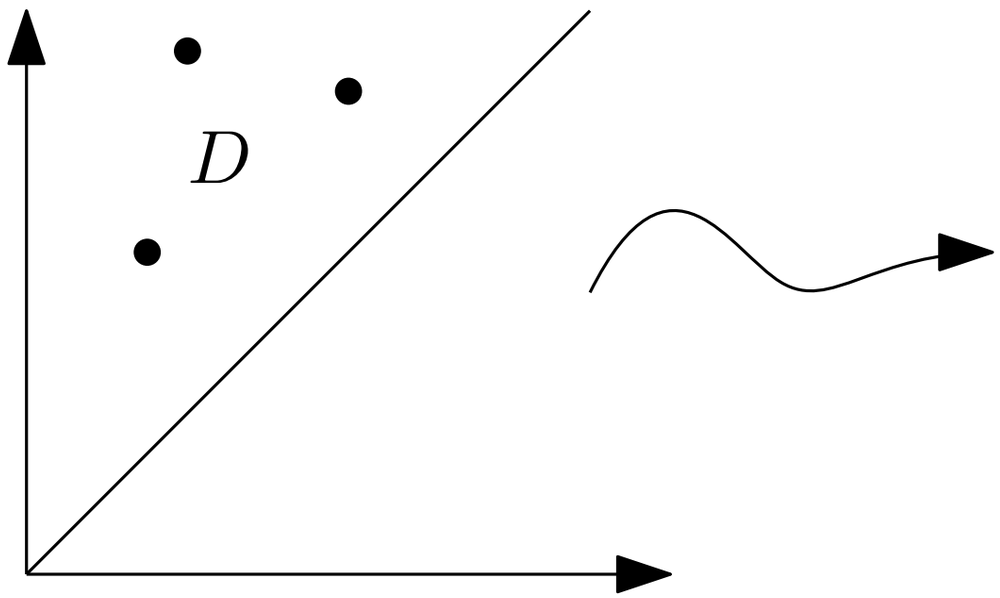
$$\Psi : \mathsf{Dgm} \in \mathcal{P} \mapsto \sum_{r \in \mathsf{Dgm}} \omega(r)\omega(\cdot)k_g(\cdot, r) \in \mathcal{H}_{k_g,\omega}$$

- The inner product $K_{k_g^\omega}$ induced by this feature map satisfies

$$\begin{aligned} K_{k_g^\omega}(\mathsf{Dgm}, \mathsf{Dgm}') &= \langle \Psi(\mathsf{Dgm}), \Psi(\mathsf{Dgm}') \rangle_{\mathcal{H}_{k_g,\omega}} \\ &= \sum_{r \in \mathsf{Dgm}, r' \in \mathsf{Dgm}'} \omega(r)\omega(r')k_g(r,r') \end{aligned}$$

- Stability results of the PWGK-induced distance w.r.t. the bottleneck distance and the 1-Wasserstein distance on persistence diagrams can be shown.

Persistence diagrams as discrete measures



$$D := \sum_{\mathbf{r} \in D} \delta_{\mathbf{r}}$$

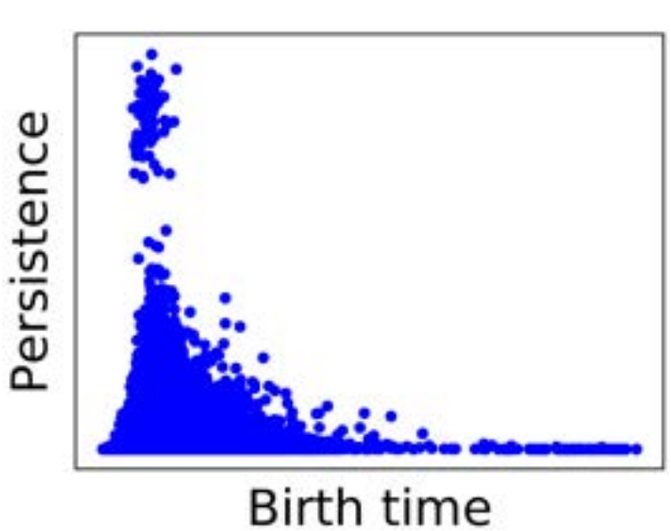
Motivations:

- The space of measures is much nicer that the space of PD.
- In the "standard" algebraic persistence theory, persistence diagrams naturally appear as discrete measures in the plane (over rectangles).
- Many persistence representations can be defined via this measure approach
- Also interesting for defining kernels for PD.

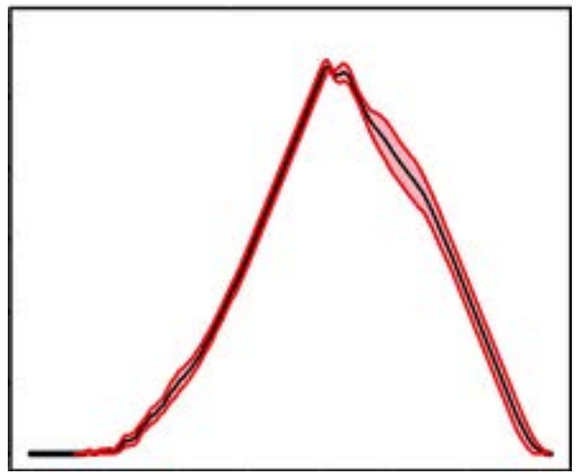
Representation of Persistence diagrams

A representation is called linear if there exists $\phi : \mathbb{R}^2_> \to \mathcal{H}$ such that

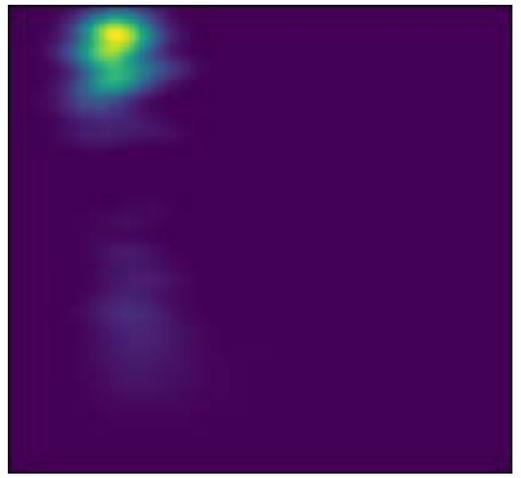
$$\Phi(\mathsf{Dgm}) = \sum_{r \in \mathsf{Dgm}} \phi(r) = \int \phi(\mathbf{r}) \; dD(\mathbf{r})$$



Distrib. of life span, total persistence,...



Persistent silhouette
[Chazal & al, 2013]

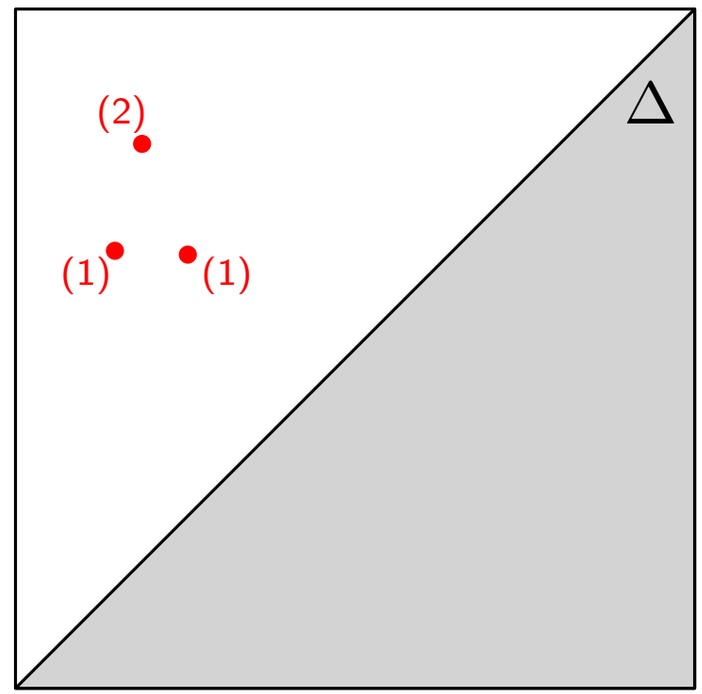


Persistent surface
[Adams & al, 2016]

($\sim$ explicit definition of the feature map)

Space of persistence diagrams

Persistence diagram $\equiv$ **finite** multiset in the open half-plane



Space of persistence diagrams

Persistence diagram $\equiv$ **finite** multiset in the open half-plane

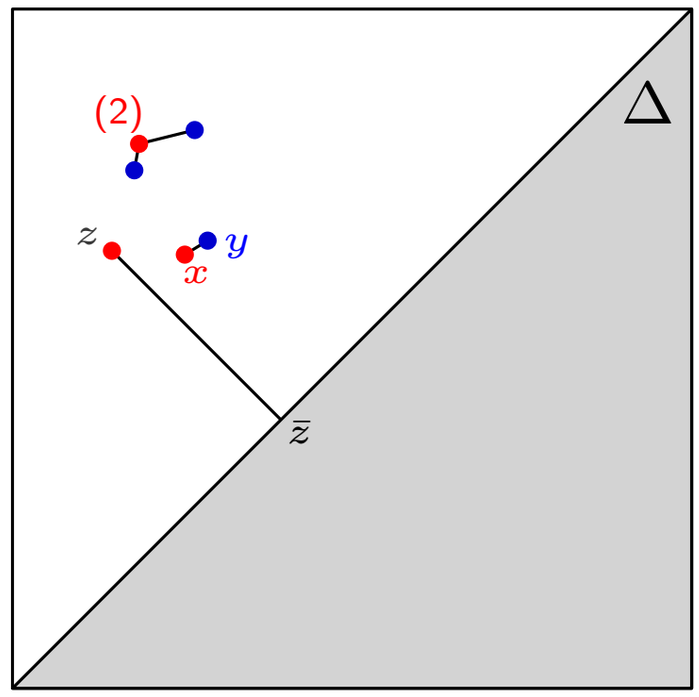
Given a **partial matching** $M : X \leftrightarrow Y$:

cost of a matched pair $(x, y) \in M$: $c_p(x, y) := \|x - y\|_\infty^p$

cost of an unmatched point $z \in X \sqcup Y$: $c_p(z) := \|z - \bar{z}\|_\infty^p$

cost of M:

$$c_p(M) := \left(\sum_{(x,\, y) \text{ matched}} c_p(x, y) \; + \sum_{z \text{ unmatched}} c_p(z) \right)^{1/p}$$



Space of persistence diagrams

Persistence diagram $\equiv$ **finite** multiset in the open half-plane

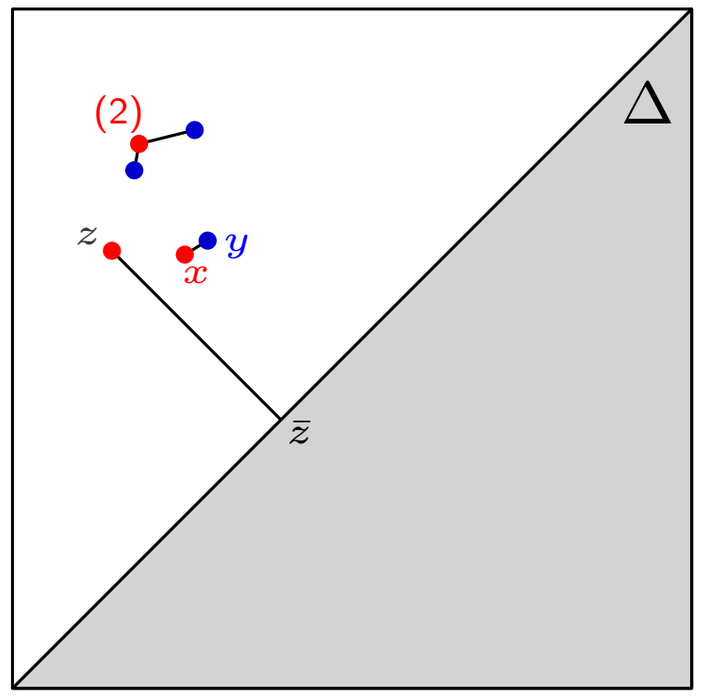
Given a **partial matching** $M : X \leftrightarrow Y$:

cost of a matched pair $(x, y) \in M$: $c_p(x, y) := \|x - y\|_\infty^p$

cost of an unmatched point $z \in X \sqcup Y$: $c_p(z) := \|z - \bar{z}\|_\infty^p$

cost of M:

$$c_p(M) := \left(\sum_{(x,\, y) \text{ matched}} c_p(x, y) \; + \sum_{z \text{ unmatched}} c_p(z) \right)^{1/p}$$



Def: p-th diagram distance (extended metric):

$$\mathrm{d}_p(X, Y) := \inf_{M : X \leftrightarrow Y} c_p(M)$$

Def: bottleneck distance:

$$\mathrm{d}_\infty(X, Y) := \lim_{p \to \infty} \mathrm{d}_p(X, Y)$$

Space of persistence diagrams

Persistence diagram $\equiv$ **finite** multiset in the open half-plane

Given a **partial matching** $M : X \leftrightarrow Y$:

cost of a matched pair $(x, y) \in M$: $c_p(x, y)$

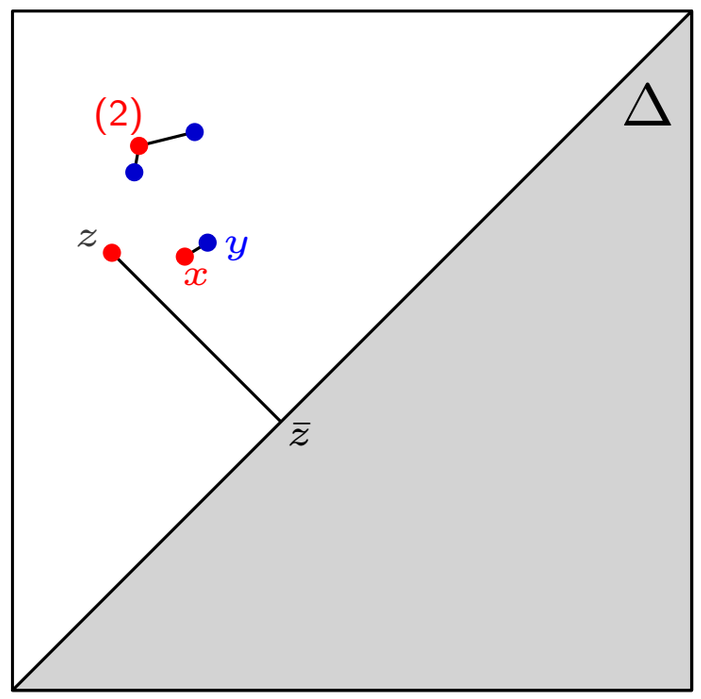
cost of an unmatched point $z \in X \sqcup Y$: $c_p(z) := \|z - \bar{z}\|_\infty^p$

d_p is **NOT** cnsd

$\Rightarrow$ previous theorem is not applicable

cost of M:

$$c_p(M) := \left(\sum_{(x,\, y) \text{ matched}} c_p(x, y) \; + \sum_{z \text{ unmatched}} c_p(z) \right)^{1/p}$$



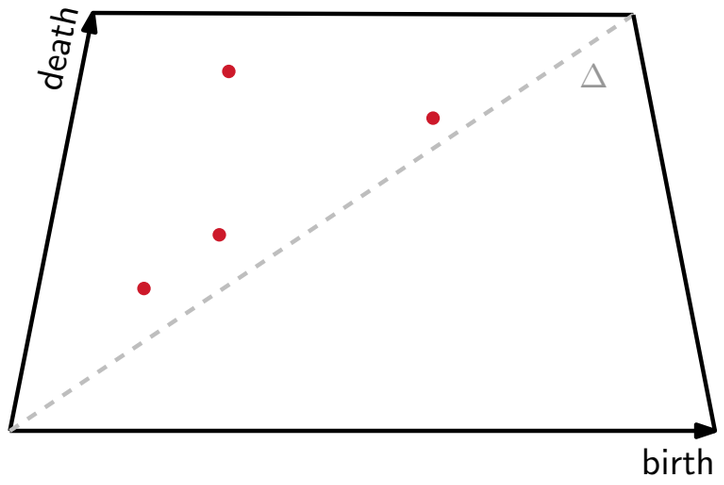
Def: p-th diagram distance (extended metric):

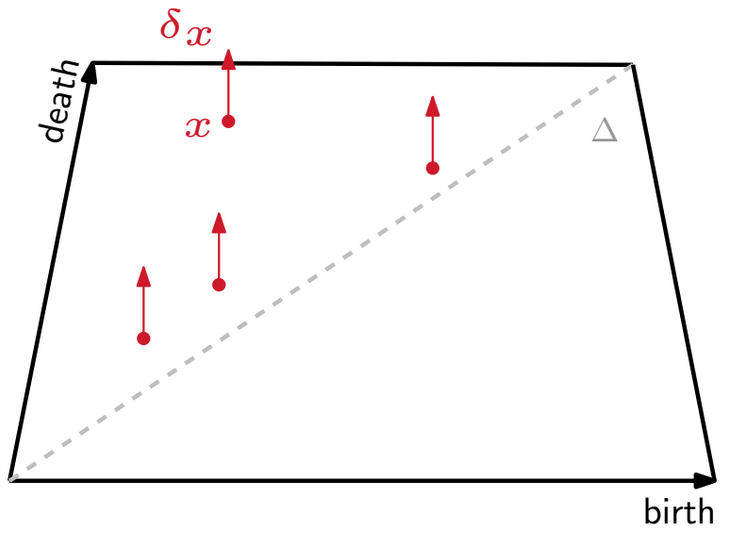
$$\mathrm{d}_p(X, Y) := \inf_{M : X \leftrightarrow Y} c_p(M)$$

Def: bottleneck distance:

$$\mathrm{d}_\infty(X, Y) := \lim_{p \to \infty} \mathrm{d}_p(X, Y)$$

Persistence diagrams as discrete measures

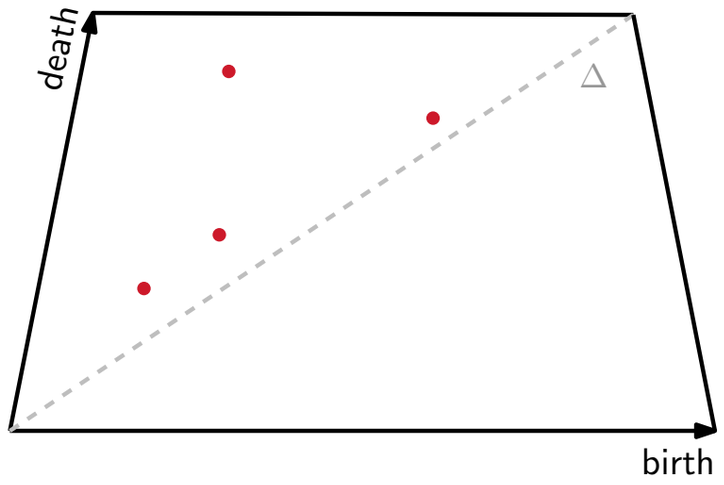


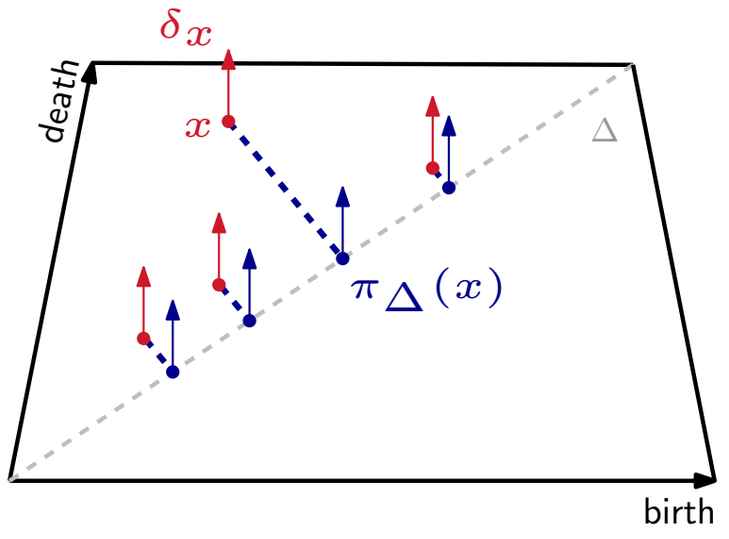


$$\mu_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x$$

Pb: $\mathrm{d}_p(\mathsf{Dgm}, \mathsf{Dgm}') \not\propto W_p(\mu_{\mathsf{Dgm}}, \mu_{\mathsf{Dgm}'})$

Persistence diagrams as discrete measures





$$\mu_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x$$

Pb: $\mathrm{d}_p(\mathsf{Dgm}, \mathsf{Dgm}') \not\propto W_p(\mu_{\mathsf{Dgm}}, \mu_{\mathsf{Dgm}'})$

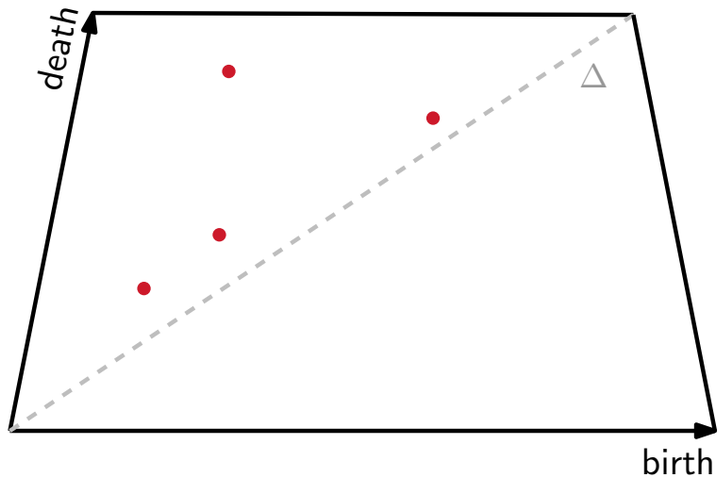
$\rightarrow$ given $\mathsf{Dgm}, \mathsf{Dgm}'$, let

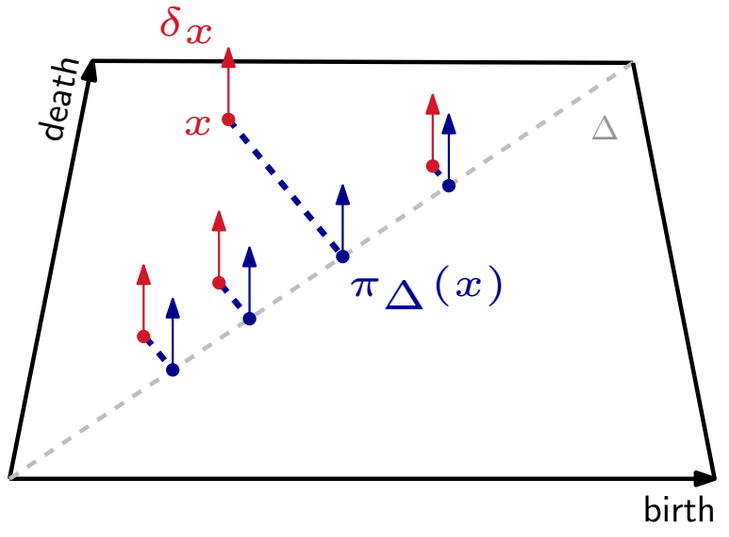
$$\bar{\mu}_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x + \sum_{y \in \mathsf{Dgm}'} \delta_{\pi_\Delta(y)}$$

$$\bar{\mu}_{\mathsf{Dgm}'} := \sum_{y \in \mathsf{Dgm}'} \delta_y + \sum_{x \in \mathsf{Dgm}} \delta_{\pi_\Delta(x)}$$

Then, $\mathrm{d}_p(\mathsf{Dgm}, \mathsf{Dgm}') \leq W_p(\bar{\mu}_{\mathsf{Dgm}}, \bar{\mu}_{\mathsf{Dgm}'}) \leq 2\,\mathrm{d}_p(\mathsf{Dgm}, \mathsf{Dgm}')$

Persistence diagrams as discrete measures





$$\mu_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x$$

Pb: $\mathrm{d}_p(\mathsf{Dgm}, \mathsf{Dgm}') \not\propto W_p(\mu_{\mathsf{Dgm}}, \mu_{\mathsf{Dgm}'})$

$\rightarrow$ given $\mathsf{Dgm}, \mathsf{Dgm}'$, let

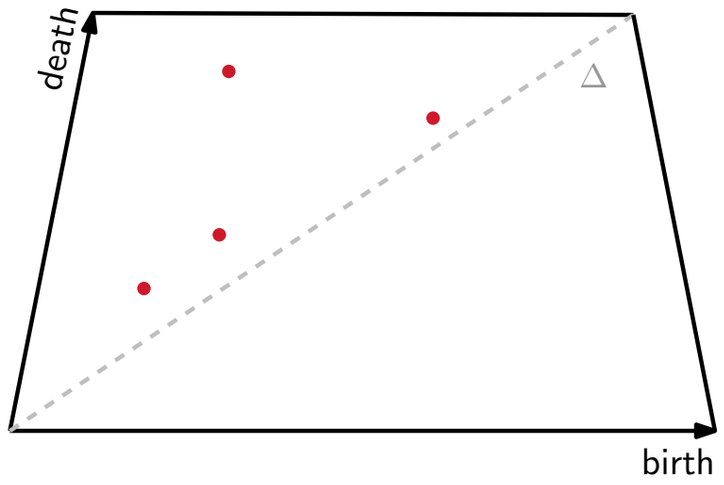
$$\bar{\mu}_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x + \sum_{y \in \mathsf{Dgm}'} \delta_{\pi_\Delta(y)}$$

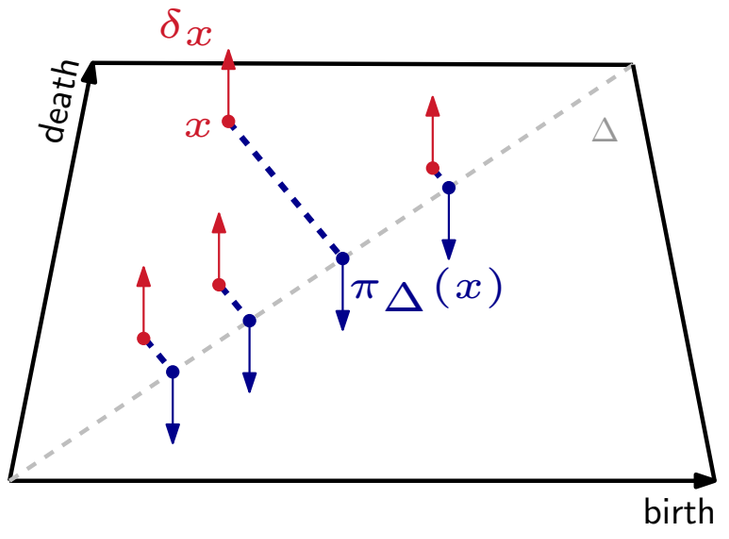
$$\bar{\mu}_{\mathsf{Dgm}'} := \sum_{y \in \mathsf{Dgm}'} \delta_y + \sum_{x \in \mathsf{Dgm}} \delta_{\pi_\Delta(x)}$$

Then, $\mathrm{d}_p(\mathsf{Dgm}, \mathsf{Dgm}') \leq W_p(\bar{\mu}_{\mathsf{Dgm}}, \bar{\mu}_{\mathsf{Dgm}'}) \leq 2\,\mathrm{d}_p(\mathsf{Dgm}, \mathsf{Dgm}')$

Pb: $\bar{\mu}_{\mathsf{Dgm}}$ depends on Dgm'

Persistence diagrams as discrete measures





$$\mu_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x$$

Pb: $\mathrm{d}_p(\mathsf{Dgm}, \mathsf{Dgm}') \not\propto W_p(\mu_{\mathsf{Dgm}}, \mu_{\mathsf{Dgm}'})$

Solution: transfer mass negatively to μ_D:

$$\tilde{\mu}_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x - \sum_{x \in \mathsf{Dgm}} \delta_{\pi_\Delta(x)} \quad \in \mathcal{M}_0(\mathbb{R}^2)$$

$\rightarrow$ signed discrete measure of total mass zero

Persistence diagrams as discrete measures

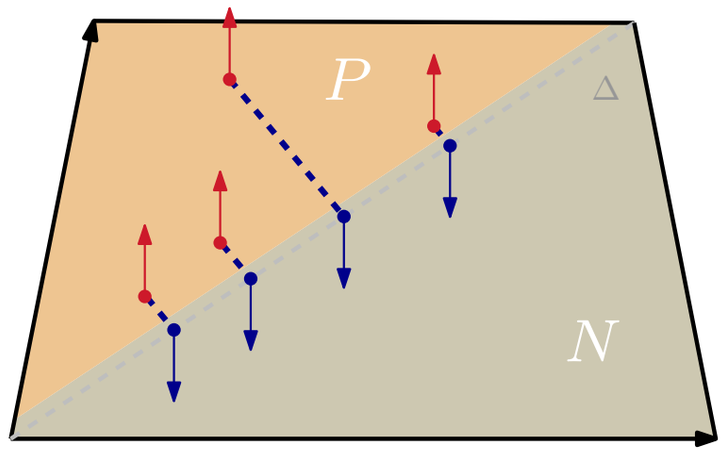
Hahn decomp. thm: For any $\mu \in \mathcal{M}_0(X, \Sigma)$ there exist measurable sets P, N such that:

(i) $P \cup N = X$ and $P \cap N = \emptyset$

(ii) $\mu(B) \geq 0$ for every measureable set $B \subseteq P$

(iii) $\mu(B) \leq 0$ for every measureable set $B \subseteq N$

Moreover, the decomposition is essentially unique.



$\forall B \in \Sigma$, let $\mu^+(B) := \mu(B \cap P)$ and $\mu^-(B) := -\mu(B \cap N) \in \mathcal{M}_+(X)$

Def: $\|\mu\|_K := W_1(\mu^+, \mu^-)$

metric: Kantorovich norm $\|\cdot\|_K$

Prop: $\forall \mu, \nu \in \mathcal{M}_0(X), \quad W_1(\mu^+ + \nu^-, \nu^+ + \mu^-) = \|\mu - \nu\|_K$

Persistence diagrams as discrete measures

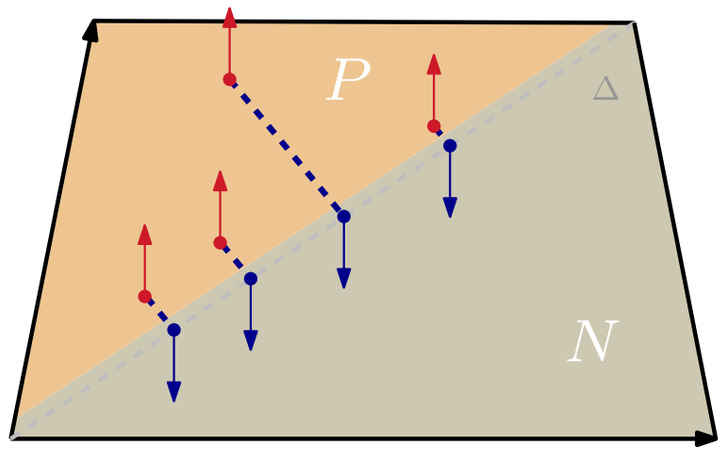
Hahn decomp. thm: For any $\mu \in \mathcal{M}_0(X, \Sigma)$ there exist measurable sets P, N such that:

(i) $P \cup N = X$ and $P \cap N = \emptyset$

(ii) $\mu(B) \geq 0$ for every measureable set $B \subseteq P$

(iii) $\mu(B) \leq 0$ for every measureable set $B \subseteq N$

Moreover, the decomposition is essentially unique.



$\forall B \in \Sigma$, let $\mu^+(B) := \mu(B \cap P)$ and $\mu^-(B) := -\mu(B \cap N) \in \mathcal{M}_+(X)$

Def: $\|\mu\|_K := W_1(\mu^+, \mu^-)$

metric: Kantorovich norm $\|\cdot\|_K$

Prop: $\forall \mu, \nu \in \mathcal{M}_0(X), \quad W_1(\underbrace{\mu^+ + \nu^-}_{\bar{\mu}_{\text{Dgm}}}, \underbrace{\nu^+ + \mu^-}_{\bar{\mu}_{\text{Dgm}'}}) = \|\mu - \nu\|_K$

for persistence diagrams:

$\tilde{\mu}_{\text{Dgm}} \quad \tilde{\mu}_{\text{Dgm}'}$

$$W_1(\bar{\mu}_{\text{Dgm}}, \bar{\mu}_{\text{Dgm}'}) = \|\tilde{\mu}_{\text{Dgm}} - \tilde{\mu}_{\text{Dgm}'}\|_K$$

A Wasserstein Gaussian kernel for PDs?

Thm:
If $d : X \times X \to \mathbb{R}_+$ symmetric is *conditionally negative semidefinite*, i.e.:

$$\forall n \in \mathbb{N},\ \forall x_1, \ldots, x_n \in X,\ \sum_{i=1}^{n} \alpha_i = 0 \Longrightarrow \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, d(x_i, x_j) \leq 0,$$

then $k(x, y) := \exp\left(-\frac{d(x,y)}{2\sigma^2}\right)$ is positive semidefinite.

Pb: W_1 is not cnsd, neither is d_1

Solutions:

- relax the measures (e.g. convolution)
- relax the metric (e.g. regularization, slicing)

Sliced Wasserstein metric

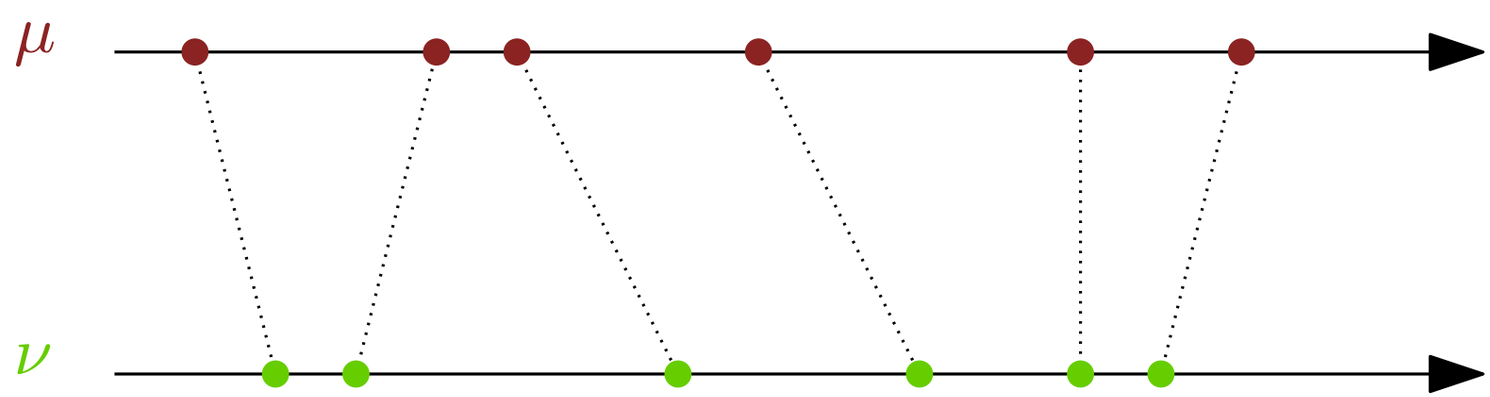
[*Sliced Wasserstein Kernel for persistence diagrams*, Carrière, Cuturi, Oudot, ICML, 2017]

Special case: $X = \mathbb{R}$, μ, ν discrete measures of mass n

$\mu := \sum_{i=1}^{n} \delta_{x_i}, \quad \nu := \sum_{i=1}^{n} \delta_{y_i}$

Sort the atoms of μ, ν along the real line: $x_i \leq x_{i+1}$ and $y_i \leq y_{i+1}$ for all i

Then: $W_1(\mu, \nu) = \sum_{i=1}^{n} |x_i - y_i| = \|(x_1, \cdots, x_n) - (y_1, \cdots, y_n)\|_1$



Sliced Wasserstein metric

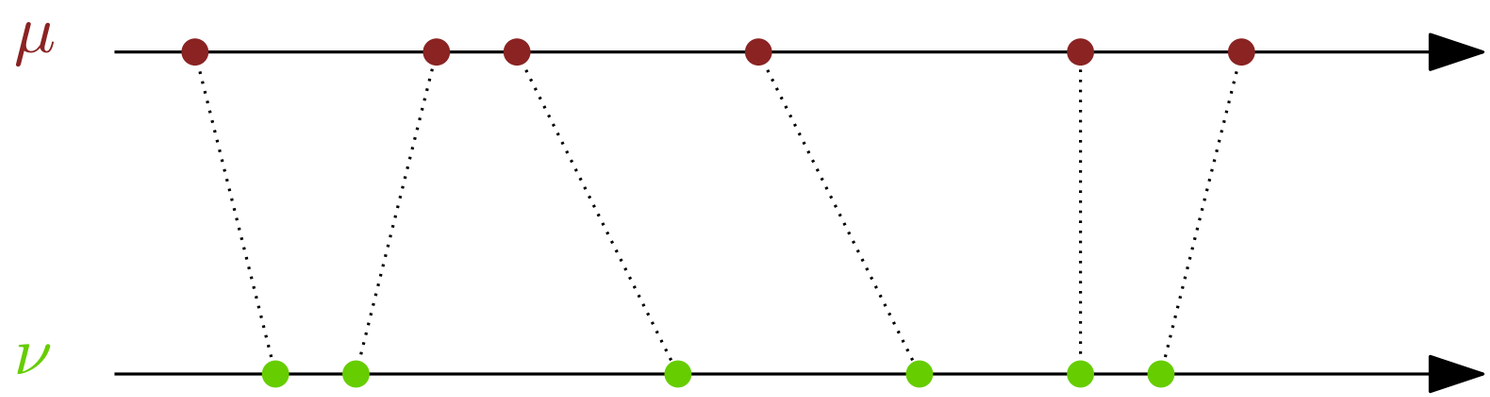
[*Sliced Wasserstein Kernel for persistence diagrams*, Carrière, Cuturi, Oudot, ICML, 2017]

Special case: $X = \mathbb{R}$, μ, ν discrete measures of mass n

$\mu := \sum_{i=1}^n \delta_{x_i}, \quad \nu := \sum_{i=1}^n \delta_{y_i}$

Sort the atoms of μ, ν along the real line: $x_i \leq x_{i+1}$ and $y_i \leq y_{i+1}$ for all i

Then: $W_1(\mu, \nu) = \sum_{i=1}^n |x_i - y_i| = \|(x_1, \cdots, x_n) - (y_1, \cdots, y_n)\|_1$



$\rightarrow W_1$ is cnsd and easy to compute (same with $\|\cdot\|_K$ for signed measures)

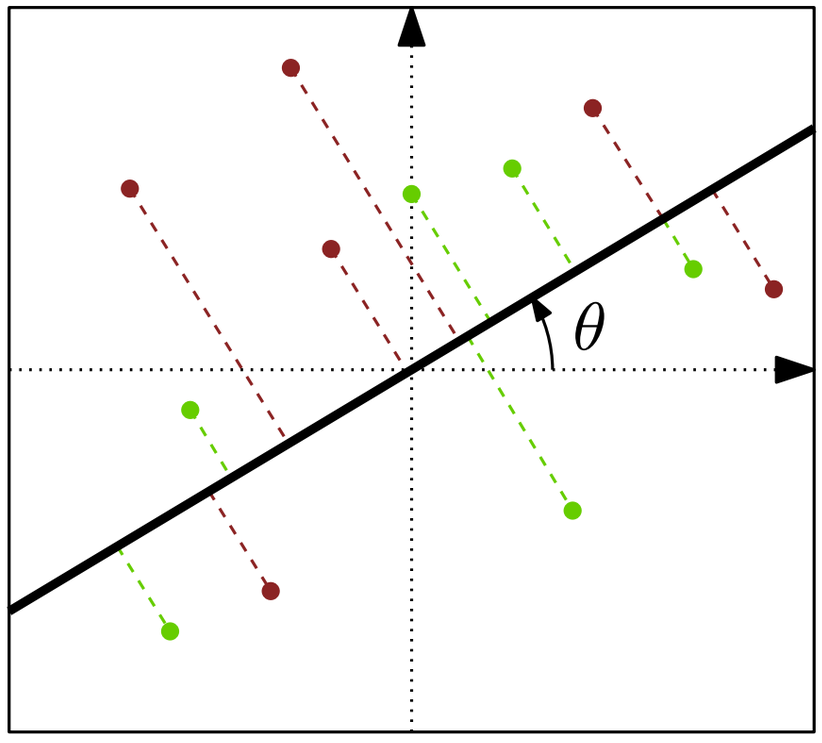
Sliced Wasserstein metric

[*Sliced Wasserstein Kernel for persistence diagrams*, Carrière, Cuturi, Oudot, ICML, 2017]

Def: (sliced Wasserstein distance) for $\mu, \nu \in \mathcal{M}_+(\mathbb{R}^2)$,

$$SW_1(\mu,\nu) := \frac{1}{2\pi} \int_{\theta \in \mathbb{S}^1} W_1(\pi_\theta \# \mu,\, \pi_\theta \# \nu)\, d\theta$$

where π_θ = orthogonal projection onto line passing through origin with angle θ.



Sliced Wasserstein metric

[*Sliced Wasserstein Kernel for persistence diagrams*, Carrière, Cuturi, Oudot, ICML, 2017]

Def: (sliced Wasserstein distance) for $\mu, \nu \in \mathcal{M}_+(\mathbb{R}^2)$,

$$SW_1(\mu, \nu) := \frac{1}{2\pi} \int_{\theta \in \mathbb{S}^1} W_1(\pi_\theta \# \mu,\, \pi_\theta \# \nu)\, d\theta$$

where π_θ = orthogonal projection onto line passing through origin with angle θ.

Props: (inherited from W_1 over $\mathbb{R}$)

- satisfies the axioms of a metric

- well-defined barycenters, fast to compute via stochastic gradient descent, etc.

- conditionally negative semidefinite

Sliced Wasserstein kernel

[*Sliced Wasserstein Kernel for persistence diagrams*, Carrière, Cuturi, Oudot, ICML, 2017]

Def: Given $\sigma > 0$, for any $\mu, \nu \in \mathcal{M}_+(\mathbb{R}^2)$:

$$k_{SW}(\mu, \nu) := \exp\left(-\frac{SW_1(\mu, \nu)}{2\sigma^2}\right)$$

Cor: (from SW cnsd)
k_{SW} is positive semidefinite.

Sliced Wasserstein kernel

[*Sliced Wasserstein Kernel for persistence diagrams*, Carrière, Cuturi, Oudot, ICML, 2017]

Def: Given $\sigma > 0$, for any $\mu, \nu \in \mathcal{M}_+(\mathbb{R}^2)$:

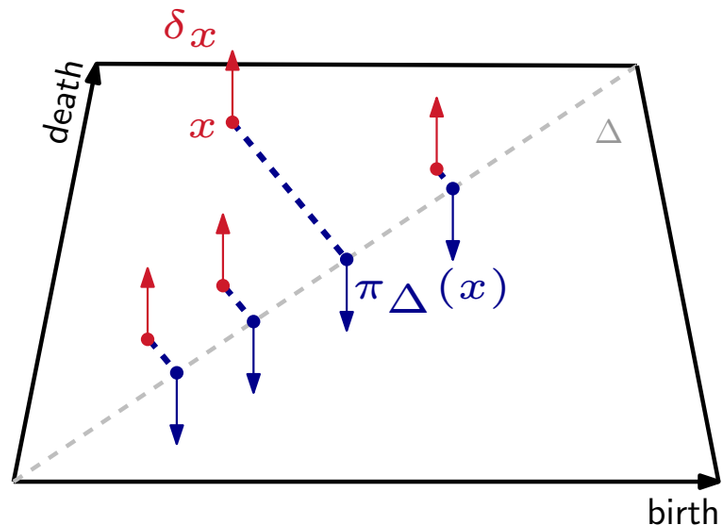
$$k_{SW}(\mu, \nu) := \exp\left(-\frac{SW_1(\mu, \nu)}{2\sigma^2}\right)$$

Cor: (from SW cnsd)
k_{SW} is positive semidefinite.

$\rightarrow$ application to persistence diagrams:

$$\mathsf{Dgm} \mapsto \mu_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x$$

$$\mapsto \tilde{\mu}_{\mathsf{Dgm}} := \mu_{\mathsf{Dgm}} - \pi_\Delta \# \mu_{\mathsf{Dgm}}$$



$$SW_1(\mathsf{Dgm}, \mathsf{Dgm}') := \int_{\theta \in \mathcal{S}^1} \|\pi_\theta \# \tilde{\mu}_{\mathsf{Dgm}} - \pi_\theta \# \tilde{\mu}_{\mathsf{Dgm}'}\|_K \, d\theta$$

$$k_{SW}(\mathsf{Dgm}, \mathsf{Dgm}') := \exp\left(-\frac{SW_1(\mathsf{Dgm}, \mathsf{Dgm}')}{2\sigma^2}\right)$$

Sliced Wasserstein kernel

[*Sliced Wasserstein Kernel for persistence diagrams*, Carrière, Cuturi, Oudot, ICML, 2017]

Def: Given $\sigma > 0$, for any $\mu, \nu \in \mathcal{M}_+(\mathbb{R}^2)$:

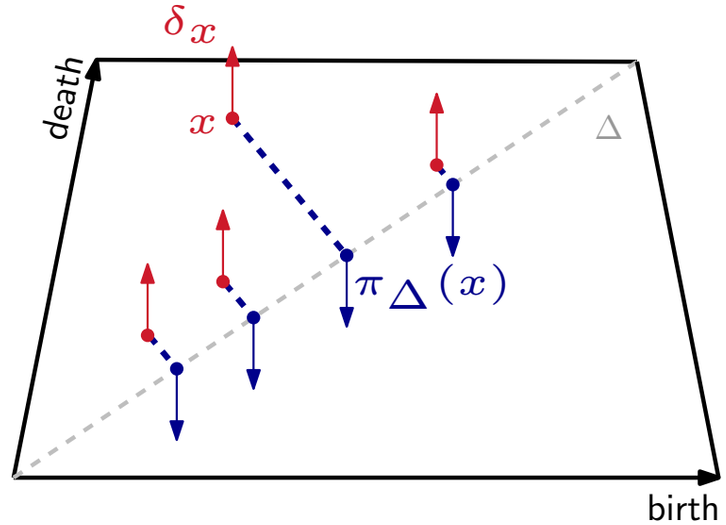
$$k_{SW}(\mu, \nu) := \exp\left(-\frac{SW_1(\mu, \nu)}{2\sigma^2}\right)$$

Cor: (from SW cnsd)
k_{SW} is positive semidefinite.

$\rightarrow$ application to persistence diagrams:

$$\mathsf{Dgm} \mapsto \mu_{\mathsf{Dgm}} := \sum_{x \in \mathsf{Dgm}} \delta_x$$

$$\mapsto \tilde{\mu}_{\mathsf{Dgm}} := \mu_{\mathsf{Dgm}} - \pi_\Delta \# \mu_{\mathsf{Dgm}}$$



$$SW_1(\mathsf{Dgm}, \mathsf{Dgm}') := \int_{\theta \in \mathcal{S}^1} \|\pi_\theta \# \tilde{\mu}_{\mathsf{Dgm}} - \pi_\theta \# \tilde{\mu}_{\mathsf{Dgm}'}\|_K \, d\theta$$

- positive semidefinite

$$k_{SW}(\mathsf{Dgm}, \mathsf{Dgm}') := \exp\left(-\frac{SW_1(\mathsf{Dgm}, \mathsf{Dgm}')}{2\sigma^2}\right)$$

- simple and fast to compute

Sliced Wasserstein kernel

[*Sliced Wasserstein Kernel for persistence diagrams*, Carrière, Cuturi, Oudot, ICML, 2017]

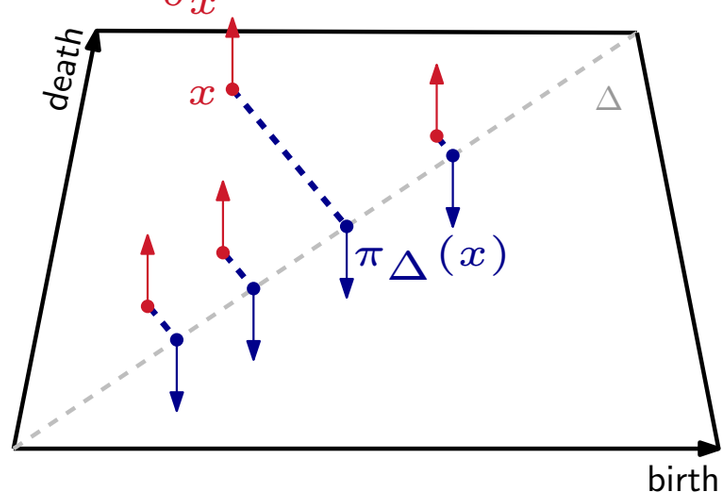
Thm:
The metrics d_1 and SW_1 on the space $\mathcal{D}_N$ of persistence diagrams of size bounded by N are strongly equivalent, namely: for $\mathsf{Dgm}, \mathsf{Dgm}' \in \mathcal{D}_N$,

$$\frac{1}{2+4N(2N-1)}\,\mathrm{d}_1(\mathsf{Dgm},\mathsf{Dgm}') \;\leq\; SW_1(\mathsf{Dgm},\mathsf{Dgm}') \;\leq\; 2\sqrt{2}\,\mathrm{d}_1(\mathsf{Dgm},\mathsf{Dgm}')$$

$\rightarrow$ application to persistence diagrams:

$$\mathsf{Dgm} \mapsto \mu_{\mathsf{Dgm}} := \sum_{x\in\mathsf{Dgm}} \delta_x$$

$$\mapsto \tilde{\mu}_{\mathsf{Dgm}} := \mu_{\mathsf{Dgm}} - \pi_\Delta \# \mu_{\mathsf{Dgm}}$$



$$SW_1(\mathsf{Dgm},\mathsf{Dgm}') := \int_{\theta\in\mathcal{S}^1} \|\pi_\theta\#\tilde{\mu}_{\mathsf{Dgm}} - \pi_\theta\#\tilde{\mu}_{\mathsf{Dgm}'}\|_K \, d\theta$$

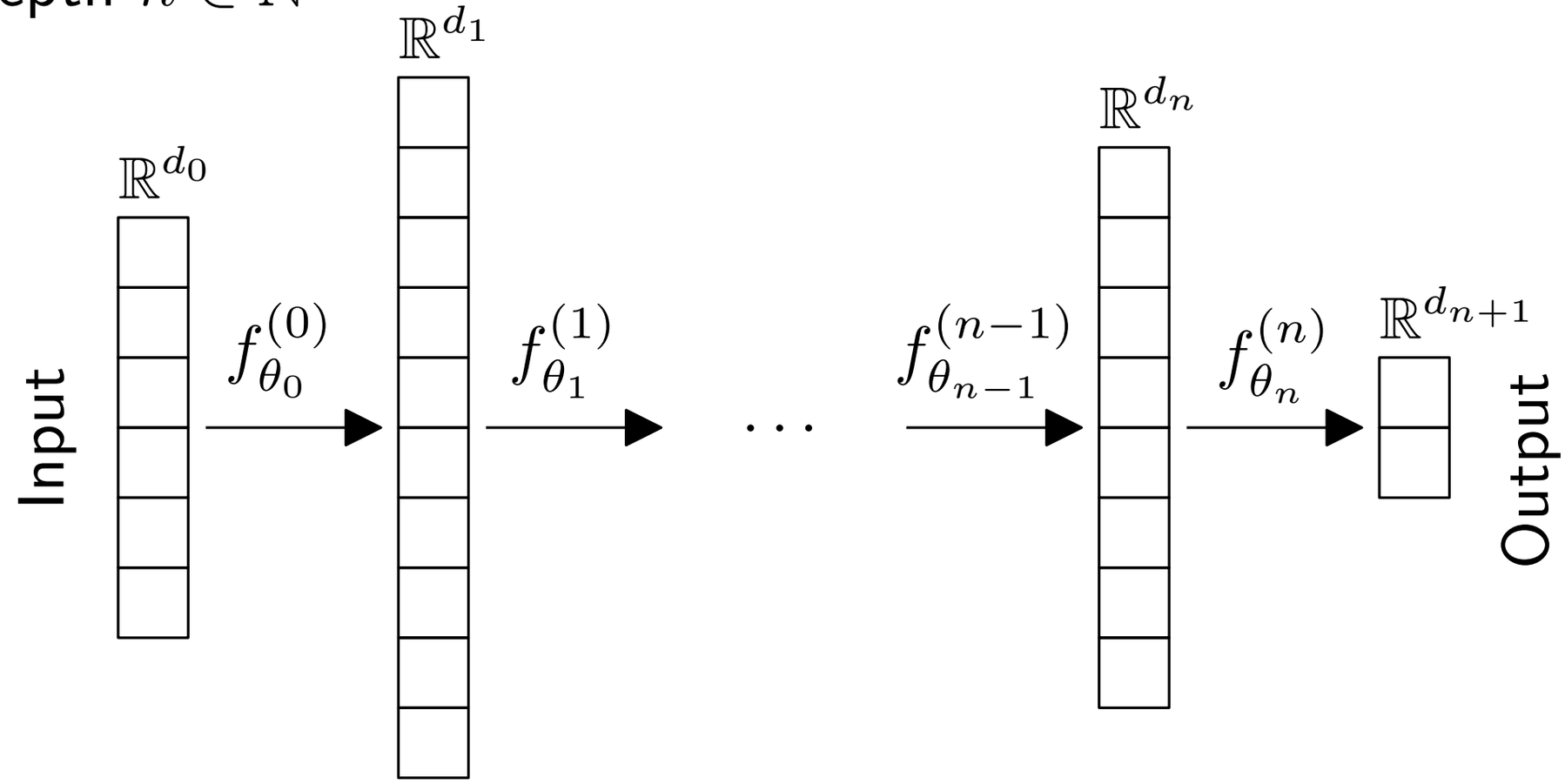
$$k_{SW}(\mathsf{Dgm},\mathsf{Dgm}') := \exp\left(-\frac{SW_1(\mathsf{Dgm},\mathsf{Dgm}')}{2\sigma^2}\right)$$

7 - Topological Data Analysis and Machine Learning

2- Perslay

Reminder: Neural Net

NN with depth $n \in \mathbb{N}^*$



$$\theta_k = (W_k \in \mathbb{R}^{d_{k+1} \times d_k},\ b_k \in \mathbb{R}^{d_{k+1}}), \quad \sigma : x \mapsto \max(0, x) \text{ or } (1 + \mathrm{e}^{-x})^{-1}$$

$$f^{(k)}_{\theta_k} : x \in \mathbb{R}^{d_k} \mapsto \sigma(W_k \cdot x + b_k) \in \mathbb{R}^{d_{k+1}}$$

Final classifier/regressor: $F_\theta = f^{(n)}_{\theta_n} \circ \cdots \circ f^{(0)}_{\theta_0}$

Deep Set Architecture

Originally defined in [Zaheer et al. 2017]

Tailored to handle sets instead of finite dimensional vectors

Input: $\{x_1, ..., x_n\} \subset \mathbb{R}^d$ instead of $x \in \mathbb{R}^d$

$\Rightarrow$ Network must be *permutation invariant*.

$$\Rightarrow F(\{x_1, ..., x_n\}) = F(\{x_{\sigma(1)}, ..., x_{\sigma(n)}\}),\ \forall \sigma$$

Deep Set Architecture

Originally defined in [Zaheer et al. 2017]

Tailored to handle sets instead of finite dimensional vectors

Input: $\{x_1, ..., x_n\} \subset \mathbb{R}^d$ instead of $x \in \mathbb{R}^d$

$\Rightarrow$ Network must be *permutation invariant*.

Universality theorem

Th: [Zaheer et al. 2017]

A function f is permutation invariant iif $f(X) = \rho\left(\sum_i \phi(x_i)\right)$ for some ρ and ϕ, whenever X is included in a *countable* space

In practice: $\phi(x_i) = W \cdot x_i + b$

PersLay: adaptation to persistence diagrams

[Carrière, C., Ike, Lacombe, Royer, Umeda 2019]

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ silhouettes $\rightarrow$ Betti curves

Using any permutation invariant operation (such as max, min, kth largest value) allows to generalize to other persistence representations.

$$\text{PersLay}(dgm) = \rho\left(\text{op}\{w(p) \cdot \phi(p)\}_{p \in dgm}\right)$$

Permutation-invariant operation

Weight function

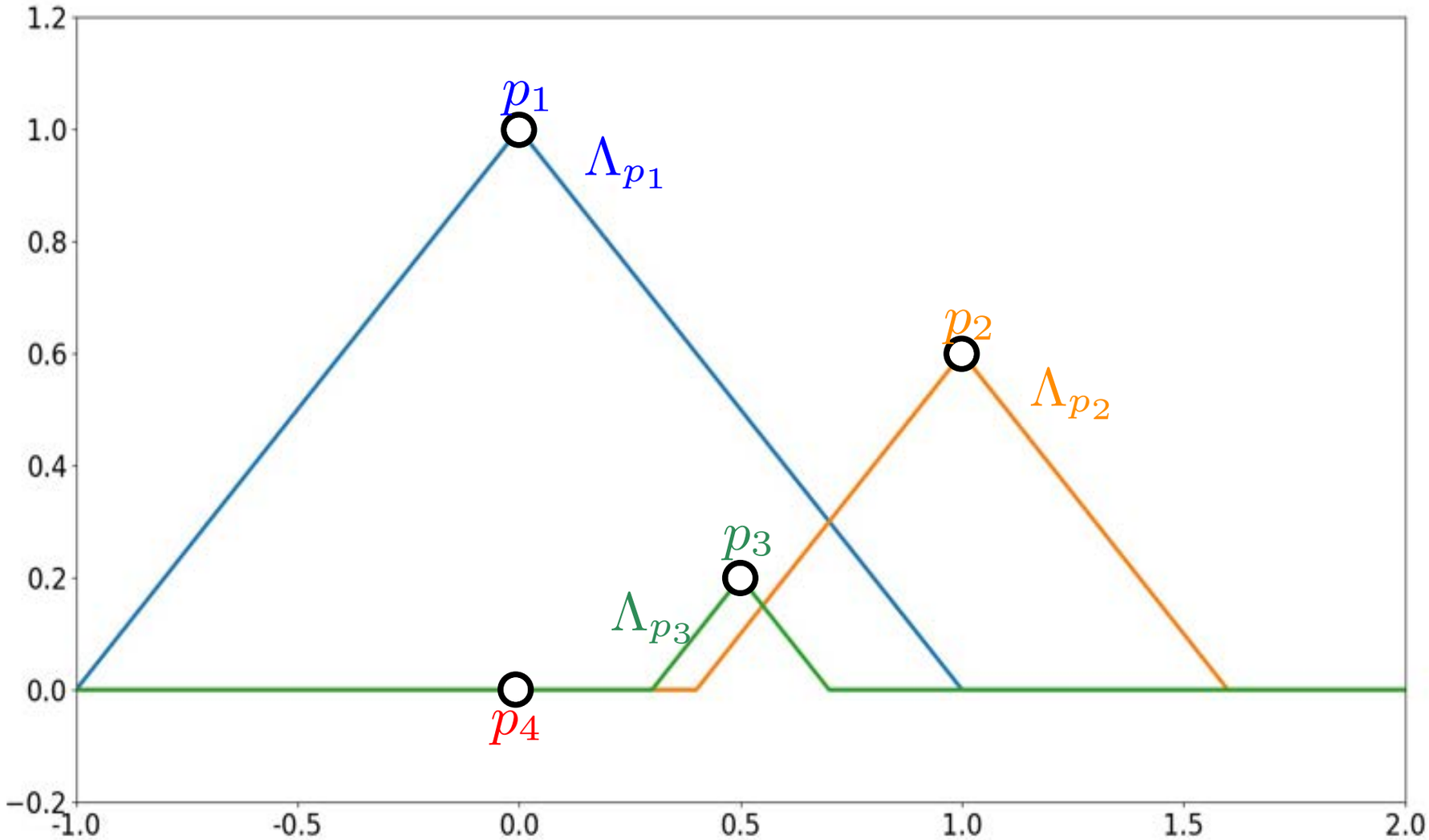
Point transformation $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^k$

PersLay: adaptation to persistence diagrams

[Carrière, C., Ike, Lacombe, Royer, Umeda 2019]

Parameters $t_1, \cdots, t_q \in \mathbb{R}$

$$w(p) = 1 \qquad \phi_\Lambda : p \mapsto \begin{bmatrix} \Lambda_p(t_1) \\ \Lambda_p(t_2) \\ \vdots \\ \Lambda_p(t_q) \end{bmatrix} \qquad \text{op} = \text{top-}k$$

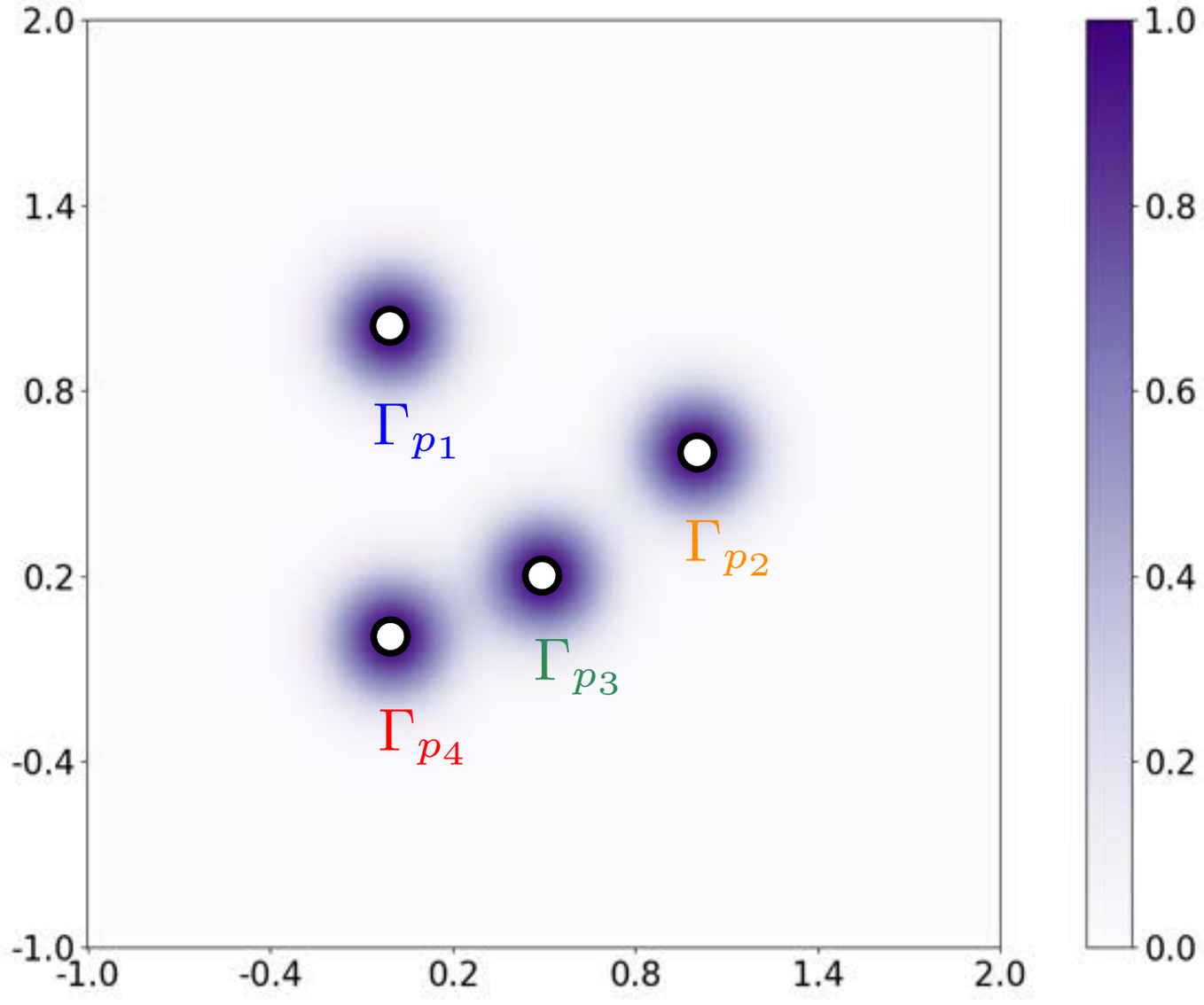


PersLay: adaptation to persistence diagrams

[Carrière, C., Ike, Lacombe, Royer, Umeda 2019]

Parameters $t_1, \cdots, t_q \in \mathbb{R}^2$

$$w(p) = w_t((x, y)) \qquad \phi_\Gamma : p \mapsto \begin{bmatrix} \Gamma_p(t_1) \\ \Gamma_p(t_2) \\ \vdots \\ \Gamma_p(t_q) \end{bmatrix} \qquad \text{op} = \text{sum}$$



PersLay: adaptation to persistence diagrams

[Carrière, C., Ike, Lacombe, Royer, Umeda 2019]

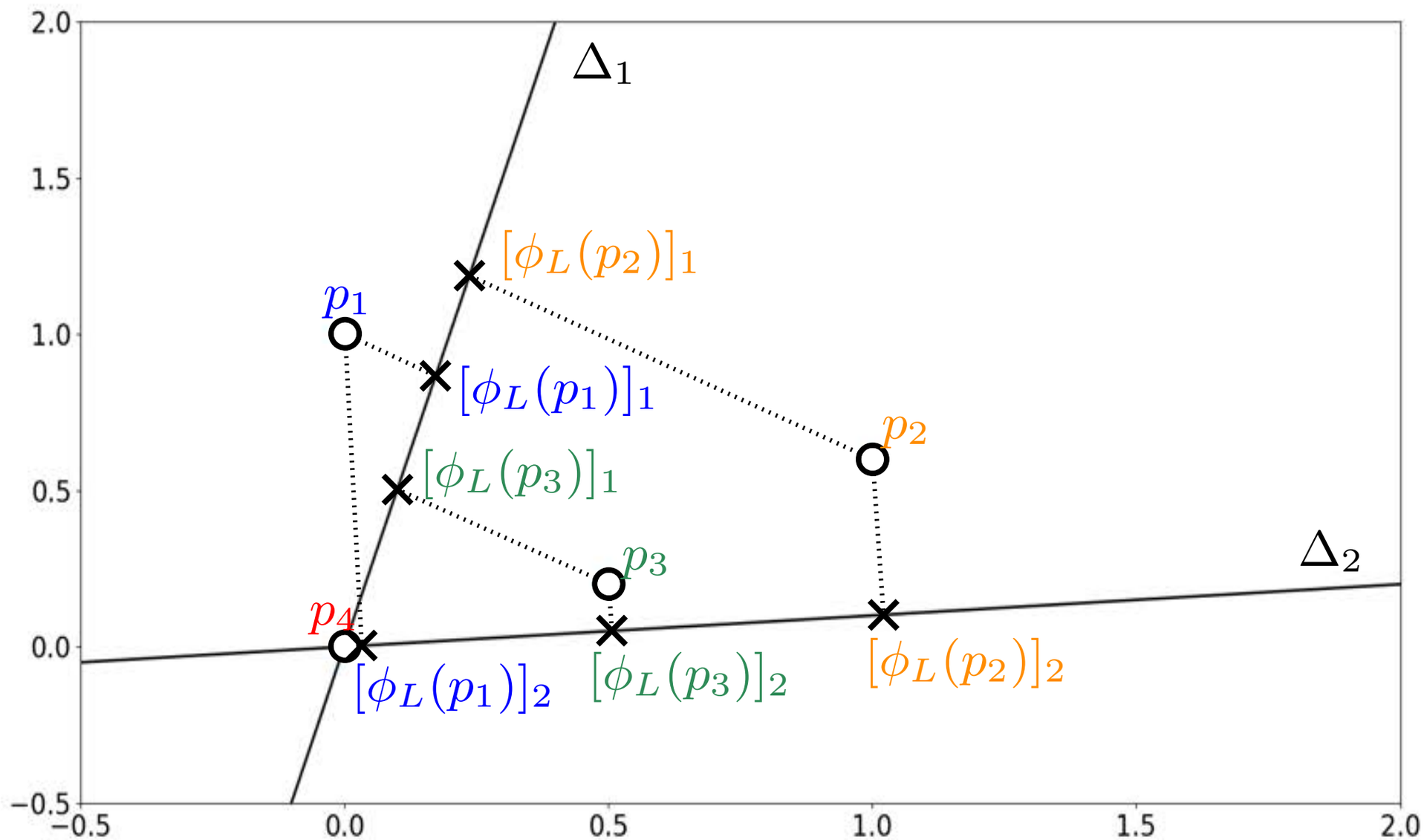
Parameters $\Delta_1, \cdots, \Delta_q \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$

$b_{\Delta_1}, \cdots, b_{\Delta_q} \in \mathbb{R}$

$$\phi_L : p \mapsto \begin{bmatrix} \langle p, e_{\Delta_1} \rangle + b_{\Delta_1} \\ \langle p, e_{\Delta_2} \rangle + b_{\Delta_2} \\ \vdots \\ \langle p, e_{\Delta_q} \rangle + b_{\Delta_q} \end{bmatrix}$$

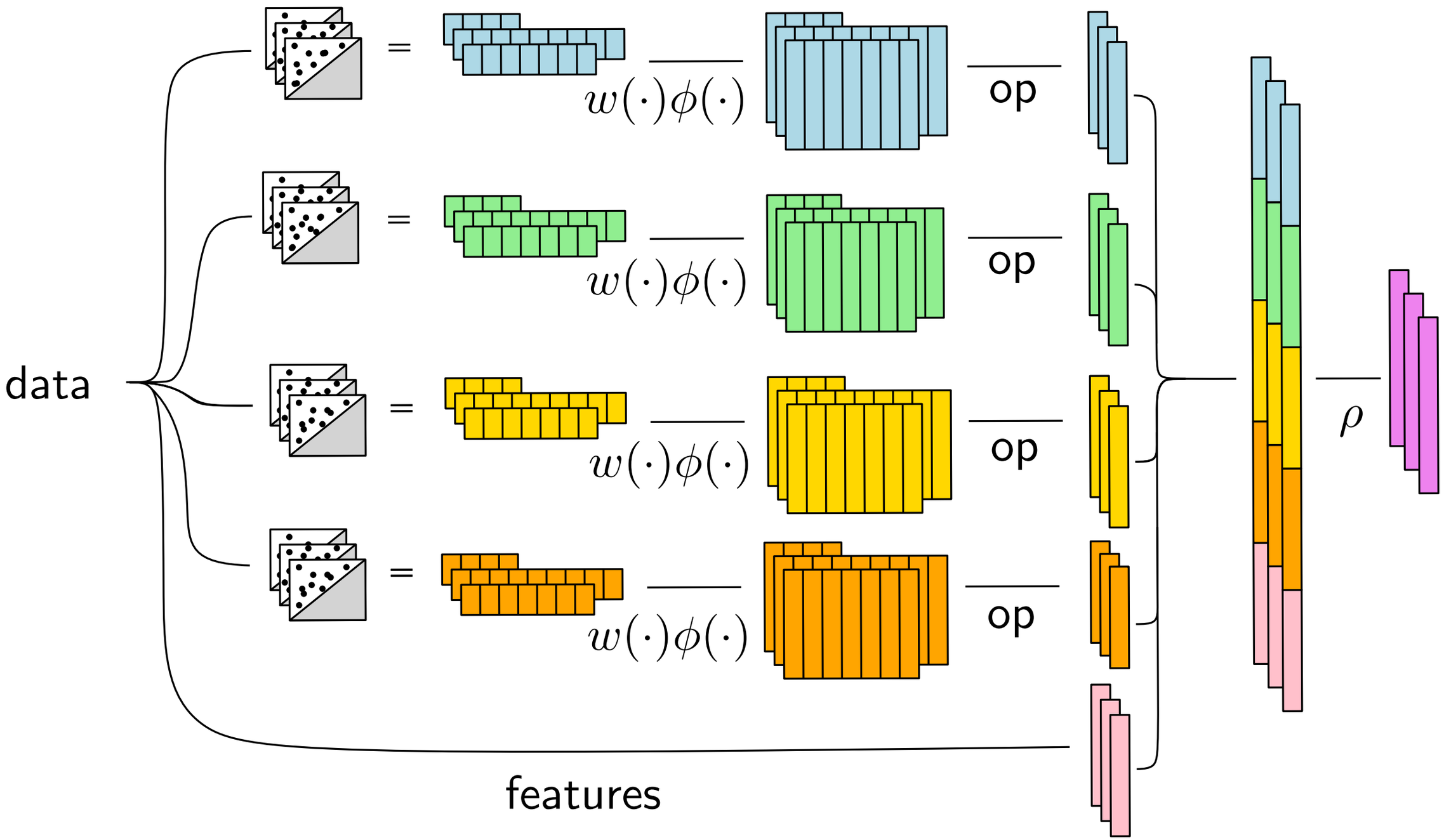
$w(p) = 1$

op $=$ top-k



PersLay: adaptation to persistence diagrams

[Carrière, C., Ike, Lacombe, Royer, Umeda 2019]



PersLay: adaptation to persistence diagrams

[Carrière, C., Ike, Lacombe, Royer, Umeda 2019]

Application to graph classification, using persistence of heat kernel signatures on graphs.

Dataset	ScaleVariant[1]	RetGK1 * [2]	RetGK11 * [2]	FGSD [3]	GCNN [4]	Spectral + HKS [5]	PersLay
REDDIT5K	—	56.1(±0.5)	55.3(±0.3)	47.8	52.9	49.7(±0.3)	**56.6**(±0.3)
REDDIT12K	—	**48.7**(±0.2)	47.1(±0.3)	—	46.6	39.7(±0.1)	47.7(±0.2)
COLLAB	—	**81.0(±0.3)**	80.6(±0.3)	80.0	79.6	67.8(±0.2)	76.4(±0.4)
IMDB-B	72.9	71.9(±1.0)	72.3(±0.6)	**73.6**	73.1	67.6(±0.6)	70.9(±0.7)
IMDB-M	50.3	47.7(±0.3)	48.7(±0.6)	**52.4**	50.3	44.5(±0.4)	48.7(±0.6)
BZR *	86.6	—	—	—	—	80.8(±0.8)	**87.2(±0.7)**
COX2 *	78.4	80.1(±0.9)	81.4(±0.6)	—	—	78.2(±1.3)	**81.6(±1.0)**
DHFR *	78.4	81.5(±0.9)	**82.5(±0.8)**	—	—	69.5(±1.0)	81.8(±0.8)
MUTAG *	88.3	90.3(±1.1)	90.1(±1.0)	**92.1**	86.7	85.8(±1.3)	89.8(±0.9)
PROTEINS *	72.6	75.8(±0.6)	75.2(±0.3)	73.4	**76.3**	73.5(±0.3)	74.8(±0.3)
NCI1 *	71.6	**84.5(±0.2)**	83.5(±0.2)	79.8	78.4	65.3(±0.2)	72.8(±0.3)
NCI109 *	70.5	—	—	**78.8**	—	64.9(±0.2)	71.7(±0.3)
FRANKENSTEIN	69.4	—	—	—	—	62.9(±0.1)	**70.7(±0.4)**

Average scores from 10 times 10-folds cross-validation

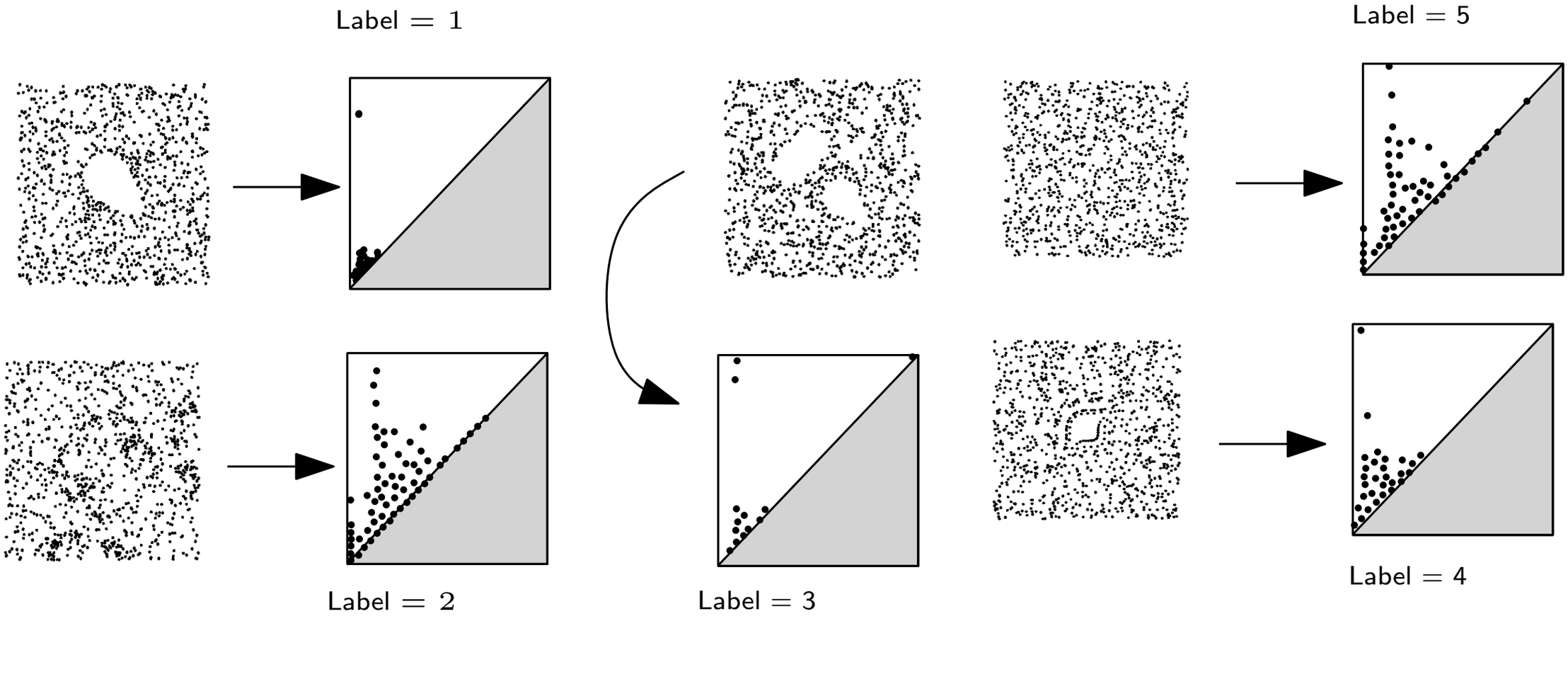
PersLay: adaptation to persistence diagrams

[Carrière, C., Ike, Lacombe, Royer, Umeda 2019]

Goal: classify orbits of *linked twisted map*

Orbits described by (depending on parameter r):

$$\begin{cases} x_{n+1} &= x_n + r\, y_n(1-y_n) \mod 1 \\ y_{n+1} &= y_n + r\, x_{n+1}(1-x_{n+1}) \mod 1 \end{cases}$$

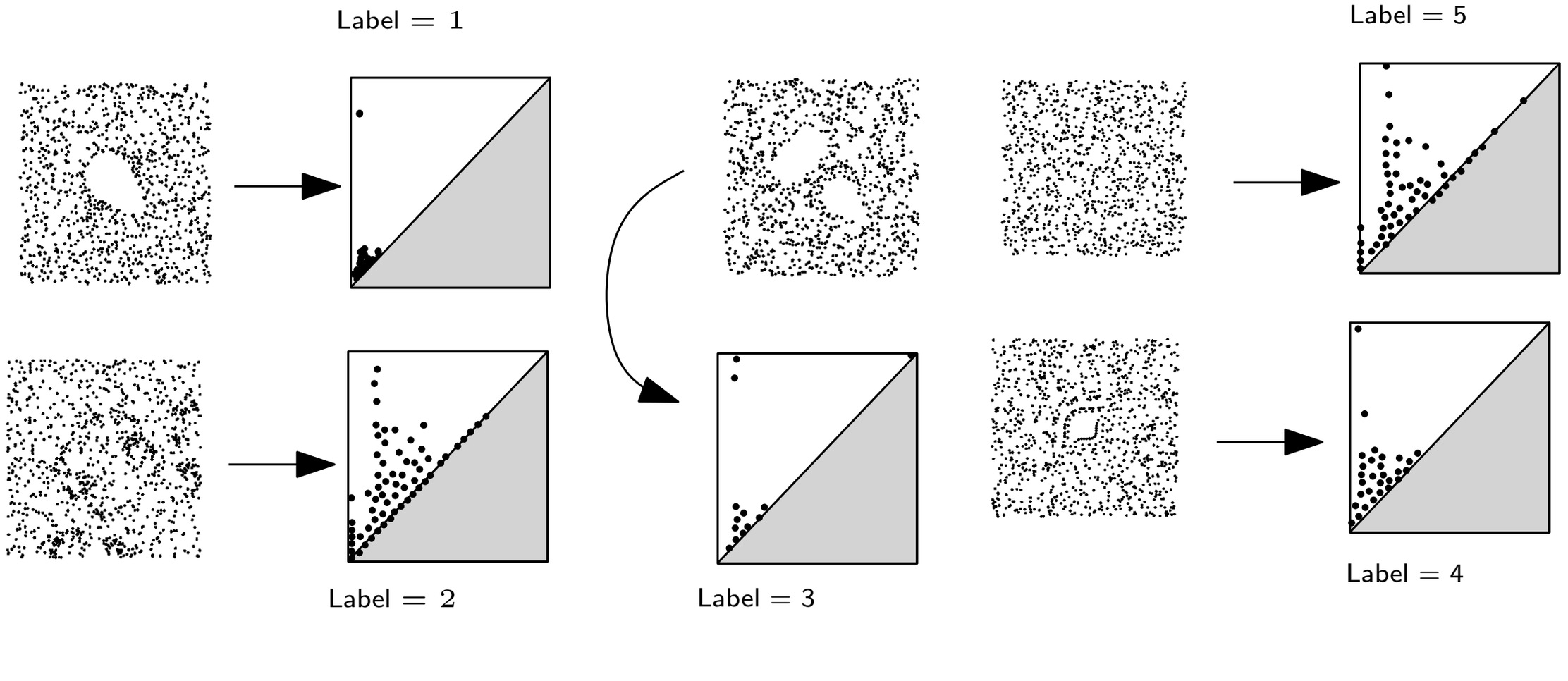


PersLay: adaptation to persistence diagrams

[Carrière, C., Ike, Lacombe, Royer, Umeda 2019]

Goal: classify orbits of *linked twisted map*

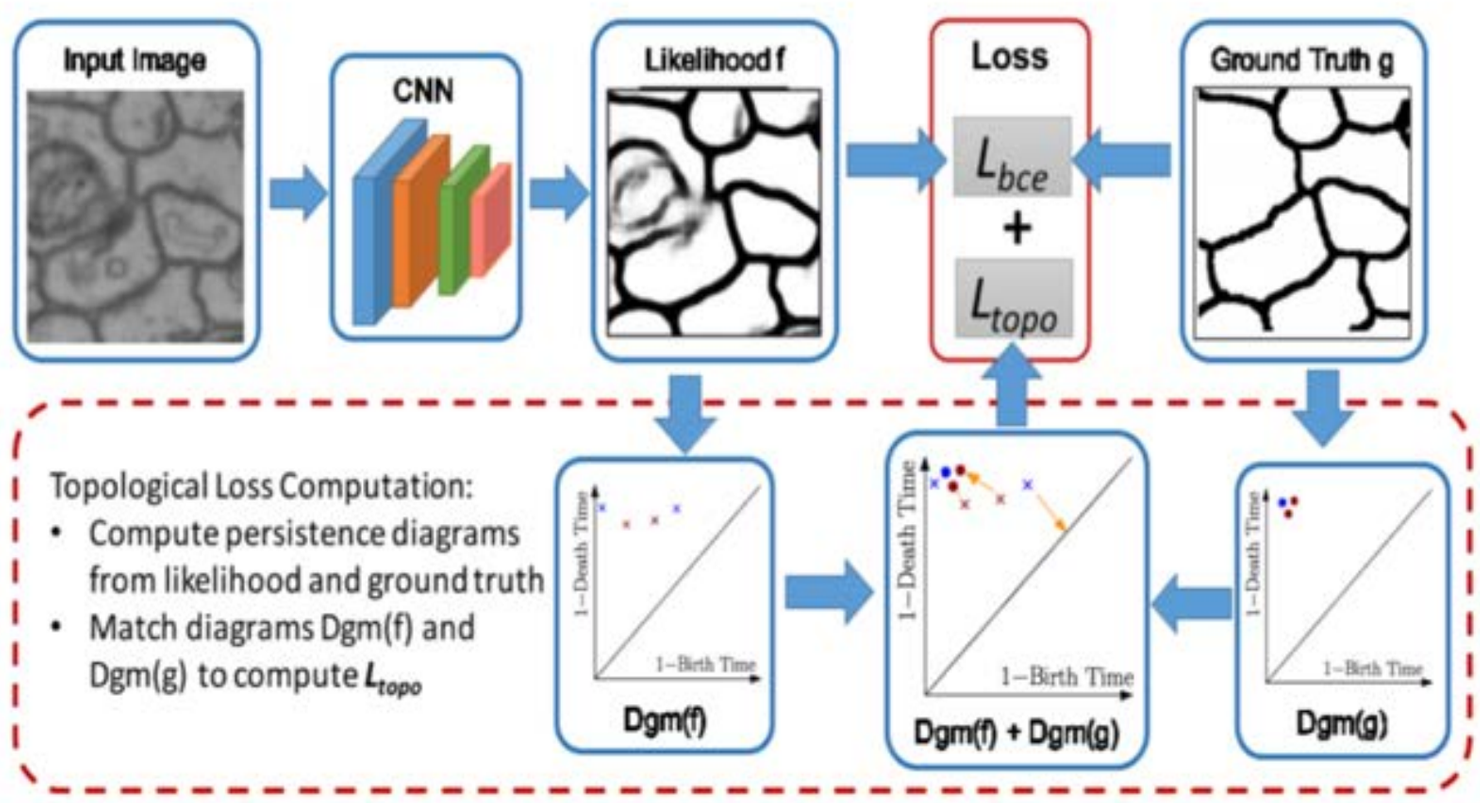
Dataset	PSS-K	PWG-K	SW-K	PF-K	PersLay
ORBIT5K	72.38($\pm$2.4)	76.63($\pm$0.7)	83.6($\pm$0.9)	85.9($\pm$0.8)	**87.7($\pm$1.0)**
ORBIT100K	—	—	—	—	**89.2($\pm$0.3)**



8 - Topological Data Analysis and Machine Learning

3- Optimizing topological loss functions

Topological Loss for Neural Networks



TopoNet (Xiaoling Hu et al. 2019) is a deep segmentation method that learns to segment with correct topology.

A topological loss enforces the segmentation results to have the same topology (persistent homology) as the ground truth.
voi aussi dans les survey de frontiers

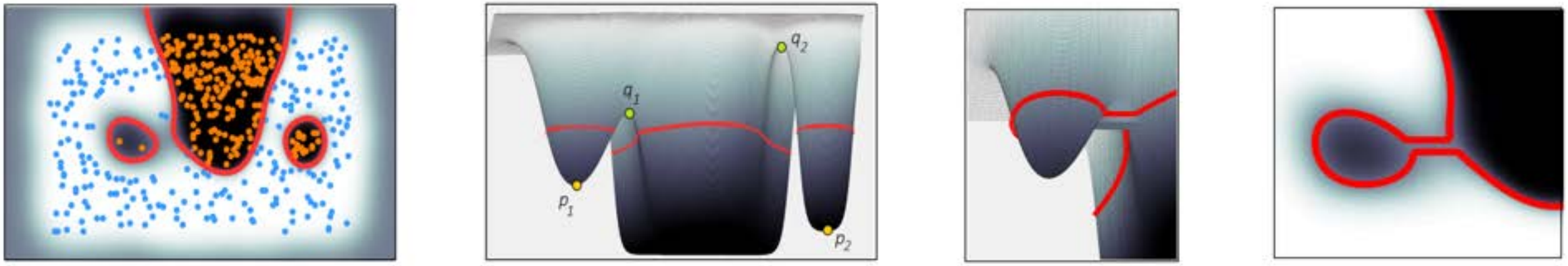
Topological Regulizer

Combine a topological penalty L_{topo} with a standard loss function

$$L(f, X_n) = \sum_{i=\ldots n} \ell(f(x_i), y_i) + \lambda L_{topo}(f)$$

where the first term could be cross-entropy loss, hinge loss and so on.

Example: Control the "topological complexity of the boundary classification of a classifier (Chen etal. AISTATS 2019).



(a) Red curve is the classification boundary S_f. (b) shows the graph of the classifier function f , with S_f marked in red. (c) Pushing the saddle q_1 down to remove this left component in S_f as shown in (d).

Computational Topology for Data Analysis, Tamal Krishna Dey and Yusu Wang, to be Published by Cambridge University Press, 2021

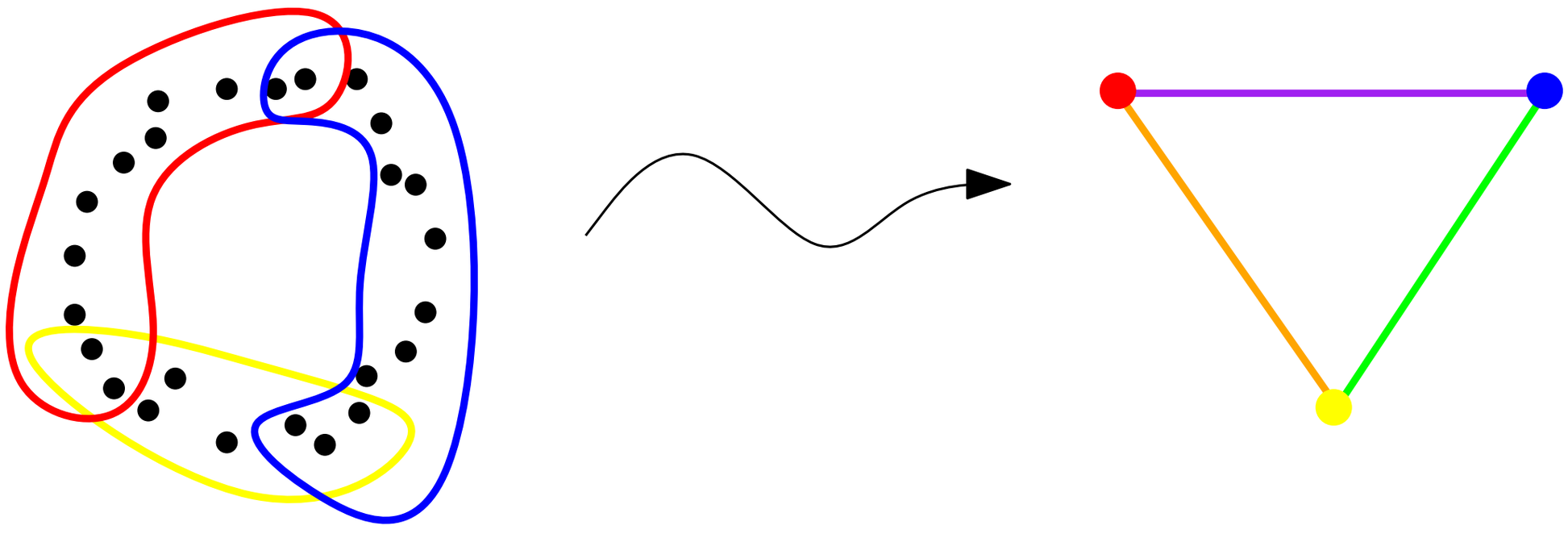
9 - Mapper

Topological exploratory data analysis

Goal: build simplicial complexes that have the same topology (homology groups, homotopy equivalence, homeomorphism, isotopy) than the data sets.

Idea:

- Define 'covers' from the data (for instance by grouping data points in 'local clusters'.
- Summarize the data through the combinatorial/topological structure of intersection patterns of these 'covers'.

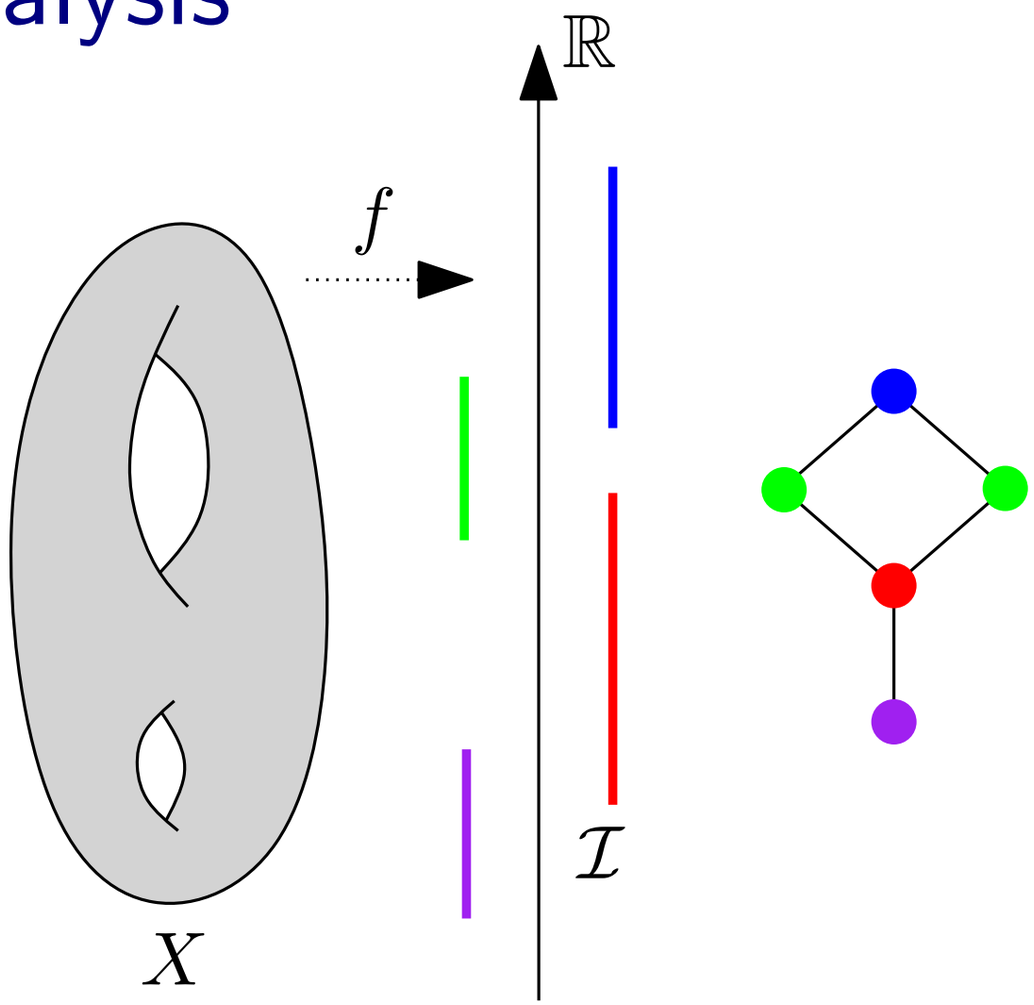


Topological exploratory data analysis

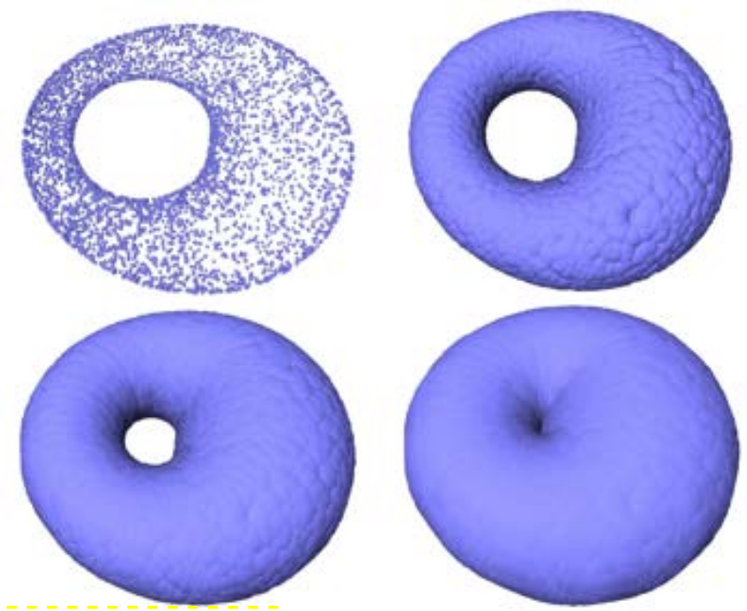
Q: How to build meaningful covers?

Two directions:

1. Using a function (lens) defined on the data:
$\rightarrow$ the Mapper algorithm
$\rightarrow$ exploratory data analysis



2. Covering data by balls:
$\rightarrow$ distance functions frameworks, persistence-based signatures,...
$\rightarrow$ geometric inference, provide a framework to establish various theoretical results in TDA.



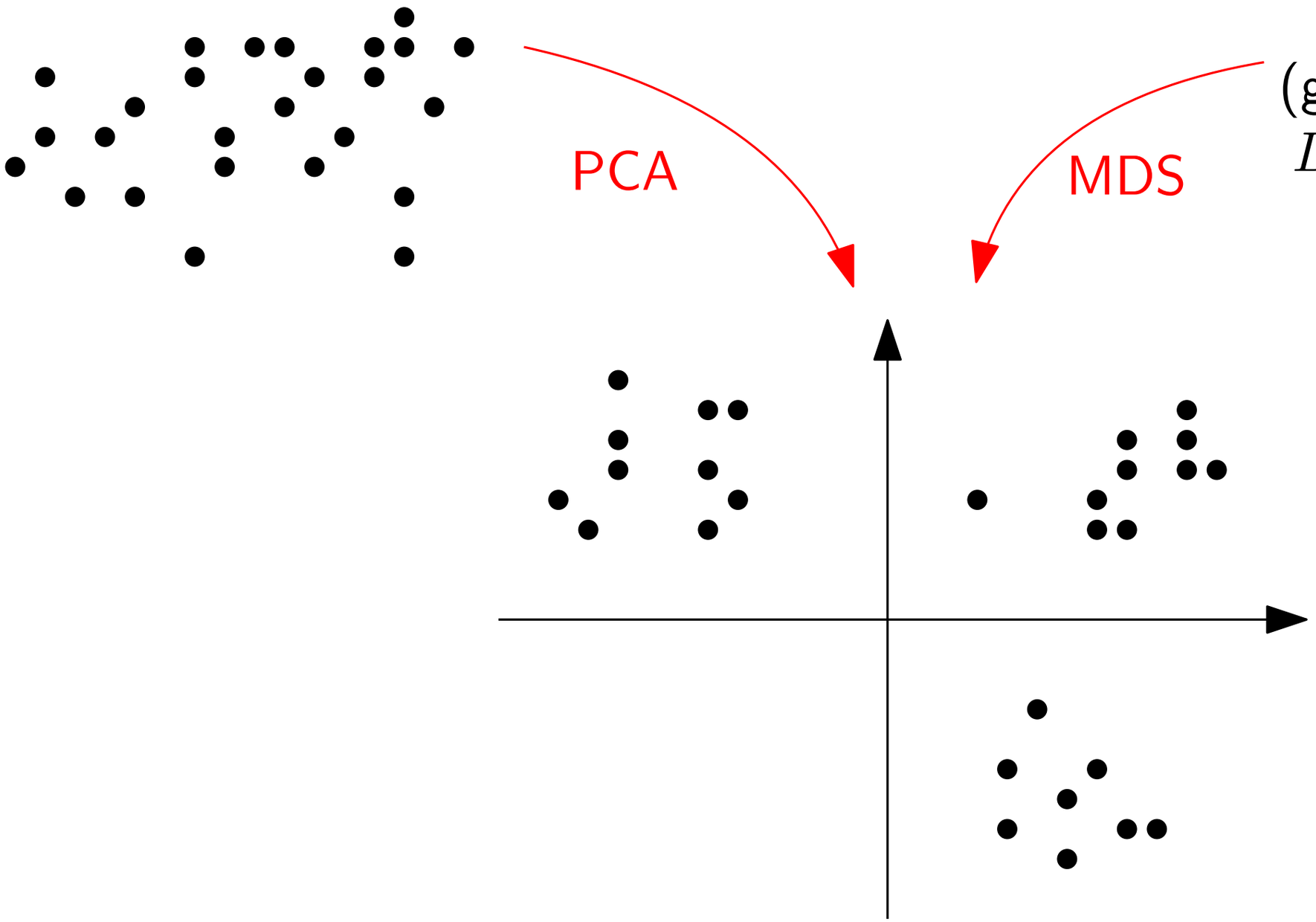
Unsupervised method for high dimensional or complex data

- Mapper: exploratory data analysis (unsupervised method) for high dimensional or complex data.
- How does Mapper compare with other unsupervised learning methods ?
- Standard approches for exploratory data analysis combine clustering and dimension reduction:
  - Dimension reduction (for regularization) then clustering
  - Clustering then dimension reduction (for visualization)
- Many approaches to clustering : hierarchical clustering, k-means clustering, distribution-based clustering, density based clustering, spectral clustering ...
- Many approaches to dimension reduction: PCA, Kernel PCA, Isomap, t-SNE, UMAP etc.

Linear dimension reduction and clustering for exploratory data analysis

Point cloud in Euclidean space

Matrix of pairwise distances
(general metric space)
$D = (d(x_i, x - j))_{i,j}$

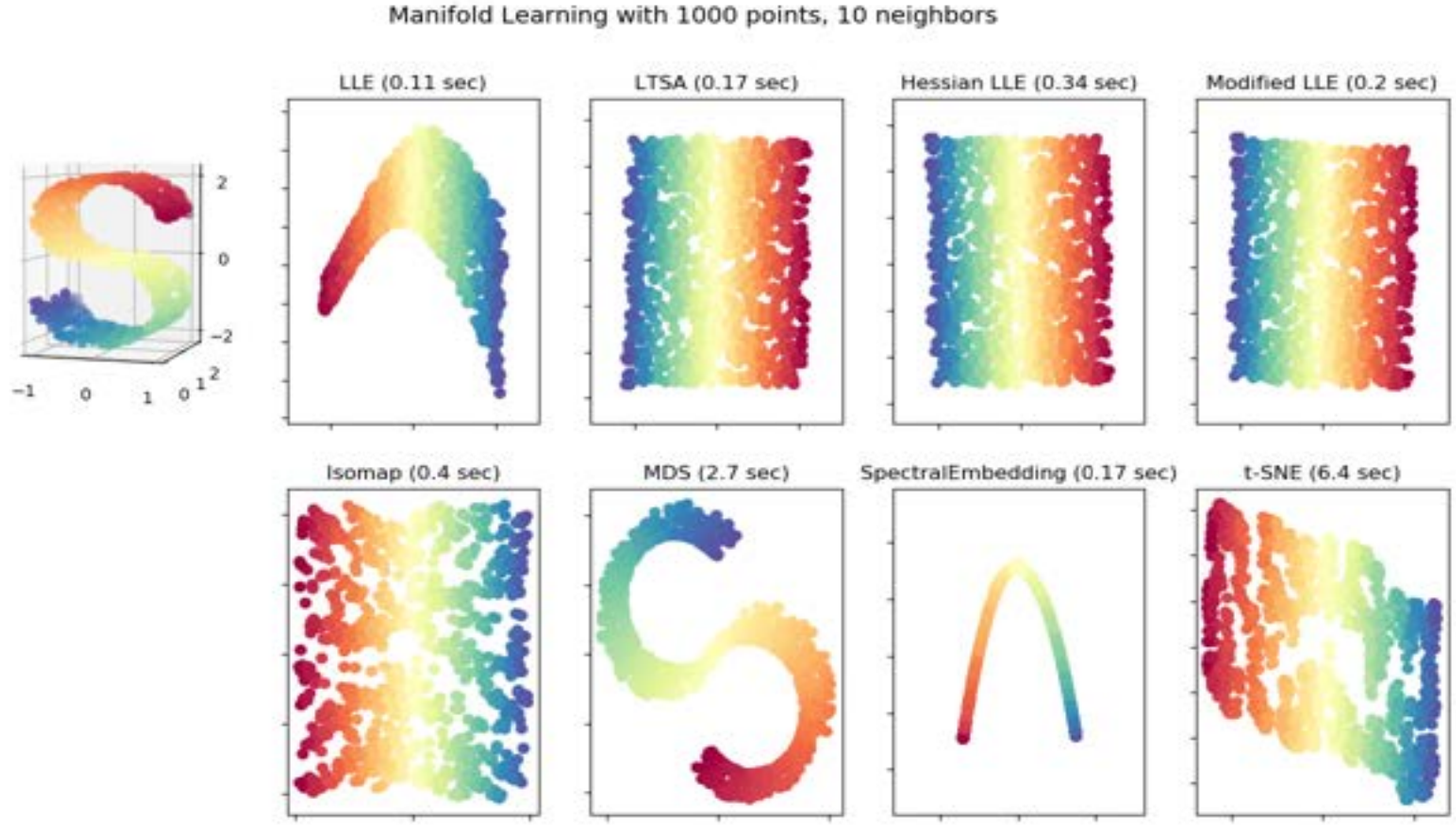


Dimension reduction (regularization) then clustering
or
Clustering then dim reduction (visualization)

Non Linear dimension reduction

- (Classical) PCA and MDS become inefficient when the data is located around highly non linear manifolds

- Many **non linear** alternative approaches

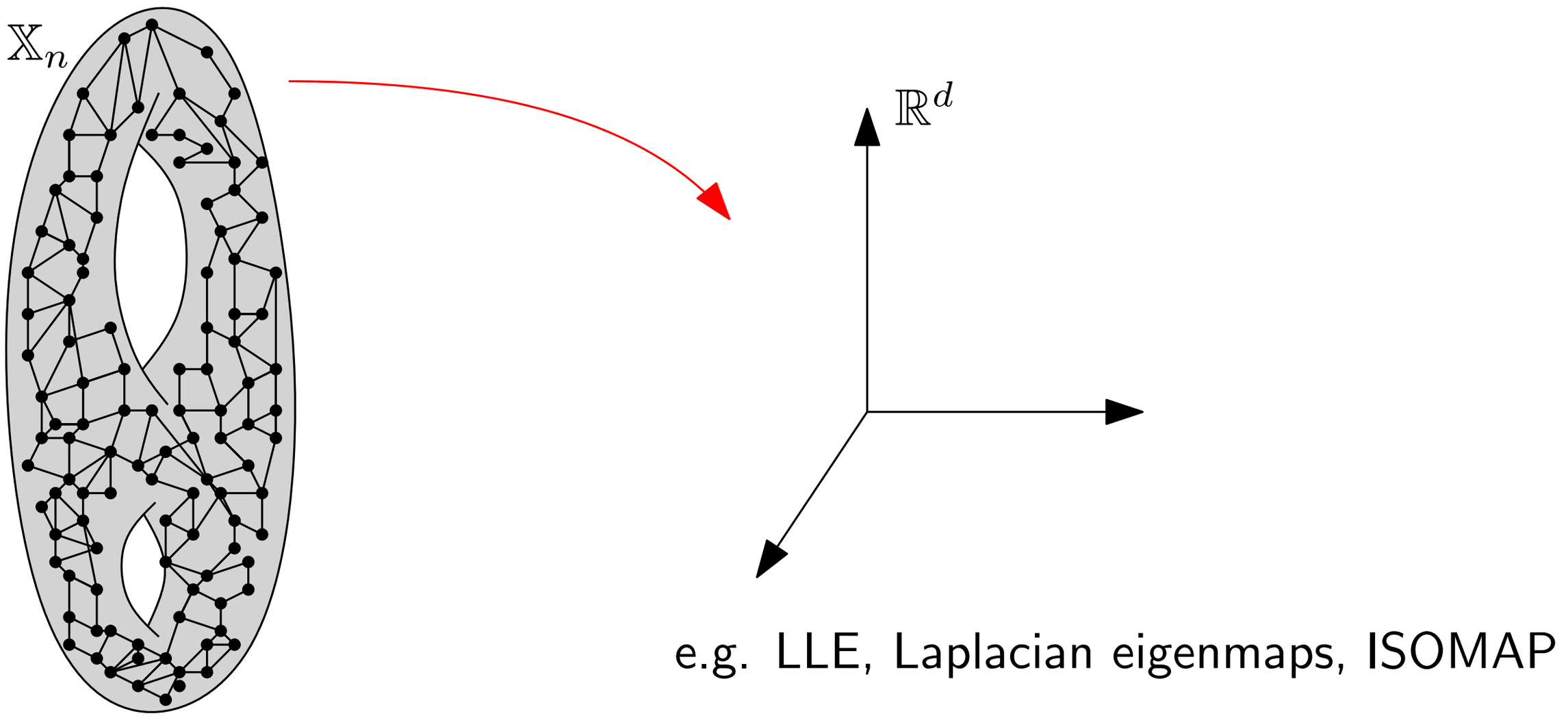


source : sckit-learn library

- Most of these methods are embedding methods: they intend to find an embedding of the data in some low dimensional space such that the "geometry" of the embedding is "as similar" as possible as the one of the initial data.

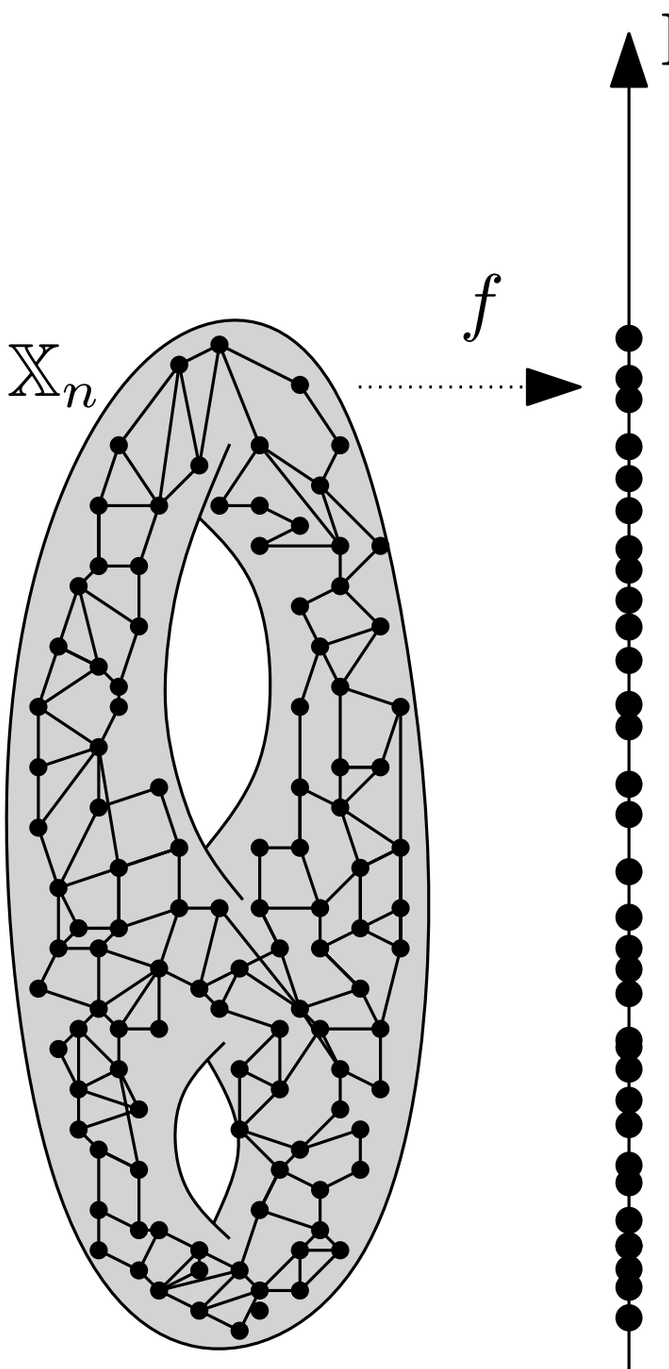
Neighborhood graphs

- Many non linear reduction methods use neighborhood graphs to approximate / represent the support of the data and then derive an embedding.



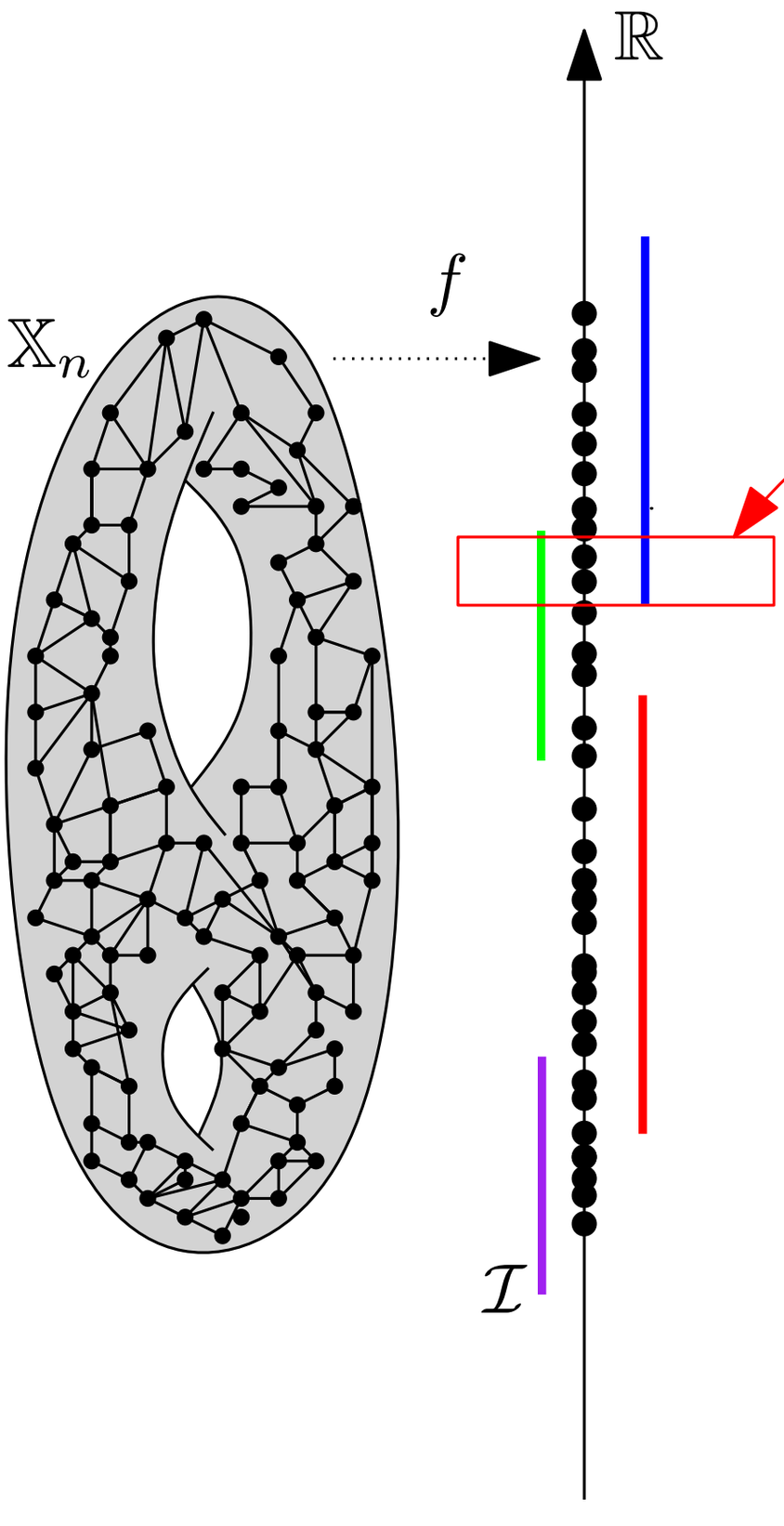
e.g. LLE, Laplacian eigenmaps, ISOMAP

A simple version of Mapper



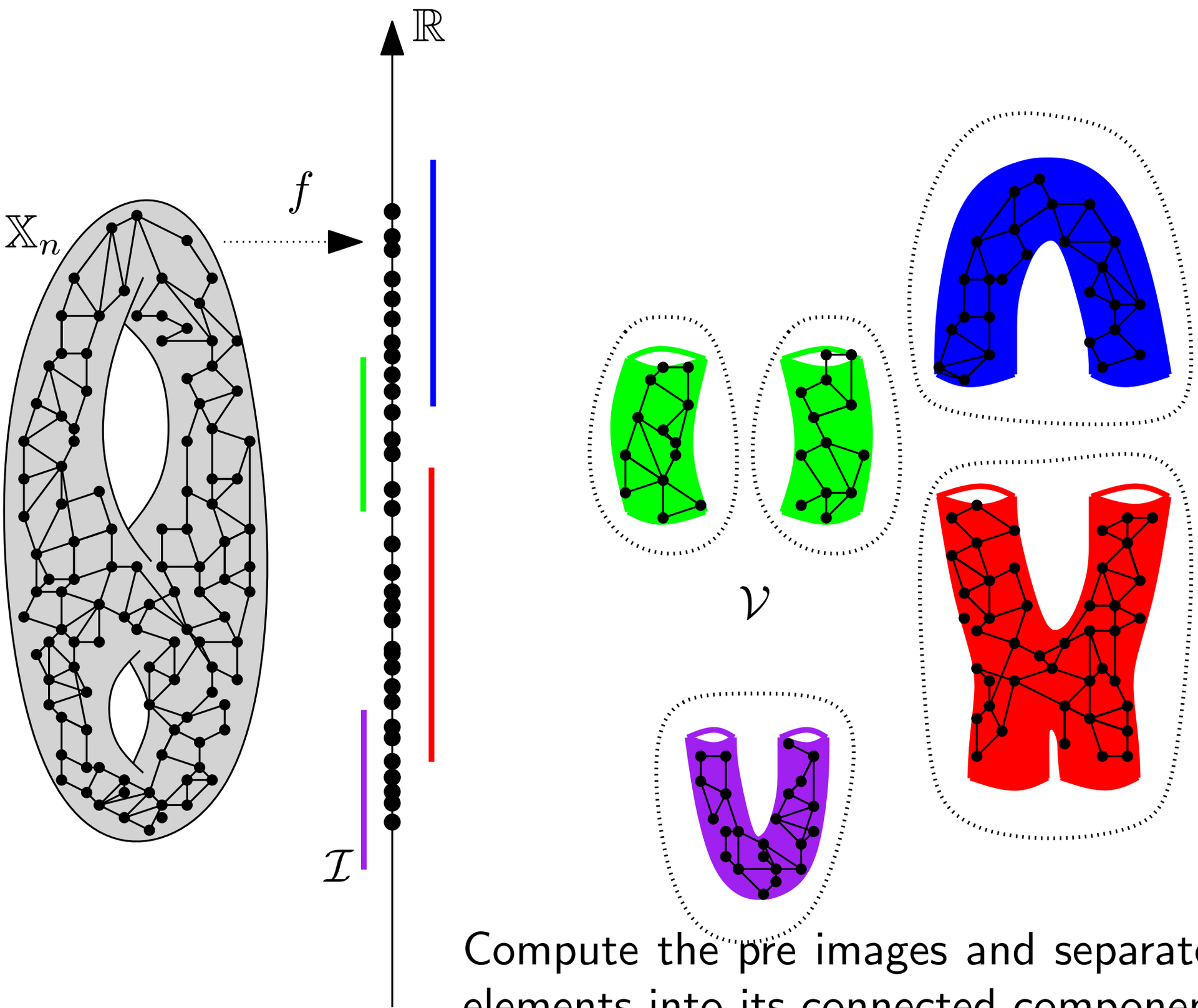
- For Mapper we observe :
  - A neighborhood graph (or metric graph) of $\mathbb{X}_n$
  - The value of a filter function f at the data points: $f : \mathbb{X}_n \mapsto \mathbb{R}$.
- Mapper does not use the graph for producing an embedding but for directly representing the data.
- It is possible to derive a clustering of the data from the graph but we also want to take into account f.
- Mapper : clustering of $\mathbb{X}_n$ based on a filter function and on the neighborhood graph (connectivity of the clusters)

A simple version of Mapper



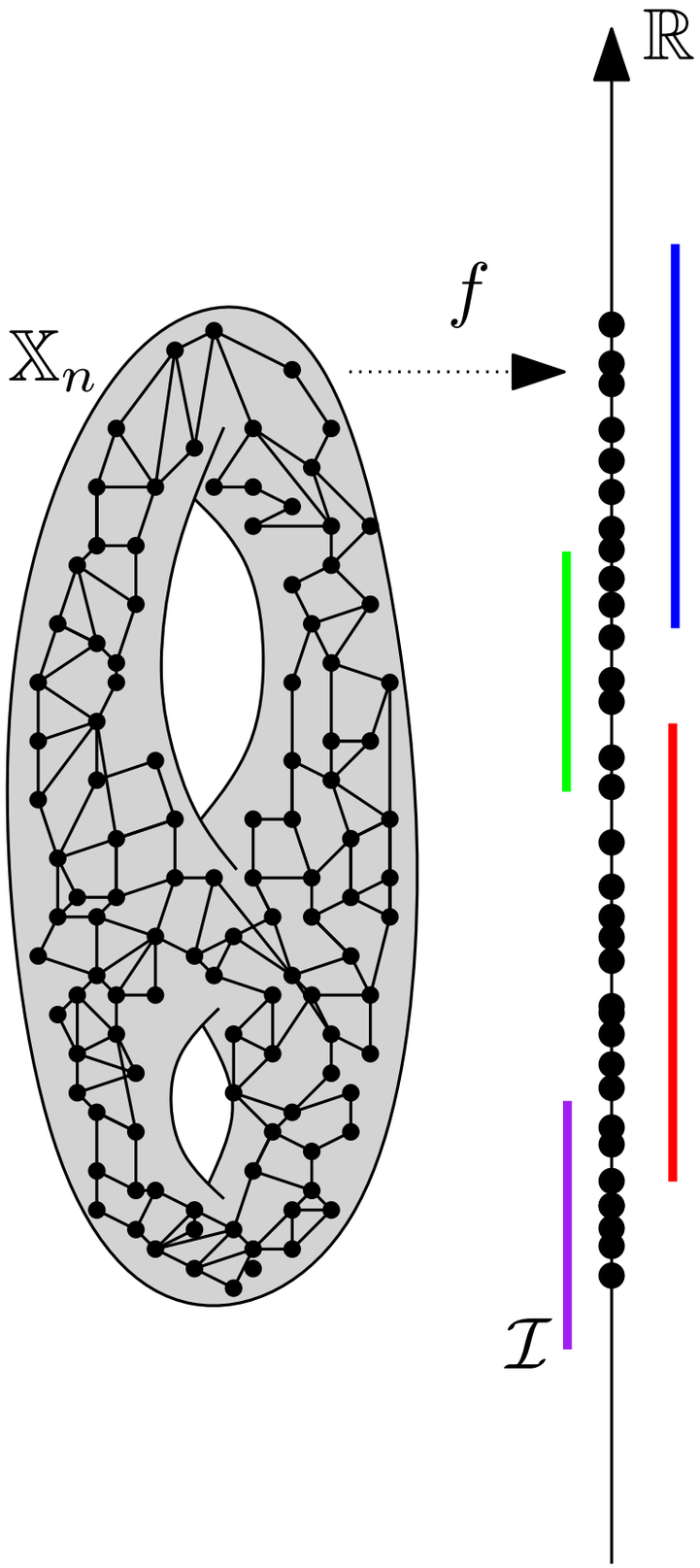
Cover $\mathrm{Im}(f)$ by open overlapping intervals: $\mathrm{Im} f \subseteq \bigcup_{I \in \mathcal{I}} I$

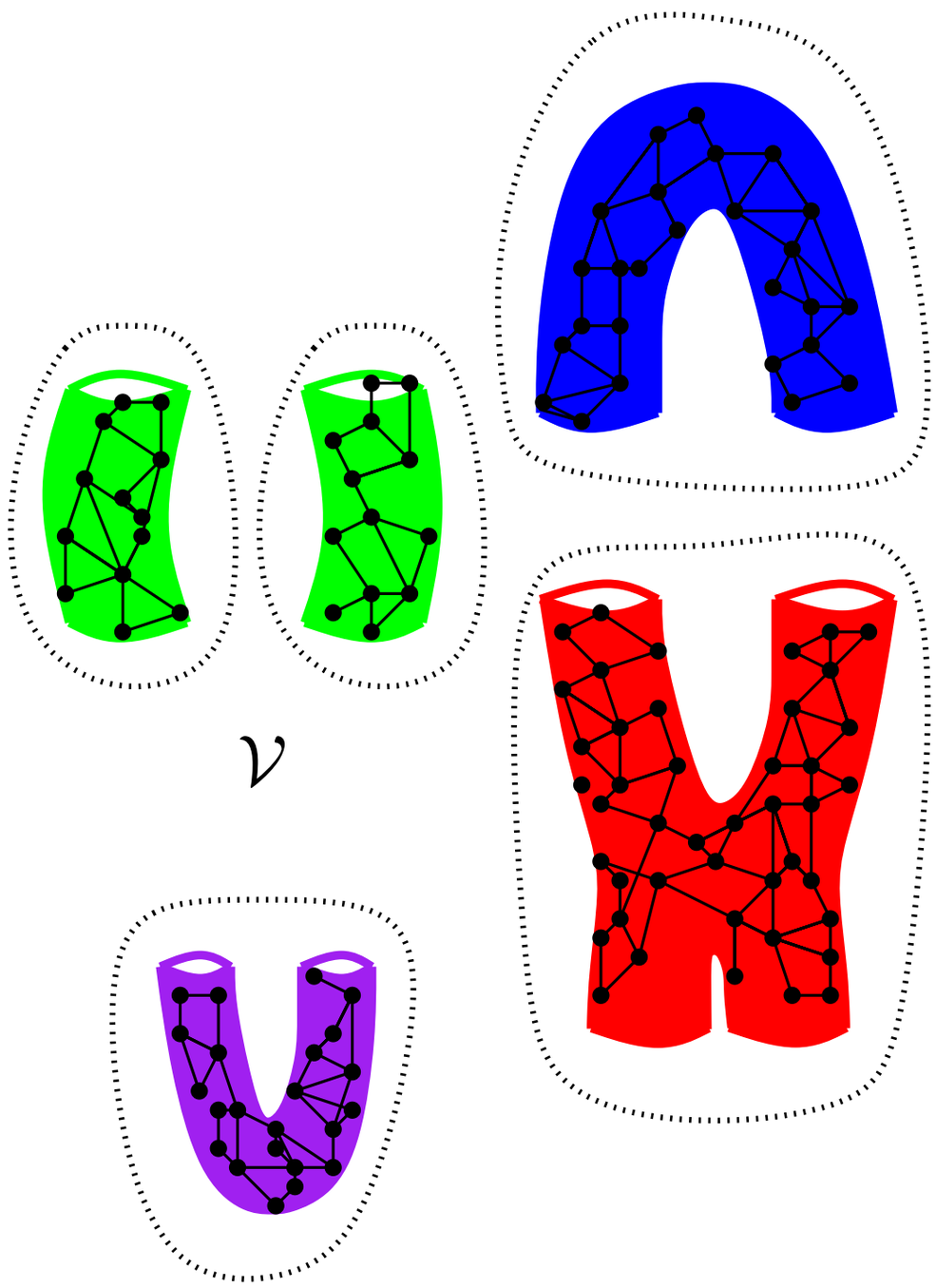
A simple version of Mapper

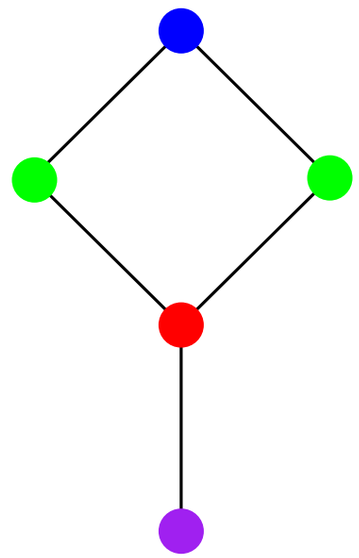


Compute the pre images and separate each of its elements into its connected components.

A simple version of Mapper







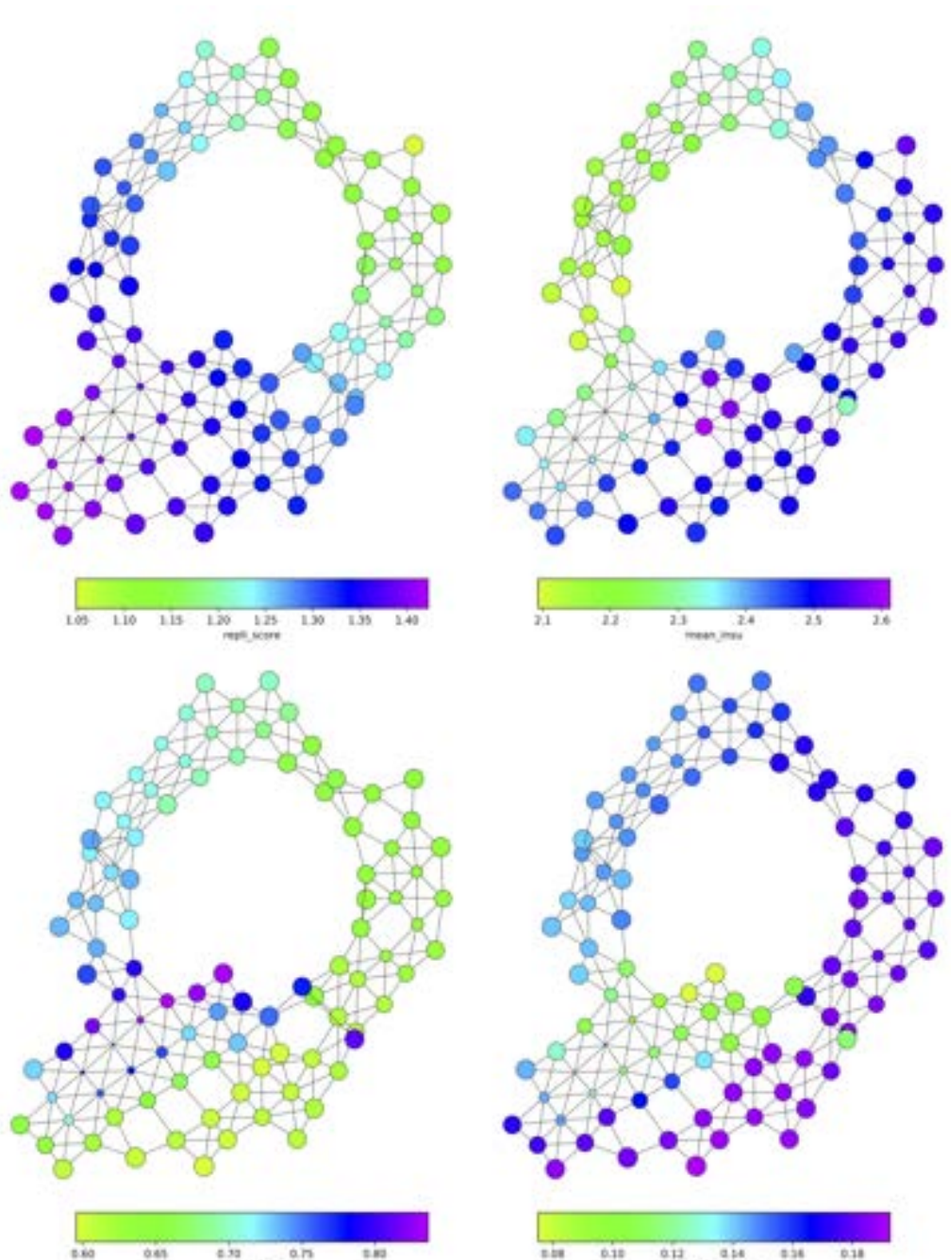
Mapper

The Mapper is the *nerve* of $\mathcal{V}$.

Applications of Mapper

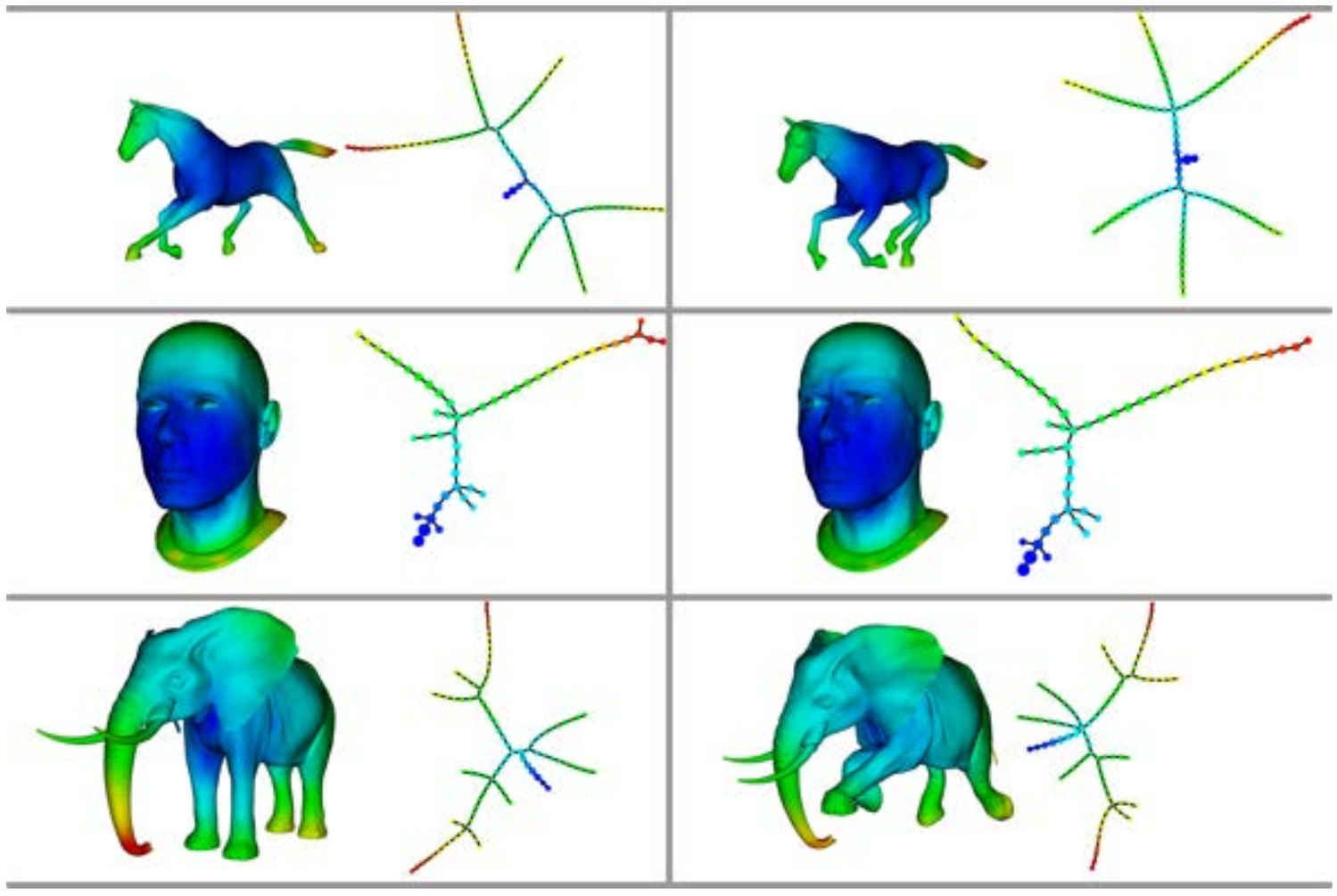
Two types of applications:

- clustering : identify statistically relevant subpopulations through patterns (flares, loops)
- feature selection based on the graph



[Topological Data Analysis of Single-cell Hi-C Contact Maps, Carriere and Rabadan 2018]

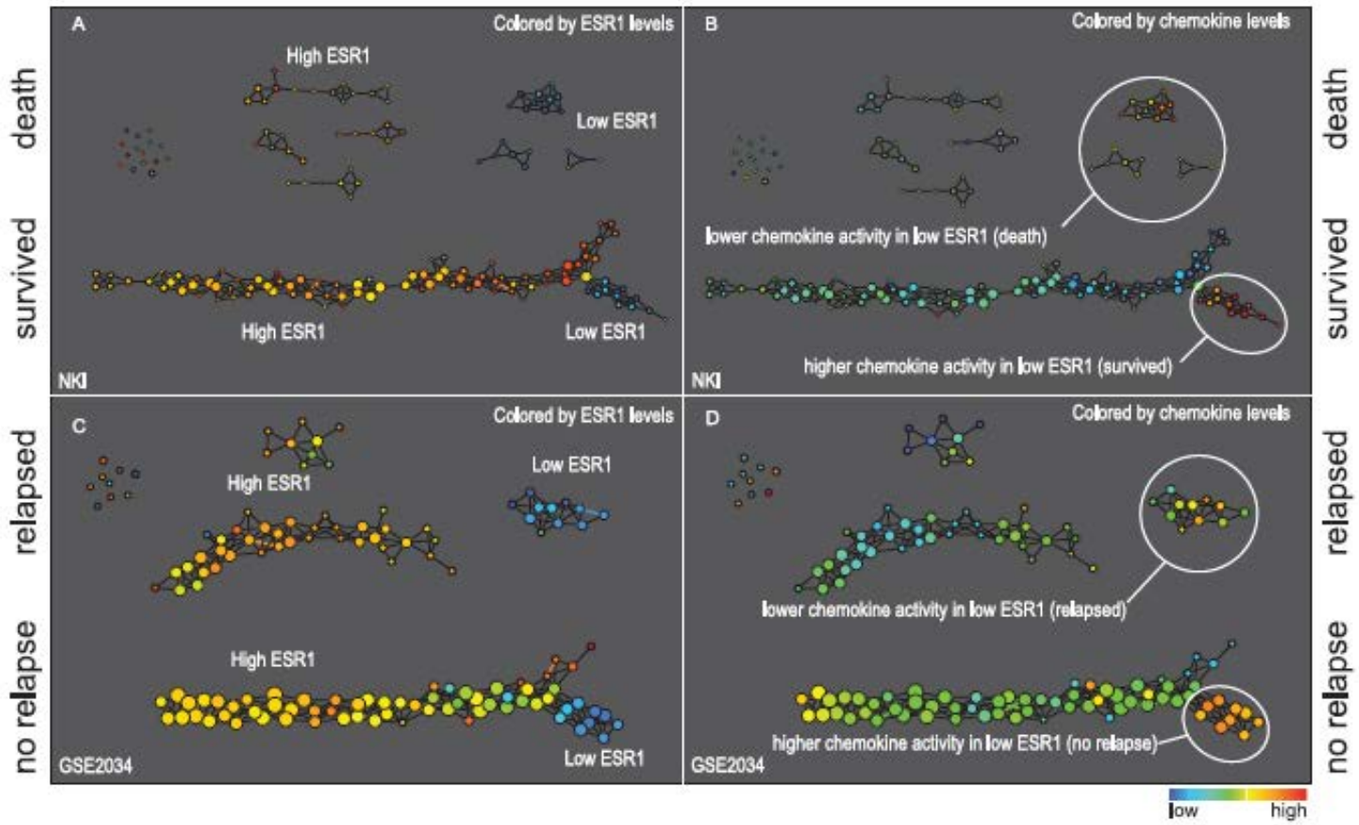
Applications of Mapper



3d shapes classification

[Singh, Mémoli, Carlsson 2007]

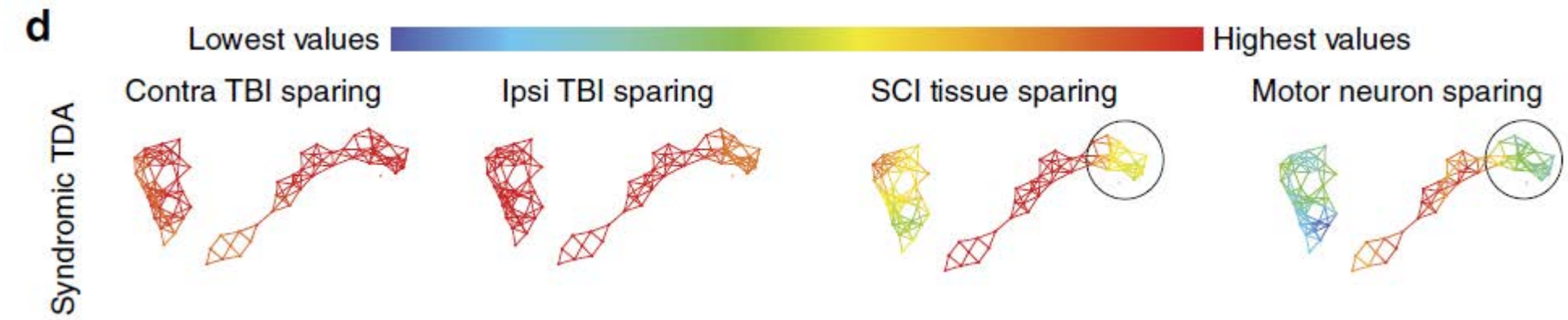
Applications of Mapper

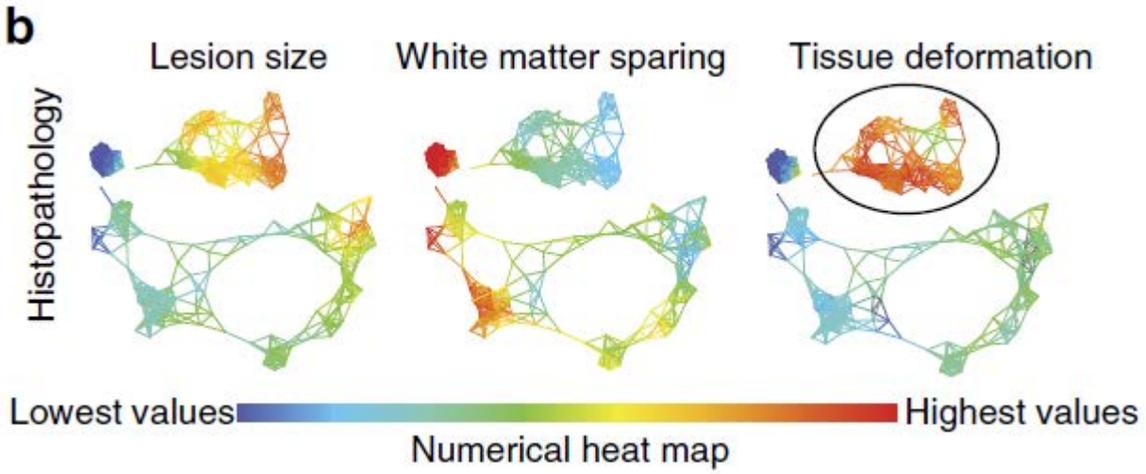


breast cancer subtype identification

[Nicolau et al. 2011]

Applications of Mapper





recovery from spinal cord injuries

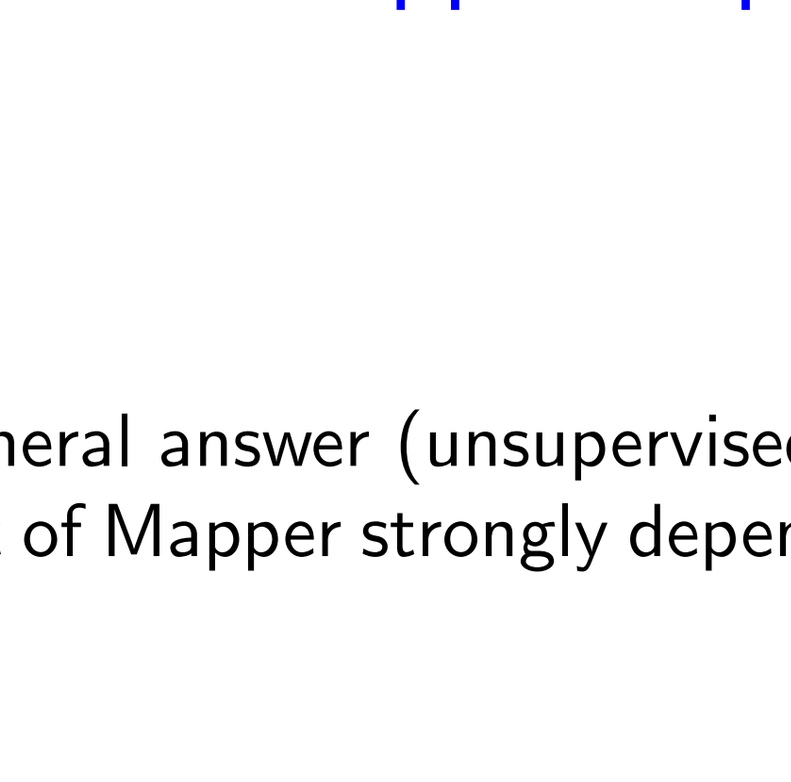
[Nielson et al. 2015]

Many other applications

- Implicit networks in the US house of representatives (Lum et al. 2014)
- Protein folding pathways (Yao et al. 2009)
- Diagnostic of pulmonary embolism (Rucco et al. 2014)
- ...

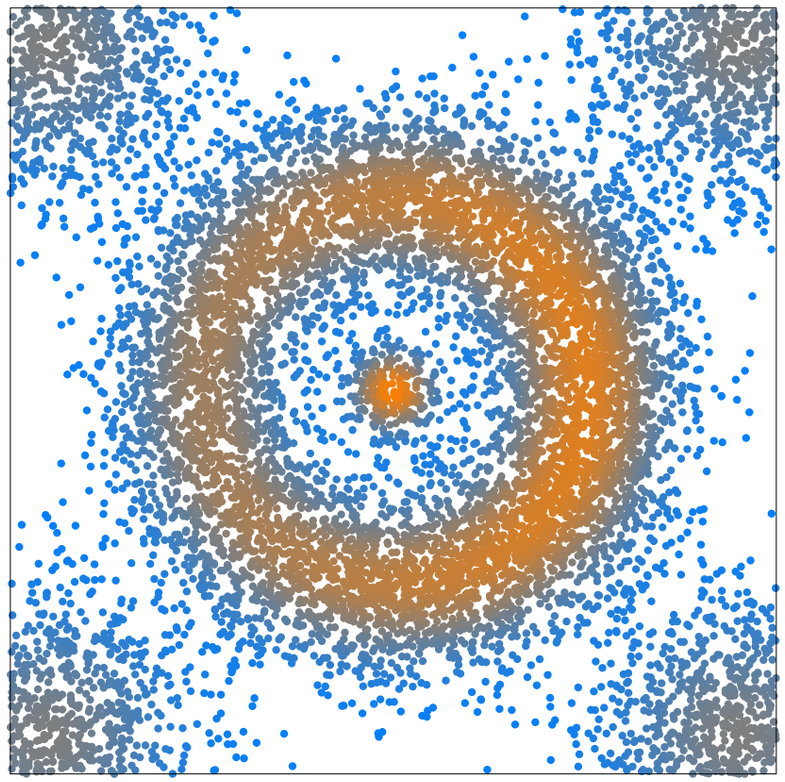
Mapper in practice: filter choice

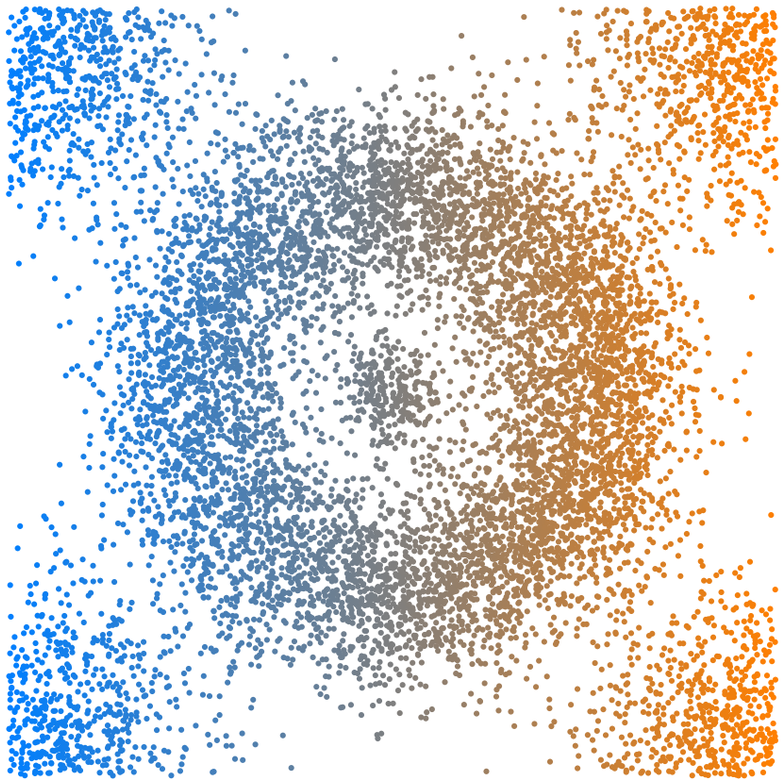
No general answer (unsupervised setting) but the output of Mapper strongly depends on this choice.



$\hat{f}$ = density estimator

f_x = x-coordinate

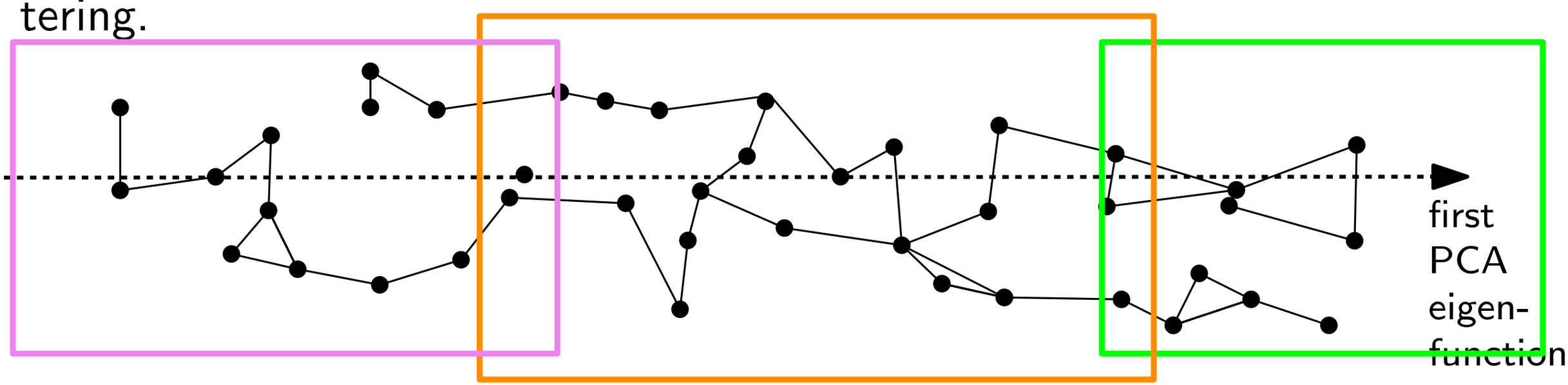




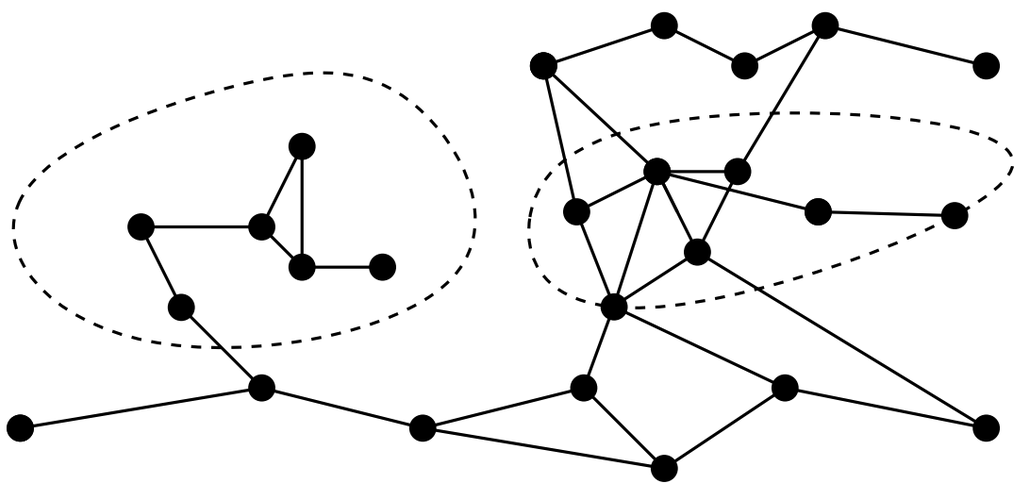
Mapper in practice: filter choice

- **Co variable filter:** focus the analysis on one particular variable. First, it finds a clustering according the value of the variable and then it refines the clustering according to the graph.

- **PCA eigenfunctions:** mix PCA based clustering and graph based clustering.



- **Density estimator filter:** mix density based clustering and graph based clustering



- **eccentricity filter**, ...

Mapper in practice: parameter choice

Parameters:

- function $f : \mathbb{X}_n \to \mathbb{R}$ ← filter

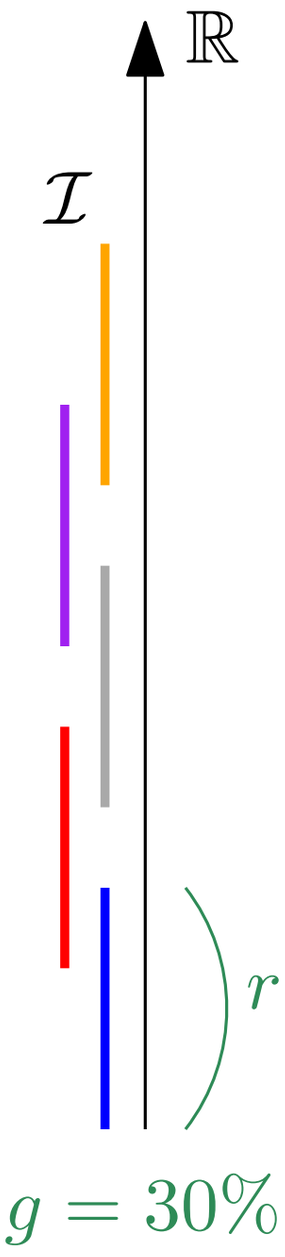
- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals ← range scale

$\to$ uniform cover $\mathcal{I}$:

  - resolution / granularity: r (diameter of intervals)

  - gain: g (percentage of overlap)

- neighborhood size δ ← geometric scale



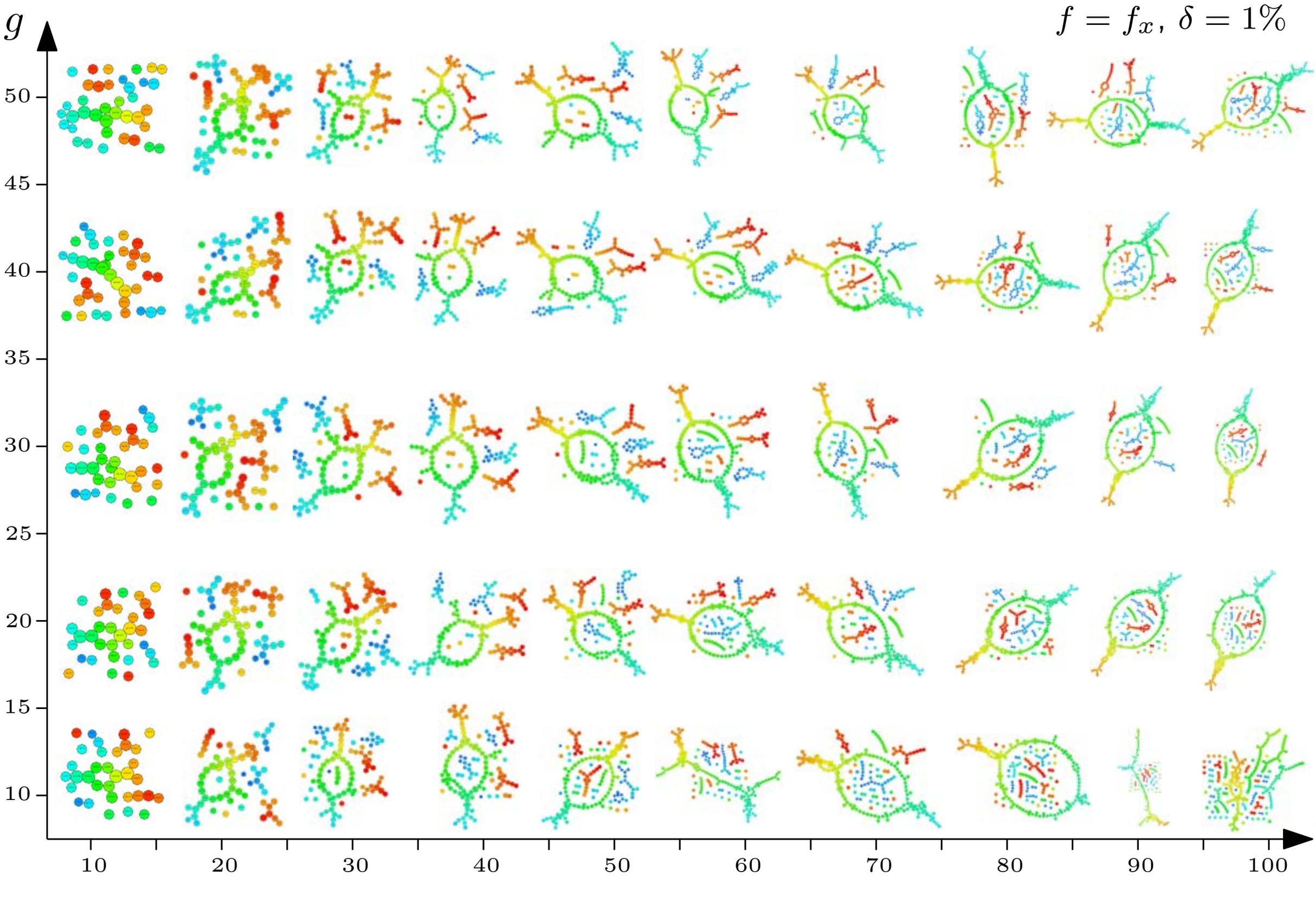
Parameters choice ?

- Cross validation ?
- "brute-force - stability" approach ?

high-dimensional data sets[40,48]. This is performed automatically within the software, by deploying an ensemble machine learning algorithm that iterates through overlapping subject bins of different sizes that resample the metric space (with replacement), thereby using a combination of the metric location and similarity of subjects in the network topology. After performing millions of iterations, the algorithm returns the most stable, consensus vote for the resulting 'golden network' (Reeb graph), representing the multidimensional data shape[12,40].

Nielson et al.: *Topological Data Analysis for Discovery in Preclinical Spinal Cord Injury and Traumatic Brain Injury*, Nature, 2015

Parameters choice: "brute-force - stability" approach ?



Analysis of the Mapper

[Carriere and Oudot 2016] : structure and stability of the Mapper in a deterministic setting :

- the topological structure of the Mapper can be described by looking at appropriate signatures : extended persistence homology
- rigorous connexion between the (underlying) Reeb graph and the Mapper
- convinient (pseudo) metrics to compare the Mapper and Reeb graph
- locally a metric [Carriere Oudot 2017]

See also [Babu 2013], [Dey and Wang 2013], [Munch and Wang 2016] for results about the convergence of the Mapper.

This work [Carriere M. Oudot 2018] :

- rigorous statistical framework for the Mapper ;
- statistical convergence of the Mapper ;
- data driven methods for tuning parameters in Mapper ;
- confidence regions for Mapper signatures.

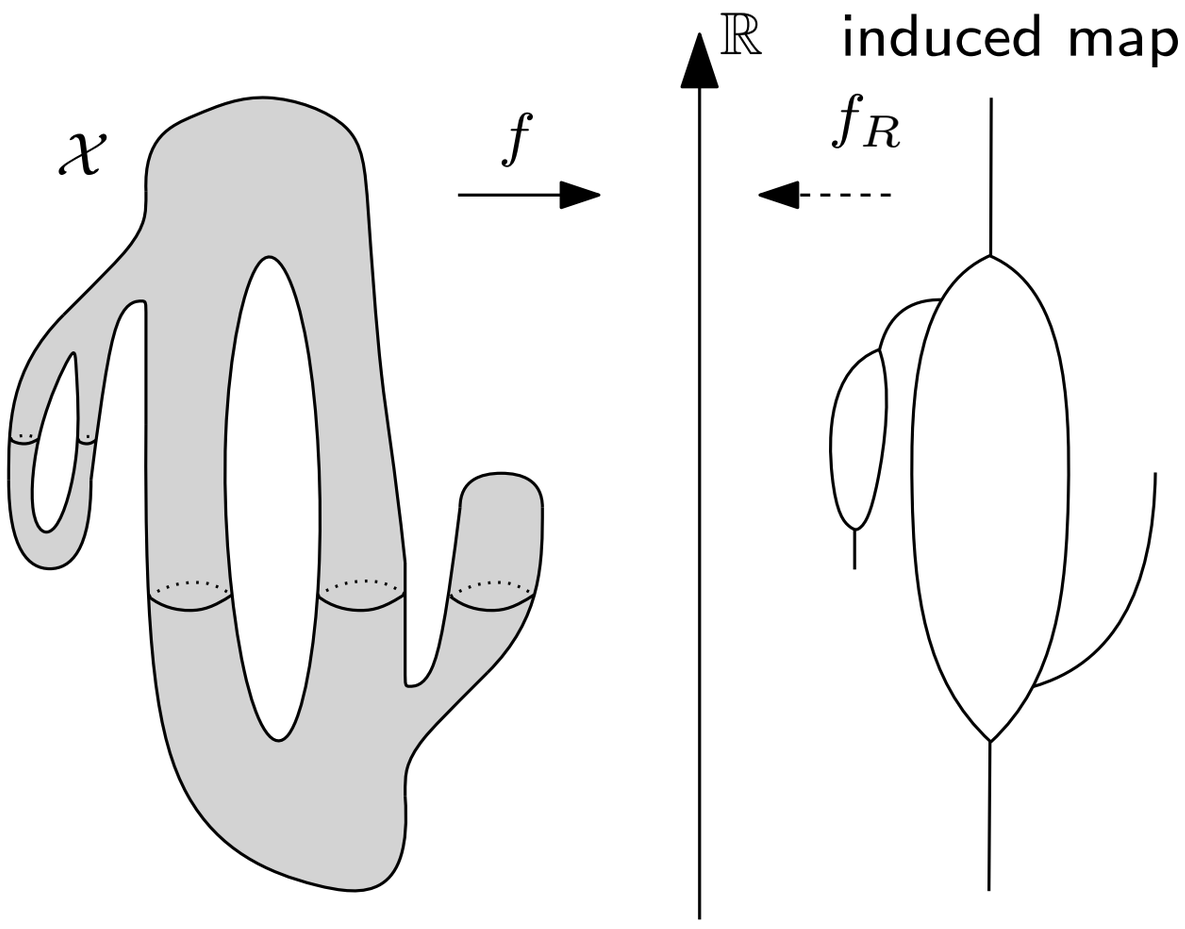
Statistical framework for Mapper ?

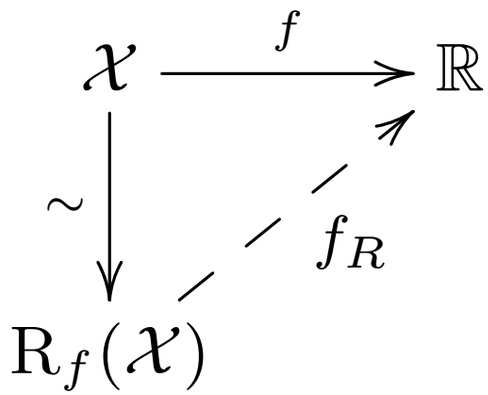
- Limit object for the Mapper : [Singh, Mémoli, Carlsson 2007] : Mapper is a *statistical version of the Reeb graph*
- We use topological signatures to compare Mapper with the Reeb Graph.
- Regularity assumptions and sampling assumptions to get the convergence.

Comparing Reeb graph with the Mapper

Reeb Graph

[Singh, Mémoli, Carlsson 2007] : Mapper is a *statistical version of the Reeb graph*



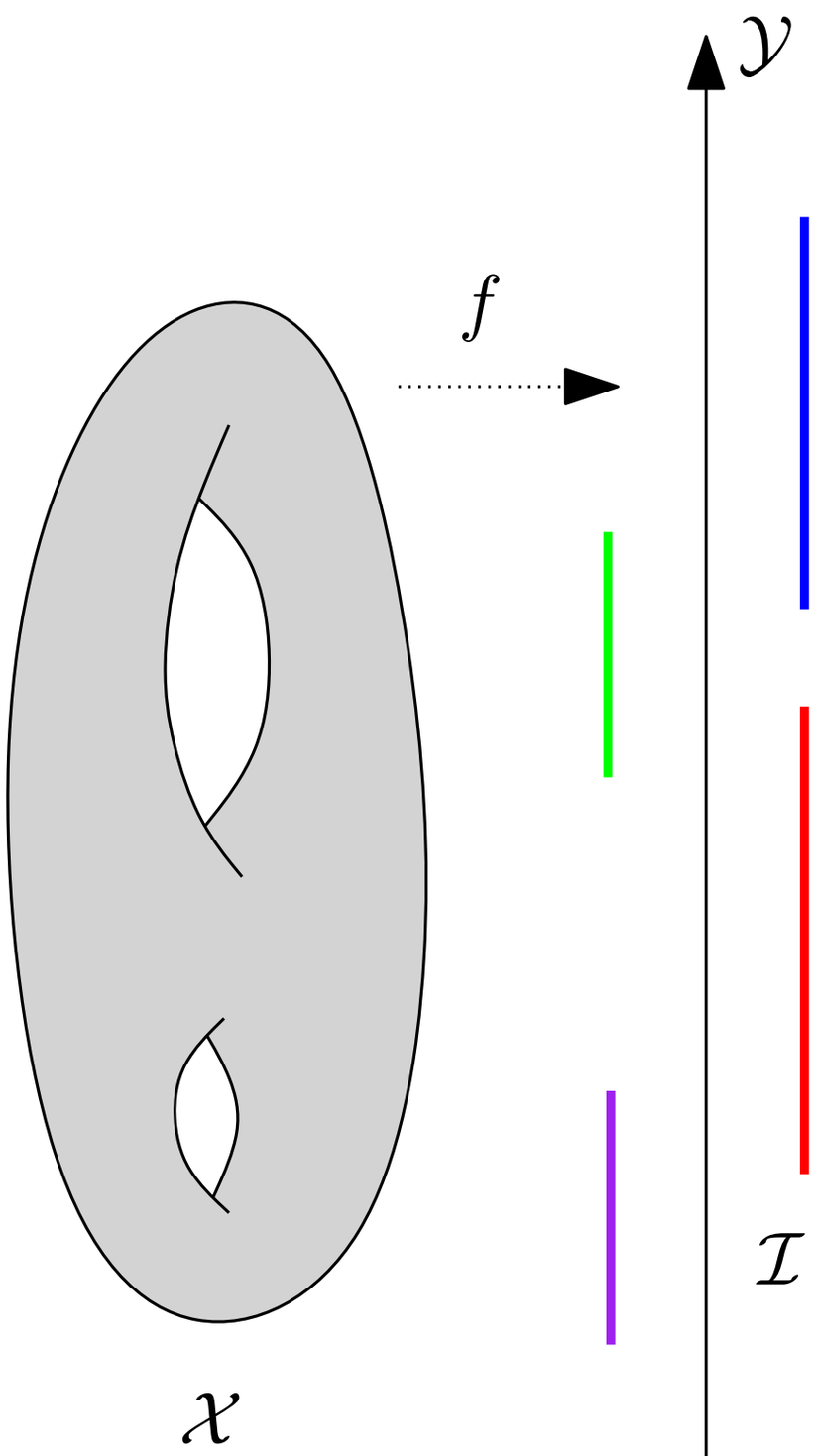


$x \sim y \iff [\ f(x) = f(y)$ and x, y belong to same cc of $f^{-1}(\{f(x)\})\ ]$

$\mathrm{R}_f(\mathcal{X}) := \mathcal{X}/\sim$ quotient space

Reeb Graph $\rightarrow$ Continuous Mapper $\rightarrow$ Empirical Mapper

Mapper in the continuous setting



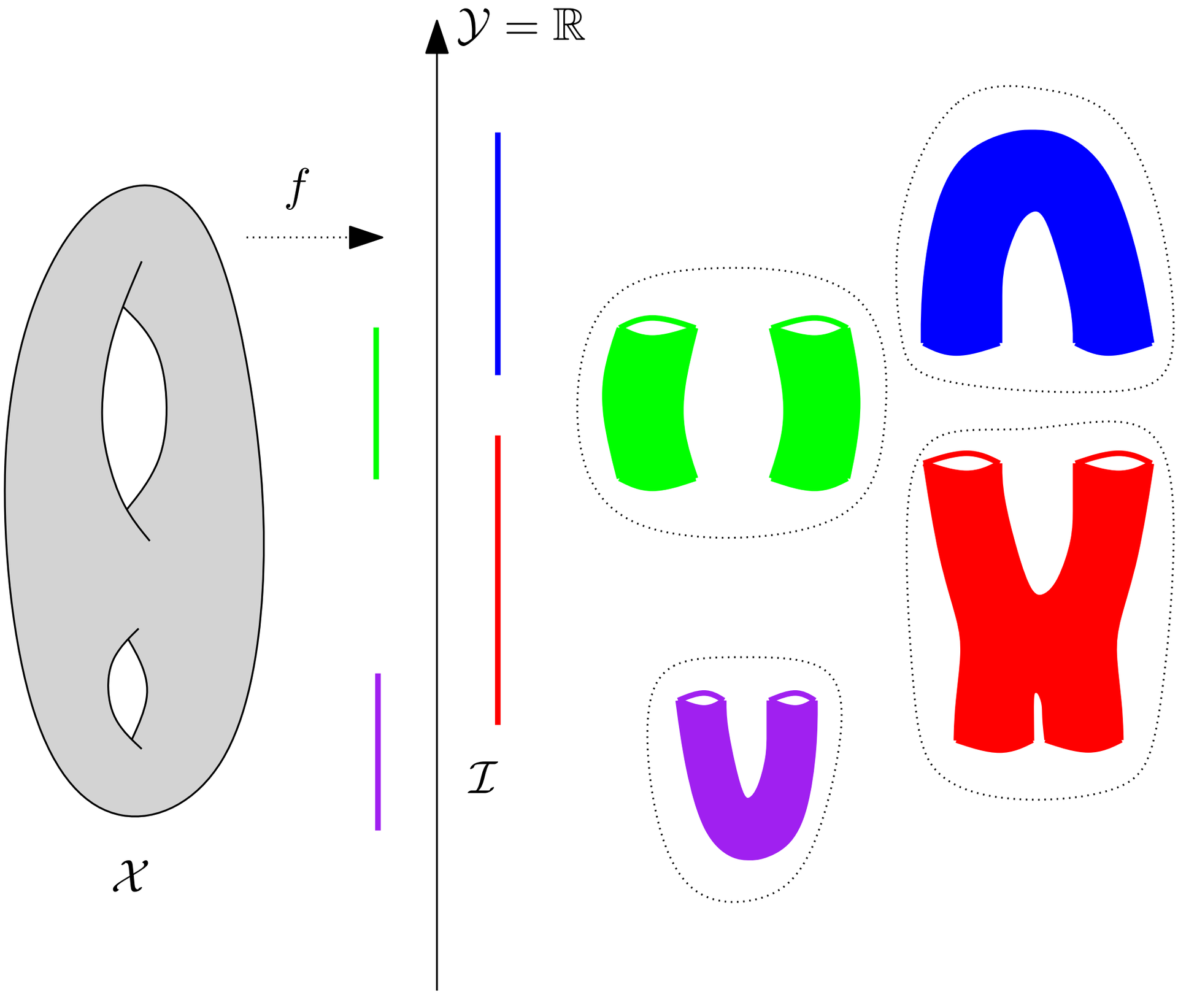
$\mathcal{X}$

f

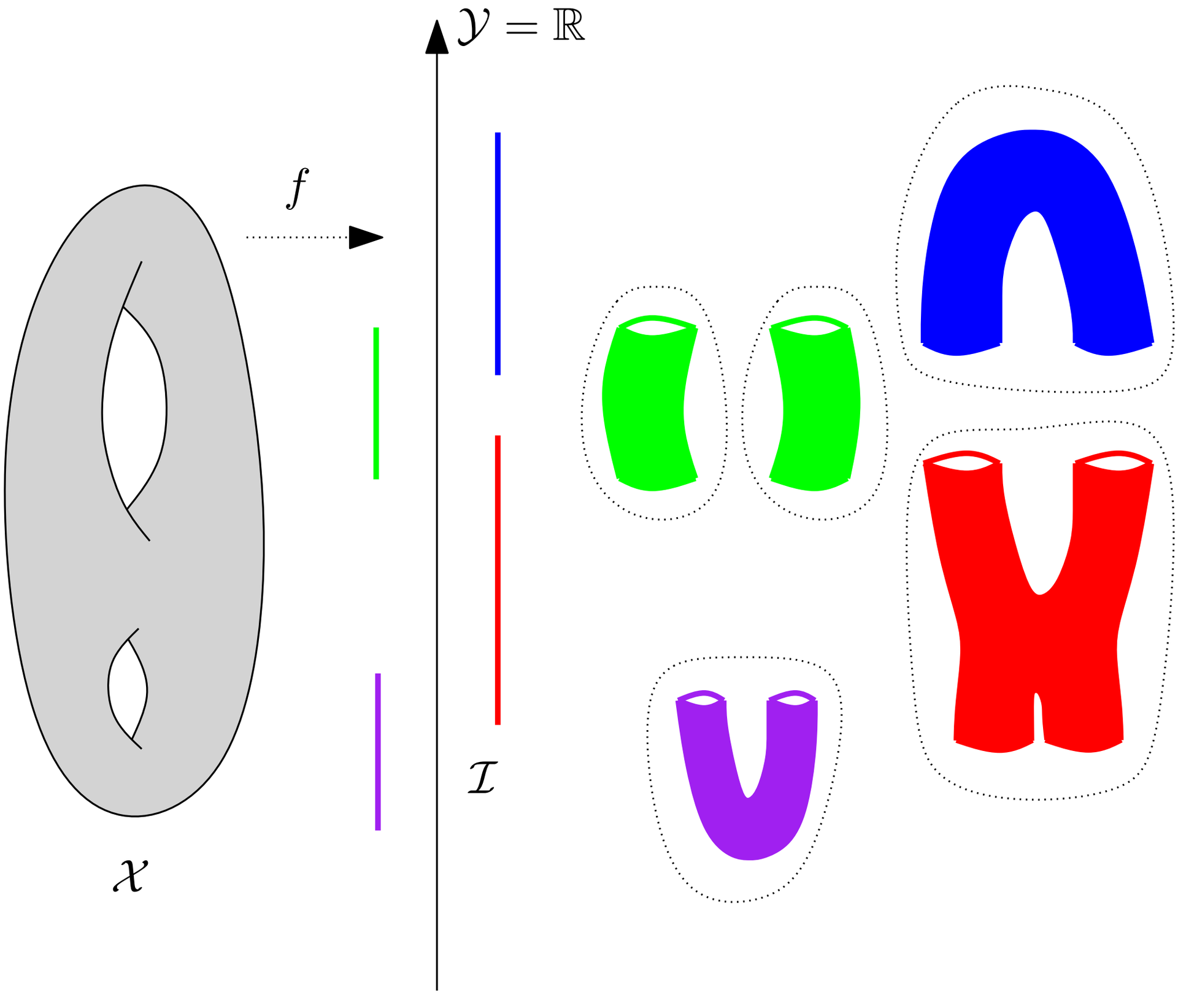
$\mathcal{Y} = \mathbb{R}$

$\mathcal{I}$

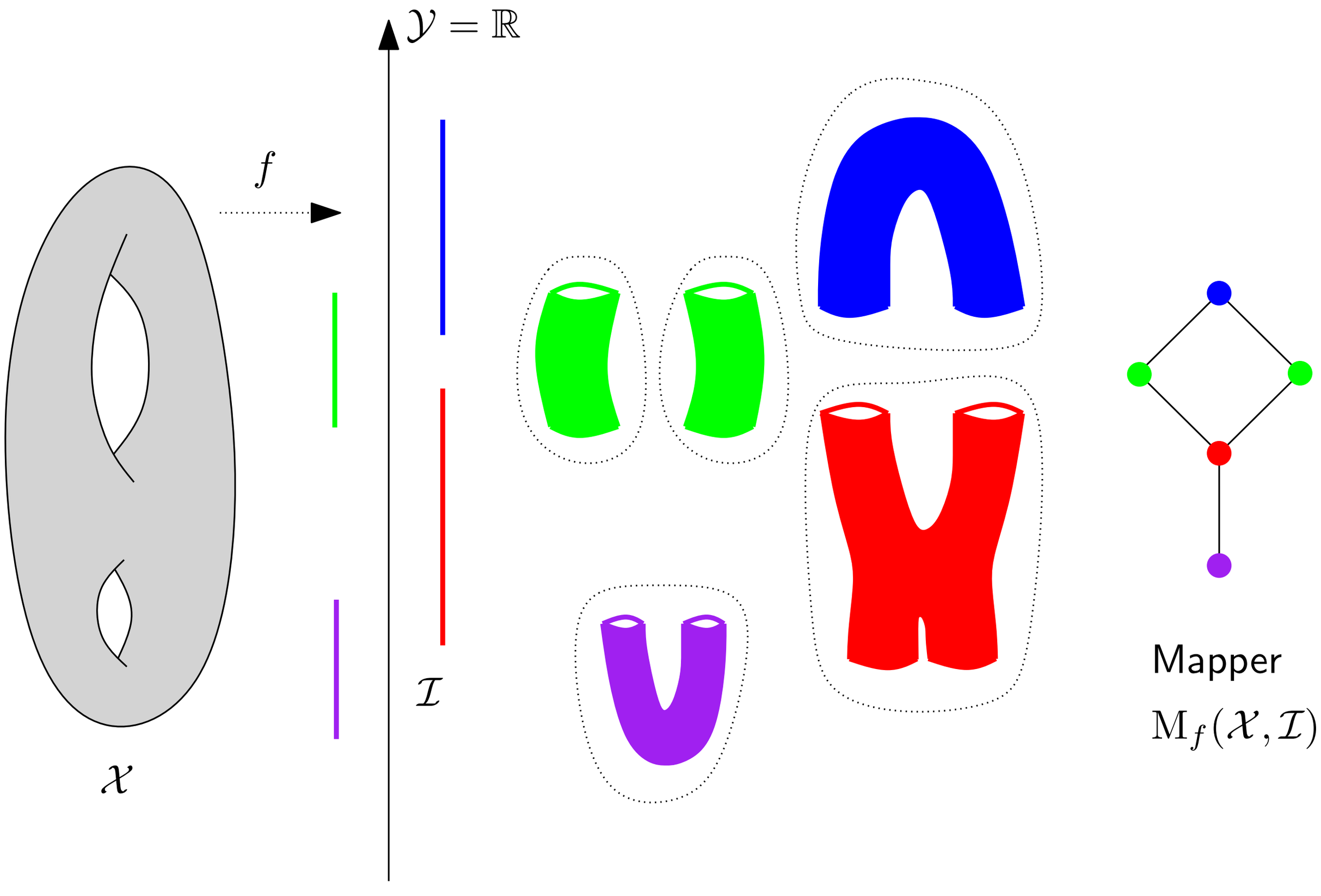
Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting



Mapper in the continuous setting

Input: - Metric space (or topological space) $\mathcal{X}$

- continuous function $f : \mathcal{X} \to \mathbb{R}$

- cover $\mathcal{I}$ of $\mathrm{im}(f)$ by open intervals: $\mathrm{im} f \subseteq \bigcup_{I \in \mathcal{I}} I$

Method:

• Compute *pullback cover* $\mathcal{U}$ of $\mathcal{X}$: $\mathcal{U} = \{f^{-1}(I)\}_{I \in \mathcal{I}}$

• Refine $\mathcal{U}$ by separating each of its elements into its various connected components in $\mathcal{X} \to$ connected cover $\mathcal{V}$

• The Mapper is the *nerve* of $\mathcal{V}$:

- 1 vertex per element $V \in \mathcal{V}$

- 1 edge per intersection $V \cap V' \neq \emptyset$, $V, V' \in \mathcal{V}$

- 1 k-simplex per $(k+1)$-fold intersection $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

Mapper in practice : our version of Mapper

Input: - point cloud $\mathbb{X}_n \subseteq \mathcal{X}$

- values of f at $\mathbb{X}_n$: $\mathbb{Y}_n := f(\mathbb{X}_n)$

- cover $\mathcal{I}$ of $\mathbb{Y}_n$ by open intervals of length r with overlapping proportion g : $\mathbb{Y}_n \subseteq \bigcup_{I \in \mathcal{I}}$

Method: • Compute δ-neighborhood graph G for $\mathbb{X}_n$

• Compute *pullback cover* $\mathcal{U}$ of $\mathbb{X}_n$: $\mathcal{U} = \{f^{-1}(I)\}_{I \in \mathcal{I}}$

• Refine $\mathcal{U}$ by separating each of its elements into its various connected components in $G \rightarrow$ connected cover $\mathcal{V}$

• The Mapper is the *nerve* of $\mathcal{V}$:

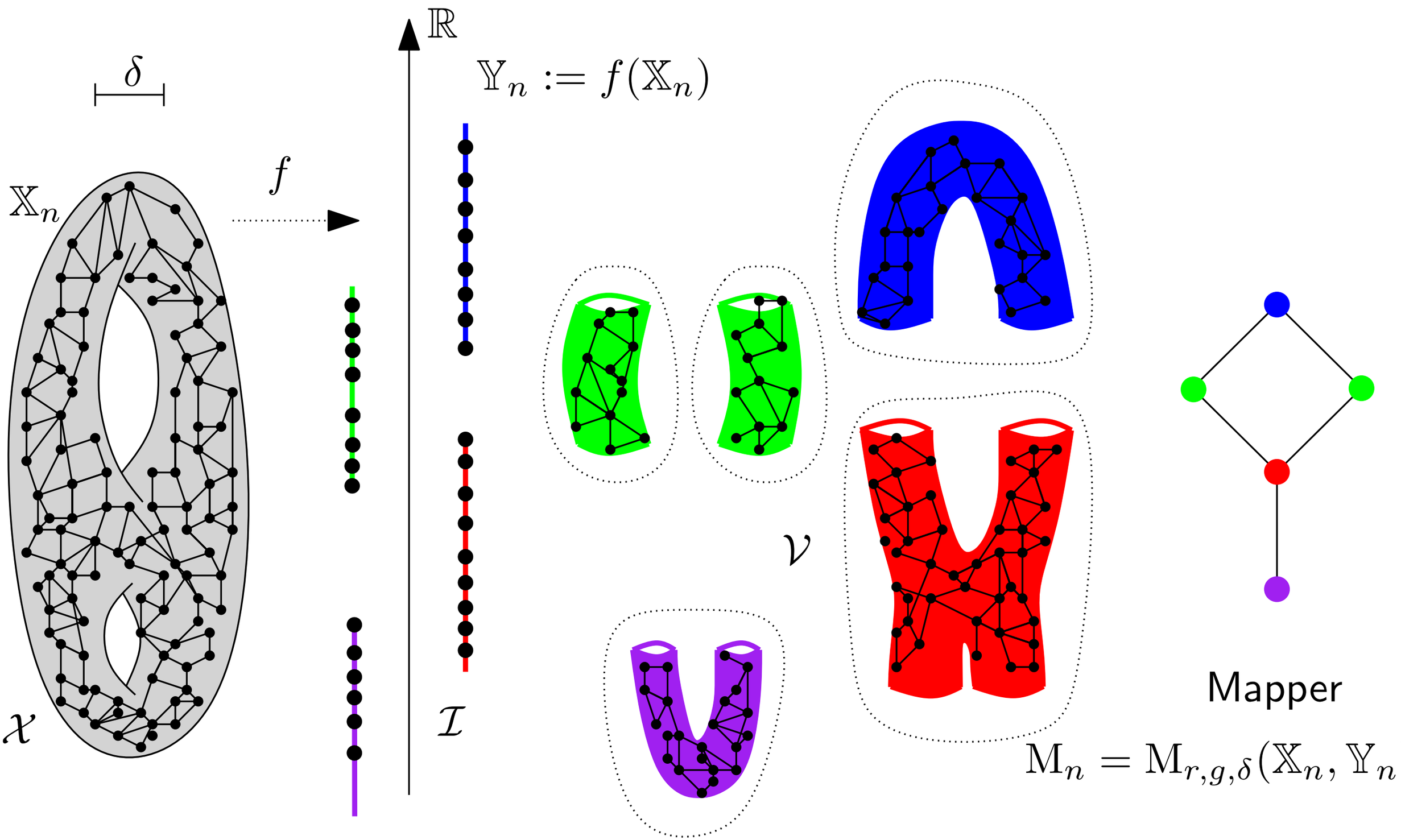
(intersections materialized by data points)

- 1 vertex per element $V \in \mathcal{V}$

- 1 edge per intersection $V \cap V' \neq \emptyset$, $V, V' \in \mathcal{V}$

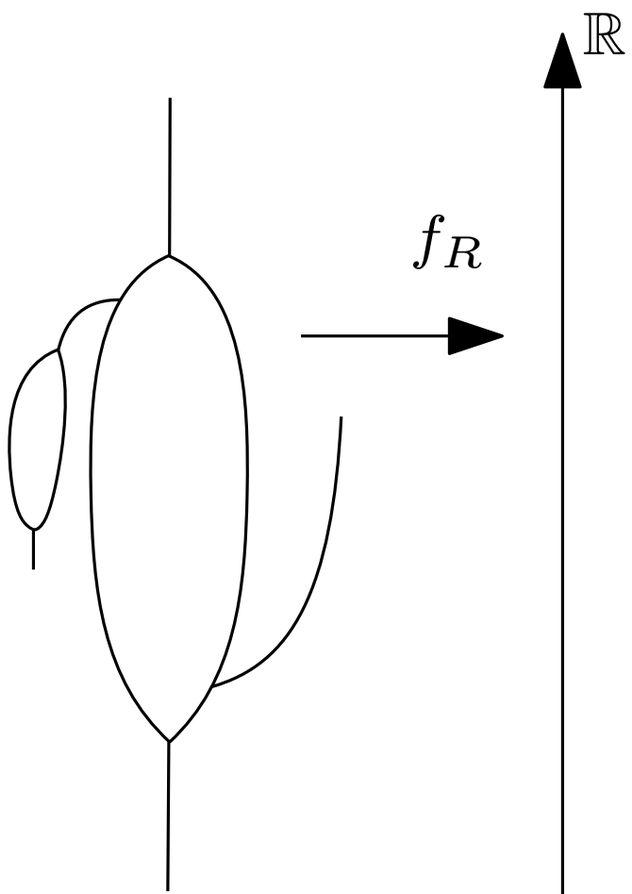
- 1 k-simplex per $(k+1)$-fold intersection $\bigcap_{i=0}^{k} V_i \neq \emptyset$, $V_0, \cdots, V_k \in \mathcal{V}$

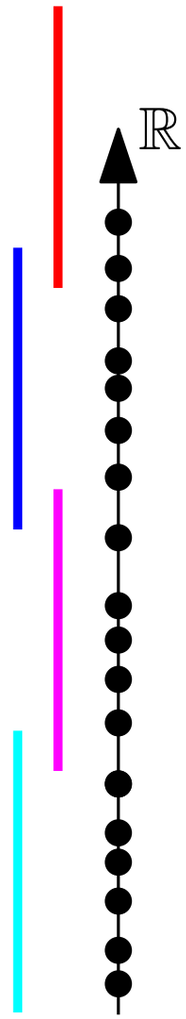
Mapper in practice : our version of Mapper

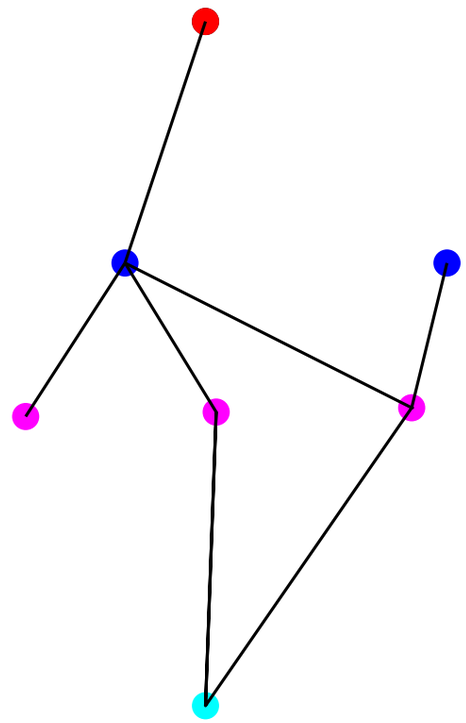


$G_\delta(\mathbb{X}_n) = \delta$-neighborhood graph of $\mathbb{X}_n$

Compare Reeb Graph with the Mapper



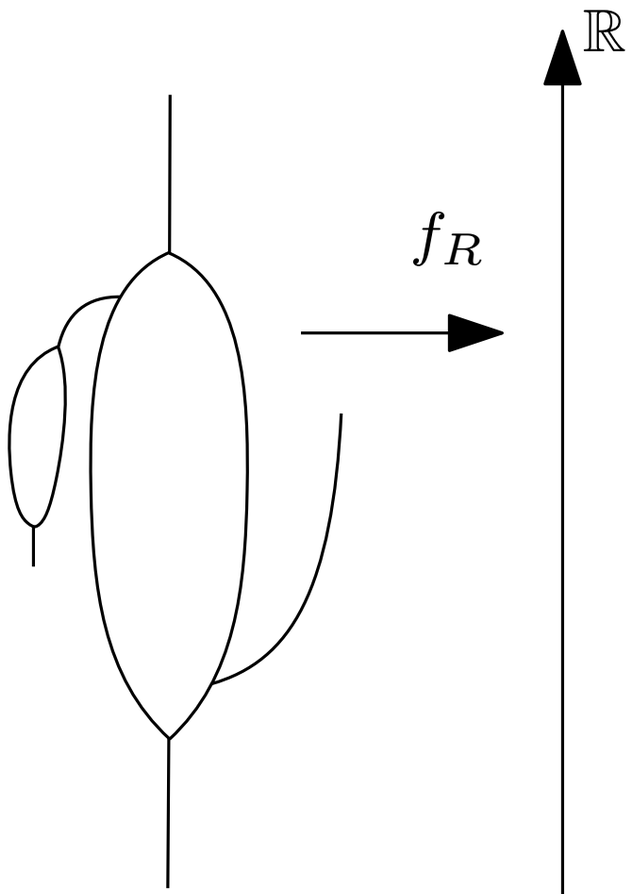


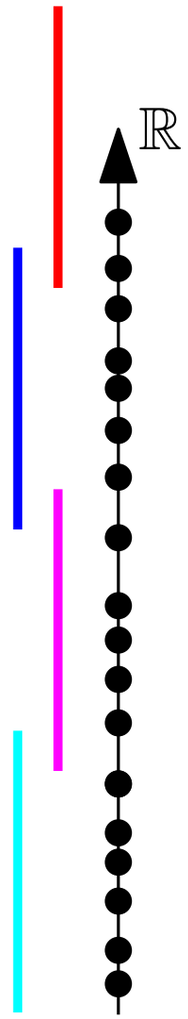


Reeb graph
$\mathrm{R}_f(\mathcal{X})$

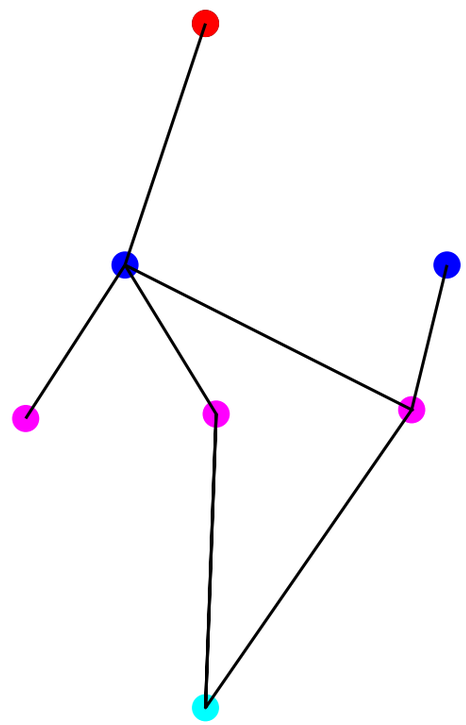
$\mathrm{M}_n = \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$

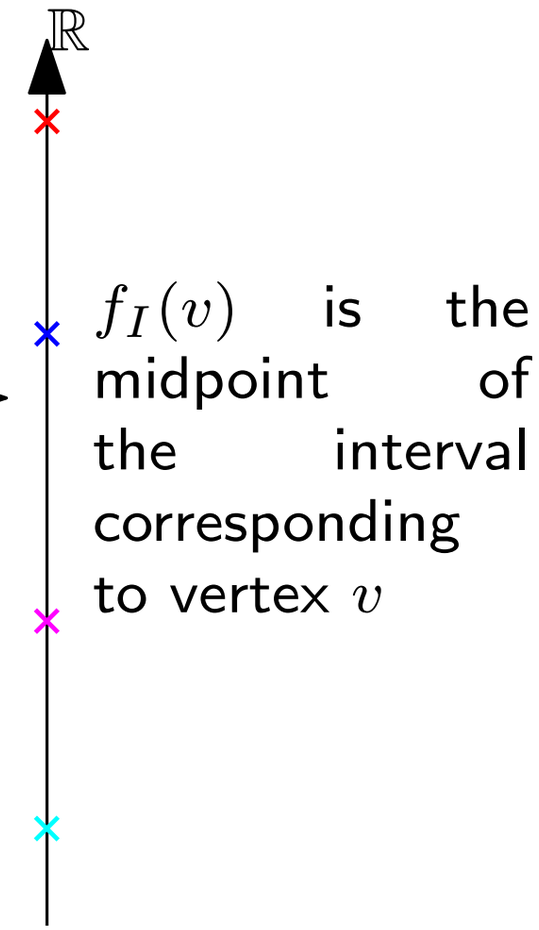
Compare Reeb Graph with the Mapper





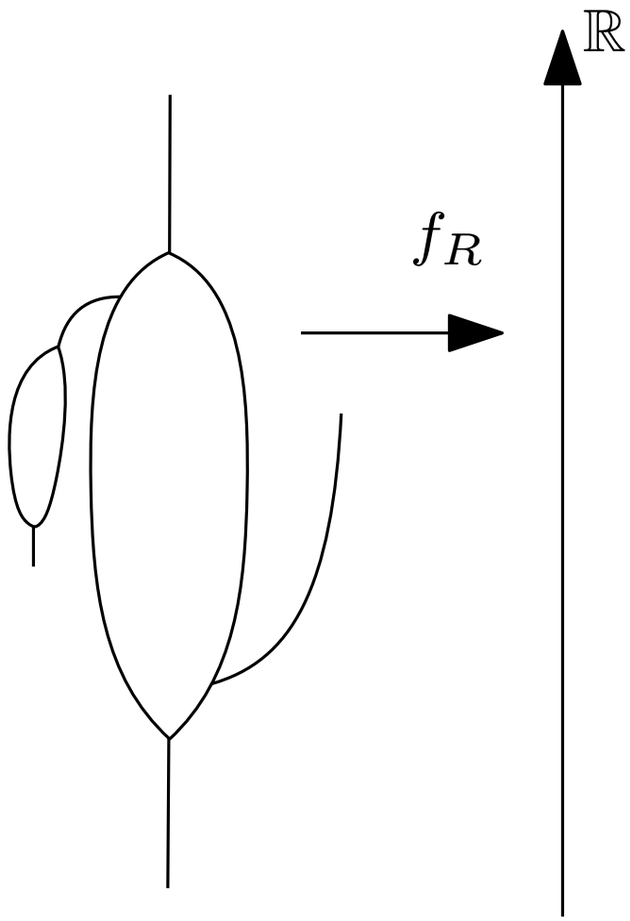
Reeb graph
$\mathrm{R}_f(\mathcal{X})$



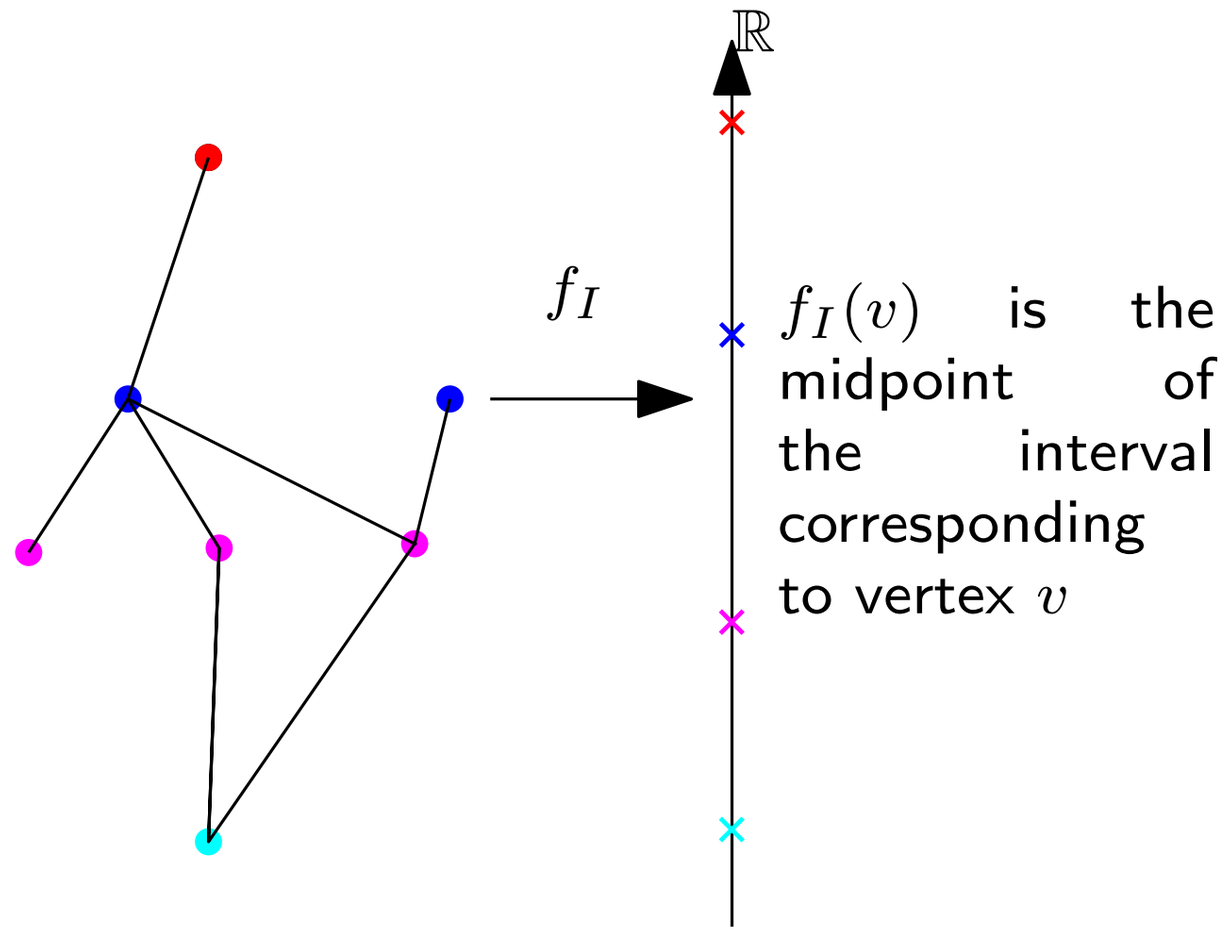


$$\mathrm{M}_n = \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$$

Compare Reeb Graph with the Mapper



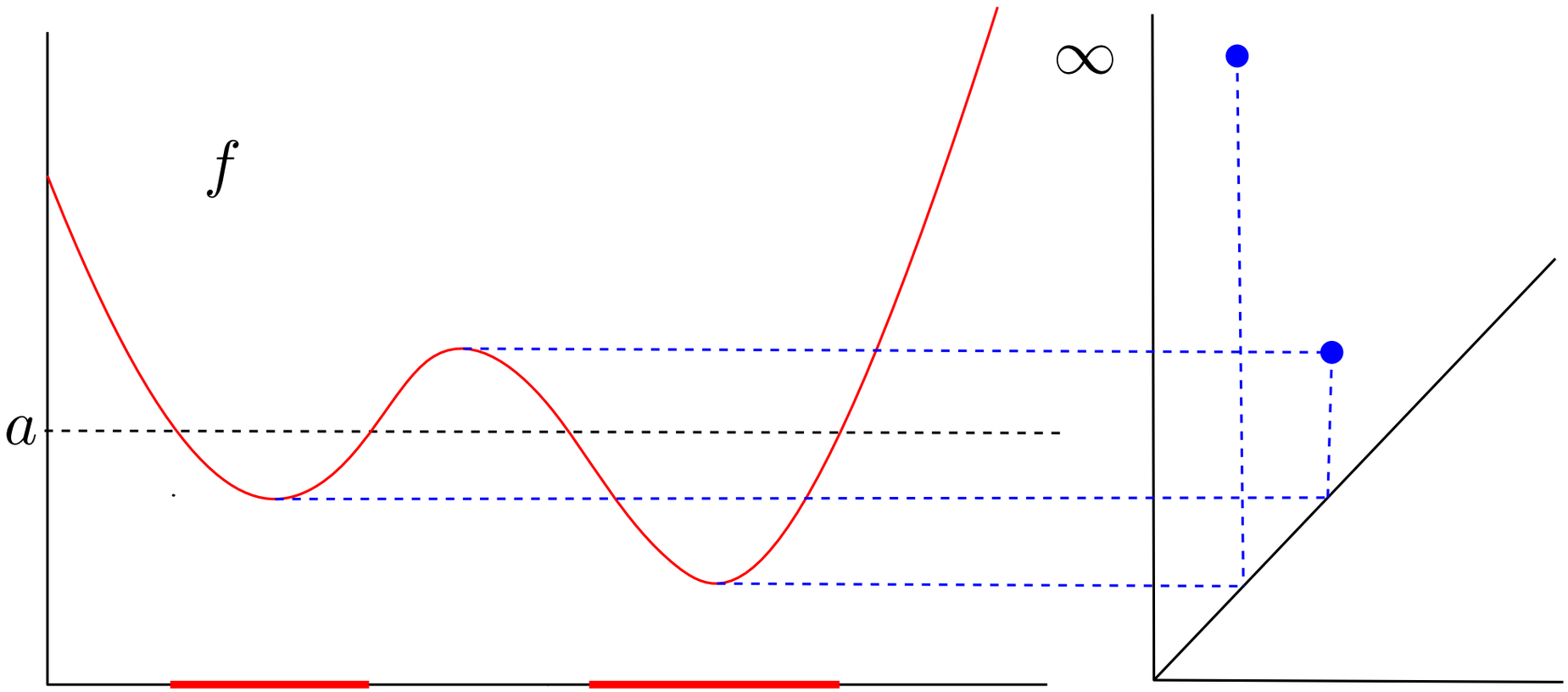
Reeb graph
$\mathrm{R}_f(\mathcal{X})$



$\mathrm{M}_n = \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$

Descriptors to compare Reeb graph and the Mapper: (extended) persistence

Sub level sets of a function



Persistence diagram

- The family $\{X^{(-\infty,\alpha]}\}_{\alpha\in\mathbb{R}}$ of sublevel sets of f defines a *filtration* w.r.t. inclusion:

$$X^{(-\infty,\alpha]} \subseteq X^{(-\infty,\beta]} \text{ for all } \alpha \leq \beta \in \mathbb{R}$$

- We can consider persistence homology for this filtration

Extended filtration of excursion sets

- The family $\{X^{[\alpha,+\infty)}\}_{\alpha\in\mathbb{R}}$ of **superlevel** sets of f is also nested but in the opposite direction: $X^{[\alpha,+\infty)} \supseteq X^{[\beta,+\infty)}$ for all $\alpha \leq \beta \in \mathbb{R}$.

- We can turn it into a filtration by reversing the real line :

  - let $\mathbb{R}^{\mathsf{op}} = \{\tilde{x} \mid x \in \mathbb{R}\}$, ordered by $\tilde{x} \leq \tilde{y} \Leftrightarrow x \geq y$.
  - Index the family of superlevel sets by $\mathbb{R}^{\mathsf{op}}$ to get a filtration: $\{X^{[\tilde{\alpha},+\infty)}\}_{\tilde{\alpha}\in\mathbb{R}^{\mathsf{op}}}$, with $X^{[\tilde{\alpha},+\infty)} \subseteq X^{[\tilde{\beta},+\infty)}$ for all $\tilde{\alpha} \leq \tilde{\beta} \in \mathbb{R}^{\mathsf{op}}$.

- Replace each superlevel set $X^{[\tilde{\alpha},+\infty)}$ by the pair of spaces $(X, X^{[\tilde{\alpha},+\infty)})$.

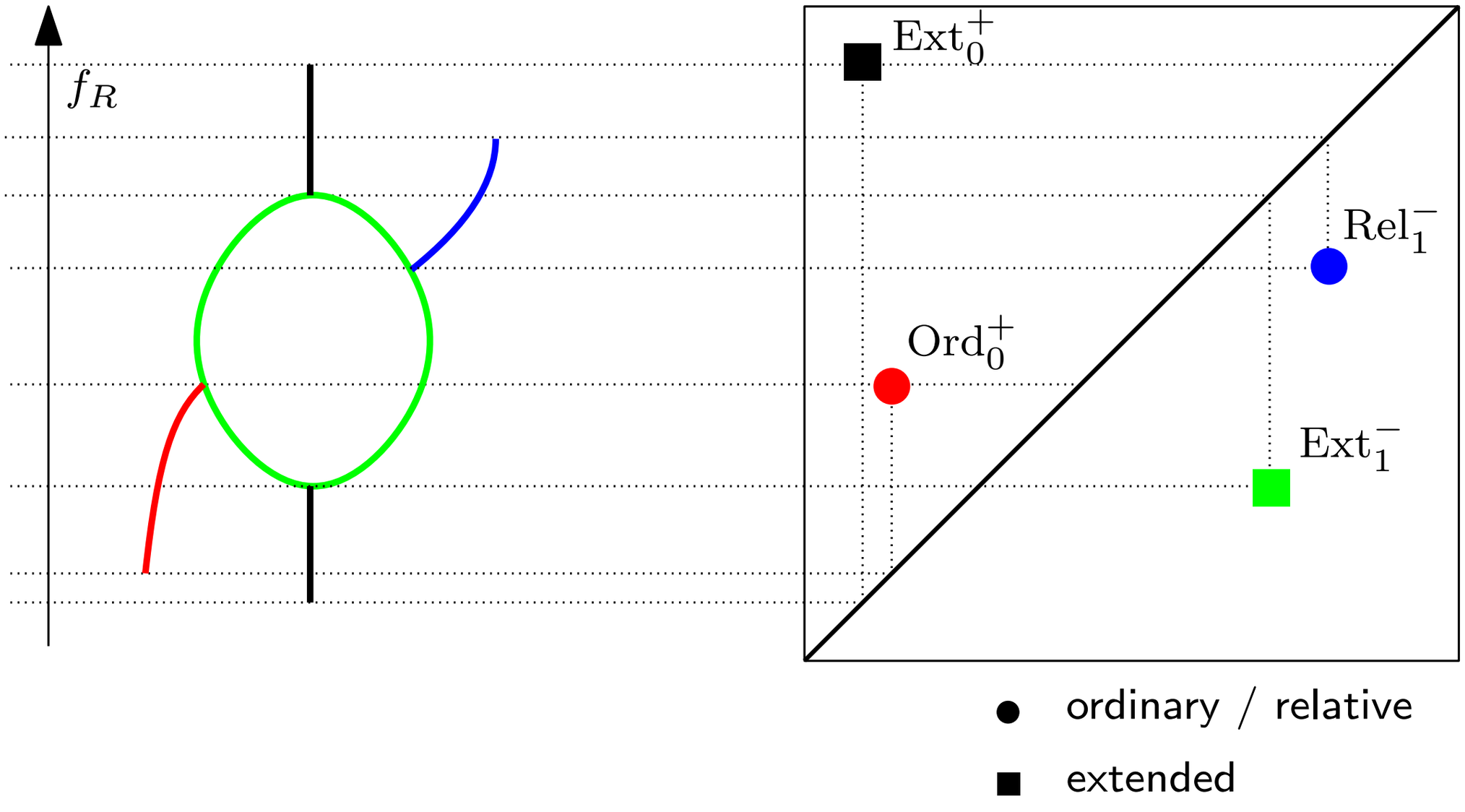
- Let $\mathbb{R}_{\mathsf{Ext}} = \mathbb{R} \cup \{+\infty\} \cup \mathbb{R}^{\mathsf{op}}$, where the order is completed by $\alpha < +\infty < \tilde{\beta}$ for all $\alpha \in \mathbb{R}$ and $\tilde{\beta} \in \mathbb{R}^{\mathsf{op}}$.

- Define the **extended filtration** of f over $\mathbb{R}_{\mathsf{Ext}}$ by:

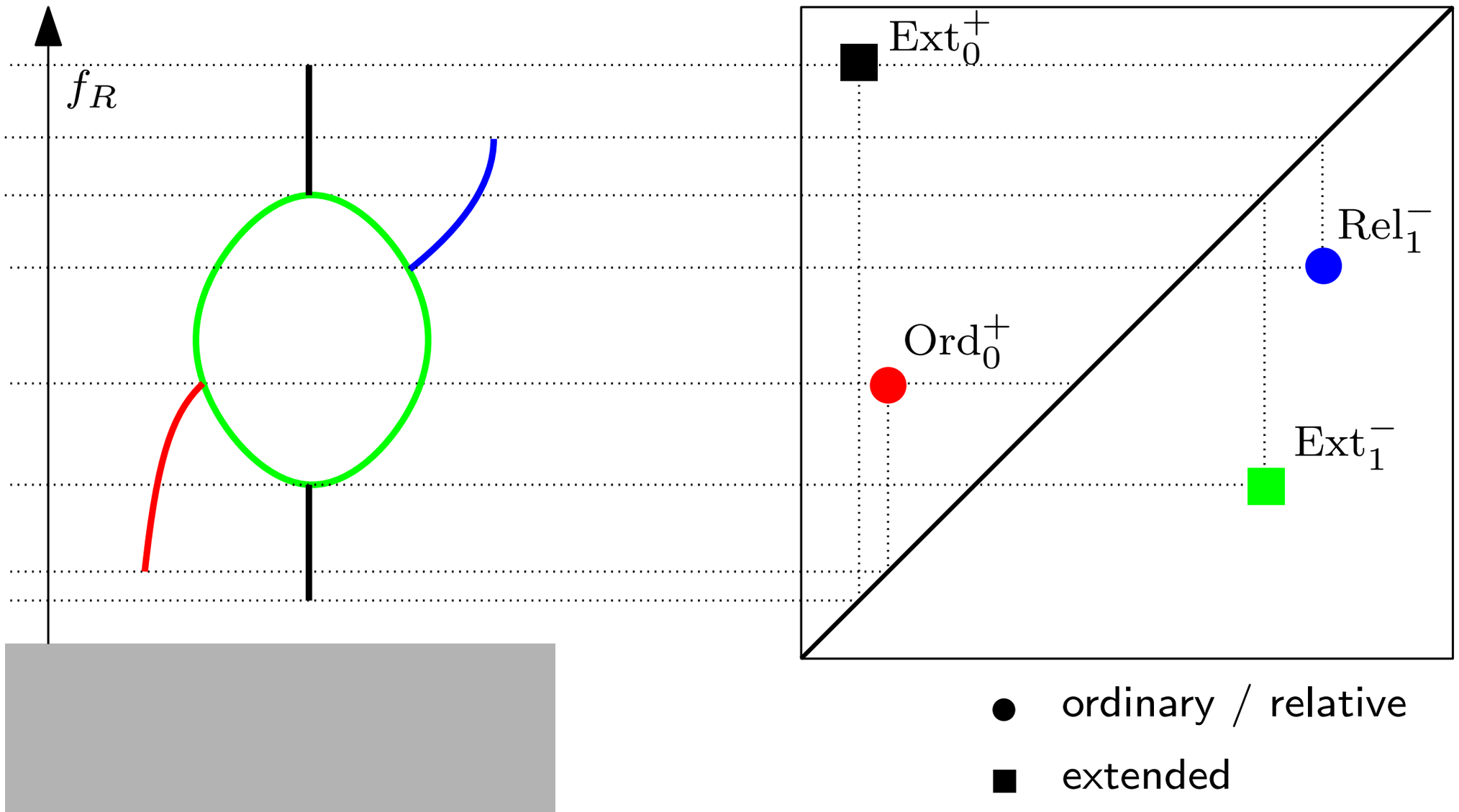
$$
\begin{array}{lcll}
F_\alpha & = & X^{(-\infty,\alpha]} & \text{for } \alpha \in \mathbb{R} \\
F_{+\infty} & = & X \equiv (X, \emptyset) & \\
F_{\tilde{\alpha}} & = & (X, X^{[\tilde{\alpha},+\infty)}) & \text{for } \tilde{\alpha} \in \mathbb{R}^{\mathsf{op}}
\end{array}
$$

Extended persistence as a descriptor for Reeb graph

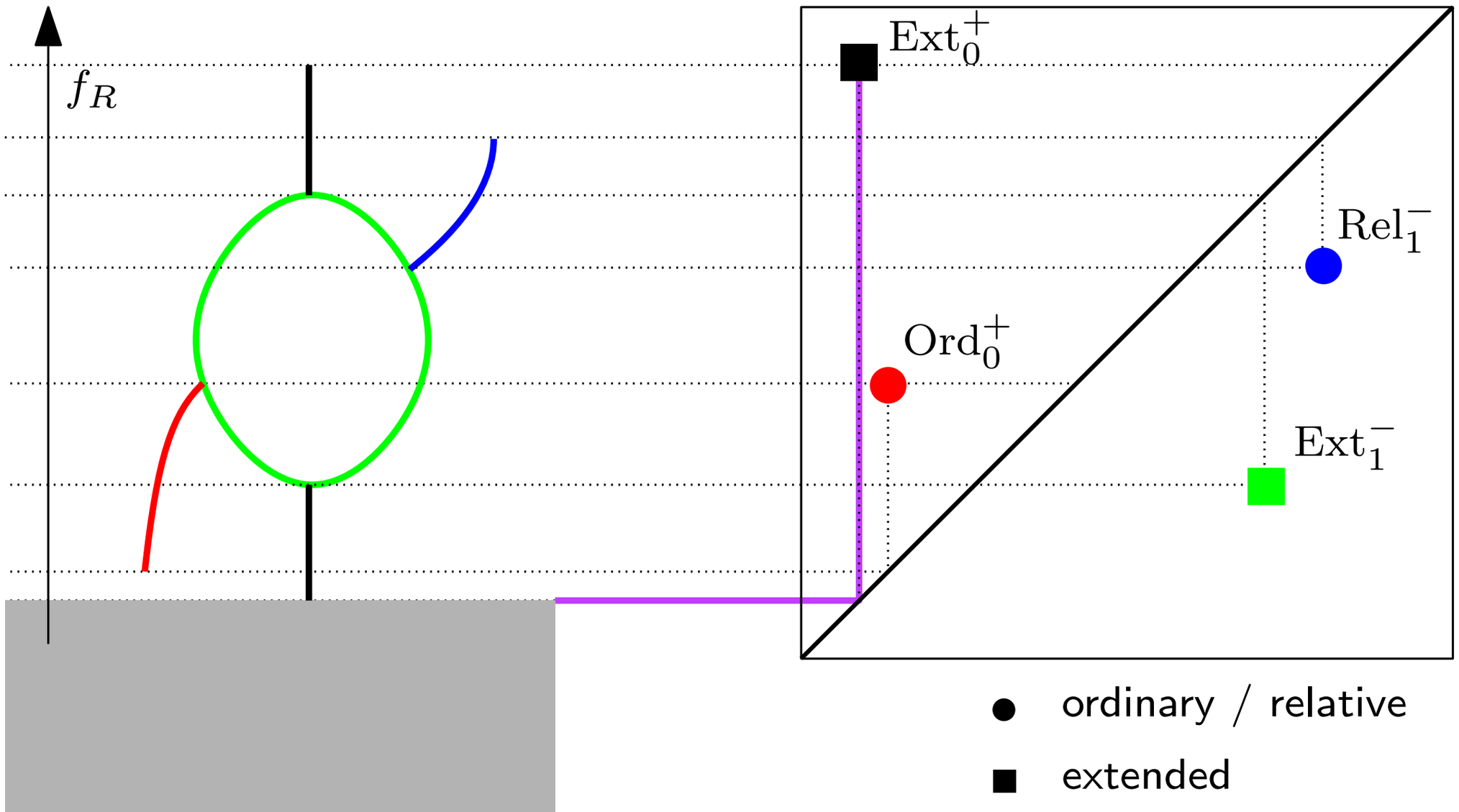
Given any graph $G = (V, E)$ and a function attached to its nodes $f : V \to \mathbb{R}$, the so-called *extended persistence diagram* is a multiset of points in the Euclidean plane $\mathbb{R}^2$ that can be computed with *extended persistence theory*. [Cohen-Steiner, Edelsbrunner, Harer 2008]



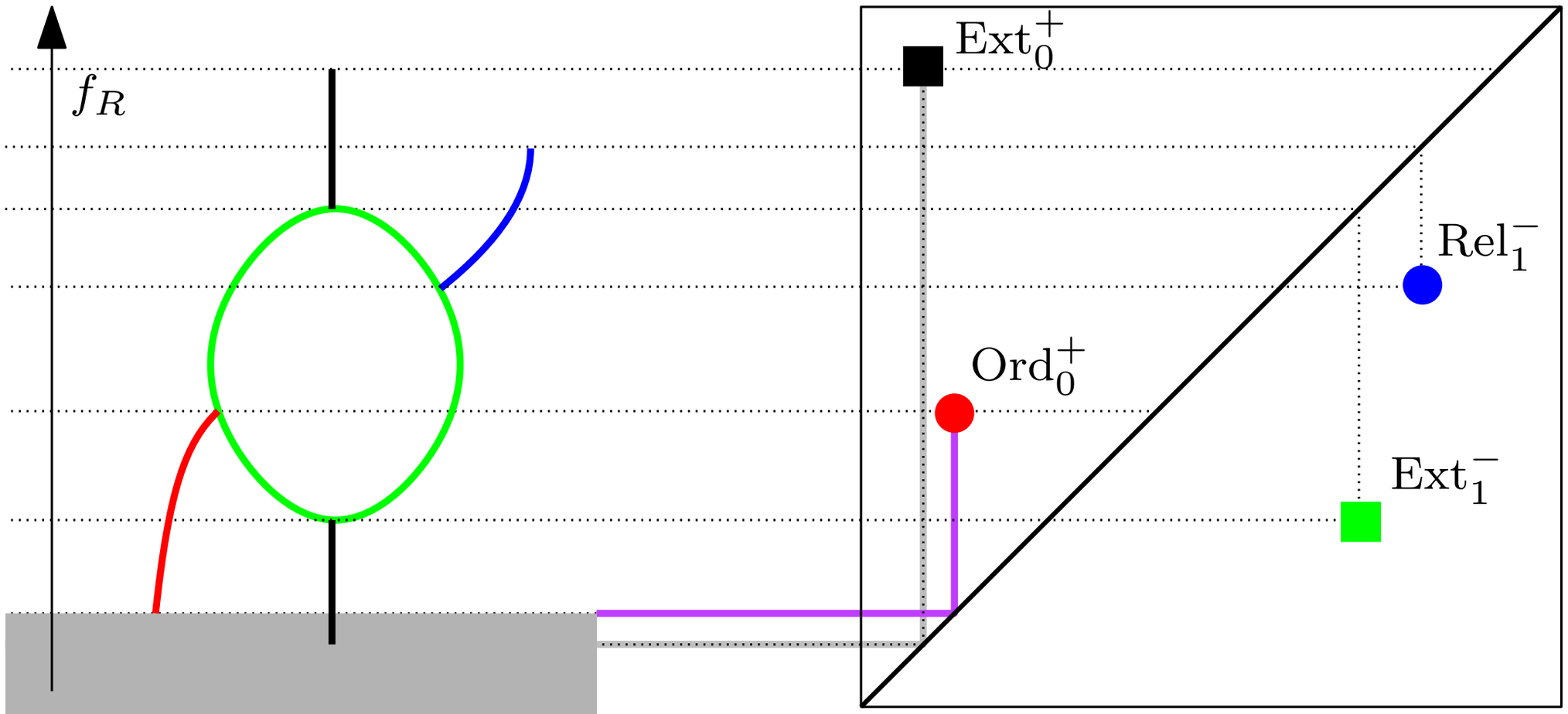
Extended persistence as a descriptor for Reeb graph



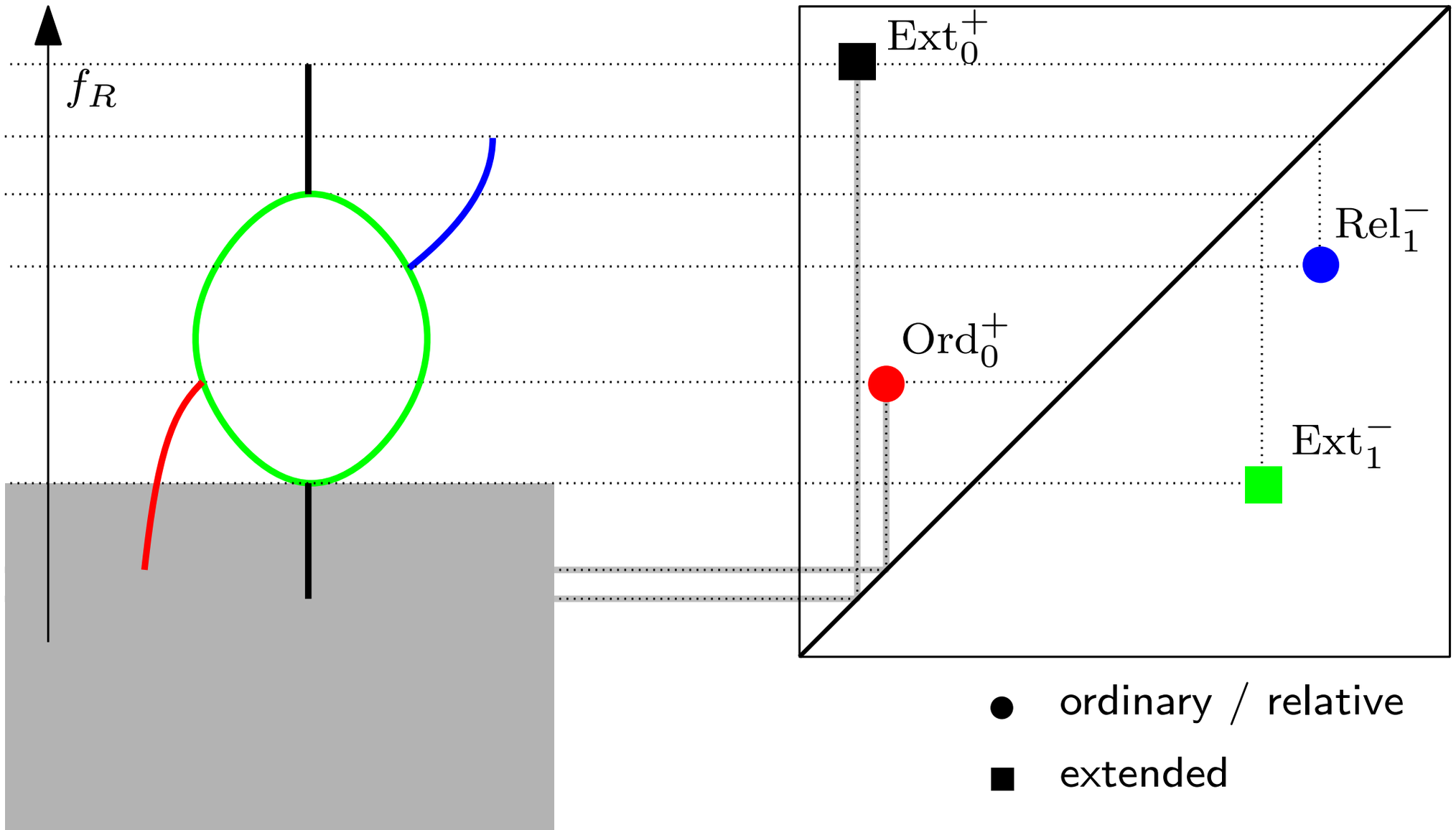
Extended persistence as a descriptor for Reeb graph



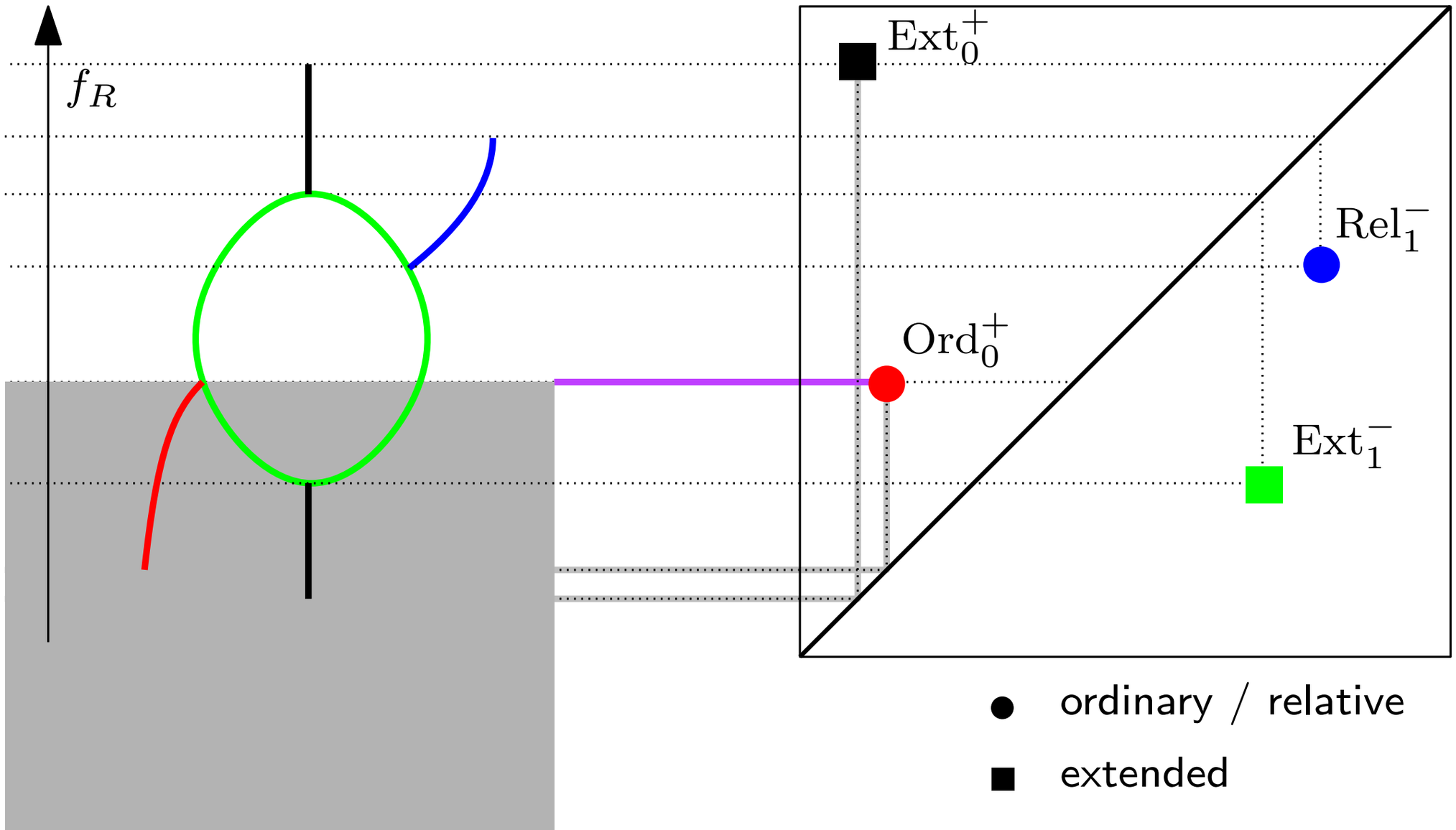
Extended persistence as a descriptor for Reeb graph



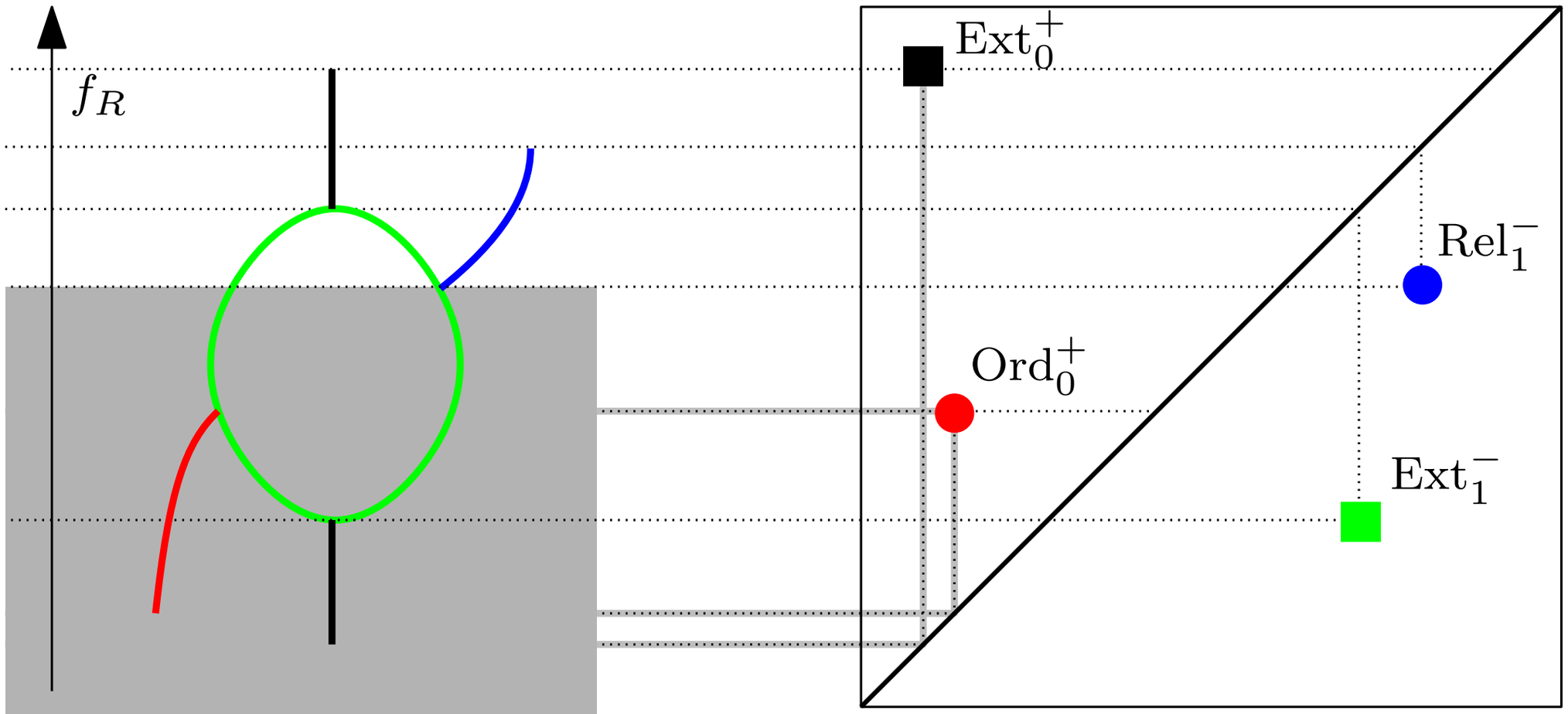
Extended persistence as a descriptor for Reeb graph



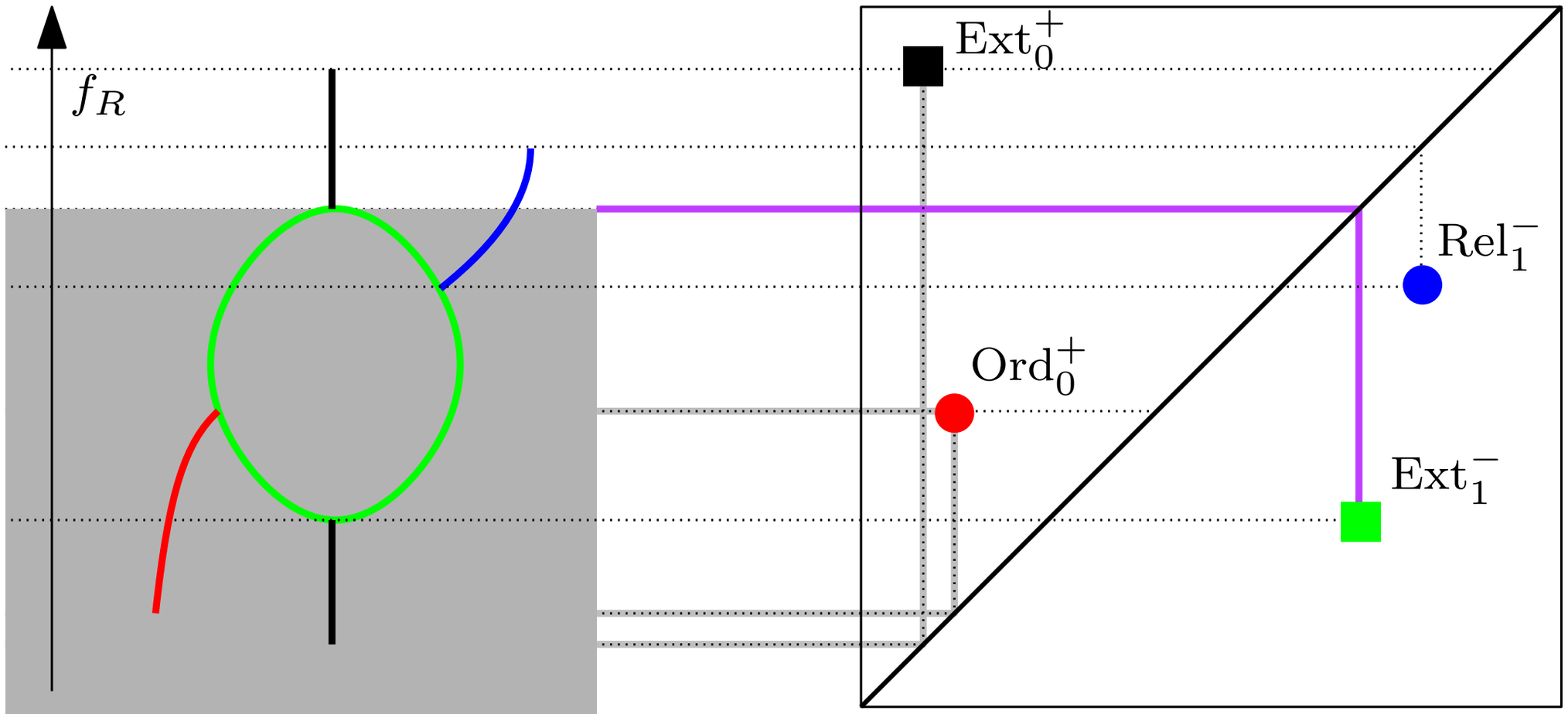
Extended persistence as a descriptor for Reeb graph



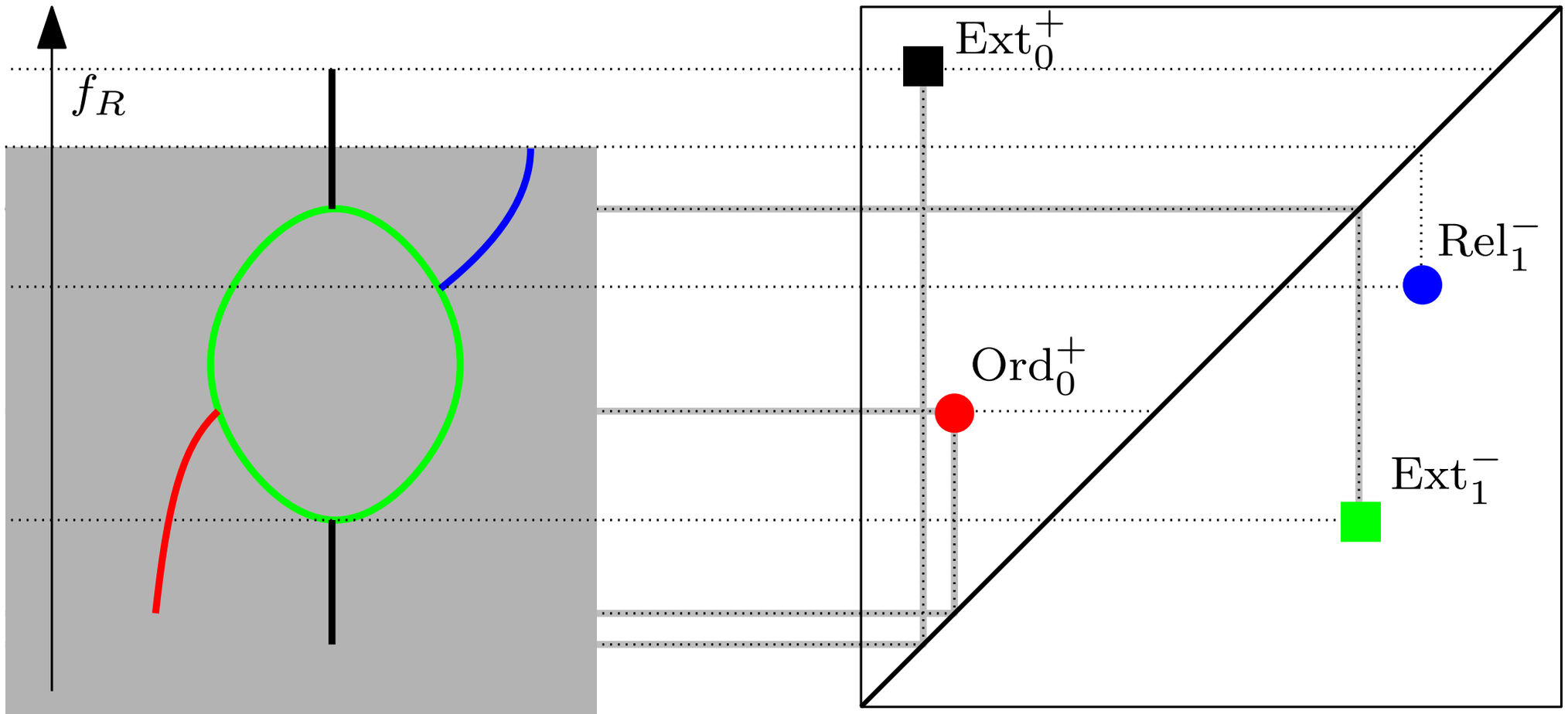
Extended persistence as a descriptor for Reeb graph



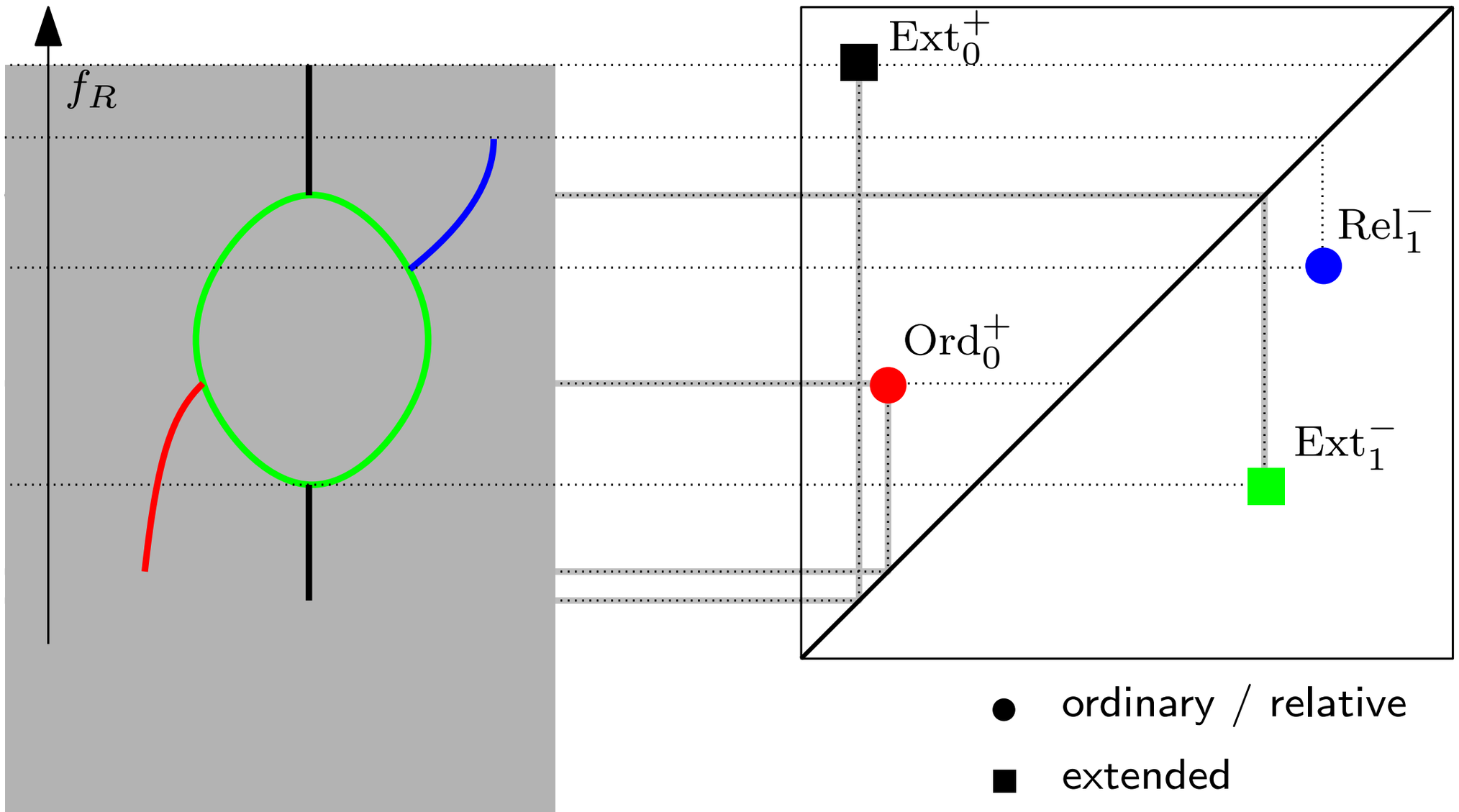
Extended persistence as a descriptor for Reeb graph



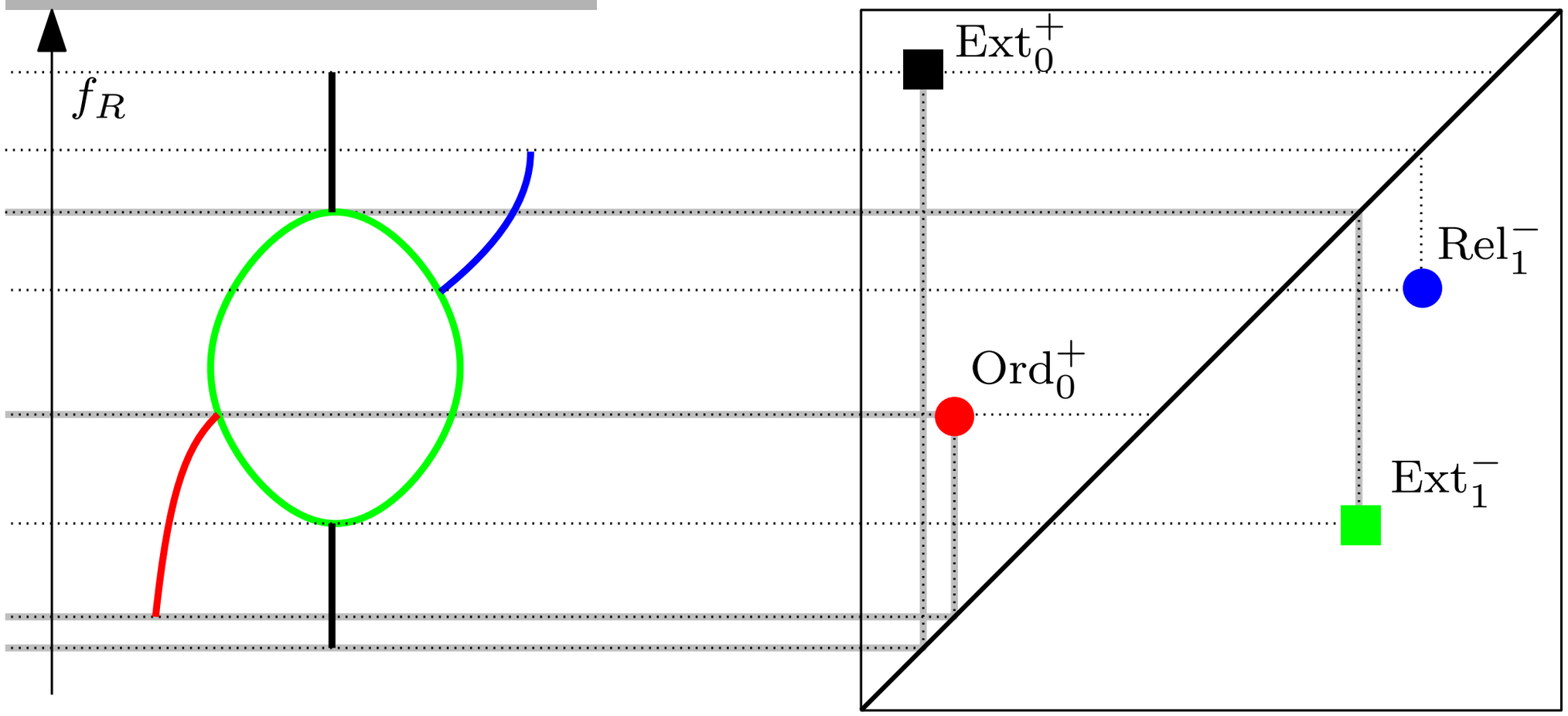
Extended persistence as a descriptor for Reeb graph



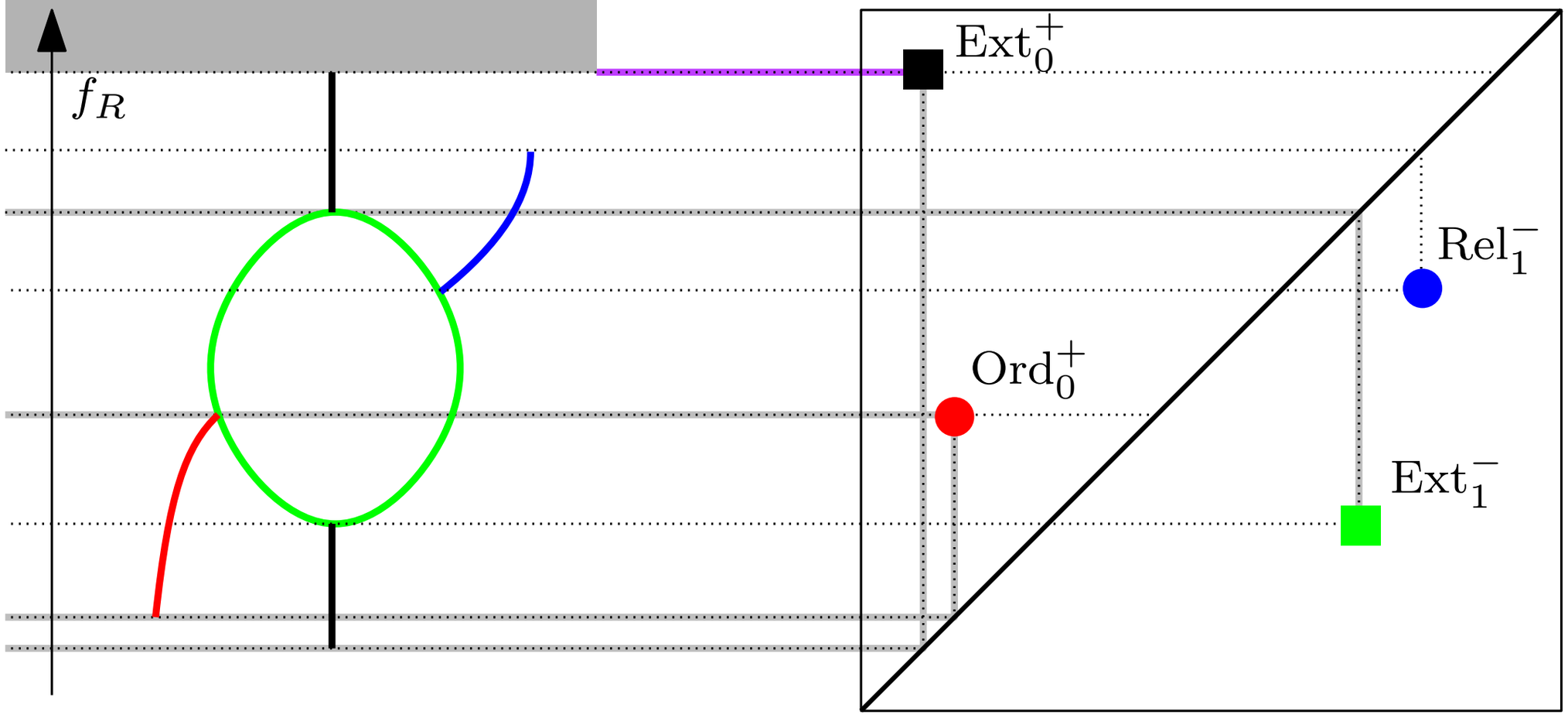
Extended persistence as a descriptor for Reeb graph



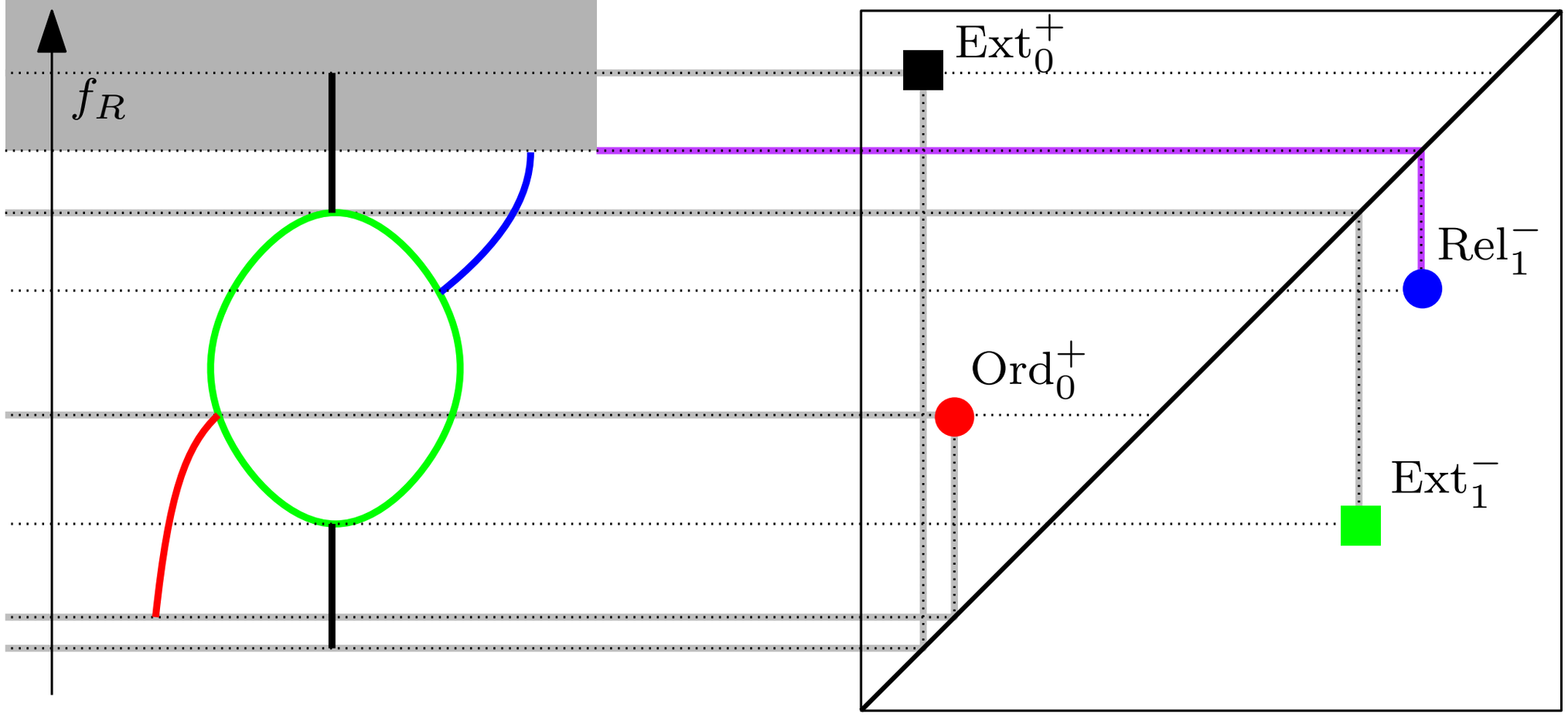
Extended persistence as a descriptor for Reeb graph



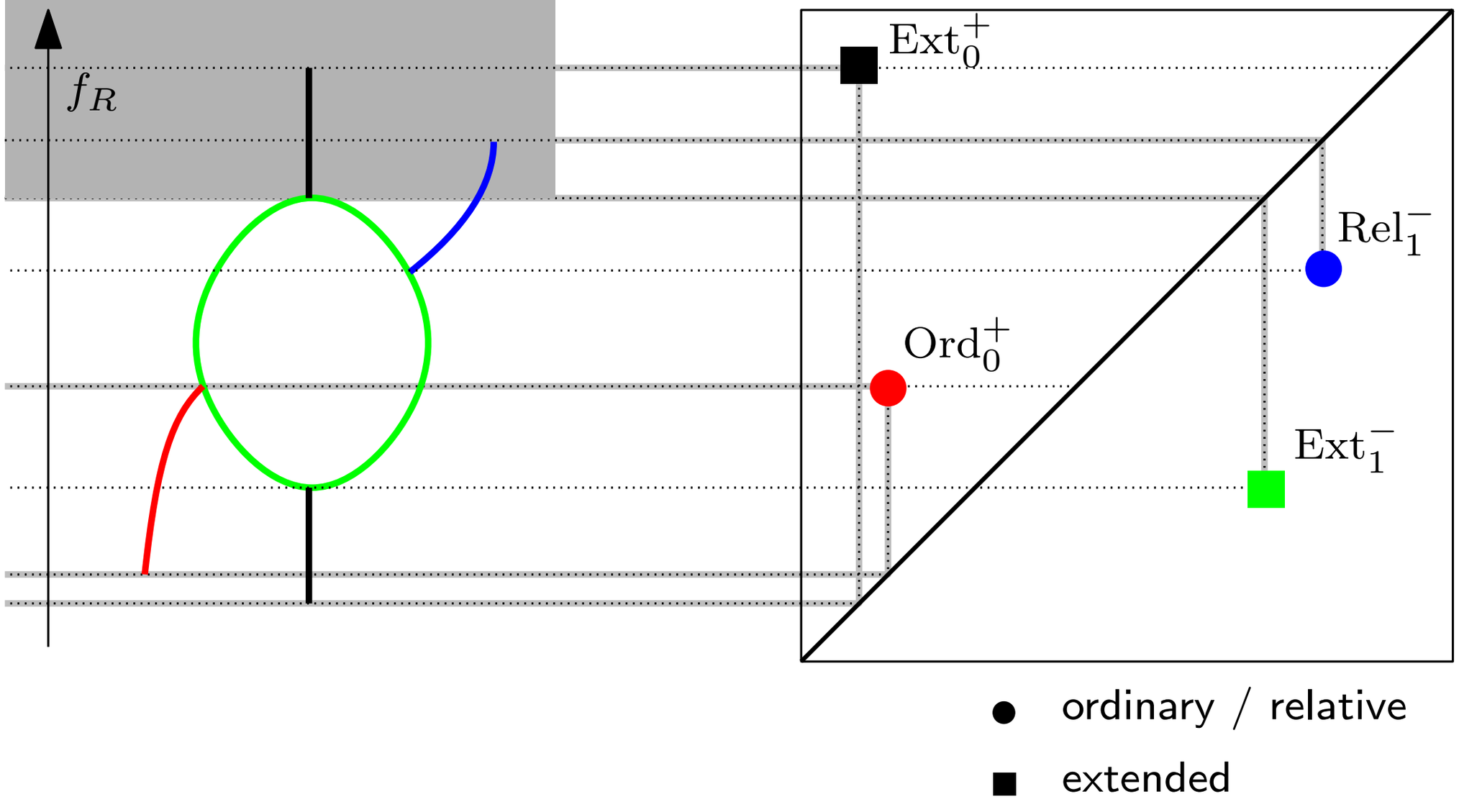
Extended persistence as a descriptor for Reeb graph



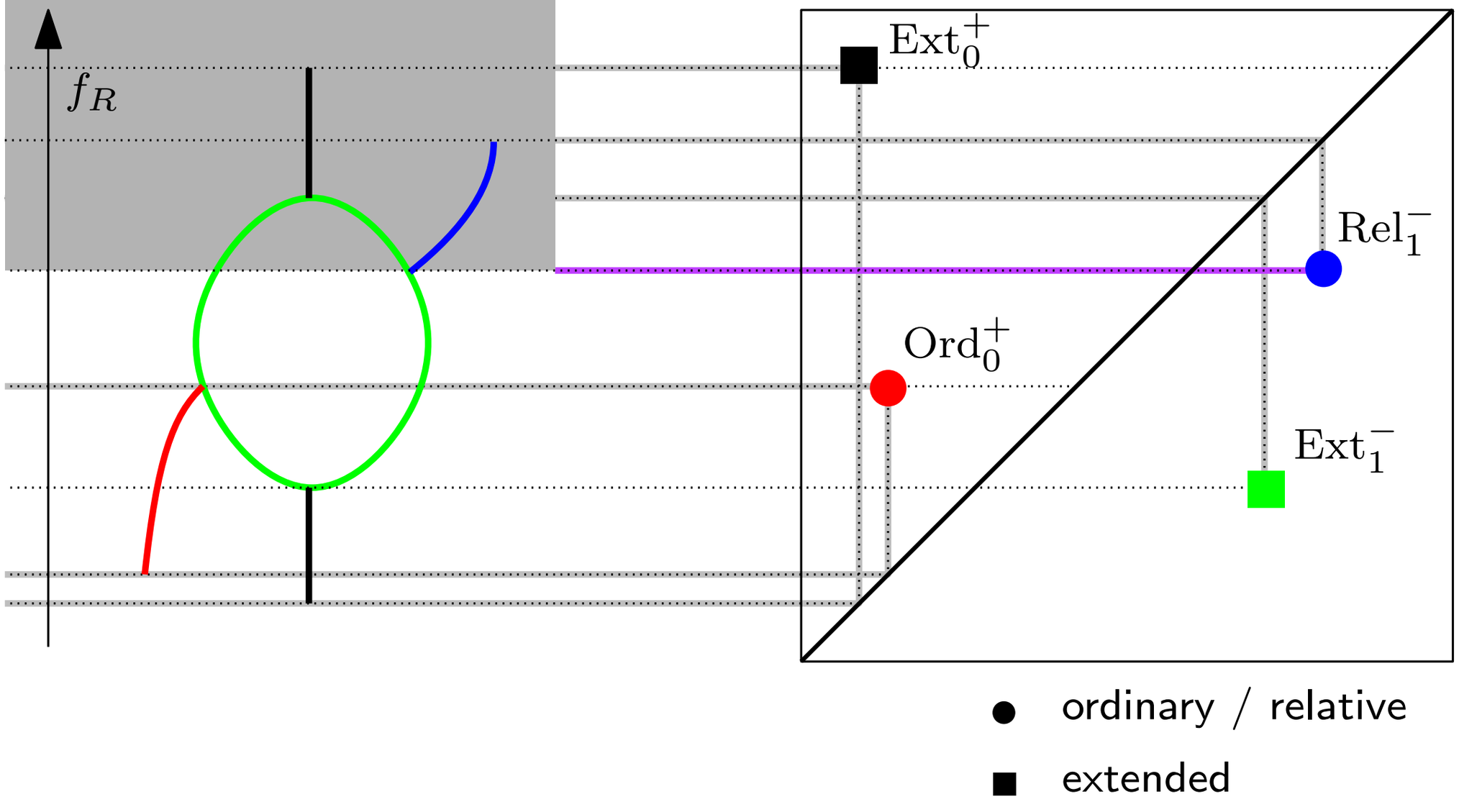
Extended persistence as a descriptor for Reeb graph



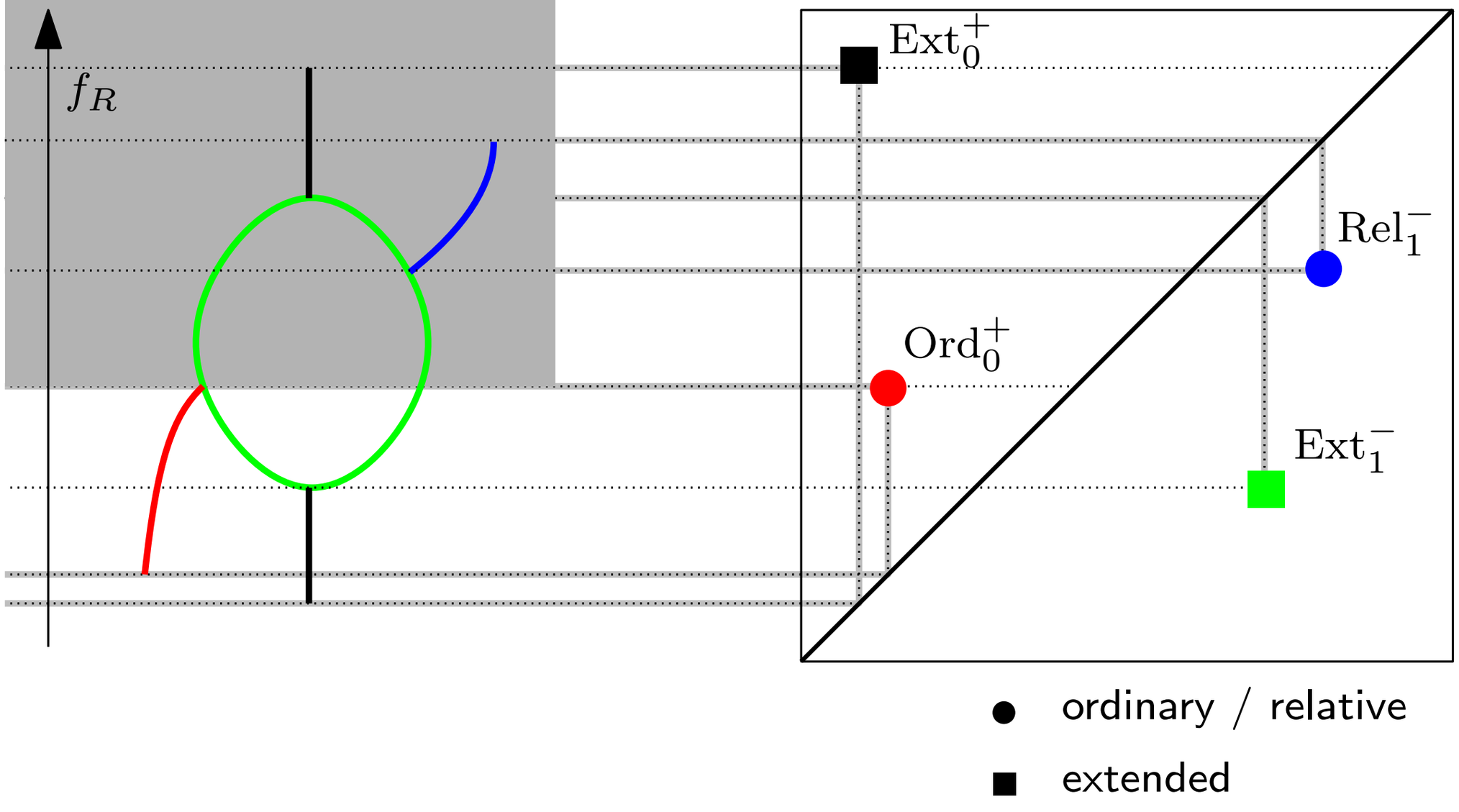
Extended persistence as a descriptor for Reeb graph



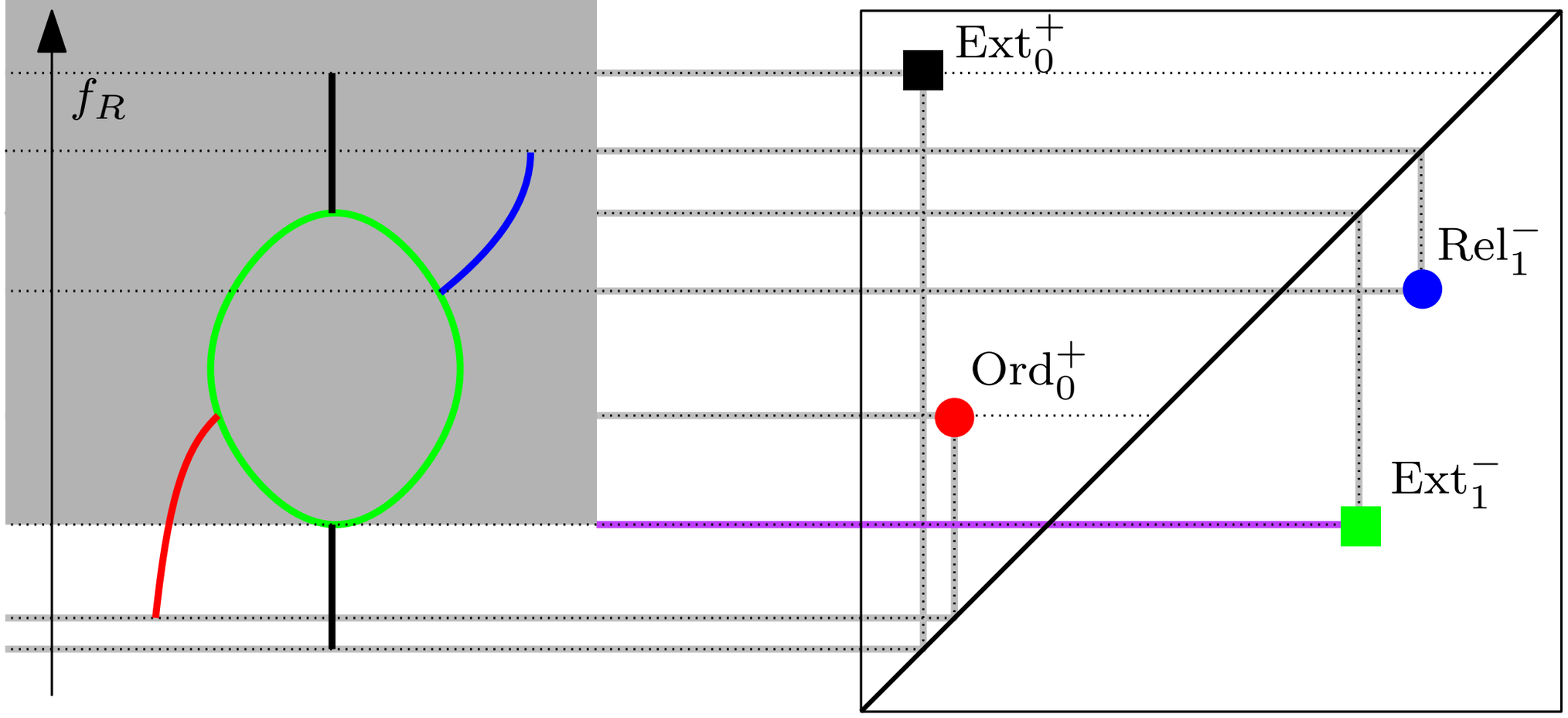
Extended persistence as a descriptor for Reeb graph



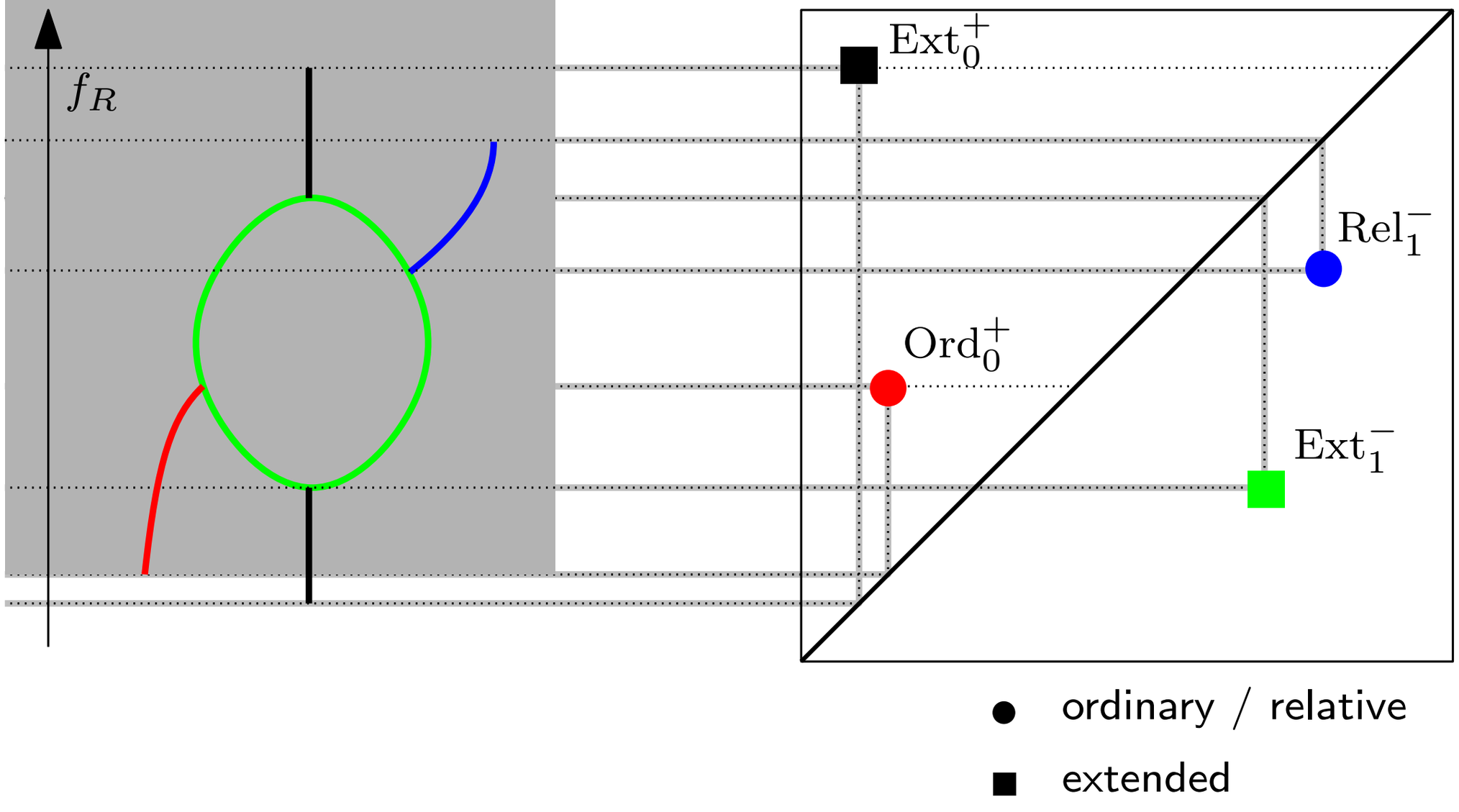
Extended persistence as a descriptor for Reeb graph



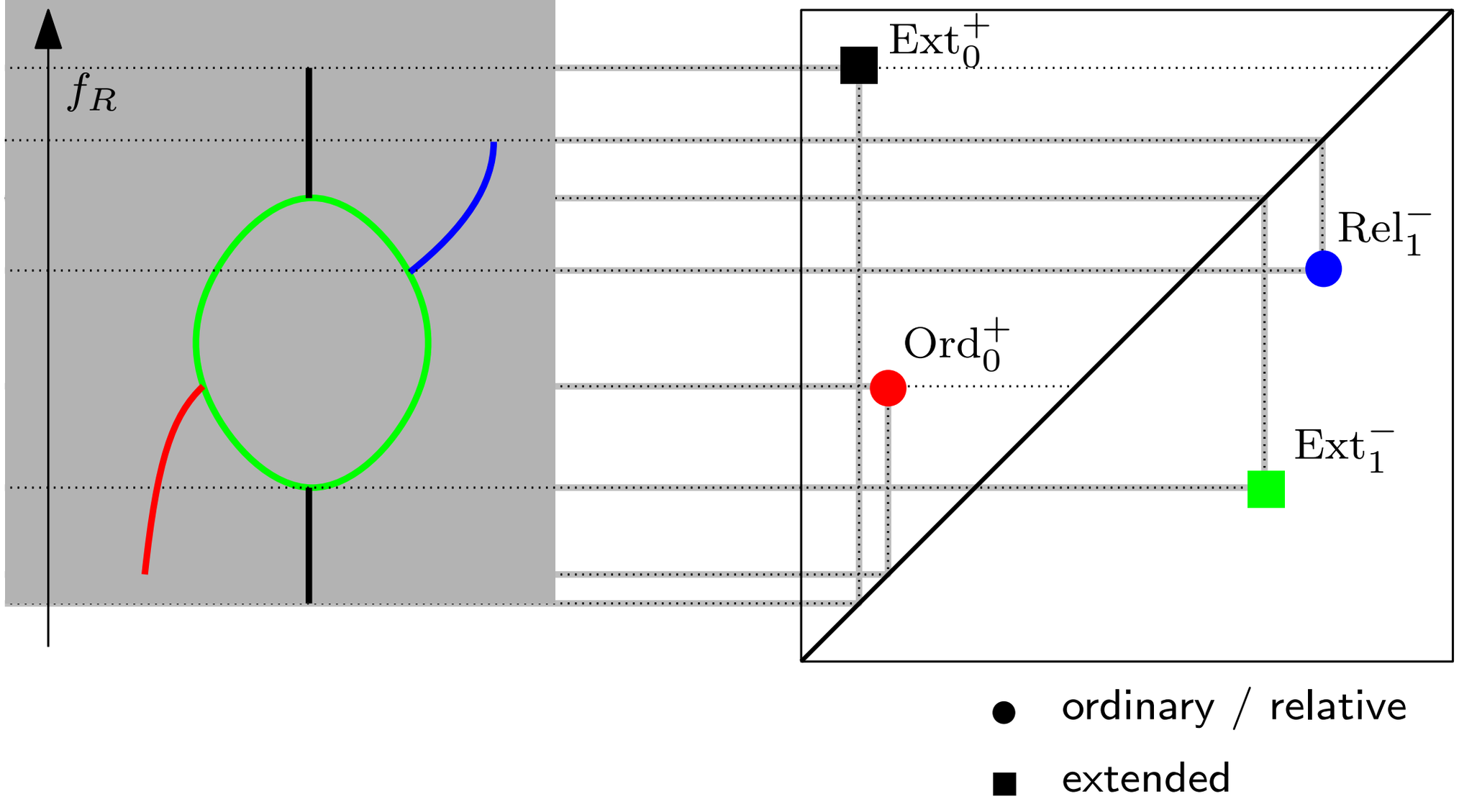
Extended persistence as a descriptor for Reeb graph



Extended persistence as a descriptor for Reeb graph

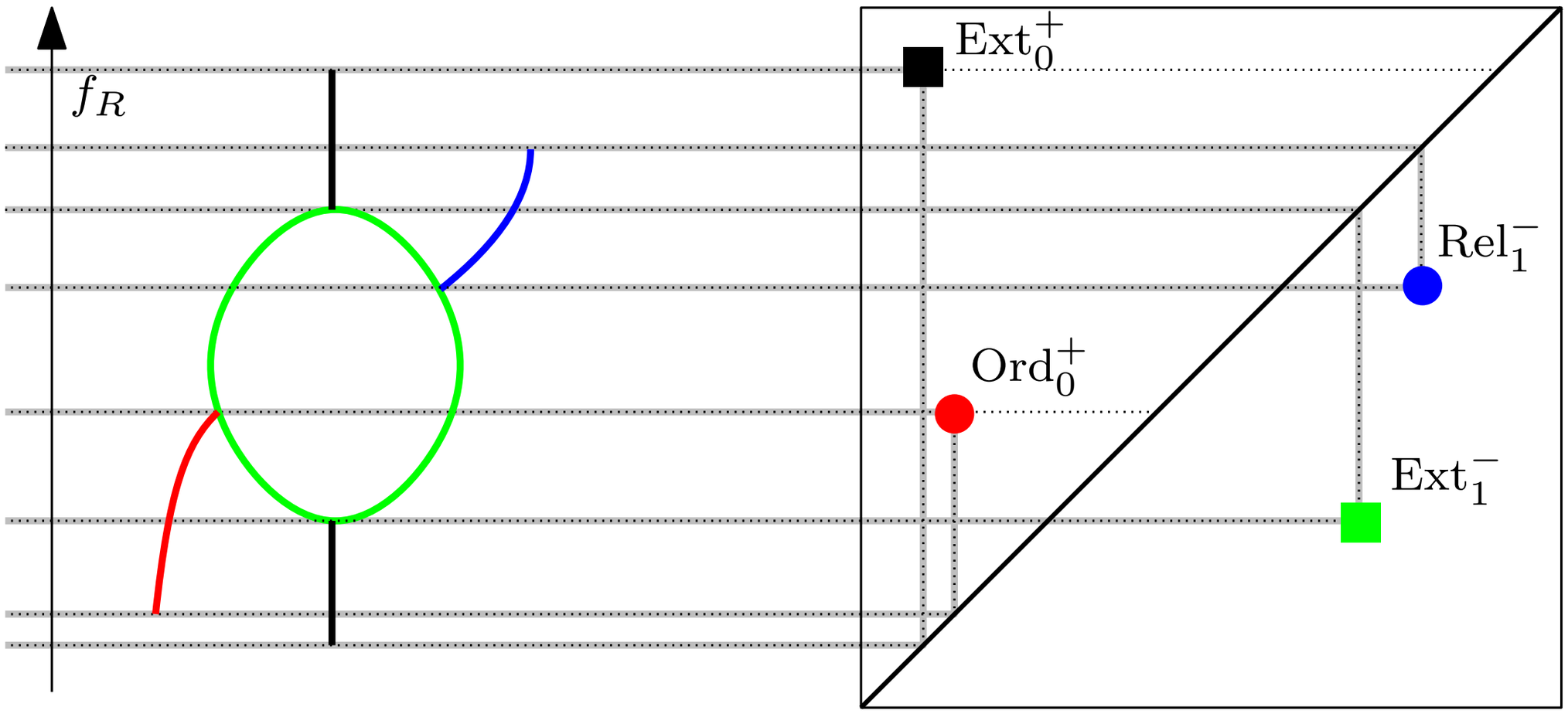


Extended persistence as a descriptor for Reeb graph



Extended persistence as a descriptor for Reeb graph

- We do the same for Mapper: DgmM_n is the extented persistence diagram of the filtration of excursion sets of f_I on M_n.
- We see DgmR_f and DgmM_n as (bag-of-features) descriptors for $\mathrm{R}_f(\mathcal{X})$ and M_n.



Rel: appears/dies in superlevels

Ext: appears in sublevels, dies in superlevels

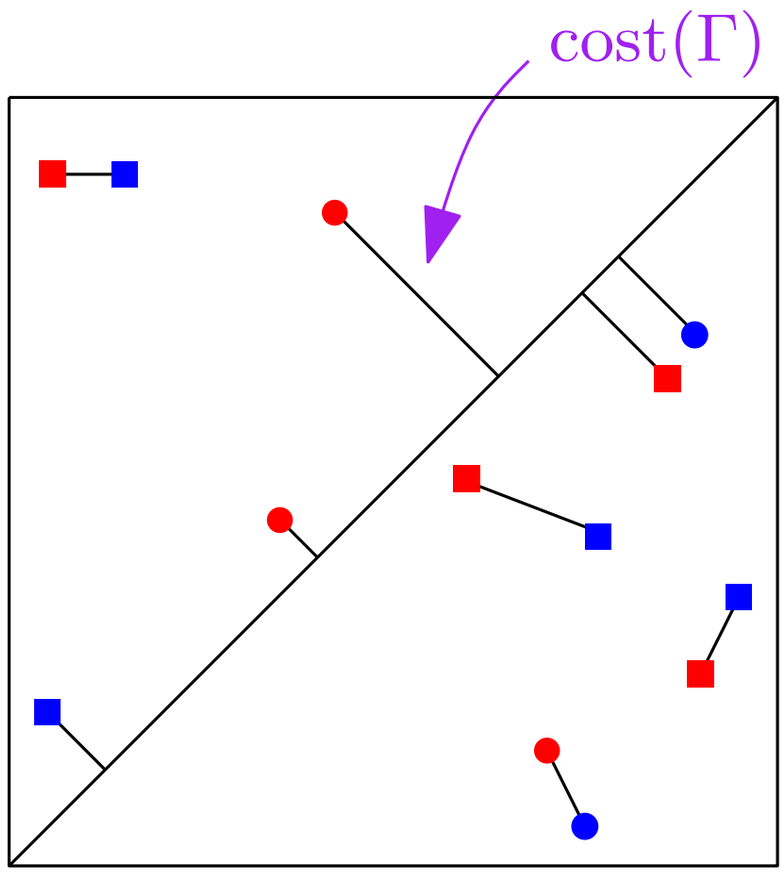
● ordinary / relative

■ extended

Extended persistence as a descriptor for Reeb graph

Partial matching Γ between two diagrams :

- match points of the same type (ordinary, relative, extended) and of the same homological dimension only.
- points can always be matched with the diagonal

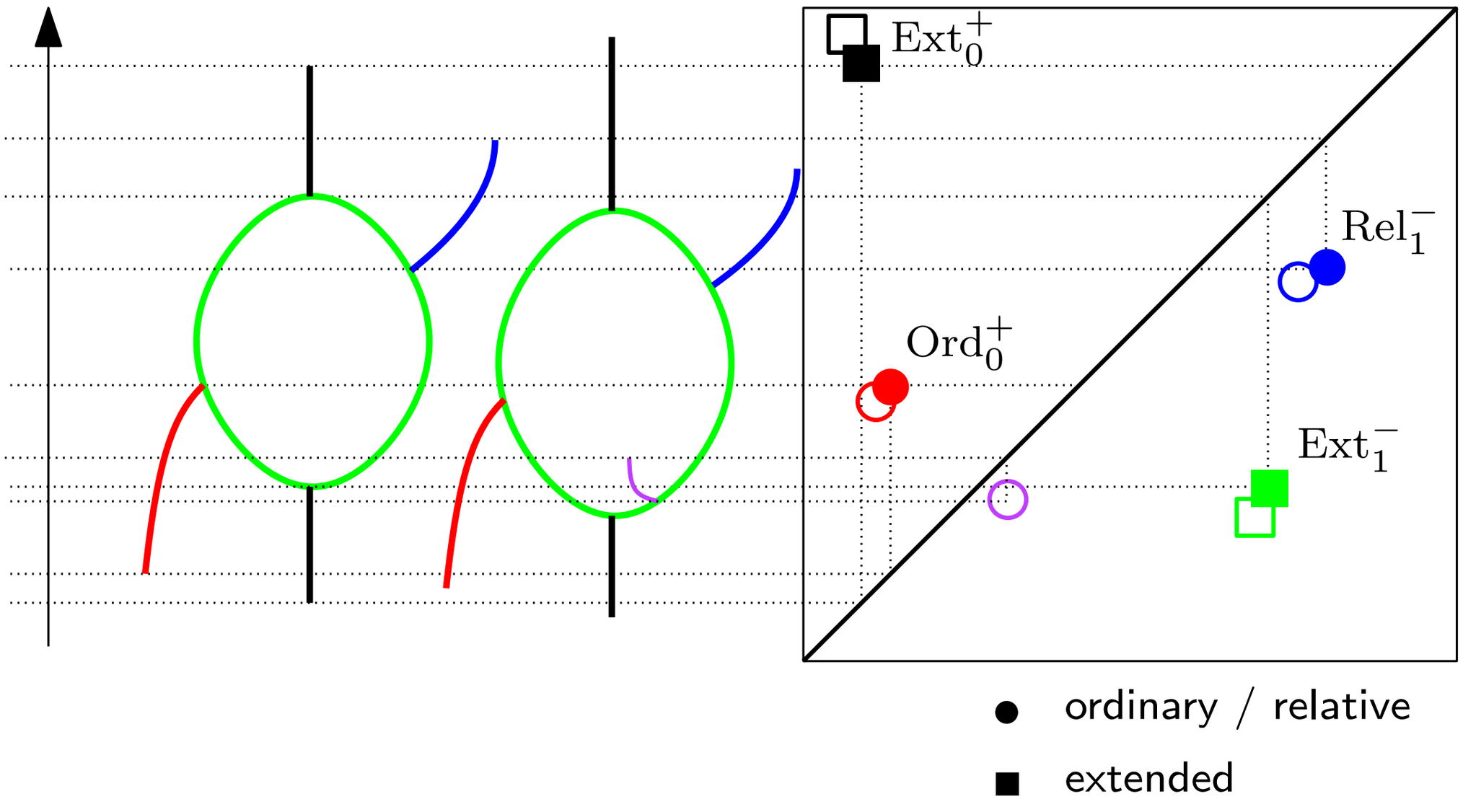


Extended persistence as a descriptor for Reeb graph

The **bottleneck distance** between Dg and Dg′ is:

$$d_B(\mathrm{Dg}, \mathrm{Dg}') = \inf_{\Gamma} \ \mathit{cost}(\Gamma),$$

where Γ ranges over all *p*artial matchings between Dg and Dg′.

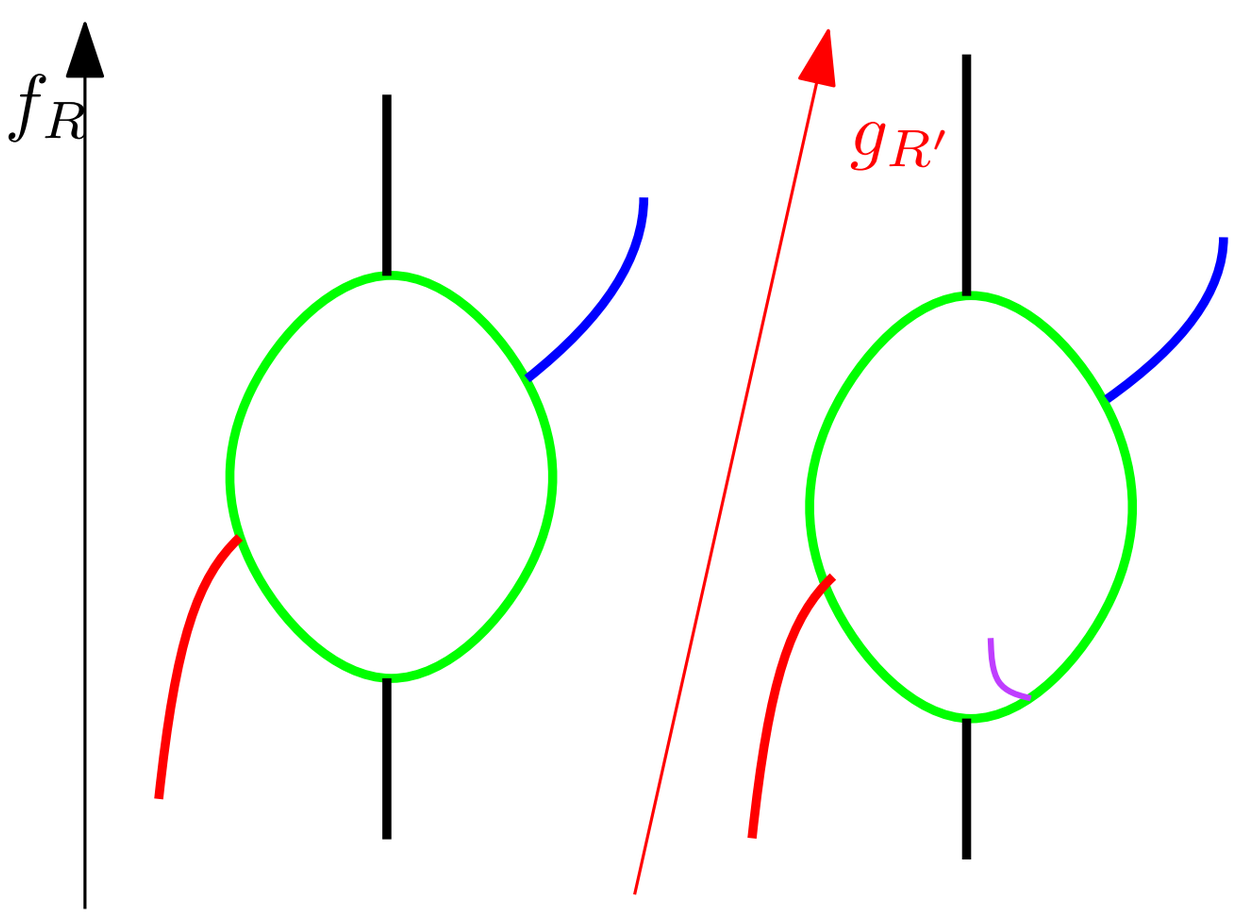


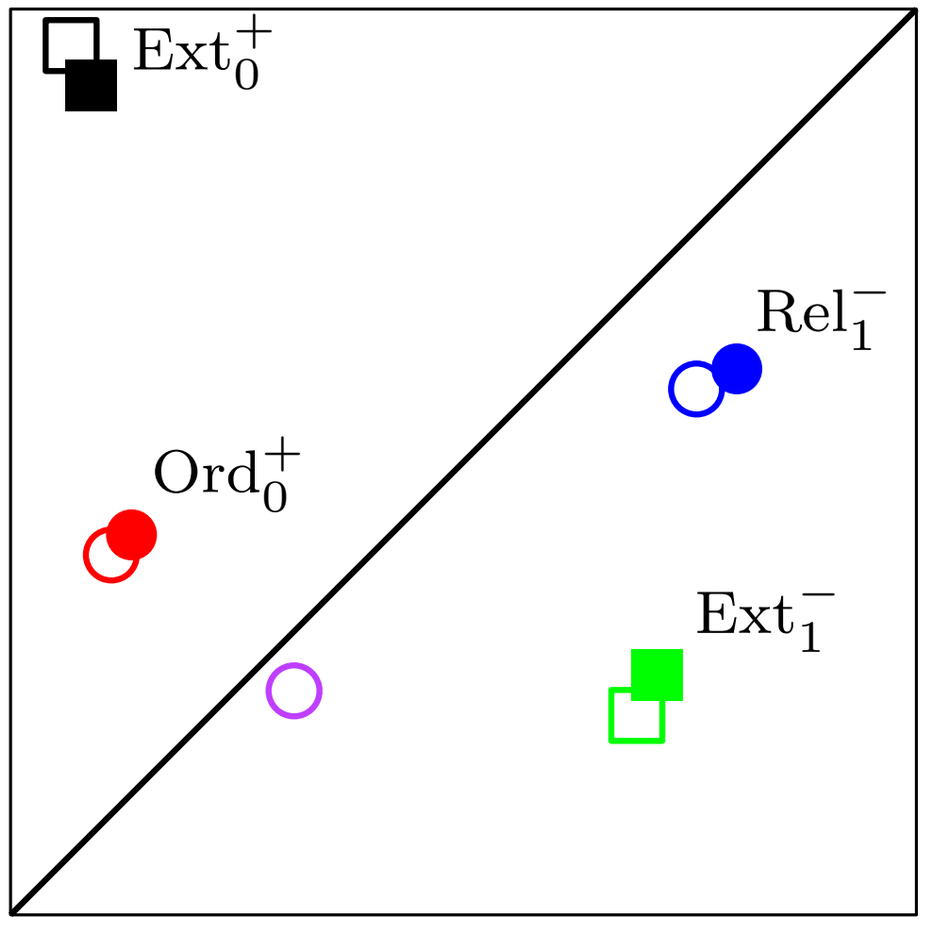
Extended persistence as a descriptor for Reeb graph

The **bottleneck distance** between Dg and Dg′ is:

$$d_B(\mathrm{Dg}, \mathrm{Dg}') = \inf_{\Gamma} \ \mathit{cost}(\Gamma),$$

where Γ ranges over all *p*artial matchings between Dg and Dg′.





● ordinary / relative

■ extended

An approximation inequality for Mapper

Regularity of the filter function

(exact) modulus of continuity of $f :\mapsto \mathbb{R}$

$$\omega_f(\delta) := \sup_{\|x-x'\|\leq\delta} |f(x) - f(x')|$$

Approximation inequality (exact filter):

Theorem : Assume that $\mathcal{X}$ is a compact submanifold of dimension d in $\mathbb{R}^D$ with positive reach rch and positive convexity radius ρ. Let f be a Morse type function defined on $\mathcal{X}$, and let $\mathbb{X}_n$ be a point cloud in $\mathcal{X}$ such that $4d_H(\mathcal{X}, \mathbb{X}_n) \leq \delta$. Then, for δ such that

$$\delta \leq \frac{1}{4}\min\{rch, \rho\},$$
$$\max\{|f(X) - f(X')| : X, X' \in \mathbb{X}_n, \|X - X'\| \leq \delta\} < gr$$

the Mapper $\mathrm{M}_n := \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, f(\mathbb{X}_n))$ is such that:

$$d_B\left(\mathsf{DgR}_f(\mathcal{X}), \mathsf{DgM}_n\right) \leq r + 2\omega(\delta).$$

An approximation inequality for Mapper

Regularity of the filter function

(exact) modulus of continuity of $f :\mapsto \mathbb{R}$

$$\omega_f(\delta) := \sup_{\|x-x'\|\leq\delta} |f(x) - f(x')|$$

Approximation inequality (exact filter):

Theorem : Assume that $\mathcal{X}$ is a compact submanifold of dimension d in $\mathbb{R}^D$ with positive reach rch and positive convexity radius ρ. Let f be a Morse type function defined on $\mathcal{X}$, and let $\mathbb{X}_n$ be a point cloud in $\mathcal{X}$ such that $4d_H(\mathcal{X}, \mathbb{X}_n) \leq \delta$. Then, for δ small enough, the Mapper $\mathrm{M}_n := \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, f(\mathbb{X}_n))$ is such that:

$$d_B\left(\mathrm{DgR}_f(\mathcal{X}), \mathrm{DgM}_n\right) \leq r + 2\omega(\delta).$$

- d_B is controlled by the Hausdorff approximation of $\mathcal{X}$, the resolution level in the codomain and the regularity of the filter.
- Linear projection (PCA) filters are one Lipschitz

An approximation inequality for Mapper: filter approximation

- Assume that the filter function $\hat{f}$ used to compute the Mapper is only an approximation of the filter function f with which the Reeb graph is computed (PCA, density estimators ...)
- In this context : the pair $(\mathbb{X}_n, \hat{f})$ appears as an approximation of the pair $(\mathcal{X}, f)$.
- Assume that

$$\max\{\max\{|f(X) - f(X')|, |\hat{f}(X) - \hat{f}(X')|\} : X, X' \in \mathbb{X}_n, \|X - X'\| \leq \delta\} < gr.$$

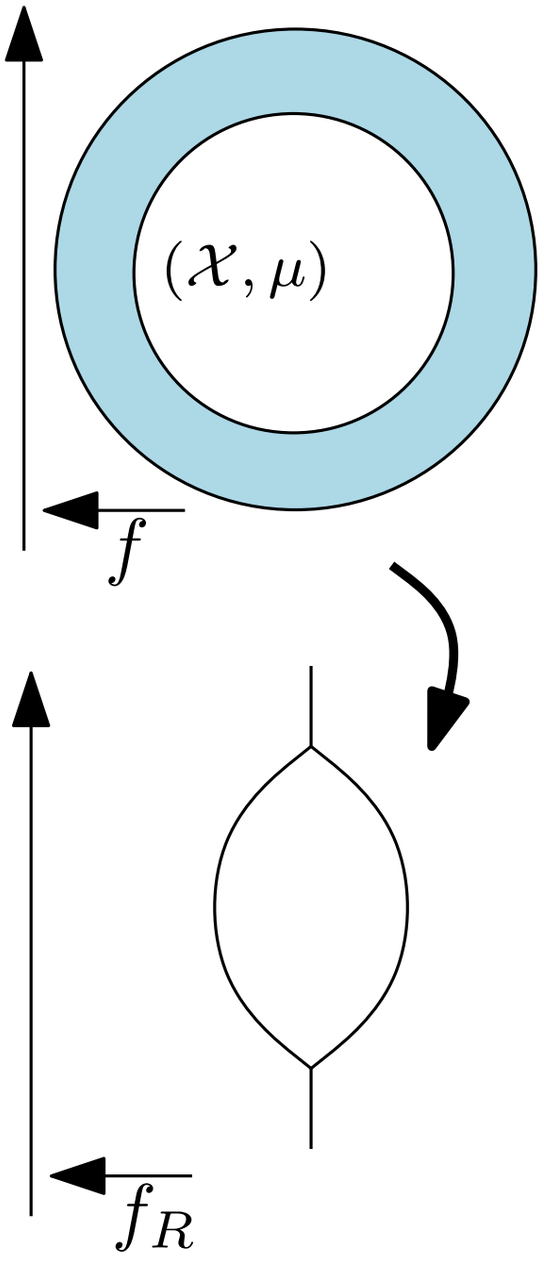
Then, the Mapper $\hat{\mathsf{M}}_n = \mathsf{M}_{r,g,\delta}(\mathbb{X}_n, \hat{f}(\mathbb{X}_n))$ built on $\mathbb{X}_n$ with filter function $\hat{f}$ and parameters r, g, δ chosen as before satisfies:

$$d_B\left(\mathsf{DgR}_f(\mathcal{X}), \mathsf{Dg}\hat{\mathsf{M}}_n\right) \leq 2r + 2\omega(\delta) + \max_{1 \leq i \leq n} |f(X_i) - \hat{f}(X_i)|.$$

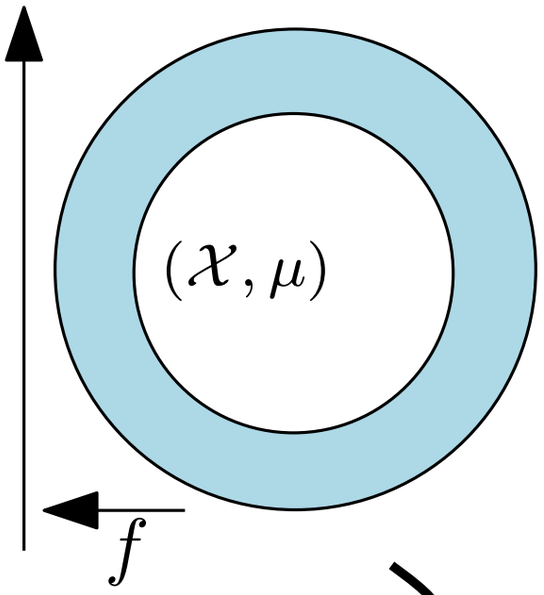
proof: stability theorem for extended persistence diagrams [Cohen Steiner et al. 2009]

Statistics for Mapper

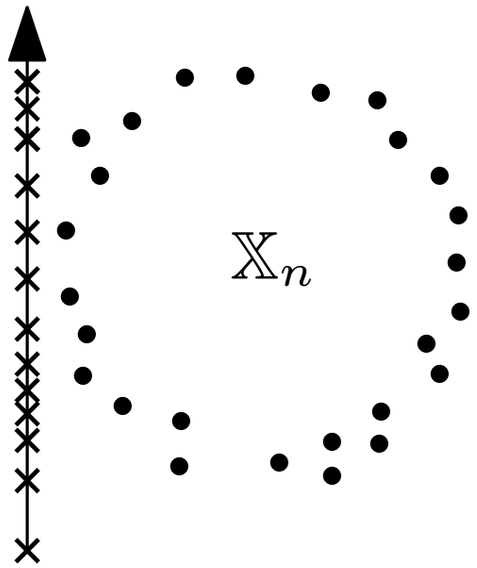
Convergence of Mapper



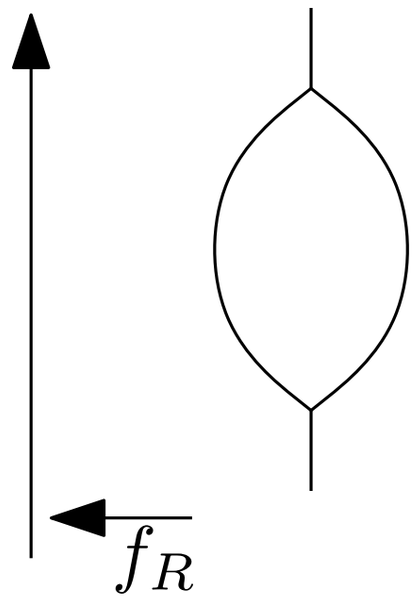
Convergence of Mapper



n points sampled
i.i.d. according to μ
μ is (a, b)-standard



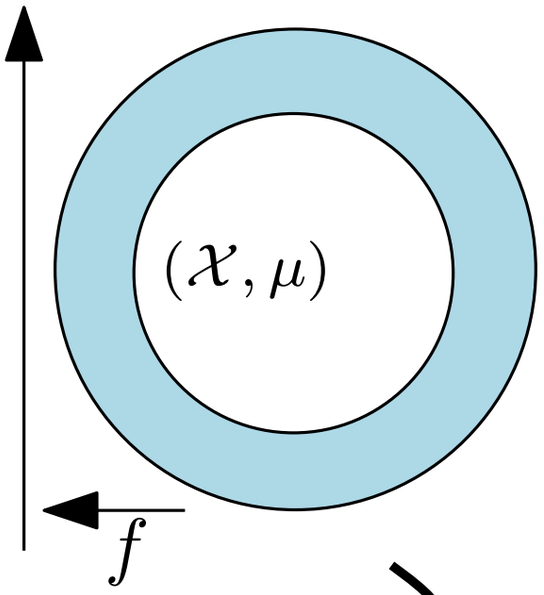
$$\mathbb{Y}_n = f(\mathbb{Y}_n) \text{ or } = \hat{f}(\mathbb{Y}_n)$$

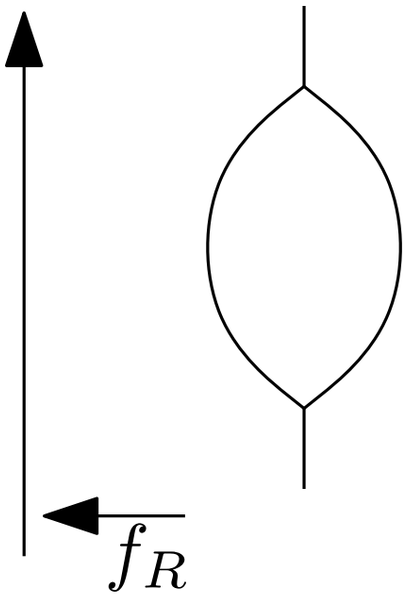


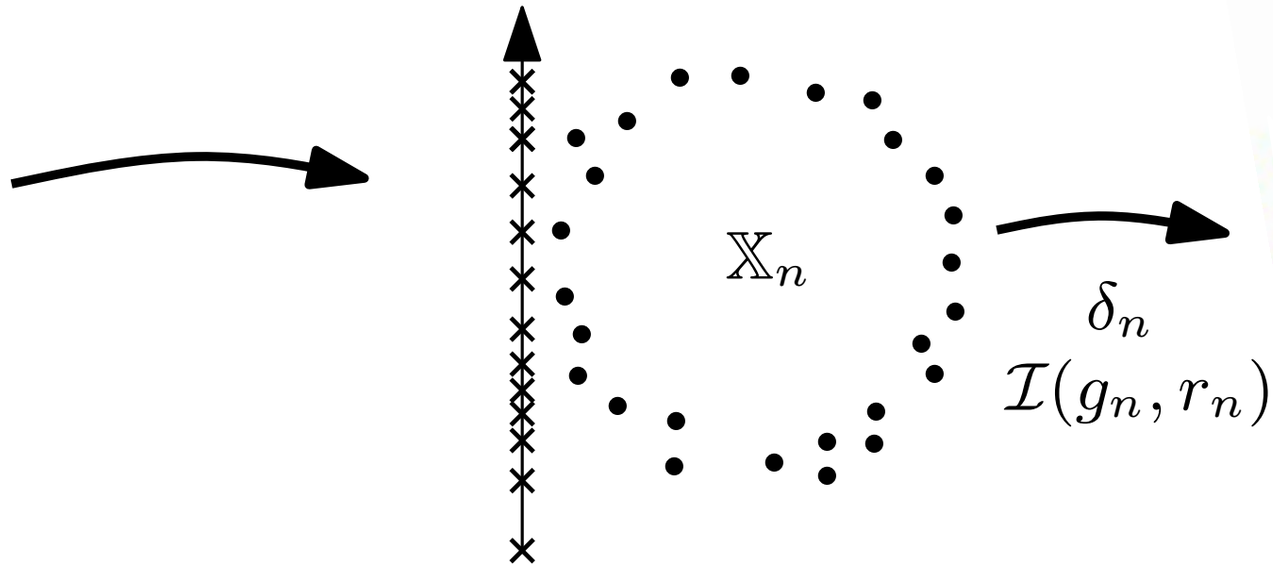
For any Euclidean ball $B(x, t)$:

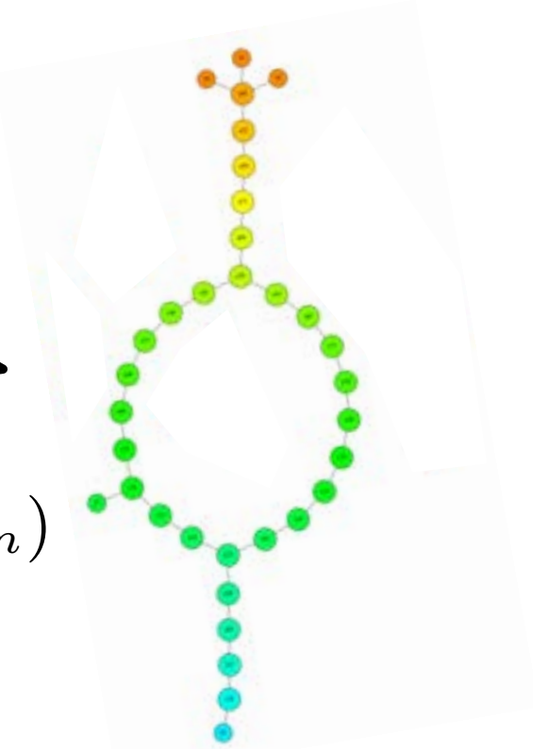
$$P(B(x,t)) \geq \min(1, at^d).$$

Convergence of Mapper









$$\mathbb{Y}_n = f(\mathbb{Y}_n) \text{ or } = \hat{f}(\mathbb{Y}_n)$$

Questions:

- Statistical properties of the estimator $\mathrm{M}_f(\mathbb{X}_n, \delta_n, \mathcal{I}(g_n, r_n))$?
- Convergence to the ground truth $\mathrm{R}_f(\mathcal{X})$ in d_∞? Deviation bounds?

Convergence of Mapper: known generative model and known filter

Theorem : *Under the assumptions of the approximation inequality*, take

$$V_n(\delta) = \max\{|f(X) - f(X')| \,:\, X, X' \in \mathbb{X}_n, \|X - X'\| \leq \delta\},$$

$$g \in \left(0, \frac{1}{2}\right), \quad \delta_n = 8\left(\frac{2\log(n)}{an}\right)^{1/b} \quad \text{and} \quad r_n = \frac{V_n(\delta_n)}{g}.$$

Let ω be a modulus of continuity of f such that $\omega(x)/x$ is a decreasing function. Let M_n the Mapper computed with parameters (r_n, g, δ_n). Then :

$$\mathbb{E}\left[d_B\left(\mathrm{R}_f(\mathcal{X}), \mathrm{M}_n\right)\right] \leq C\omega(\delta_n)$$

- Lipschitz filter functions : same rate of convergence as for persistence diagram inference and support estimation for the Hausdorff metric.
- Filters with concave modulus of continuity: the "distortion" created by the filter function slows down the convergence.
- Minimax rate of convergence:

$$\sup_{P \in \mathcal{P}_{a,b}} \mathbb{E}\left[\sup_{f \in \mathcal{F}(P,\omega)} d_B\left(\mathrm{R}_f(\mathcal{X}), \mathsf{M}_n\right)\right] \leq C\,\omega\left(\frac{2 \cdot 8^b}{a}\,\frac{\log(n)}{n}\right)^{1/b}$$

Convergence of Mapper: unknown generative model, known filter

- Framework:
  - Exact values $\mathbb{Y}_n = f(\mathbb{X}_n)$ of the filter on the point could
  - Unknown parameters a and b $\rightarrow$ alternative definition for δ_n is needed.
- We adapt a subsampling approach from [Fasy et.al 2014] to tune δ_n:
  - let $\beta > 0$ and let $s(n) := \frac{n}{\log(n)^{1+\beta}}$
  - Let $\delta_n := \mathrm{d_H}(\hat{X}_n^{s(n)}, \hat{X}_n)$ where $\hat{X}_n^{s(n)}$ is a random subset of $\hat{X}_n$ of size $s(n)$
  - Take

$$g \in \left(0, \frac{1}{2}\right) \text{ and } \quad r_n = \frac{V_n(\delta_n)}{g}.$$

- Let M_n computed with parameters (r_n, g, δ_n), then

$$\mathbb{E}\left[d_B\left(\mathrm{R}_f(\mathcal{X}), \mathrm{M}_n\right)\right] \leq C\omega\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b}$$

Convergence of Mapper: unknown generative model, estimated filter

- Mapper can be computed with any filter function, including estimated filter functions such as PCA eigenfunctions, eccentricity functions, Laplacian eigenfunctions, density estimators, alternative distance functions ...

- **Approximation for estimated filter:**
  the Mapper $\hat{\mathsf{M}}_n = \mathsf{M}_{r,g,\delta}(\mathbb{X}_n, \hat{f}(\mathbb{X}_n))$ built on $\mathbb{X}_n$ with filter function $\hat{f}$ and parameters r, g, δ satisfies:

$$d_B\left(\mathsf{DgR}_f(\mathcal{X}), \mathsf{Dg}\hat{\mathsf{M}}_n\right) \leq 2r + 2\omega(\delta) + \max_{1\leq i\leq n} |f(X_i) - \hat{f}(X_i)|.$$

Convergence of Mapper: unknown generative model, estimated filter

Application to PCA eigenfunction

- Π_1 : proj. onto first principal direction of the covariance operator
- $\hat{\Pi}_1$: proj. onto first principal direction of the emp. covariance operator

Let ω_1 be a **known** upper bound on $\omega(f)$ (ok for Lip function)

$$\hat{V}_n(\delta_n) = \max\{|\hat{f}(X) - \hat{f}(X')| \, : \, X, X' \in \mathbb{X}_n, \|X - X'\| \leq \delta_n\}$$

$$r_n = \frac{\max\{\omega_1(\delta_n), \hat{V}_n(\delta_n)\}}{g}.$$

Using [Biau et. al. 2012] :

$$\mathbb{E}\left[d_B\left(\mathsf{DgR}_{\Pi_1}(\mathcal{X}), \mathsf{DgM}_{r_n,g,\delta_n}(\mathbb{X}_n, \hat{\Pi}_1(\mathbb{X}_n))\right)\right] \lesssim \left(\frac{(\log(n))^{2+\beta}}{n}\right)^{1/b} \vee \frac{1}{\sqrt{n}}.$$

Confidence sets for extended persistence diagrams

- We can evaluate how accurate M_n is by providing confidence sets.
- For $\alpha \in (0,1)$, we look for some value $\eta_{n,\alpha}$ such that

$$Pd_B(\mathsf{DgM}_n, \mathsf{DgR}_f(\mathcal{X})) \geq \eta_{n,\alpha} \leq \alpha$$

  or at least such that

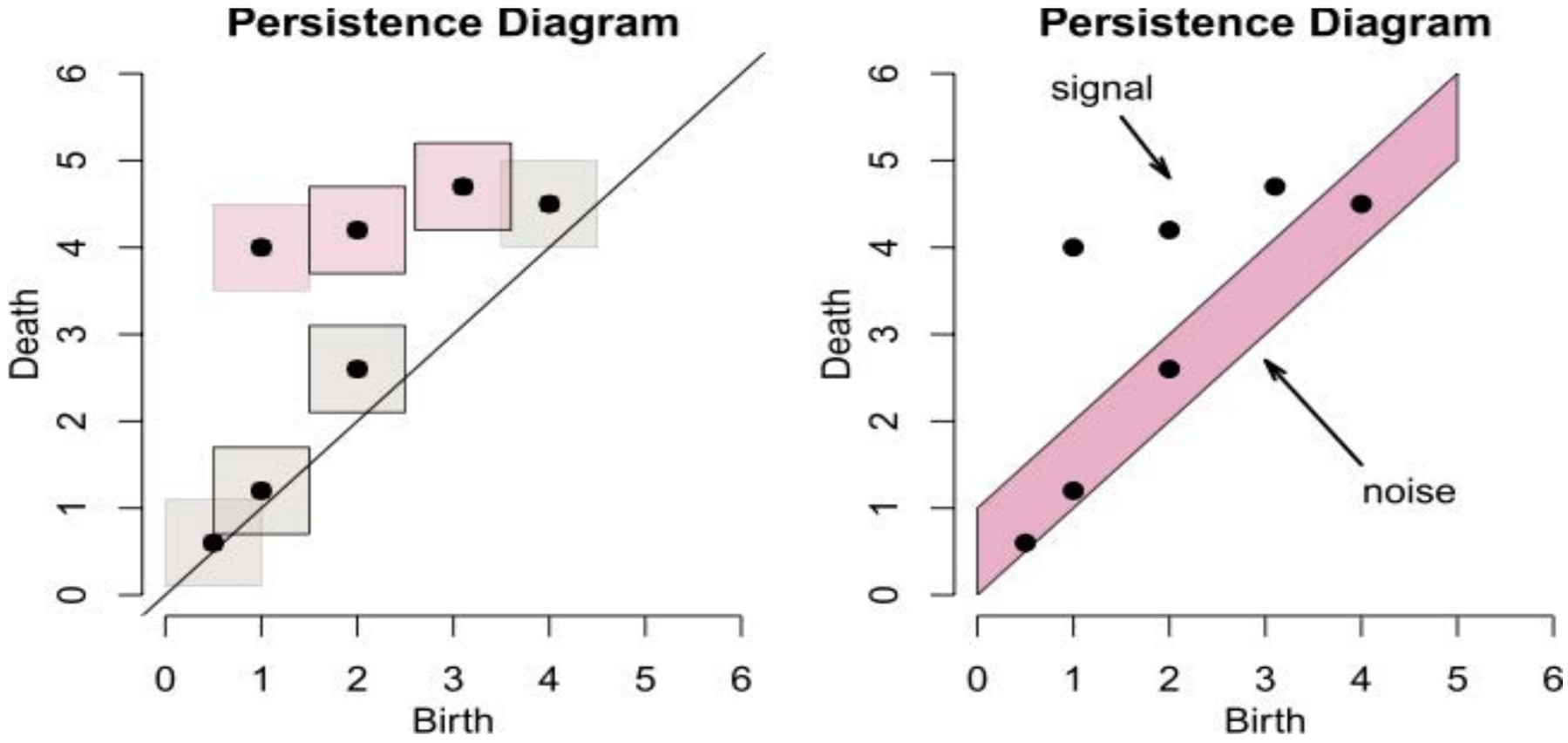
$$\limsup_{n\to\infty} Pd_B(\mathsf{DgM}_n, \mathsf{DgR}_f(\mathcal{X})) \geq \eta_{n,\alpha} \leq \alpha.$$

- $\mathcal{M}_\alpha = \{\mathrm{R} \in \mathcal{R} : d_B(\mathsf{DgM}_n, \mathsf{DgR}) \leq \alpha\}$: closed ball of radius α in the bottleneck distance and centered at the Mapper M_n in the space of Reeb graphs
- We can visualize the signatures of the points belonging to this ball in various ways [Fasy et al. 2014].

Confidence sets for extended persistence diagrams

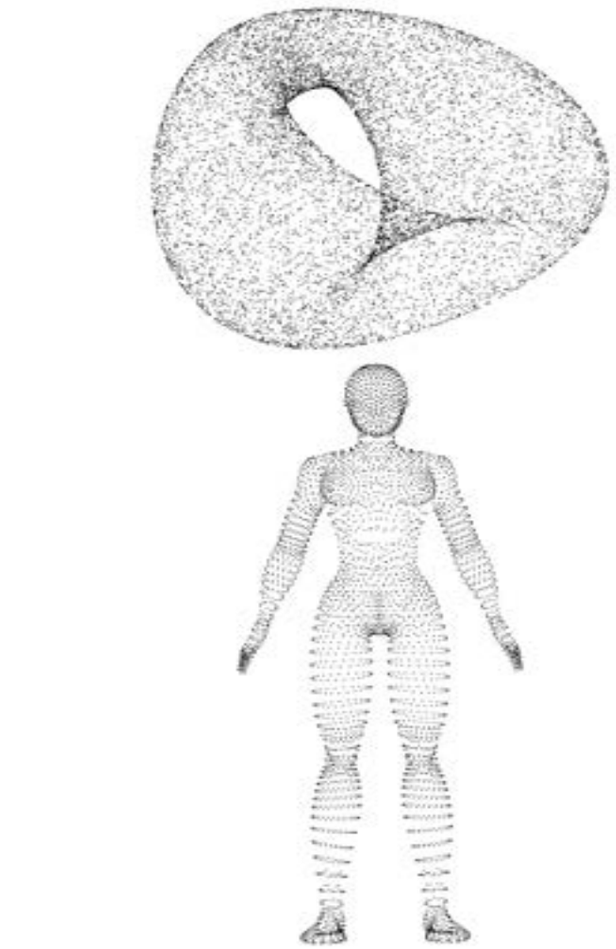
$$Pd_B(\mathrm{DgM}_n, \mathrm{DgR}_f(\mathcal{X})) \geq \eta_{n,\alpha} \leq \alpha \quad ?$$

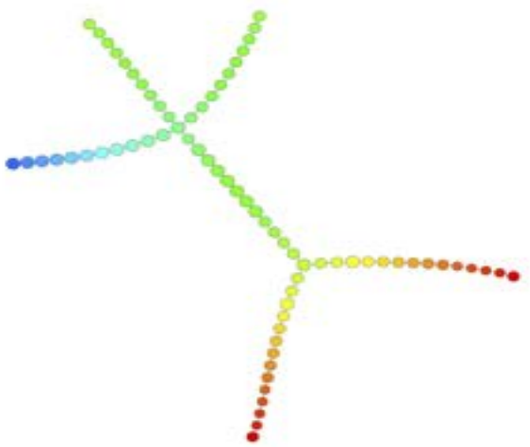
- First option: center a box of side length 2α at each point of the extended persistence diagram of M_n.
- Second option: visualize the confidence set by adding a band at (vertical) distance 2α from the diagonal (the bottleneck distance being defined for the ℓ_∞ norm).

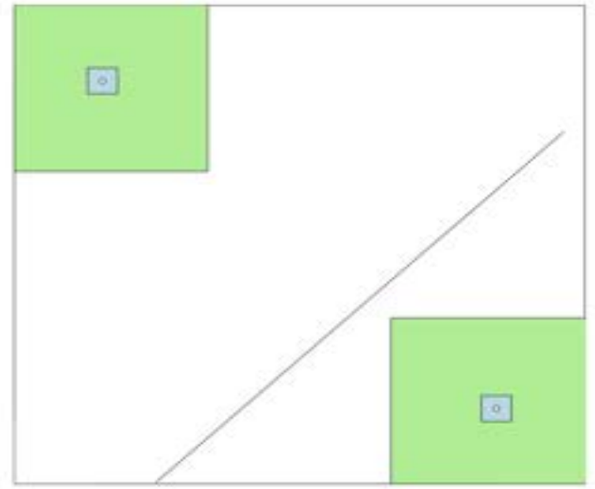


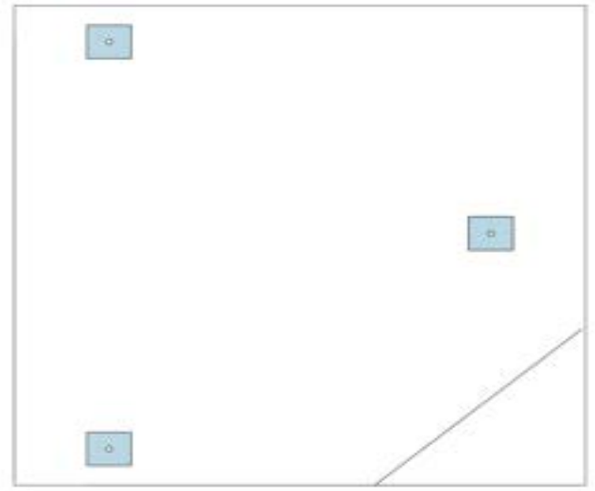
[Fasy et al. 2014]

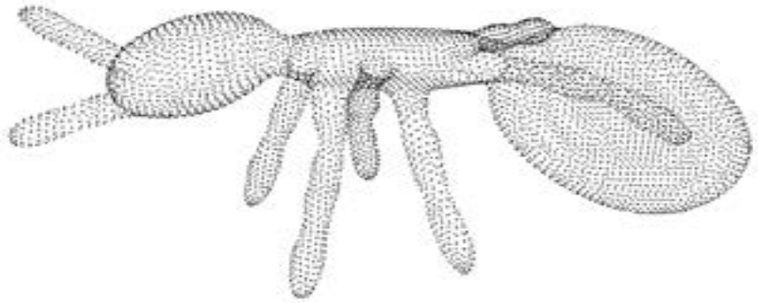
Experiments

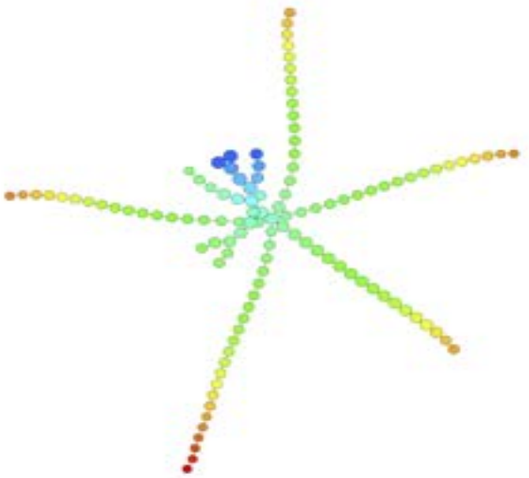


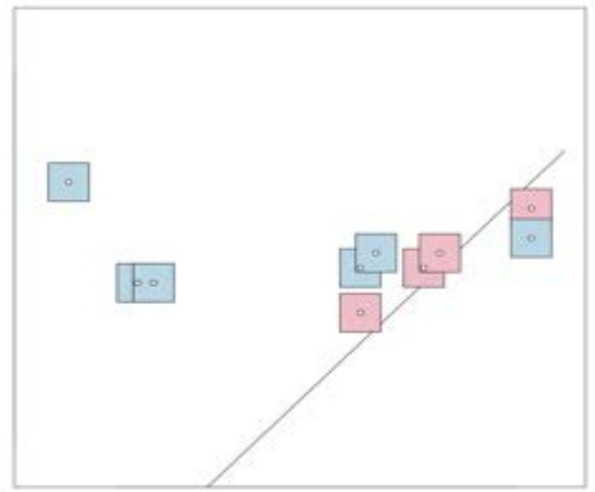




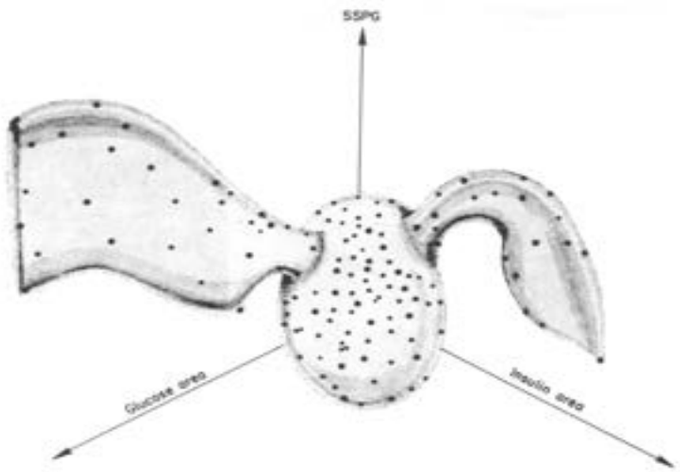


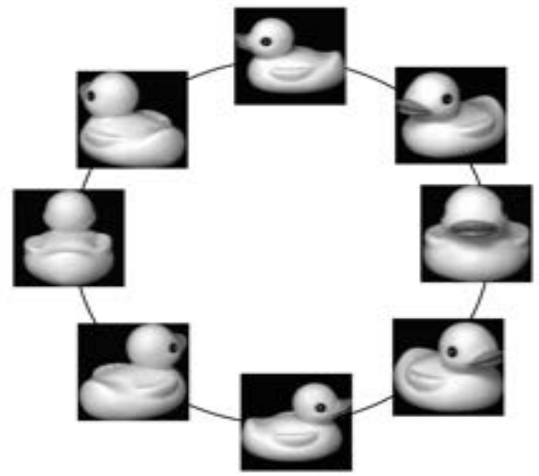


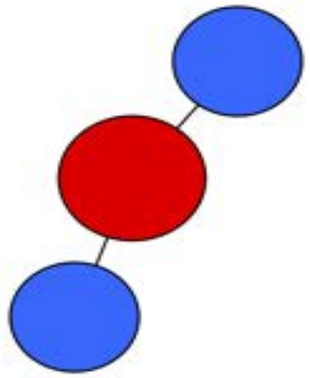


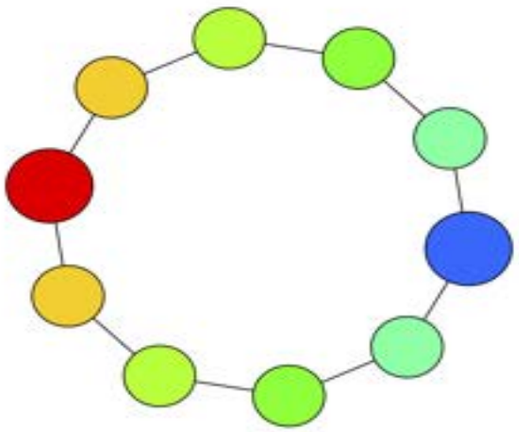


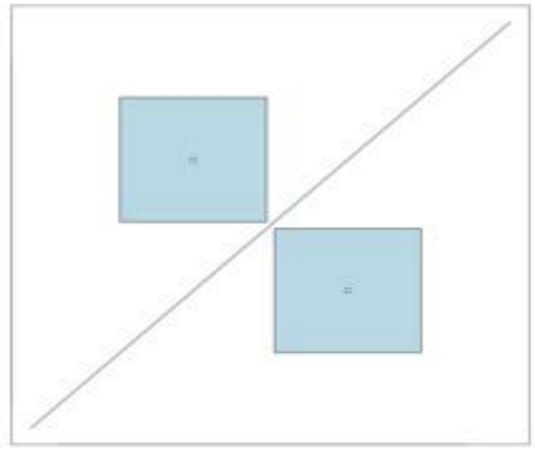
Experiments

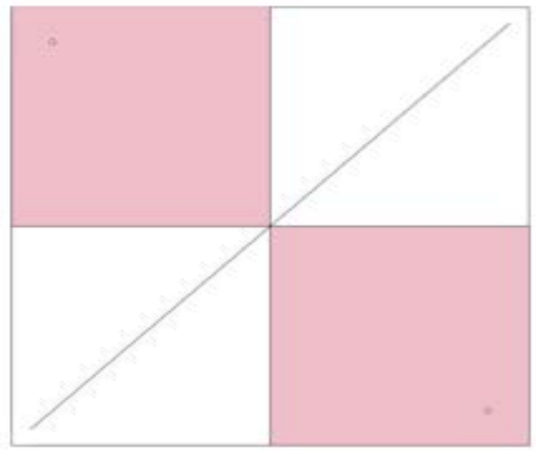


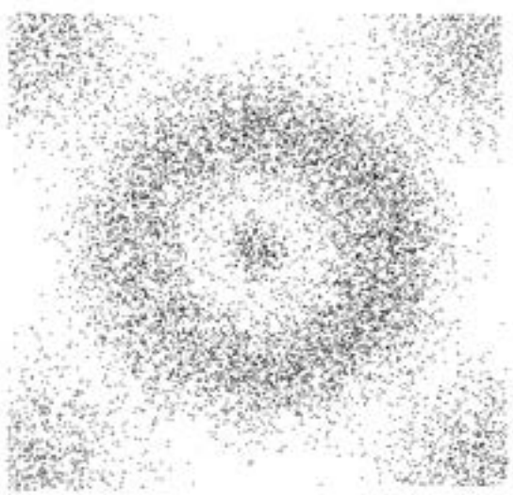


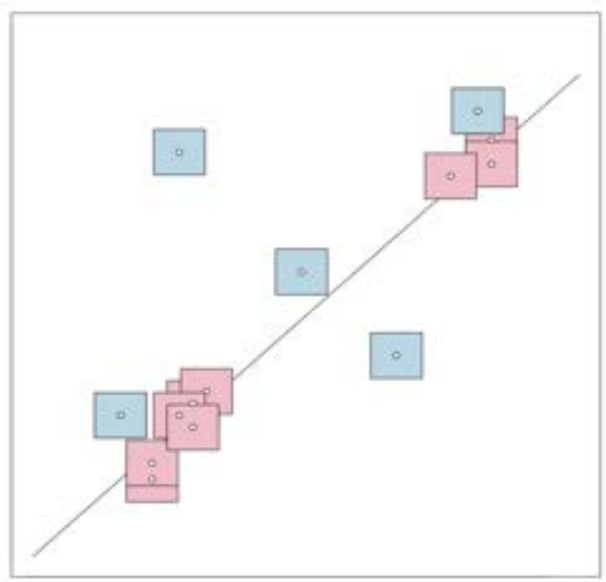




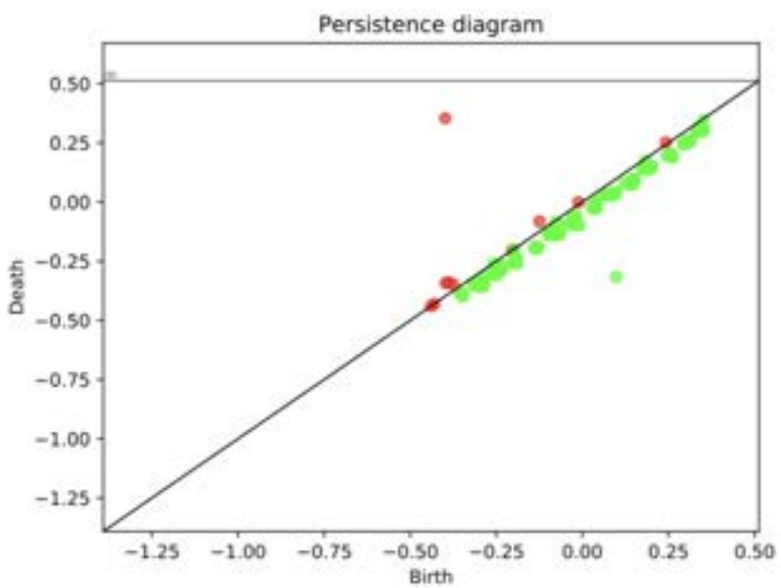








TDA of Single-cell Hi-C Contact Maps (Carriere and Rabadan 18)



- Castaneus mouse embryonic stem cells
- Consider the loci of the genome that are spatially close and interacting in the nucleus of the cell.
- For each cell : compute a contact map, which is a symmetric matrix, whose rows and columns represent small loci, or bins, of the genome.
- Stratum-adjusted correlation coefficient SCC for comparing Hi-C matrices. Provide a 1171 x 1171 similarity matrix between cells.
- Biological factors correlated with the cell cycle phases can be validated on the Mapper