GESDA Introductory School
5-9 September 2022
IESC Cargèse, Corsica

Topological Inference and Machine Learning

Mathieu Carrière
INRIA Sophia-Antipolis
mathieu.carriere@inria.fr

I. Topological Inference on Persistence Diagrams for Point Clouds

II. Topological Vectorizations for Inference and Machine Learning

(III. A Gaussian Kernel for Persistence Diagrams)

IV. Deep Learning for Persistence Diagrams

I. Topological Inference on Persistence Diagrams for Point Clouds

II. Topological Vectorizations for Inference and Machine Learning

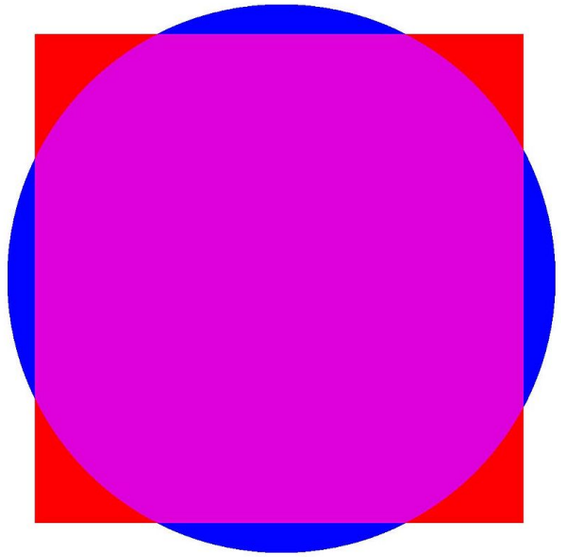
(III. A Gaussian Kernel for Persistence Diagrams)

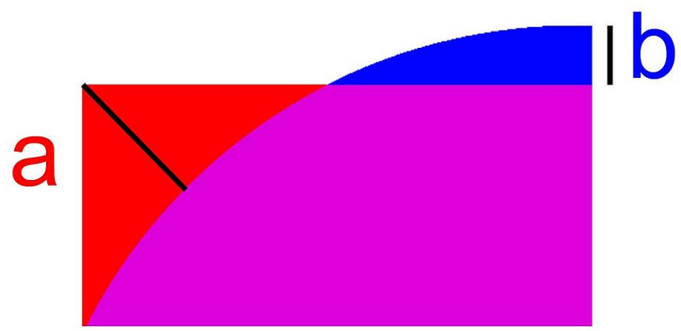
IV. Deep Learning for Persistence Diagrams

Starting point: stability properties for point clouds

Def: The Hausdorff distance between two subspaces X, Y of a common metric space (Z, d) is:

$$d_H(X,Y) = \max\{\sup_{y\in Y} d(y,X), \sup_{x\in X} d(x,Y)\}$$
$$= \max\{\sup_{y\in Y}\inf_{x\in X} d(y,x), \sup_{x\in X}\inf_{y\in Y} d(x,y)\}$$





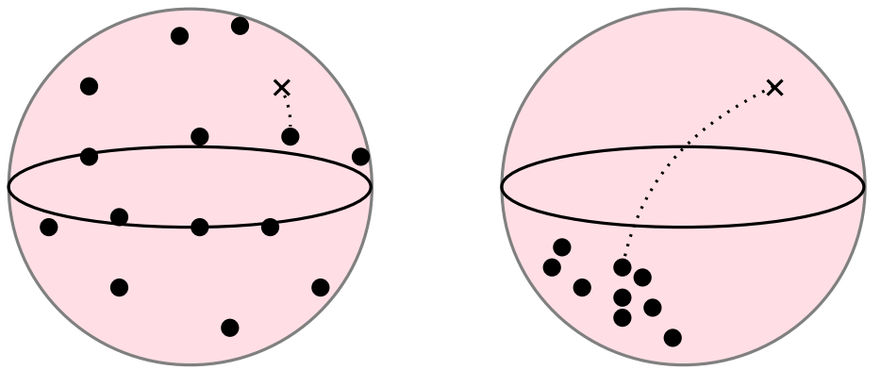
$$d_H(X,Y) = \max\{a,b\}$$

Starting point: stability properties for point clouds

Def: The Hausdorff distance between two subspaces X, Y of a common metric space (Z, d) is:

$$d_H(X,Y) = \max\{\sup_{y\in Y} d(y,X), \sup_{x\in X} d(x,Y)\}$$
$$= \max\{\sup_{y\in Y}\inf_{x\in X} d(y,x), \sup_{x\in X}\inf_{y\in Y} d(x,y)\}$$

Given a sampling $\hat{X}_n \subseteq X$, $d_H(\hat{X}_n, X)$ is a measure of sampling quality.



Starting point: stability properties for point clouds

Def: The Hausdorff distance between two subspaces X, Y of a common metric space (Z, d) is:

$$d_H(X,Y) = \max\{\sup_{y\in Y} d(y,X), \sup_{x\in X} d(x,Y)\}$$
$$= \max\{\sup_{y\in Y}\inf_{x\in X} d(y,x), \sup_{x\in X}\inf_{y\in Y} d(x,y)\}$$

Def: The Gromov-Hausdorff distance between metric spaces $(X, d_X), (Y, d_Y)$ is the Hausdorff distance of the best common isometric embedding:

$$d_{GH}((X,d_X),(Y,d_Y)) = \inf_\gamma d_H(\gamma(X),\gamma(Y)),$$

where $d(\gamma(x),\gamma(x')) = d_X(x,x')$ and $d(\gamma(y),\gamma(y')) = d_X(y,y')$.

Starting point: stability properties for point clouds

Def: The Hausdorff distance between two subspaces X, Y of a common metric space (Z, d) is:

$$d_H(X,Y) = \max\{\sup_{y\in Y} d(y,X), \sup_{x\in X} d(x,Y)\}$$
$$= \max\{\sup_{y\in Y}\inf_{x\in X} d(y,x), \sup_{x\in X}\inf_{y\in Y} d(x,y)\}$$

Def: The Gromov-Hausdorff distance between metric spaces $(X, d_X), (Y, d_Y)$ is metric distortion of the best correspondence:

$$d_{GH}((X,d_X),(Y,d_Y)) = \inf_{\mathcal{C}} \sup_{(x,y),(x',y')\in\mathcal{C}} |d_X(x,x') - d_Y(y,y')|,$$

where $\mathcal{C} \subseteq X \times Y$ s.t. $\forall x$, $\exists y_x \in Y$ s.t. $(x, y_x) \in \mathcal{C}$ (and vice-versa).

Starting point: stability properties for point clouds

Def: The Hausdorff distance between two subspaces X, Y of a common metric space (Z, d) is:

$$d_H(X,Y) = \max\{\sup_{y\in Y} d(y,X), \sup_{x\in X} d(x,Y)\}$$
$$= \max\{\sup_{y\in Y}\inf_{x\in X} d(y,x), \sup_{x\in X}\inf_{y\in Y} d(x,y)\}$$

Def: The Gromov-Hausdorff distance between metric spaces $(X, d_X), (Y, d_Y)$ is metric distortion of the best correspondence:

$$d_{GH}((X,d_X),(Y,d_Y)) = \inf_{\mathcal{C}} \sup_{(x,y),(x',y')\in\mathcal{C}} |d_X(x,x') - d_Y(y,y')|,$$

where $\mathcal{C} \subseteq X \times Y$ s.t. $\forall x$, $\exists y_x \in Y$ s.t. $(x, y_x) \in \mathcal{C}$ (and vice-versa).

Thm: If X and Y are common subspaces of a common metric space (Z, d), then

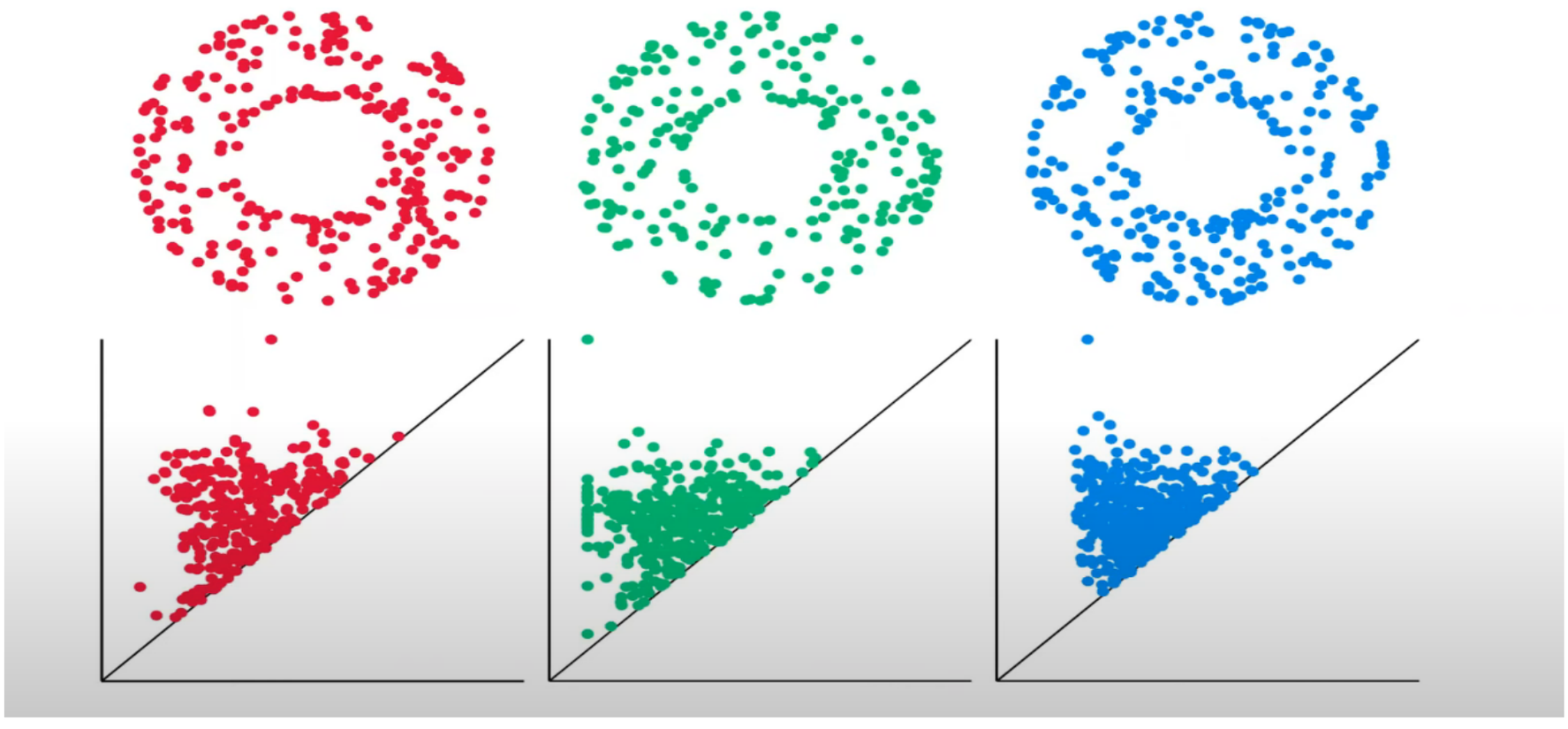
$$d_b(D_{\text{Cech}}(X), D_{\text{Cech}}(Y)) \leq d_H(X,Y).$$

Starting point: stability properties for point clouds

[*Persistence stability for geometric complexes*, Chazal, de Silva, Oudot, Geom. Dedicata, 2013].

Thm: If X and Y are pre-compact metric spaces, then

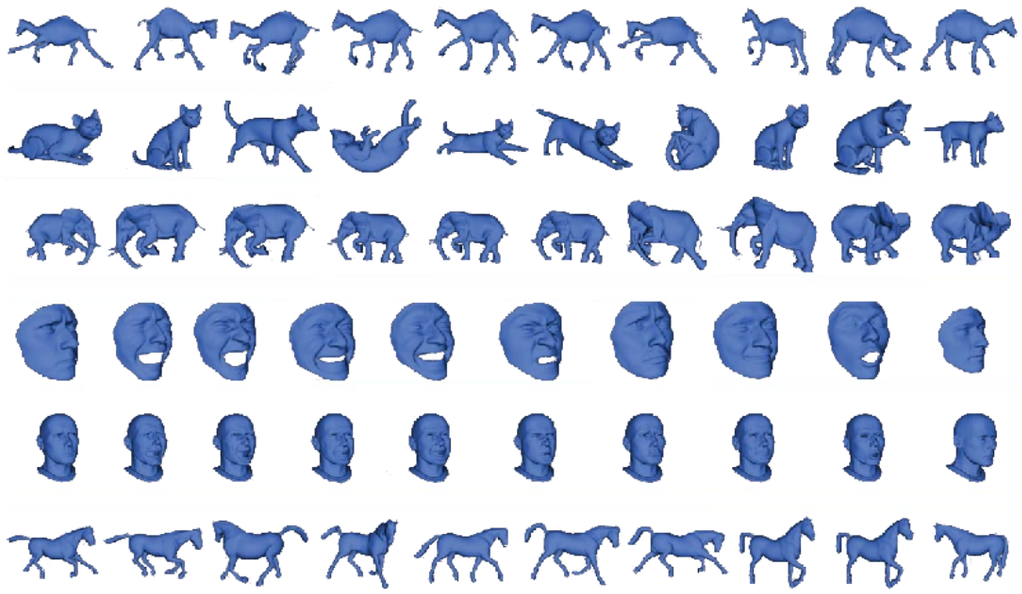
$$d_b(D_{\mathrm{Rips}}(X), D_{\mathrm{Rips}}(Y)) \leq d_{GH}(X, Y).$$

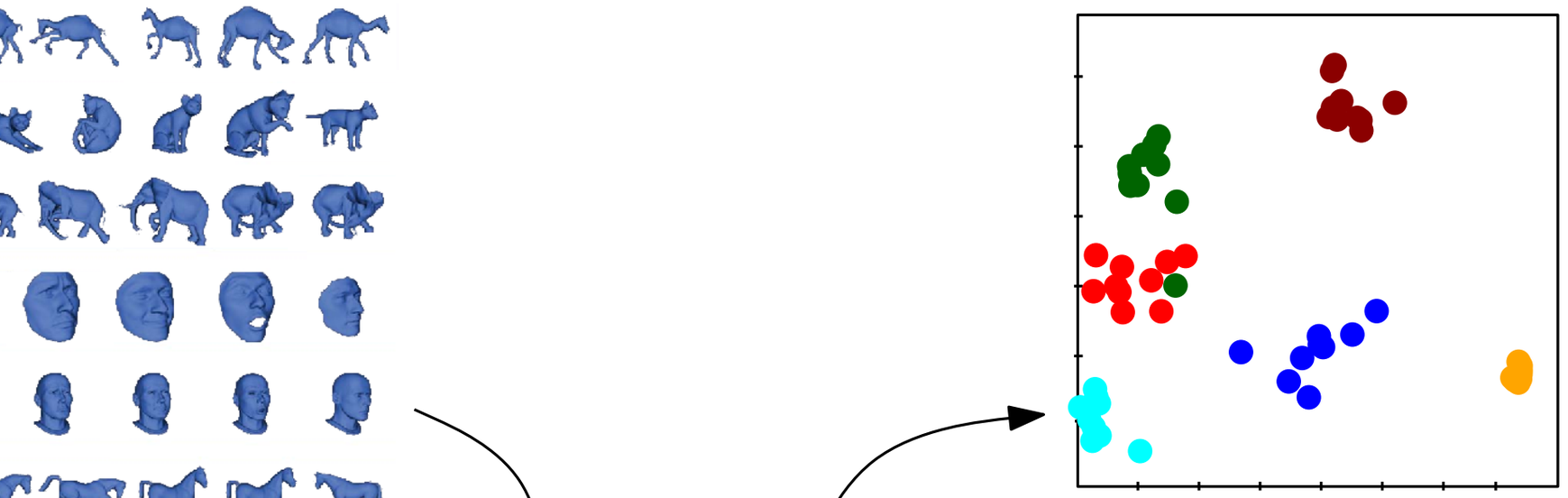


Rem: This result also holds for Čech and other families of filtrations (particular case of a more general theorem).

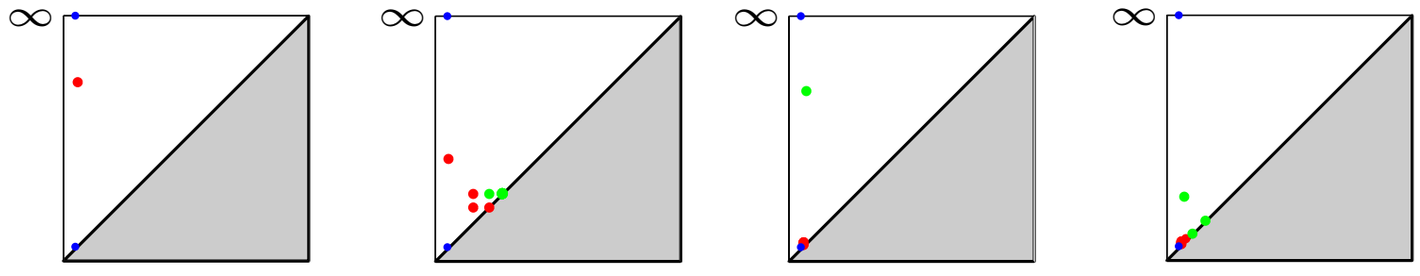
Application: non rigid shape classification

[*Gromov-Hausdorff Stable Signatures for Shapes using Persistence*, Chazal et al., Symp. Geom. Process., 2009]

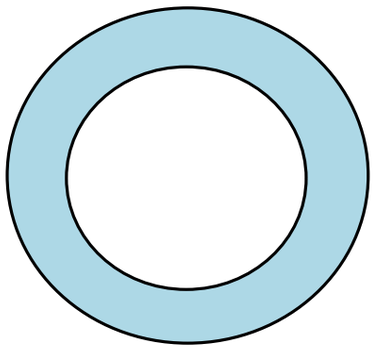




MDS using bottleneck distance.



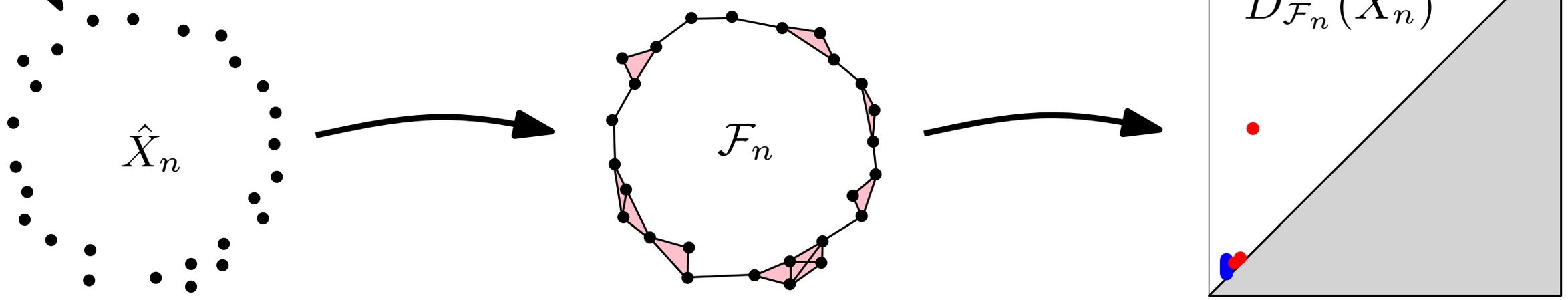
Application: Statistics on Persistence Diagrams



(X, d) metric space
μ probability measure with compact support X_μ

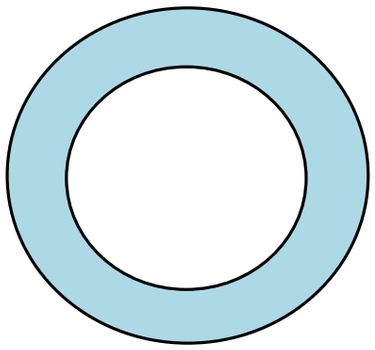
Examples:
- $\mathcal{F}_n = \mathrm{Rips}(\hat{X}_n)$
- $\mathcal{F}_n = \check{\mathrm{C}}\mathrm{ech}(\hat{X}_n)$
- $\mathcal{F}_n =$ sublevelset filtration of $d(., X_\mu)$.

Sample n points according to μ.



Questions: Statistical properties of $D_{\mathcal{F}_n}(\hat{X}_n)$? $D_{\mathcal{F}_n}(\hat{X}_n) \to ?$ as $n \to +\infty$?

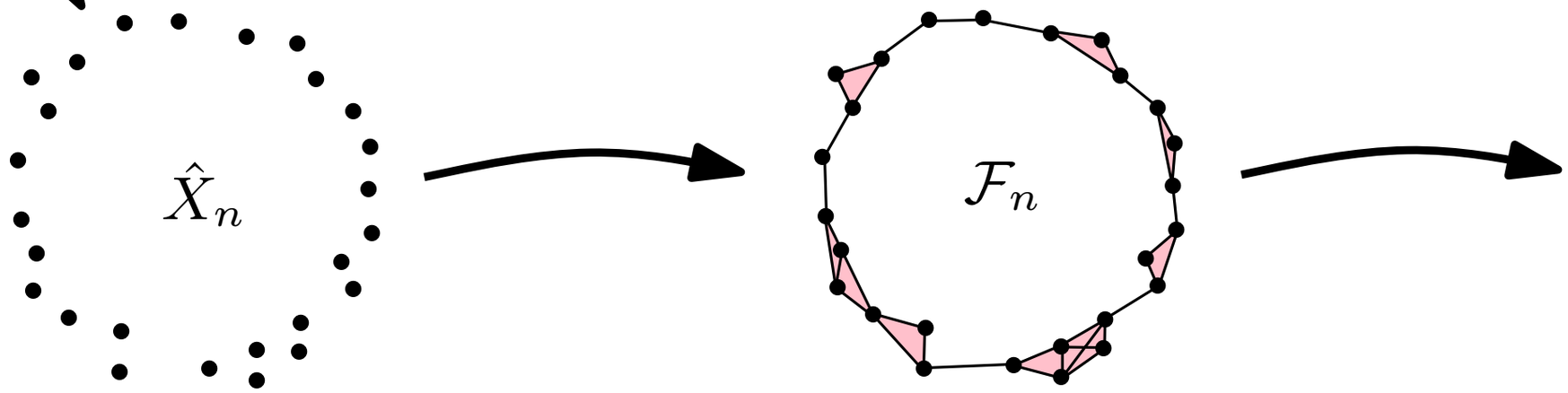
Application: Statistics on Persistence Diagrams



(X, d) metric space
μ probability measure with compact support X_μ

Sample n points according to μ.

Examples:
- $\mathcal{F}_n = \mathrm{Rips}(\hat{X}_n)$
- $\mathcal{F}_n = \check{\mathrm{C}}\mathrm{ech}(\hat{X}_n)$
- $\mathcal{F}_n =$ sublevelset filtration of $d(., X_\mu)$.



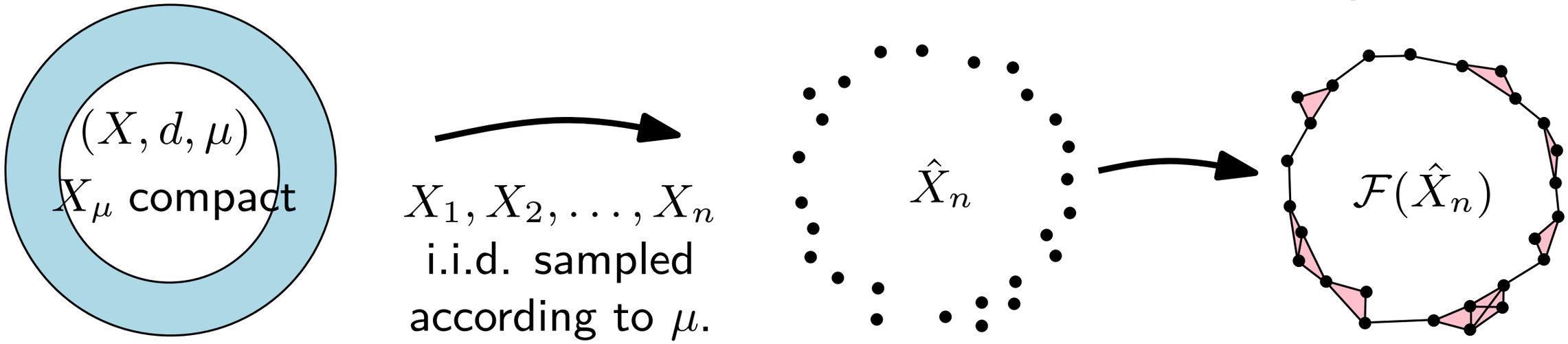
Stability thm: $d_b(D_{\mathcal{F}}(X_\mu), D_{\mathcal{F}_n}(\hat{X}_n)) \leq 2 d_{GH}(X_\mu, \hat{X}_n)$

So, for any $\varepsilon > 0$,

$$\mathbb{P}\left(d_b\left(D_{\mathcal{F}}(X_\mu), D_{\mathcal{F}_n}(\hat{X}_n)\right) > \varepsilon\right) \leq \mathbb{P}\left(d_{GH}(X_\mu, \hat{X}_n) > \frac{\varepsilon}{2}\right)$$

Deviation inequality

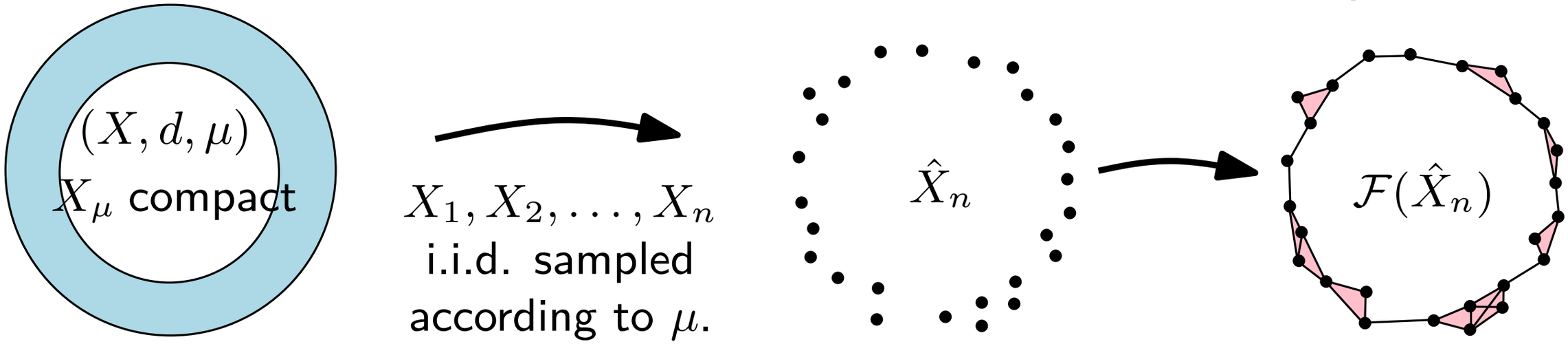
[*Convergence rates for persistence diagram estimation in Topological Data Analysis*, Chazal, Glisse, Labruère, Michel ICML, 2014]



For $a, b > 0$, μ satisfies the (a, b)-standard assumption if for any $x \in X_\mu$ and any $r > 0$, we have $\mu(B(x, r)) \geq \min(ar^b, 1)$.

Deviation inequality

[*Convergence rates for persistence diagram estimation in Topological Data Analysis*, Chazal, Glisse, Labruère, Michel ICML, 2014]



For $a, b > 0$, μ satisfies the (a, b)-standard assumption if for any $x \in X_\mu$ and any $r > 0$, we have $\mu(B(x, r)) \geq \min(ar^b, 1)$.

Thm: If μ satisfies the (a, b)-standard assumption, then for any $\varepsilon > 0$:

$$\mathbb{P}\left(d_b\left(D_{\mathcal{F}}(X_\mu), D_{\mathcal{F}_n}(\hat{X}_n)\right) > \varepsilon\right) \leq \min\left\{\frac{8^b}{a\varepsilon^b}\exp\left(-na\varepsilon^b\right), 1\right\}.$$

Moreover, $\lim_{n\to\infty}\ \mathbb{P}\left(d_b(D_{\mathcal{F}}(X_\mu), D_{\mathcal{F}_n}(\hat{X}_n)) \leq C\left(\frac{\log n}{n}\right)^{1/b}\right) = 1$, where C is a constant that only depends on a and b.

Minimax rate of convergence

[*Convergence rates for persistence diagram estimation in Topological Data Analysis*, Chazal, Glisse, Labruère, Michel ICML, 2014]

Let $\mathcal{P}$ be the set of all the probability measures on the metric space (X, d) satisfying the (a, b)-standard assumption on X.

Minimax rate of convergence

[*Convergence rates for persistence diagram estimation in Topological Data Analysis*, Chazal, Glisse, Labruère, Michel ICML, 2014]

Let $\mathcal{P}$ be the set of all the probability measures on the metric space (X, d) satisfying the (a, b)-standard assumption on X.

Thm: One has the following:

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[d_b(D_{\mathcal{F}}(X_\mu), D_{\mathcal{F}_n}(\hat{X}_n))\right] \leq C\left(\frac{\log n}{n}\right)^{1/b}$$

where the constant C depends only on a and b. Assume moreover that there exists a non isolated point x in X and let $\{x_n\}_n$ be a sequence in $X \setminus \{x\}$ such that $d(x, x_n) \sim (an)^{-1/b}$. Then for any estimator $\hat{D}_n$ of $D_{\mathcal{F}}(X_\mu)$:

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[d_b(D_{\mathcal{F}}(X_\mu), \hat{D}_n)\right] \geq C' d(x, x_n)$$

where C' is an absolute constant.

Minimax rate of convergence

[*Convergence rates for persistence diagram estimation in Topological Data Analysis*, Chazal, Glisse, Labruère, Michel ICML, 2014]

Let $\mathcal{P}$ be the set of all the probability measures on the metric space (X, d) satisfying the (a, b)-standard assumption on X.

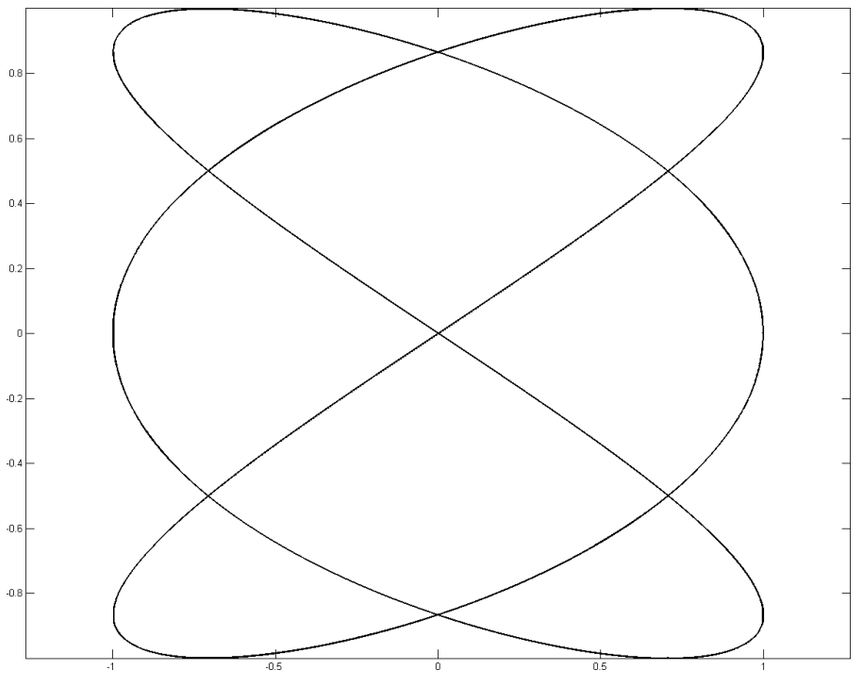
Proof: Apply Le Cam's lemma:

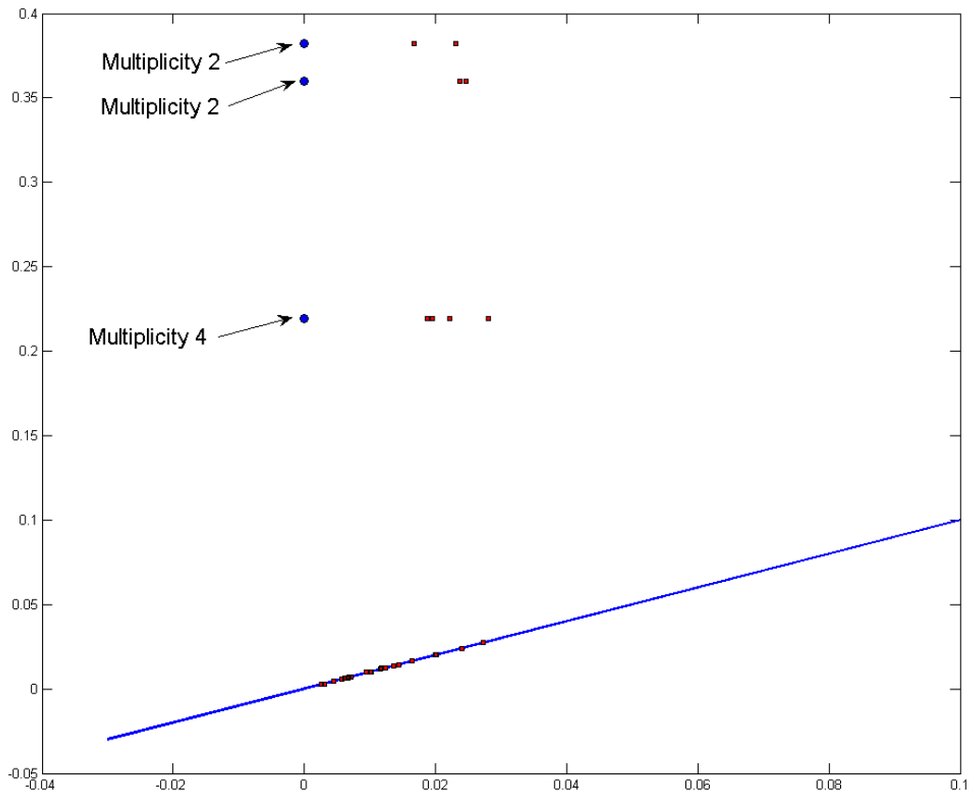
$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[d_b(D_{\mathcal{F}}(X_\mu), \hat{D}_n)\right] \geq \frac{1}{8} d_b(D_{\mathcal{F}}(X_{\mu_0}), D_{\mathcal{F}}(X_{\mu_1}))(1 - \mathrm{TV}(\mu_0, \mu_1))^{2n}$$

with $\mu_0 = \delta_x$ and $\mu_1 = \frac{1}{n}\delta_{x_n} + (1 - \frac{1}{n})\delta_x$.

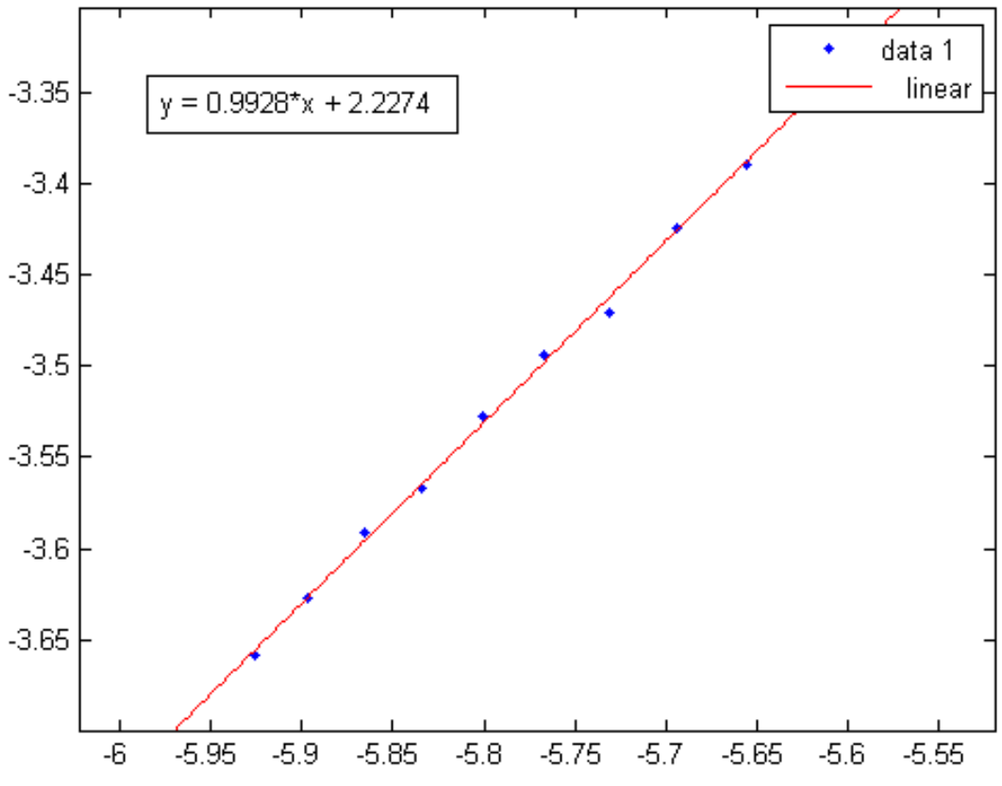
Numerical illustrations

[*Convergence rates for persistence diagram estimation in Topological Data Analysis*, Chazal, Glisse, Labruère, Michel ICML, 2014]



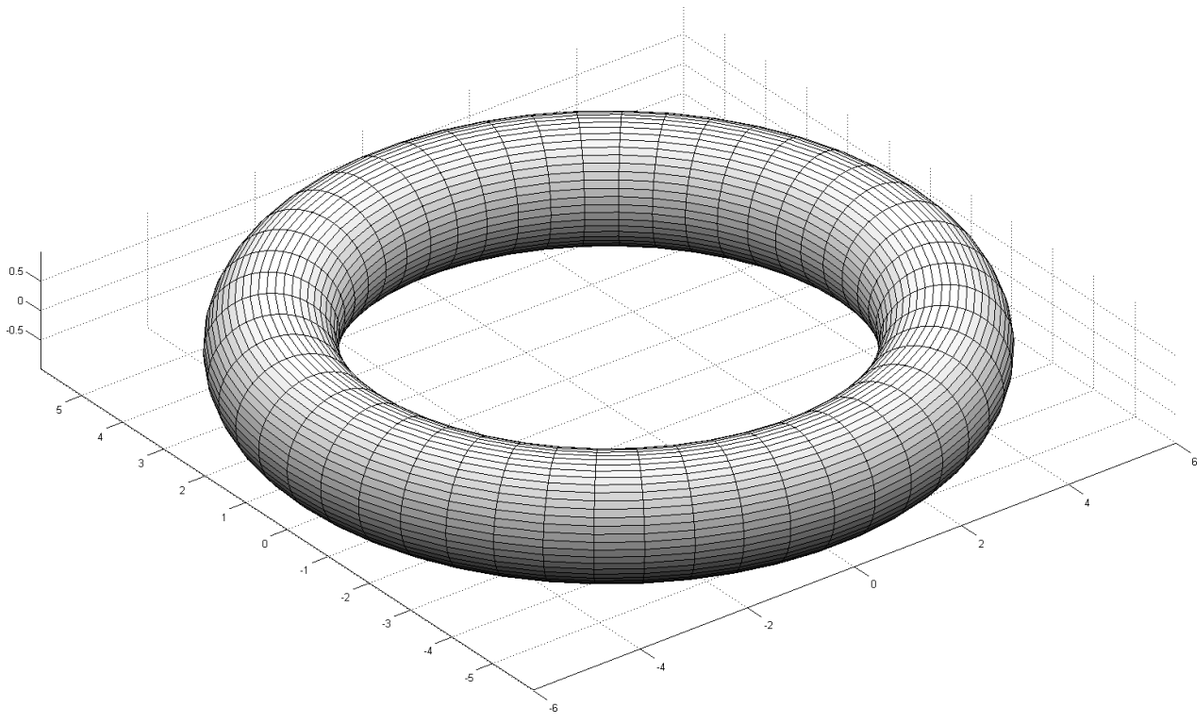


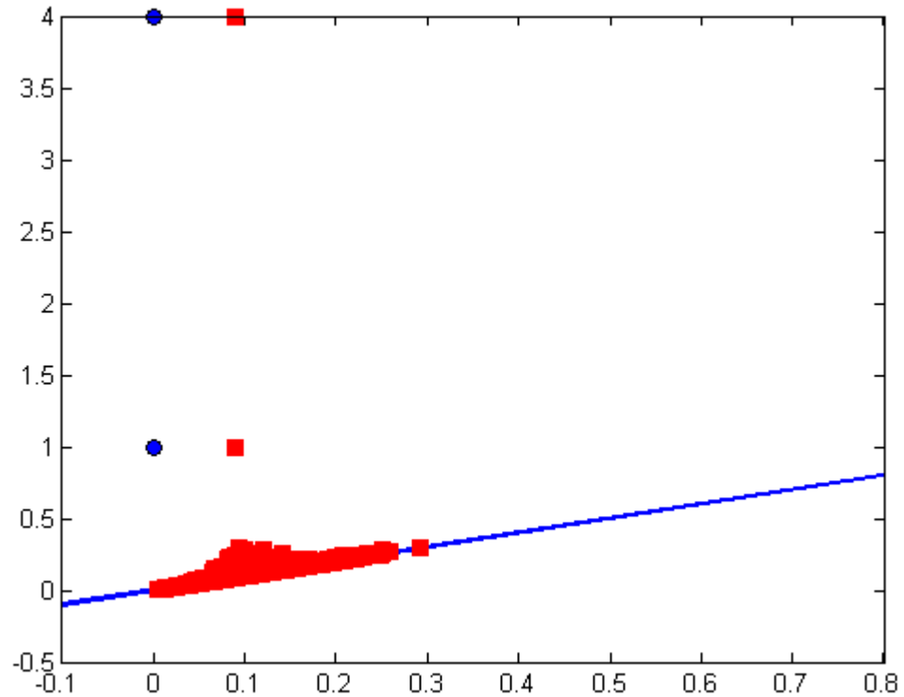
- μ: unif. measure on Lissajous curve X_μ.
- $\mathcal{F}$: distance to X_μ in $\mathbb{R}^2$.
- sample 300 sets of n points for various n.
- compute $\hat{\mathbb{E}}_n = \hat{\mathbb{E}}[d_b(D_{\mathcal{F}}(X_\mu), D_{\mathcal{F}_n}(\hat{X}_n))]$.
- plot $\log(\hat{\mathbb{E}}_n)$ as a function of $\log(\log(n)/n)$.



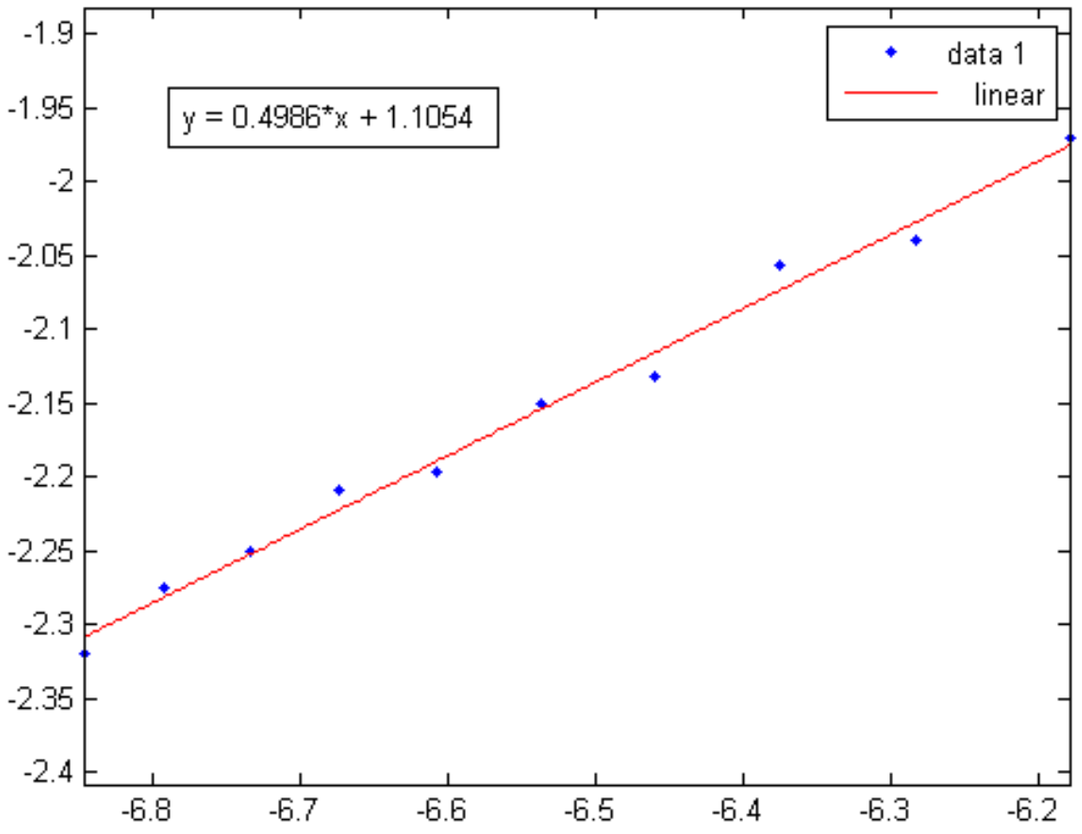
Numerical illustrations

[*Convergence rates for persistence diagram estimation in Topological Data Analysis*, Chazal, Glisse, Labruère, Michel ICML, 2014]





- μ: unif. measure on a torus X_μ.
- $\mathcal{F}$: distance to X_μ in $\mathbb{R}^3$.
- sample 300 sets of n points for various n.
- compute $\hat{\mathbb{E}}_n = \hat{\mathbb{E}}[d_b(D_{\mathcal{F}}(X_\mu), D_{\mathcal{F}_n}(\hat{X}_n))]$.
- plot $\log(\hat{\mathbb{E}}_n)$ as a function of $\log(\log(n)/n)$.



Limitations

Thm: If X and Y are pre-compact metric spaces, then

$$d_b(D_{\mathrm{Rips}}(X), D_{\mathrm{Rips}}(Y)) \leq d_{GH}(X, Y).$$

1. Vietoris-Rips (or Čech, witness) filtrations become quickly prohibitively large as the size of the data increases: $O(|X|^d)$, making the practical computation of persistence almost impossible.

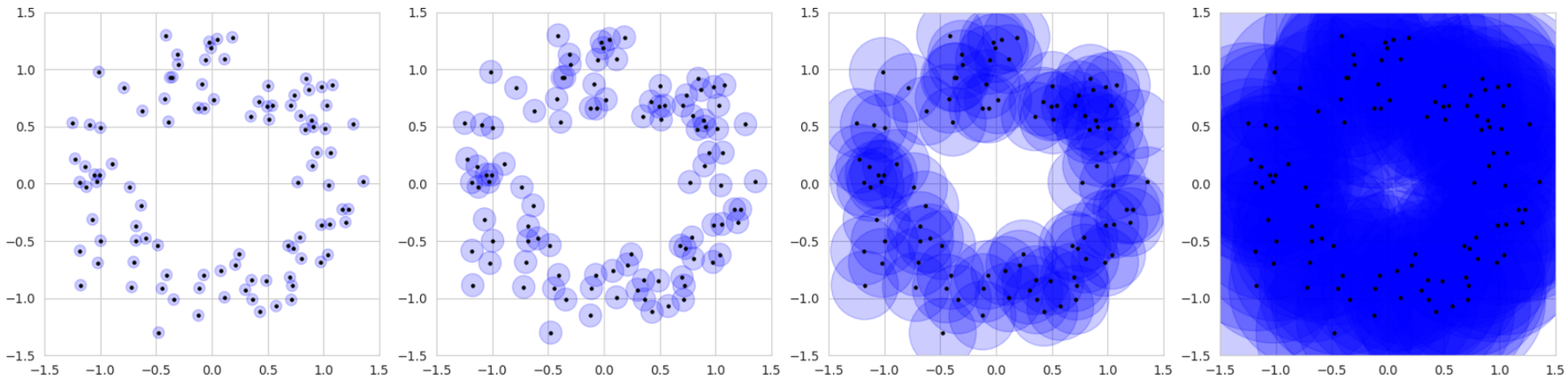
Limitations

Thm: If X and Y are pre-compact metric spaces, then

$$d_b(D_{\mathrm{Rips}}(X), D_{\mathrm{Rips}}(Y)) \leq d_{GH}(X, Y).$$

1. Vietoris-Rips (or Čech, witness) filtrations become quickly prohibitively large as the size of the data increases: $O(|X|^d)$, making the practical computation of persistence almost impossible.

2. Persistence diagrams of Vietoris-Rips (as well as Čech, witness,..) filtrations and Gromov-Hausdorff distance are very sensitive to noise and outliers.



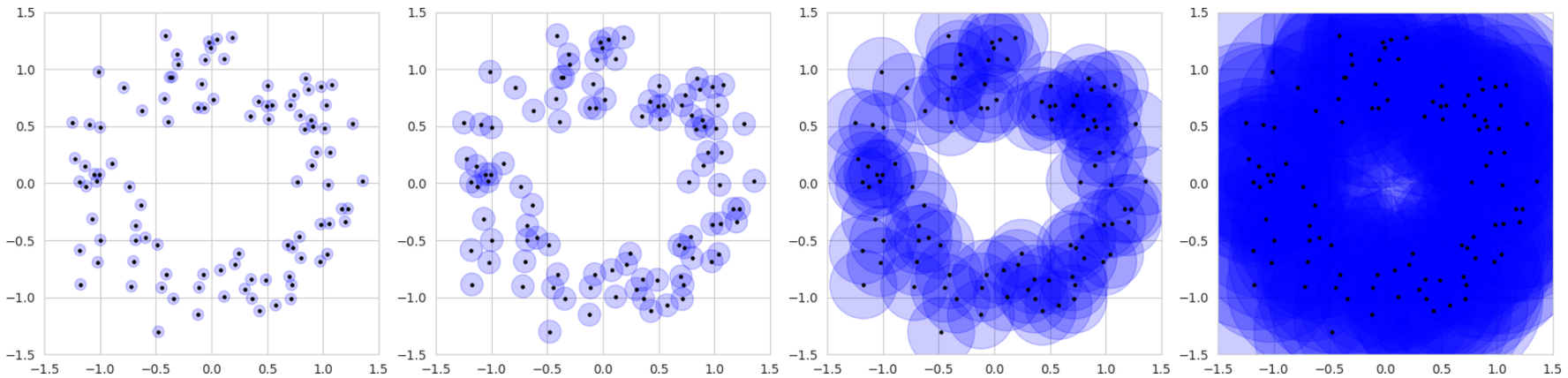
Limitations

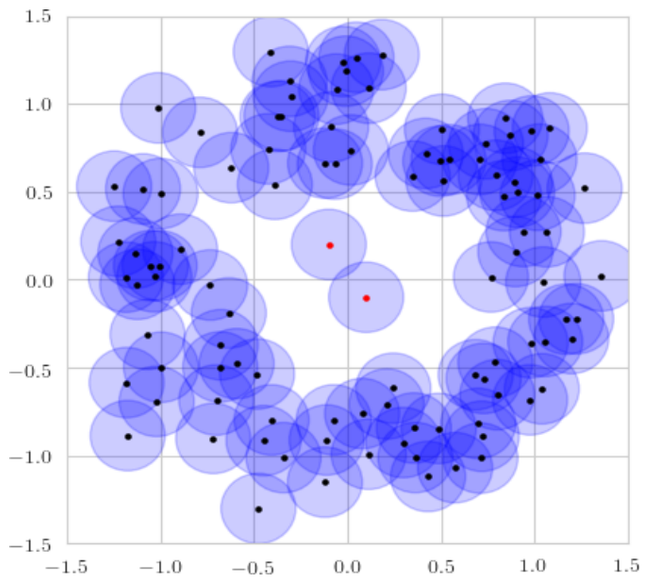
Thm: If X and Y are pre-compact metric spaces, then

$$d_b(D_{\mathrm{Rips}}(X), D_{\mathrm{Rips}}(Y)) \leq d_{GH}(X, Y).$$

1. Vietoris-Rips (or Čech, witness) filtrations become quickly prohibitively large as the size of the data increases: $O(|X|^d)$, making the practical computation of persistence almost impossible.

2. Persistence diagrams of Vietoris-Rips (as well as Čech, witness,..) filtrations and Gromov-Hausdorff distance are very sensitive to noise and outliers.





The Distance To Measure (DTM)

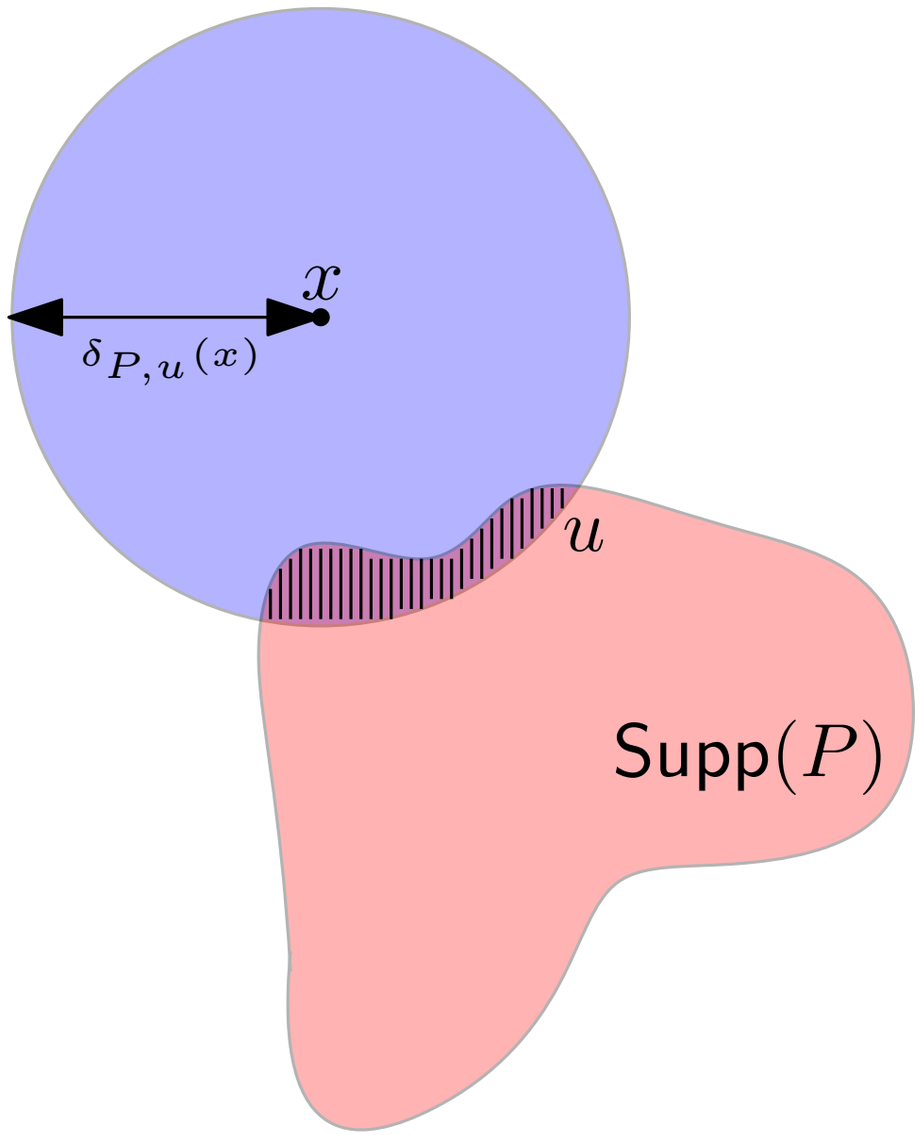
[*Geometric inference for probability measures*, Chazal, Cohen-Steiner, Mérigot, Found. Comput. Math., 2011]

The Distance To Measure (DTM)

[*Geometric inference for probability measures*, Chazal, Cohen-Steiner, Mérigot, Found. Comput. Math., 2011]

Preliminary distance function to a measure P: let $u \in]0,1[$ be a positive mass, and P a probability measure on $\mathbb{R}^d$:

$$\delta_{P,u}(x) = \inf\{r > 0 : P(B(x,r)) \geq u\}$$

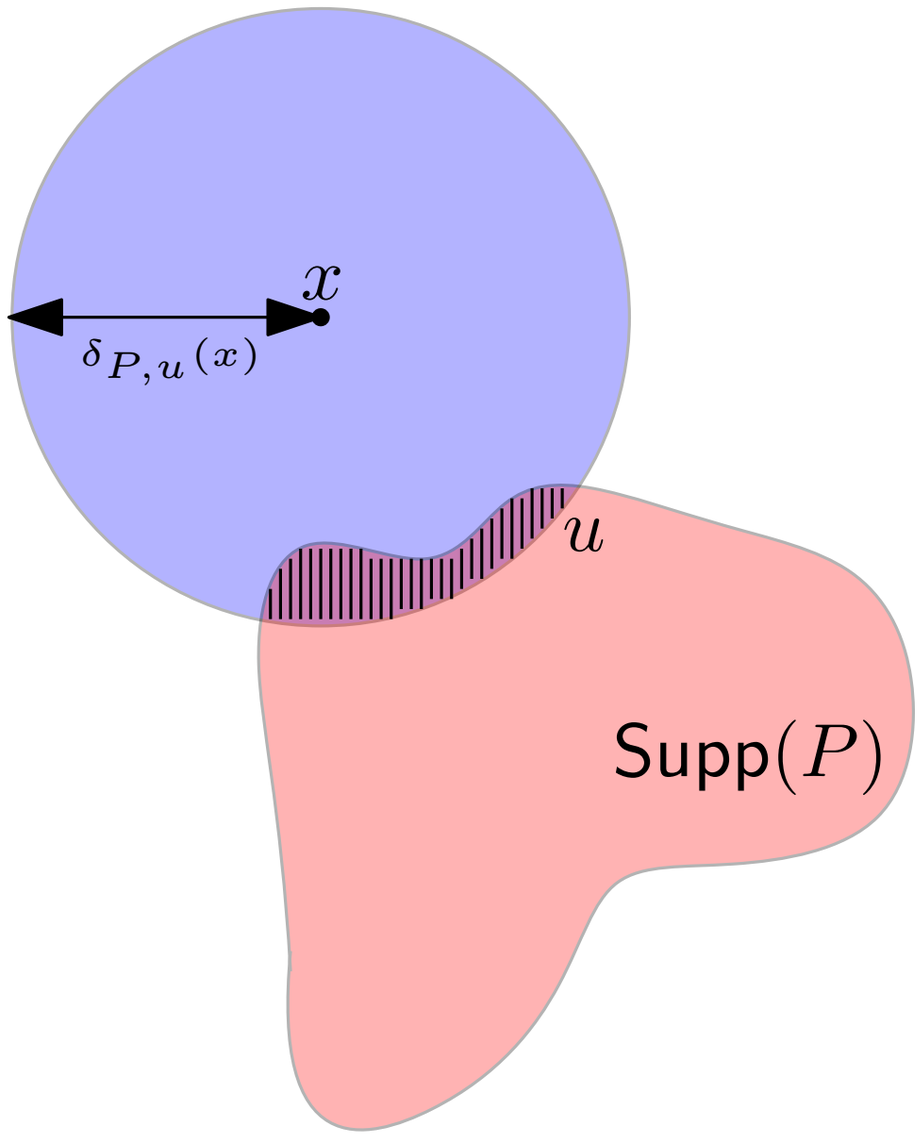


The Distance To Measure (DTM)

[*Geometric inference for probability measures*, Chazal, Cohen-Steiner, Mérigot, Found. Comput. Math., 2011]

Preliminary distance function to a measure P: let $u \in]0,1[$ be a positive mass, and P a probability measure on $\mathbb{R}^d$:

$$\delta_{P,u}(x) = \inf\{r > 0 : P(B(x,r)) \geq u\}$$



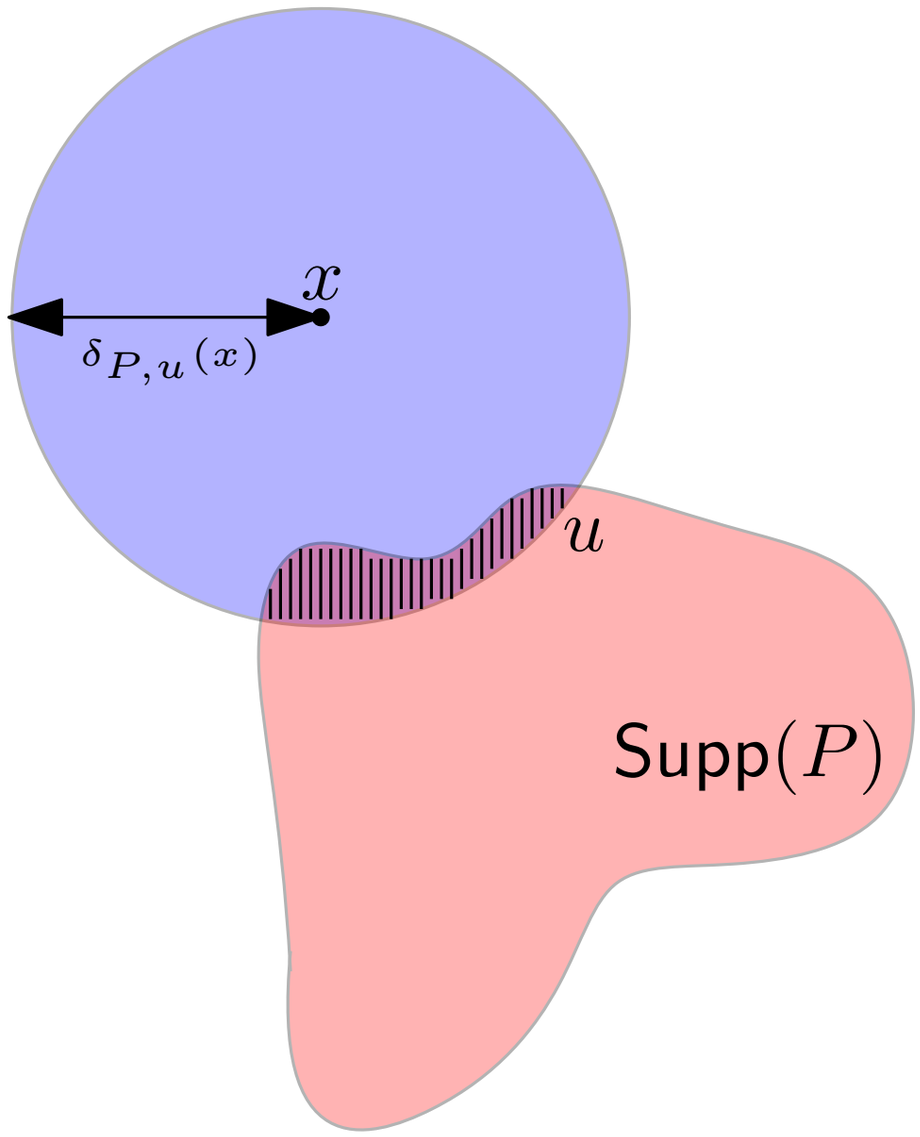
$\delta_{P,u}$ is the smallest distance needed to capture a mass of at least u.

The Distance To Measure (DTM)

[*Geometric inference for probability measures*, Chazal, Cohen-Steiner, Mérigot, Found. Comput. Math., 2011]

Preliminary distance function to a measure P: let $u \in]0,1[$ be a positive mass, and P a probability measure on $\mathbb{R}^d$:

$$\delta_{P,u}(x) = \inf\{r > 0 : P(B(x,r)) \geq u\}$$



$\delta_{P,u}$ is the smallest distance needed to capture a mass of at least u.

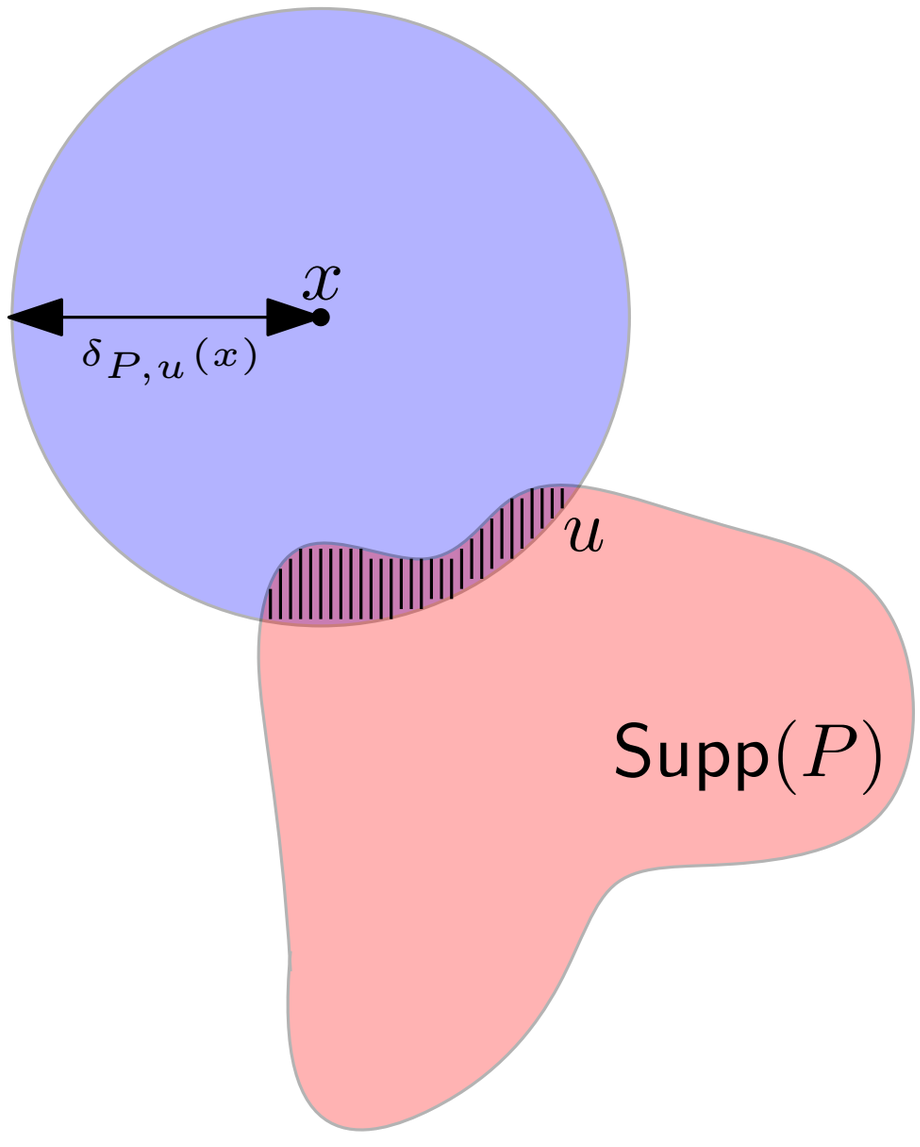
$\delta_{P,u}$ is the quantile function at u of the r.v. $\|x - X\|$ where $X \sim P$.

The Distance To Measure (DTM)

[*Geometric inference for probability measures*, Chazal, Cohen-Steiner, Mérigot, Found. Comput. Math., 2011]

Preliminary distance function to a measure P: let $u \in]0,1[$ be a positive mass, and P a probability measure on $\mathbb{R}^d$:

$$\delta_{P,u}(x) = \inf\{r > 0 : P(B(x,r)) \geq u\}$$



Def: Given a probability measure P on $\mathbb{R}^d$ and $m > 0$, the distance function to the measure P (DTM) is defined by

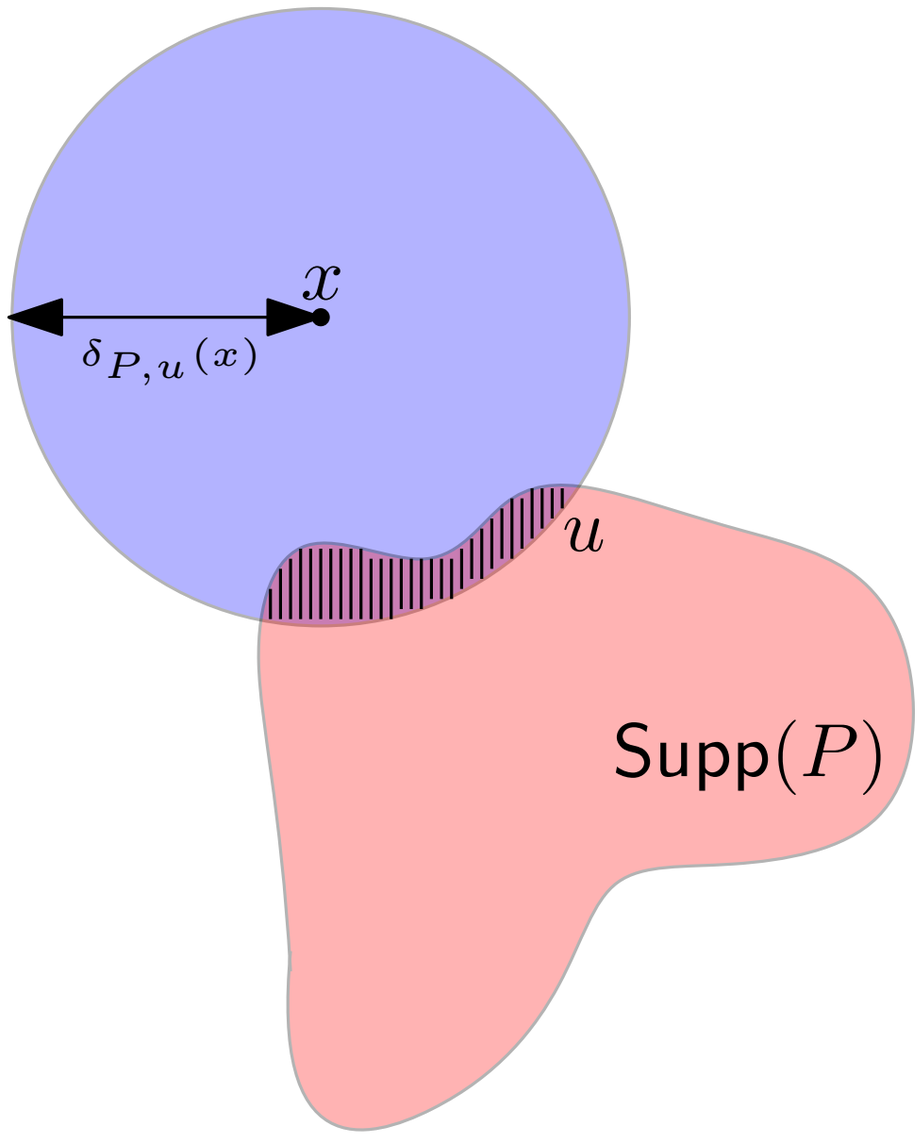
$$d_{P,m} : x \in \mathbb{R} \mapsto \left(\frac{1}{m}\int_0^m \delta_{P,u}^2(x)\mathrm{d}u\right)^{1/2}$$

The Distance To Measure (DTM)

[*Geometric inference for probability measures*, Chazal, Cohen-Steiner, Mérigot, Found. Comput. Math., 2011]

Preliminary distance function to a measure P: let $u \in]0,1[$ be a positive mass, and P a probability measure on $\mathbb{R}^d$:

$$\delta_{P,u}(x) = \inf\{r > 0 : P(B(x,r)) \geq u\}$$



Def: Given a probability measure P on $\mathbb{R}^d$ and $m > 0$, the distance function to the measure P (DTM) is defined by

$$d_{P,m} : x \in \mathbb{R} \mapsto \left(\frac{1}{m}\int_0^m \delta_{P,u}^2(x)\mathrm{d}u\right)^{1/2}$$

The DTM is robust, i.e., stable under Wasserstein perturbations:

$$\|d_{P,m} - d_{Q,m}\|_\infty \leq \frac{1}{\sqrt{m}} W_2(P,Q)$$

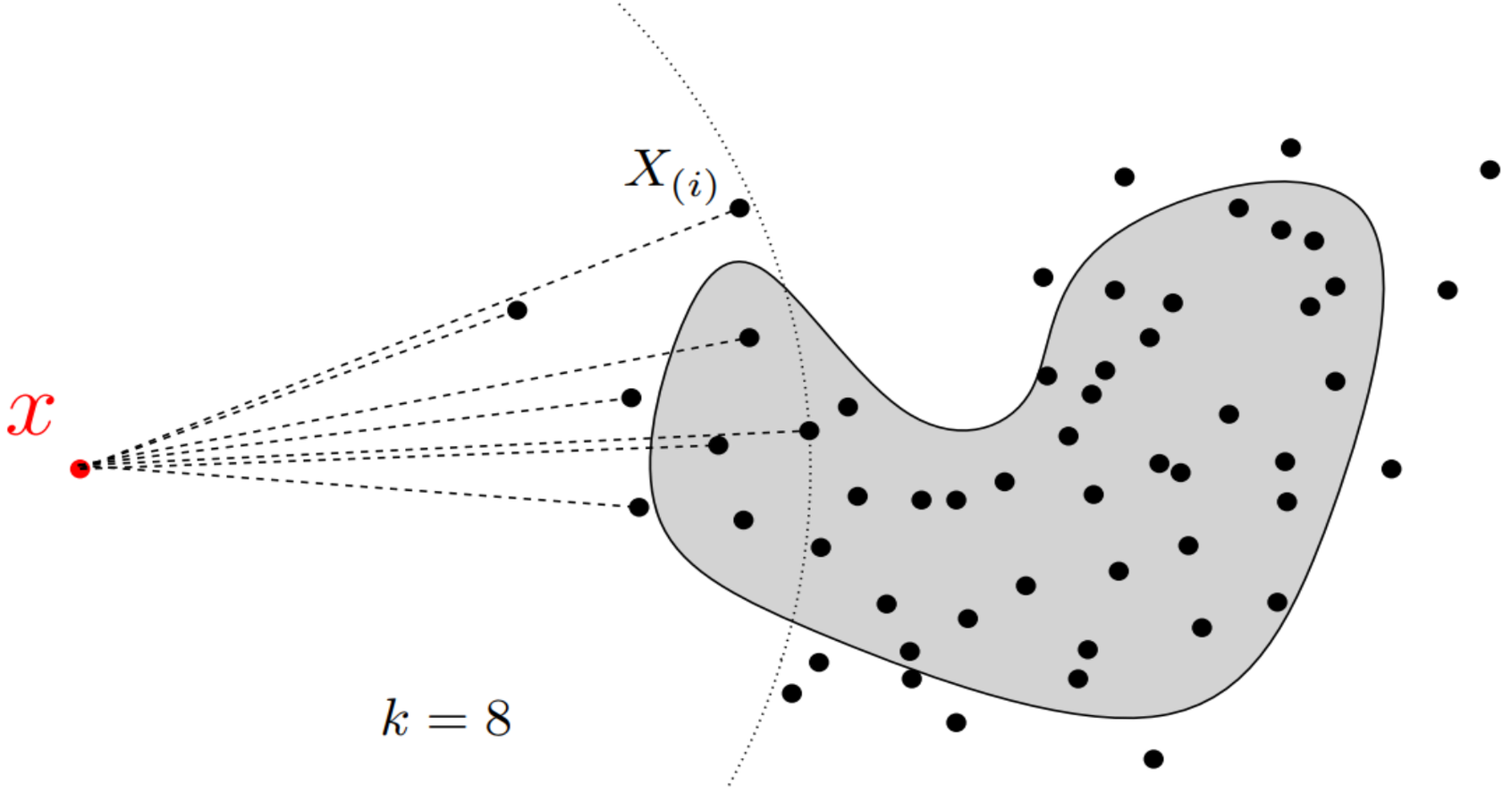
The Distance To Measure (DTM)

[*Geometric inference for probability measures*, Chazal, Cohen-Steiner, Mérigot, Found. Comput. Math., 2011]

Def: Let $X_1, \ldots, X_n$ sampled according to P and let P_n be the empirical measure. Then

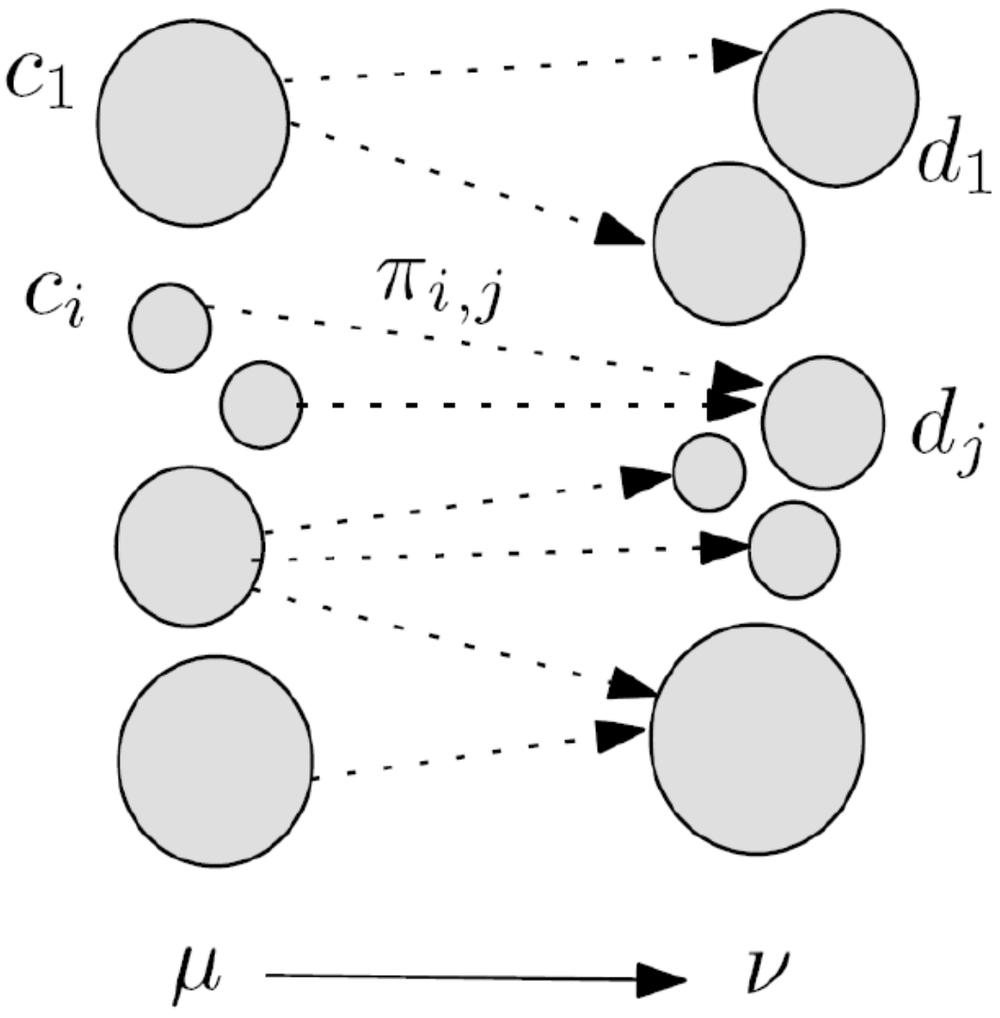
$$d_{P_n,k/n}(x) = \frac{1}{k} \sum_{i=1}^{k} \|x - X_{(i)}\|^2,$$

where $\|X_{(1)} - x\| \leq \|X_{(2)} - x\| \leq \cdots \leq \|X_{(k)} - x\| \leq \cdots \leq \|X_{(n)} - x\|$.



The Wasserstein distance

Let (X, d) be a metric space and let μ, ν be probability measures on X with finite p-moments ($p \geq 1$). The Wasserstein distance $W_p(\mu, \nu)$ quantifies the optimal cost of pushing μ onto ν, the cost of moving a small mass dx from x to y being $d(x, y)^p dx$.



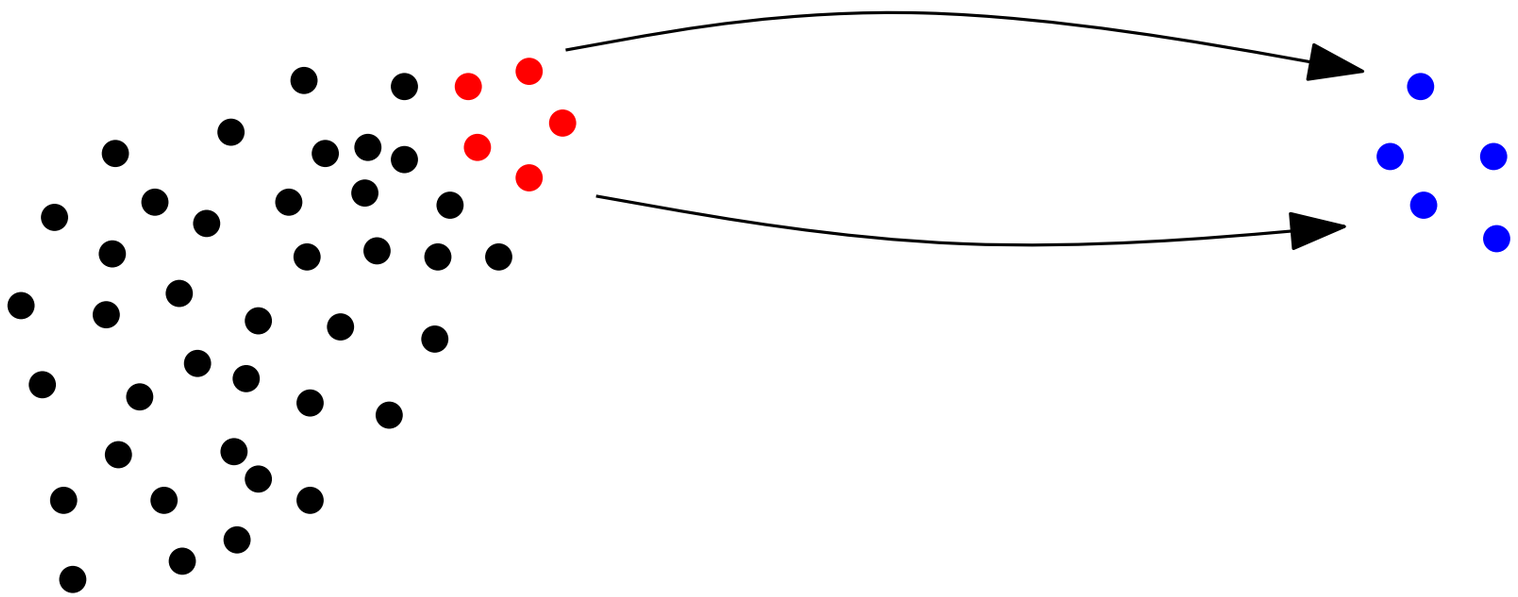
- Transport plan: Π a probability measure on $X \times X$ s.t. $\Pi(A \times \mathbb{R}^d) = \mu(A)$ and $\Pi(\mathbb{R}^d \times B) = \nu(B)$ for any borelian sets $A, B \subseteq X$.
- Cost of a transport plan:

$$C(\Pi) = \left(\int_{X \times X} d(x, y)^p d\Pi(x, y) \right)^{\frac{1}{p}}$$

- $W_p(\mu, \nu) = \inf_{\Pi} C(\Pi)$.

The Wasserstein distance

The Wasserstein distance is much less sensitive to outliers!



Ex: If $P = \{p_1, \ldots, p_n\}$ is a point cloud, and $P' = \{p_1, \ldots, p_{n-k}\} \cup \{o_1, \ldots, o_k\}$ with $d(o_i, P) = R$, then $d_H(P, P') \geq R$ but

$$W_2(\mu_P, \mu_{P'}) \leq \sqrt{\frac{k}{n}}(R + \operatorname{diam}(P)).$$

DTM-based filtrations

[*DTM-based filtrations*, Anai et al., Symp. Comp. Geom., 2019]

Def: Let V be a point cloud (in a metric space). The DTM-based complex $W(V)$ is the filtered simplicial complex indexed by $\mathbb{R}$ whose vertex set is V and whose other simplices are defined with

$$\sigma = [p_0, p_1 \ldots, p_k] \in W(V, \alpha) \iff \cap_{i=0}^{k} B(p_i, r_{p_i}(\alpha)) \neq \emptyset$$

where $r_p(\alpha) = 0$ if $\alpha \leq d_{P_n,k/n}(p)$ and $|\alpha^q - d_{P_n,k/n}(p)^q|^{1/q}$ otherwise.

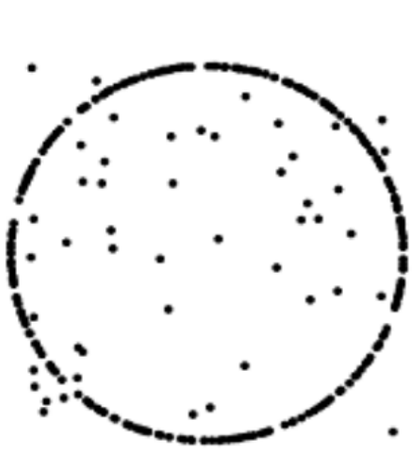
DTM-based filtrations

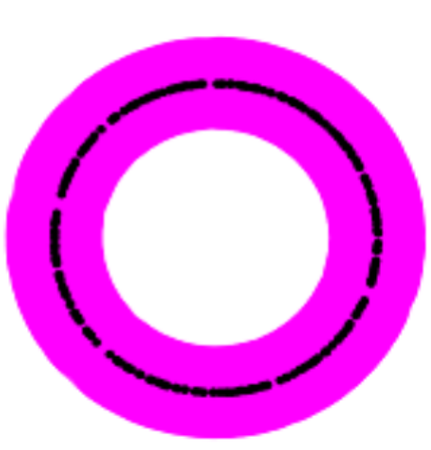
[*DTM-based filtrations*, Anai et al., Symp. Comp. Geom., 2019]

Def: Let V be a point cloud (in a metric space). The DTM-based complex $W(V)$ is the filtered simplicial complex indexed by $\mathbb{R}$ whose vertex set is V and whose other simplices are defined with

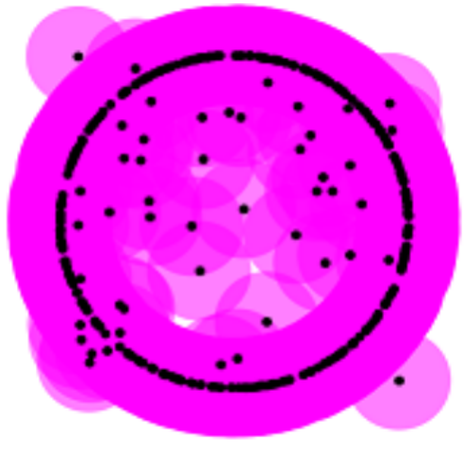
$$\sigma = [p_0, p_1 \ldots, p_k] \in W(V, \alpha) \iff \cap_{i=0}^{k} B(p_i, r_{p_i}(\alpha)) \neq \emptyset$$

where $r_p(\alpha) = 0$ if $\alpha \leq d_{P_n,k/n}(p)$ and $|\alpha^q - d_{P_n,k/n}(p)^q|^{1/q}$ otherwise.

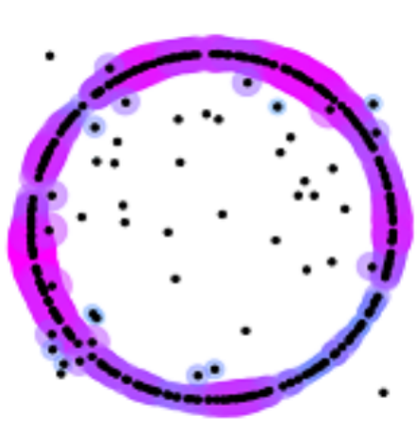




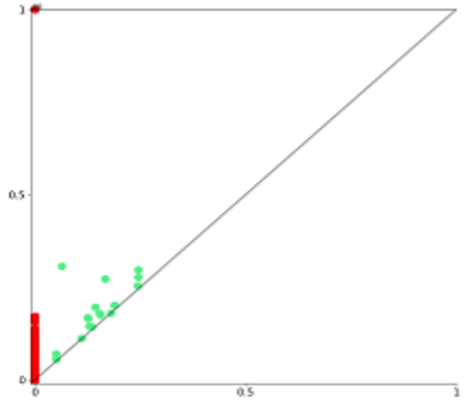
Rips

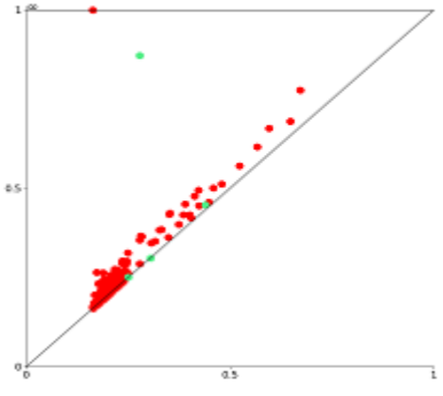


Rips



DTM-based





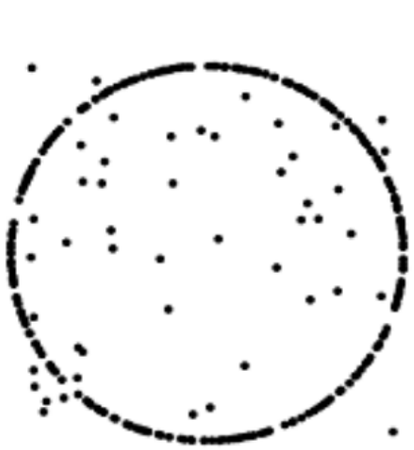
DTM-based filtrations

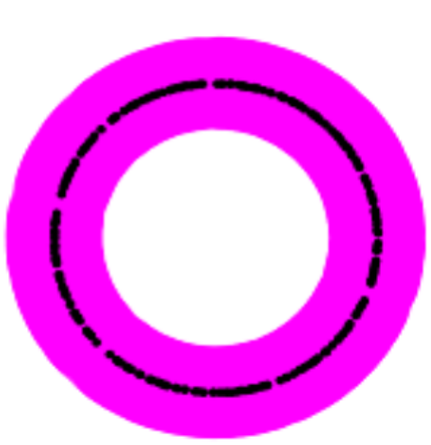
[*DTM-based filtrations*, Anai et al., Symp. Comp. Geom., 2019]

Def: Let V be a point cloud (in a metric space). The DTM-based complex $W(V)$ is the filtered simplicial complex indexed by $\mathbb{R}$ whose vertex set is V and whose other simplices are defined with

$$\sigma = [p_0, p_1 \ldots, p_k] \in W(V, \alpha) \iff \cap_{i=0}^{k} B(p_i, r_{p_i}(\alpha)) \neq \emptyset$$

where $r_p(\alpha) = 0$ if $\alpha \leq d_{P_n,k/n}(p)$ and $|\alpha^q - d_{P_n,k/n}(p)^q|^{1/q}$ otherwise.



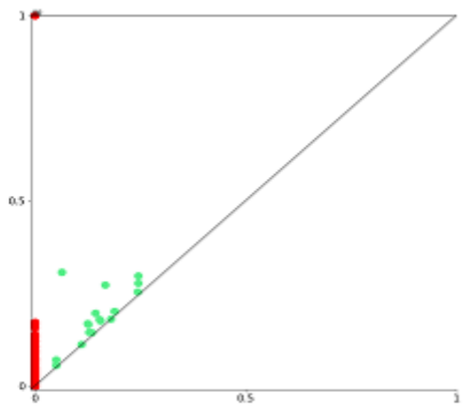


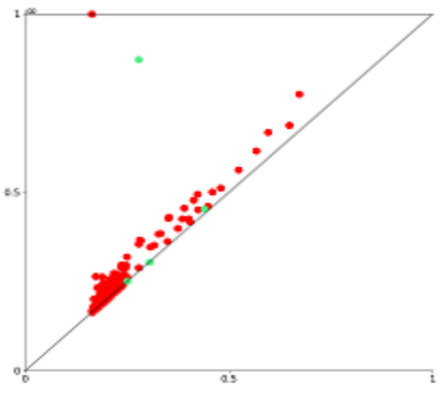
Rips

Rips

DTM-based

Thm: One has: $d_b(D_W(X), D_W(Y)) \leq \sqrt{\frac{n}{k}}(W_2(\Omega, X) + W_2(\Omega, Y)) + 2C_{\Omega,n,k}$.





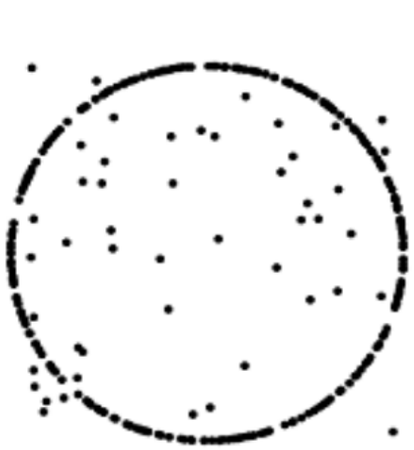
DTM-based filtrations

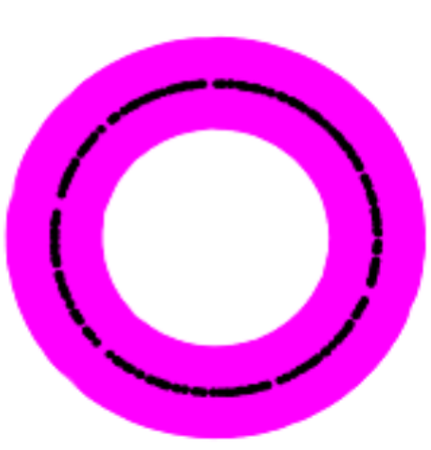
[*DTM-based filtrations*, Anai et al., Symp. Comp. Geom., 2019]

Def: Let V be a point cloud (in a metric space). The DTM-based complex $W(V)$ is the filtered simplicial complex indexed by $\mathbb{R}$ whose vertex set is V and whose other simplices are defined with

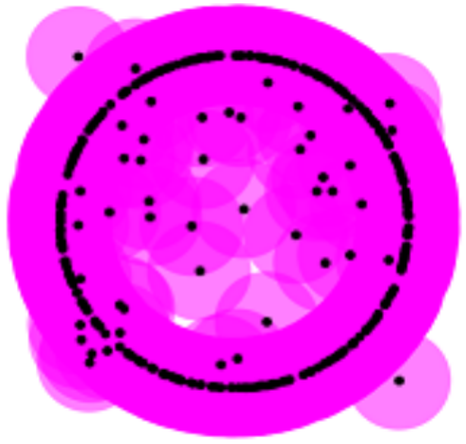
$$\sigma = [p_0, p_1 \ldots, p_k] \in W(V, \alpha) \iff \cap_{i=0}^{k} B(p_i, r_{p_i}(\alpha)) \neq \emptyset$$

where $r_p(\alpha) = 0$ if $\alpha \leq d_{P_n,k/n}(p)$ and $|\alpha^q - d_{P_n,k/n}(p)^q|^{1/q}$ otherwise.

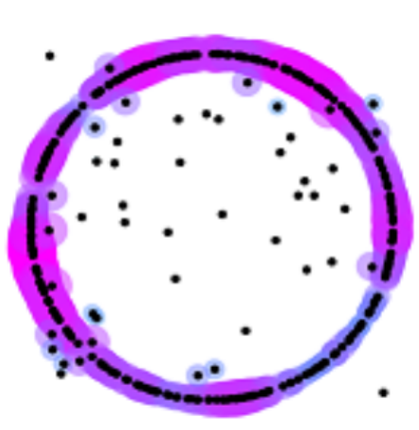




Rips



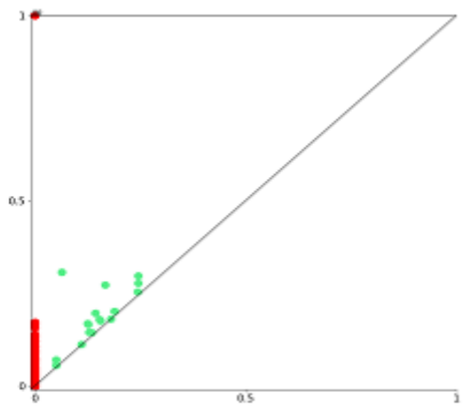
Rips

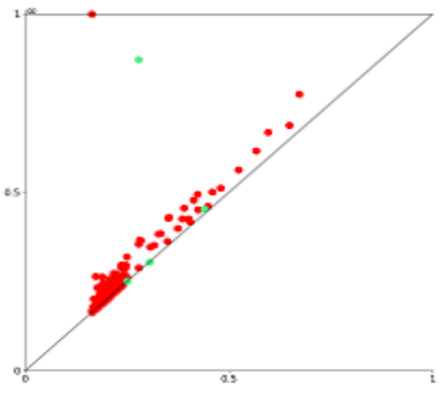


DTM-based

Thm: One has: $d_b(D_W(X), D_W(Y)) \leq \sqrt{\frac{n}{k}}(W_2(\Omega, X) + W_2(\Omega, Y)) + 2C_{\Omega,n,k}$.

"Clean subset" without outliers





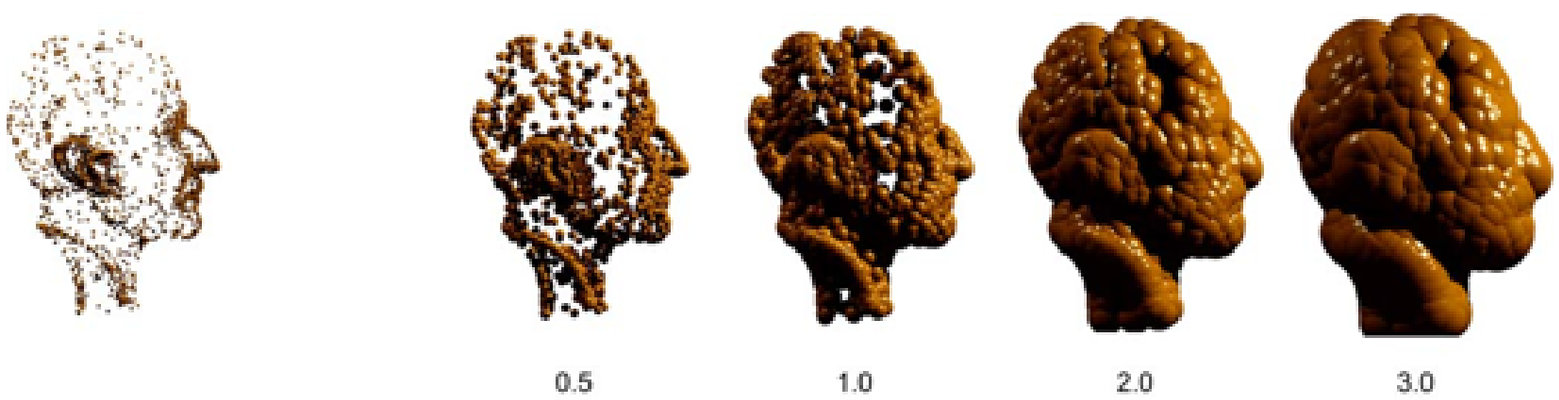
I. Topological Inference on Persistence Diagrams for Point Clouds

II. Topological Vectorizations for Inference and Machine Learning

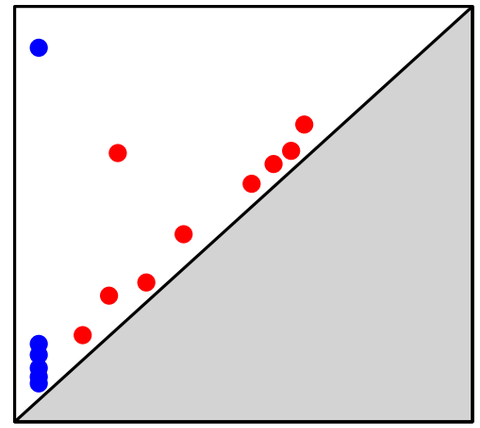
(III. A Gaussian Kernel for Persistence Diagrams)

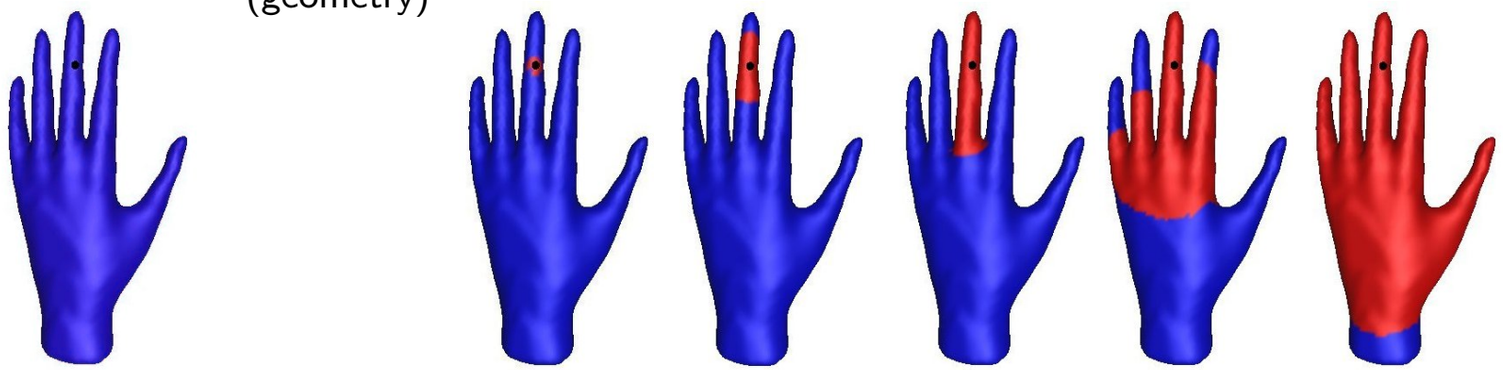
IV. Deep Learning for Persistence Diagrams

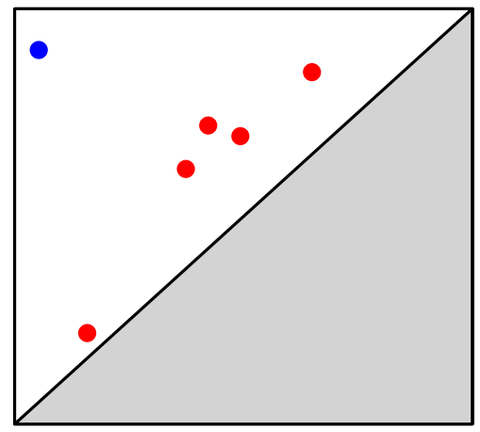
Persistence diagrams as descriptors for data



dataset $\Longrightarrow$ (geometry) 1-parameter family of spaces $\Longrightarrow$ (algebraic topology) persistence diagram







Pros:

- strong invariance and stability:

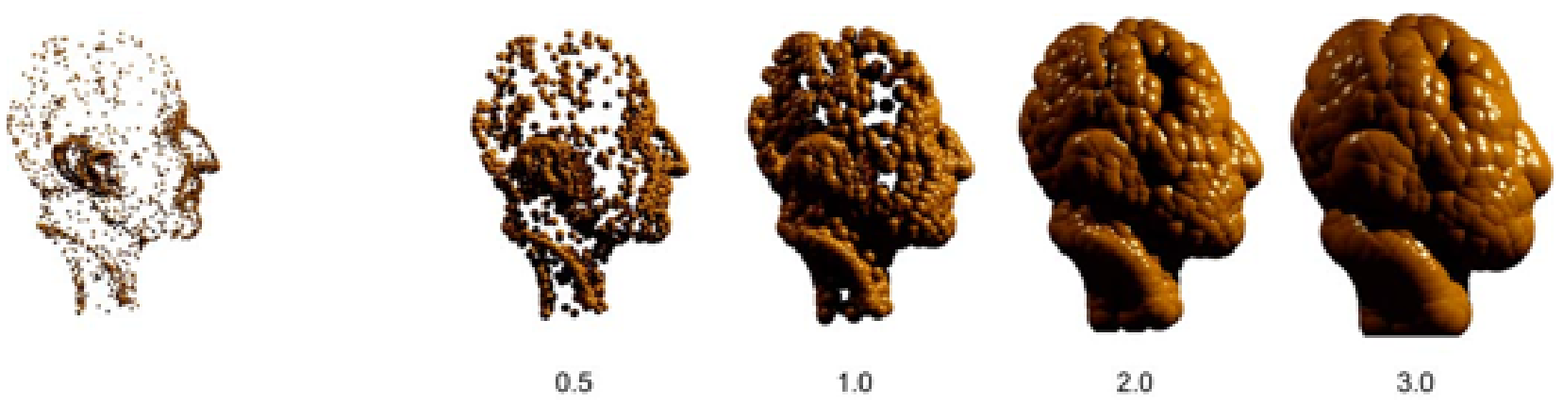
$$d_b(D_{\mathrm{Rips}}(X), D_{\mathrm{Rips}}(Y)) \leq d_{GH}(X, Y)$$

- information of a different nature
- flexible and versatile

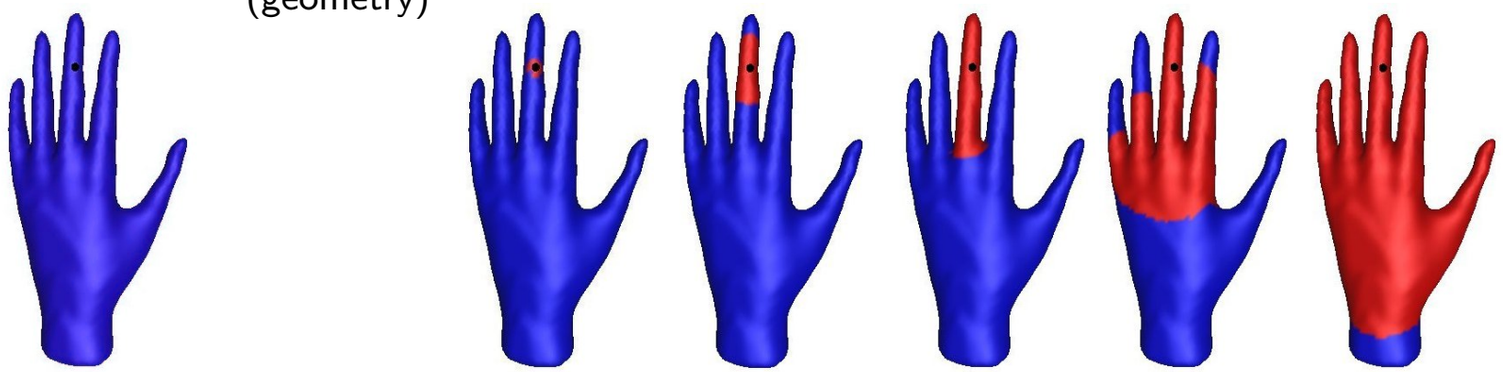
Cons:

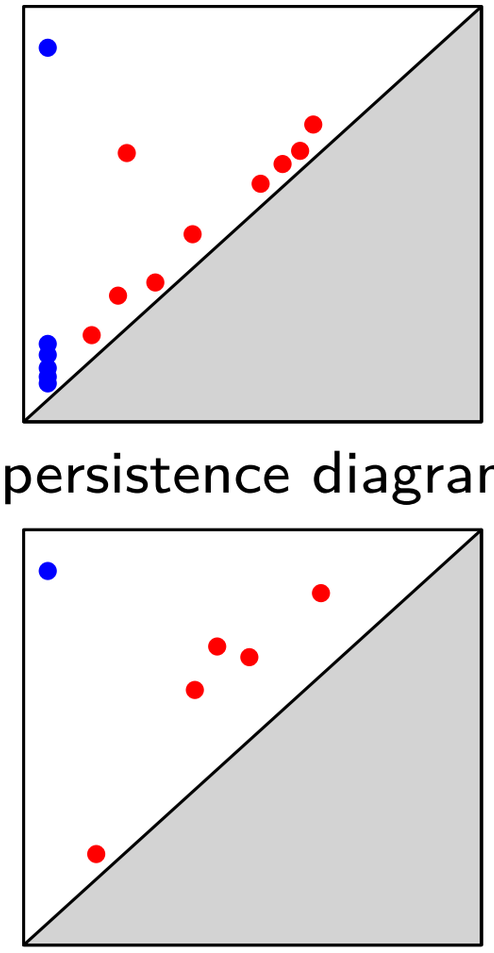
- slow to compare
- space of diagrams is not linear
- positive intrinsic curvature

Persistence diagrams as descriptors for data

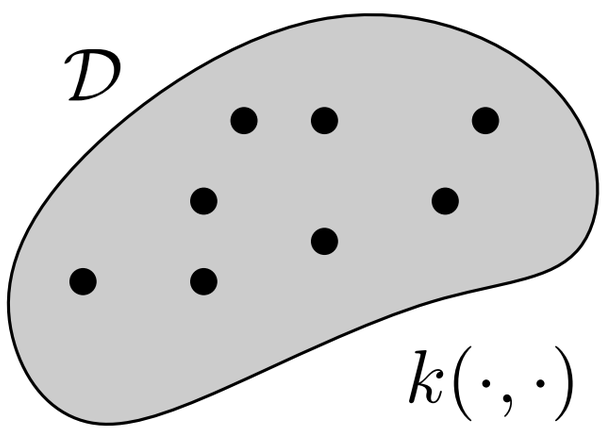


dataset ⟹ (geometry) 1-parameter family of spaces ⟹ (algebraic topology) persistence diagram



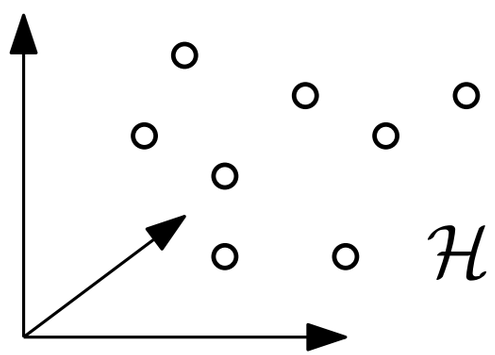


A solution: map diagrams to Hilbert space and use kernel trick



Φ

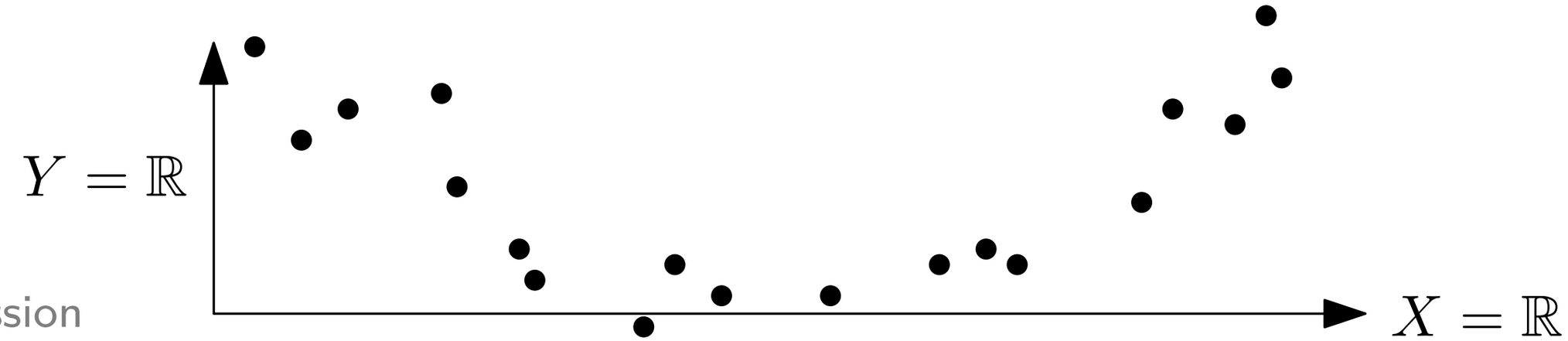
(ideally: Φ quasi-isometry)



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

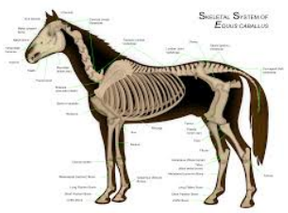
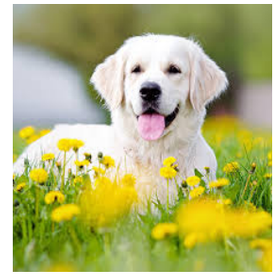
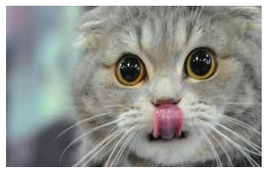
Supervised Machine Learning

Input: n observations + responses $(x_1, y_1), \ldots, (x_n, y_n) \in X \times Y$



regression

classification

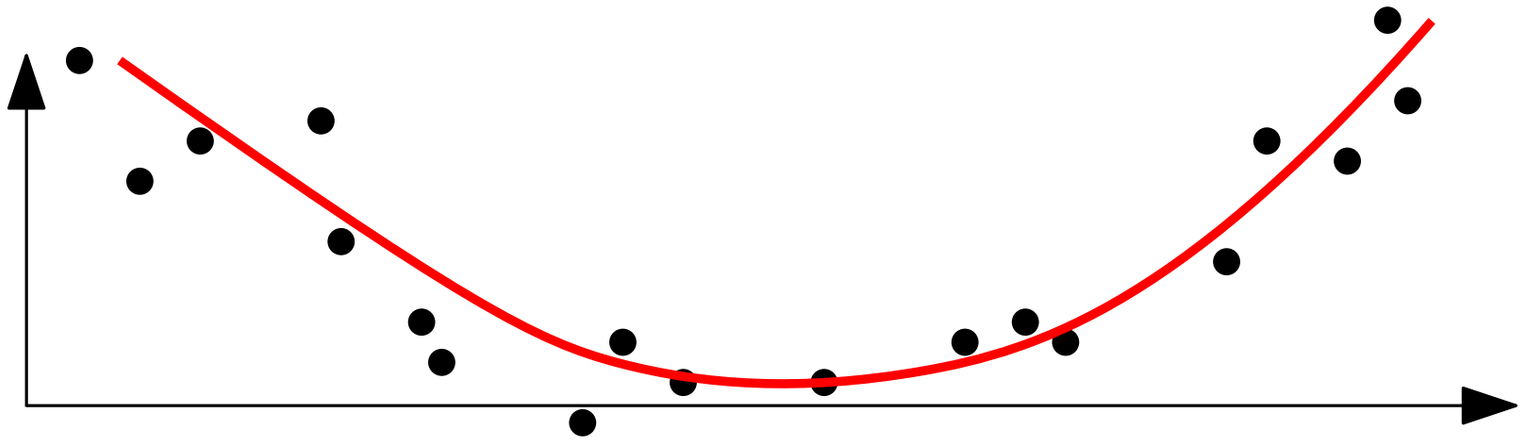




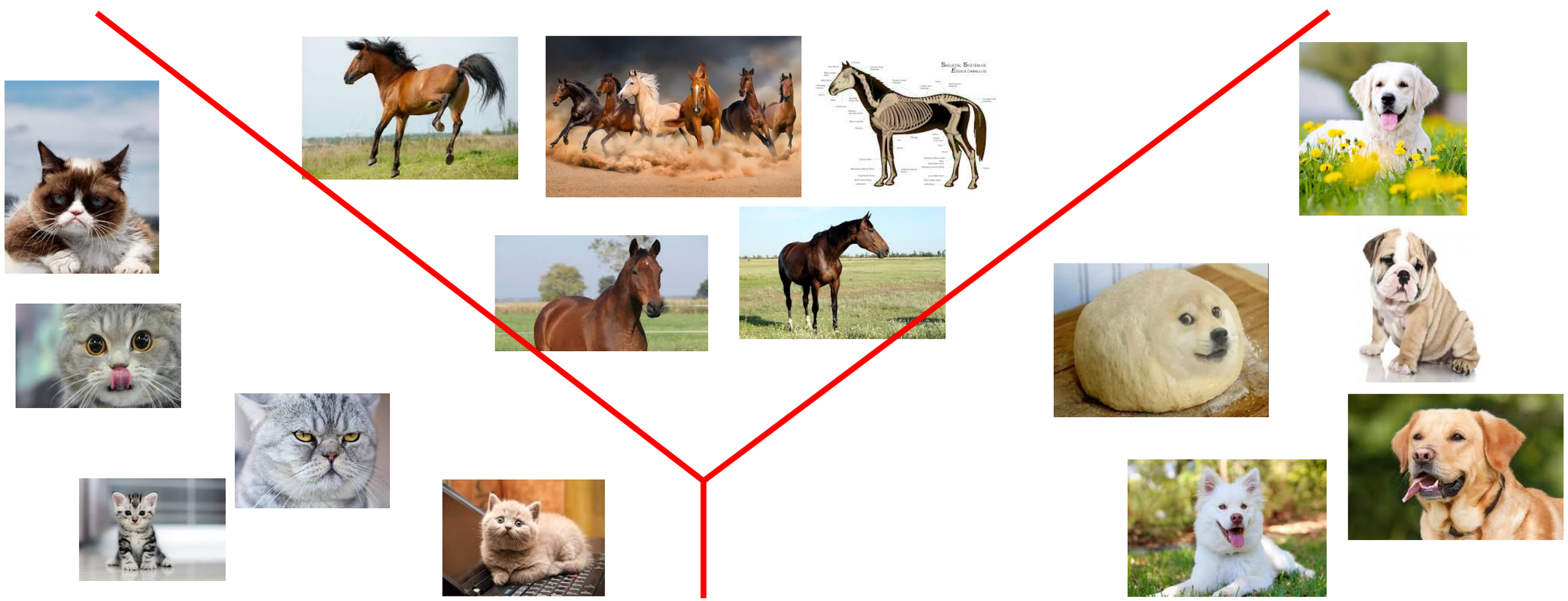
$X =$ images,
$Y = \{$cat, dog, horse$\}$

Supervised Machine Learning

Input: n observations + responses $(x_1, y_1), \ldots, (x_n, y_n) \in X \times Y$

Goal: build a predictor $f : X \to Y$ from $(x_1, y_1), \ldots, (x_n, y_n)$





Empirical Risk Minimization

Optimization problem (supervised regression / classification):

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(f)$$

$\mathcal{F}$ is the **class of predictors**

$L : X \times X \to \mathbb{R}$ is the **loss function**

$\Omega : \mathcal{F} \to \mathbb{R}$ is the **regularizer**

Empirical Risk Minimization

Optimization problem (supervised regression / classification):

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(f)$$

$\mathcal{F}$ is the **class of predictors**

$L : X \times X \to \mathbb{R}$ is the **loss function**

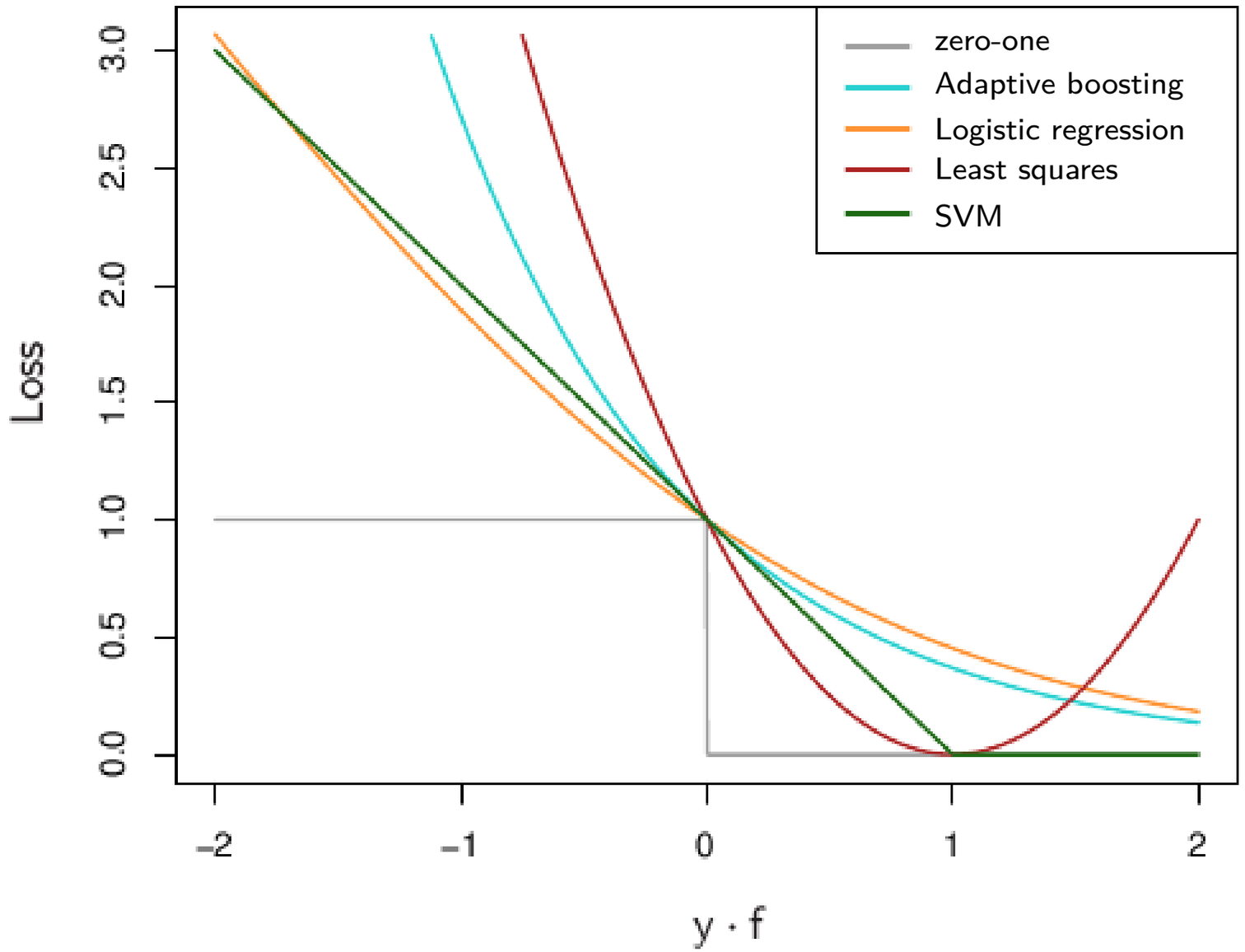
$\Omega : \mathcal{F} \to \mathbb{R}$ is the **regularizer**

$L(y_i, f(x_i))$	Name	
$\mathbb{1}_{y_i \neq f(x_i)}$	zero-one	$\to$ Bayes
$\max\{0, 1 - y_i f(x_i)\}$	hinge	$\to$ Support Vector Machines
$\exp(-y_i f(x_i))$	exponential	$\to$ Adaptive boosting
$\log(1 + \exp(-y_i f(x_i)))$	logistic	$\to$ Logistic regression
$(y_i - f(x_i))^2$	squared	$\to$ Least squares

Empirical Risk Minimization

Optimization problem (supervised regression / classification):

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(f)$$



Empirical Risk Minimization

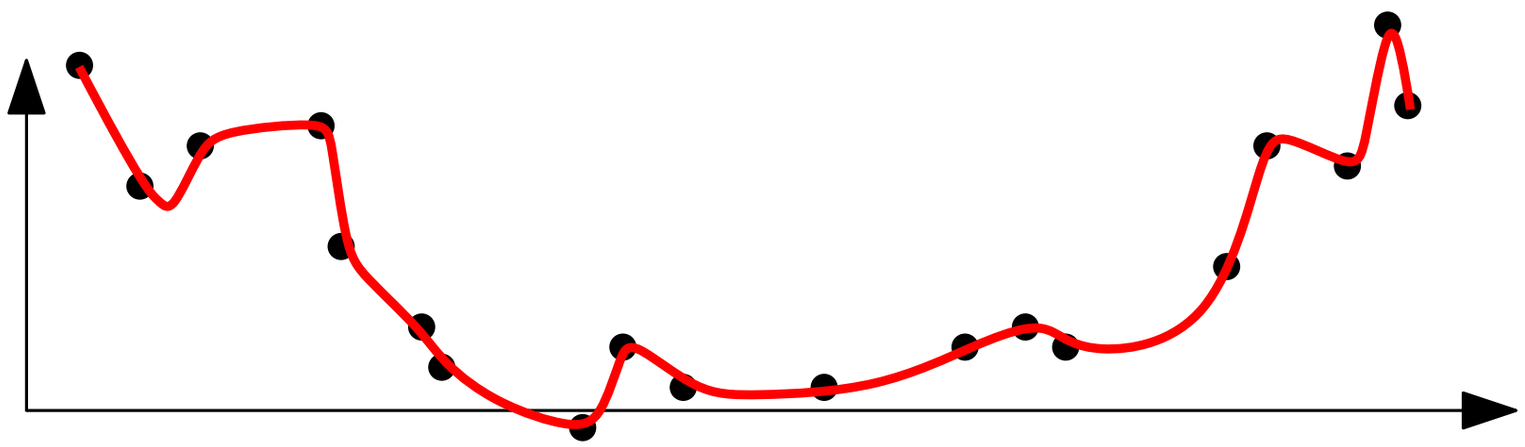
Optimization problem (supervised regression / classification):

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(f)$$

$\mathcal{F}$ is the **class of predictors**

$L : X \times X \to \mathbb{R}$ is the **loss function**

$\Omega : \mathcal{F} \to \mathbb{R}$ is the **regularizer**



Regularizer helps preventing overfitting.

Empirical Risk Minimization

Optimization problem (supervised regression / classification):

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(f)$$

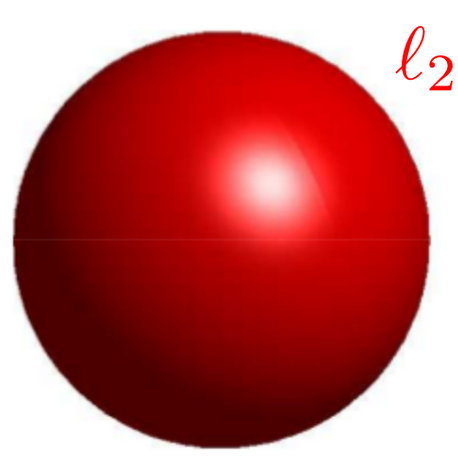
$\mathcal{F}$ is the **class of predictors**

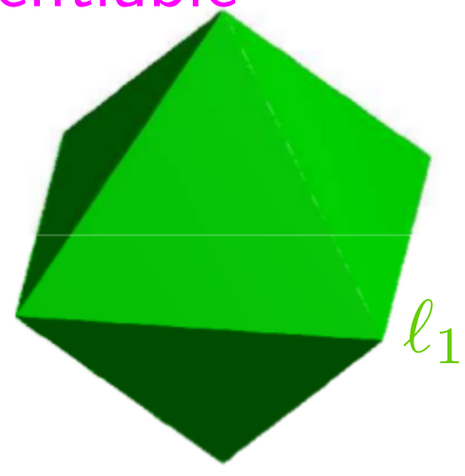
$L : X \times X \to \mathbb{R}$ is the **loss function**

$\Omega : \mathcal{F} \to \mathbb{R}$ is the **regularizer**

$$\mathcal{F} = \{f_w : w \in \mathbb{R}^d\}$$

$\Omega(w)$	Name
$\|w\|_2^2$	ℓ_2 (Tikhonov) $\to$ differentiable
$\|w\|_1$	ℓ_1 (LASSO) $\to$ sparse
$\alpha\|w\|_2^2 + (1-\alpha)\|w\|_1$	elastic net





Empirical Risk Minimization

Optimization problem (supervised regression / classification):

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(f)$$

$\mathcal{F}$ is the **class of predictors**

$L : X \times X \to \mathbb{R}$ is the **loss function**

$\Omega : \mathcal{F} \to \mathbb{R}$ is the **regularizer**

Complexity of the minimization grows with the one of $\mathcal{F}$.

However, it is easy to control when $\mathcal{F}$ is a Reproducing Kernel Hilbert Space.

Reproducing Kernel Hilbert Space

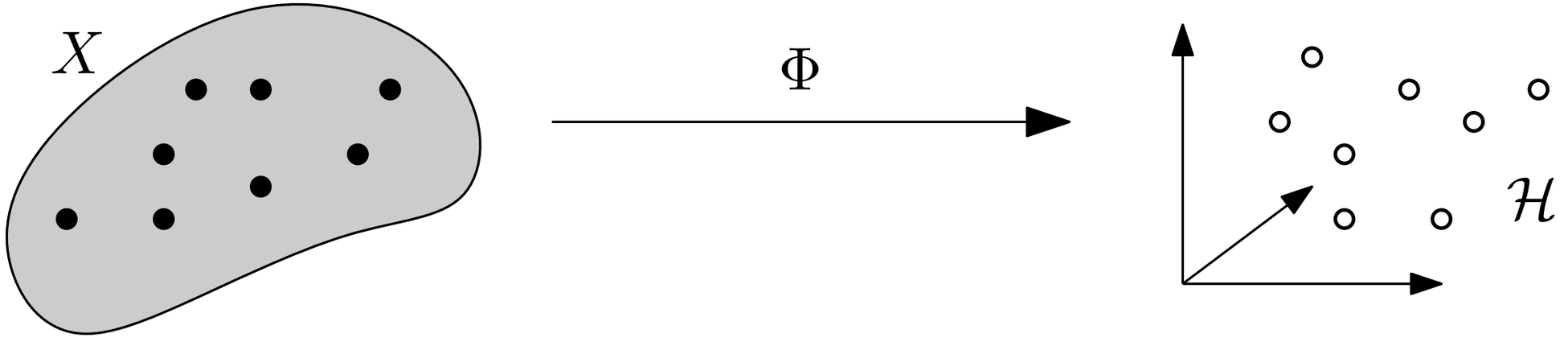
Def: Let $\mathcal{H} \subset \mathbb{R}^X$ Hilbert, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
Then, $\mathcal{H}$ is a **RKHS** on X if $\exists \Phi : X \to \mathcal{H}$ s.t.:

$$\forall x \in X, \forall f \in \mathcal{H}, \; f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$$

reproducing property

Terminology:

- **feature space** $\mathcal{H}$, **feature map** Φ
- **feature vector** $\Phi(x)$
- **kernel** $k = \langle \Phi(\cdot), \Phi(\cdot) \rangle_{\mathcal{H}} : X \times X \to \mathbb{R}$



Reproducing Kernel Hilbert Space

Def: Let $\mathcal{H} \subset \mathbb{R}^X$ Hilbert, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
Then, $\mathcal{H}$ is a **RKHS** on X if $\exists \Phi : X \to \mathcal{H}$ s.t.:

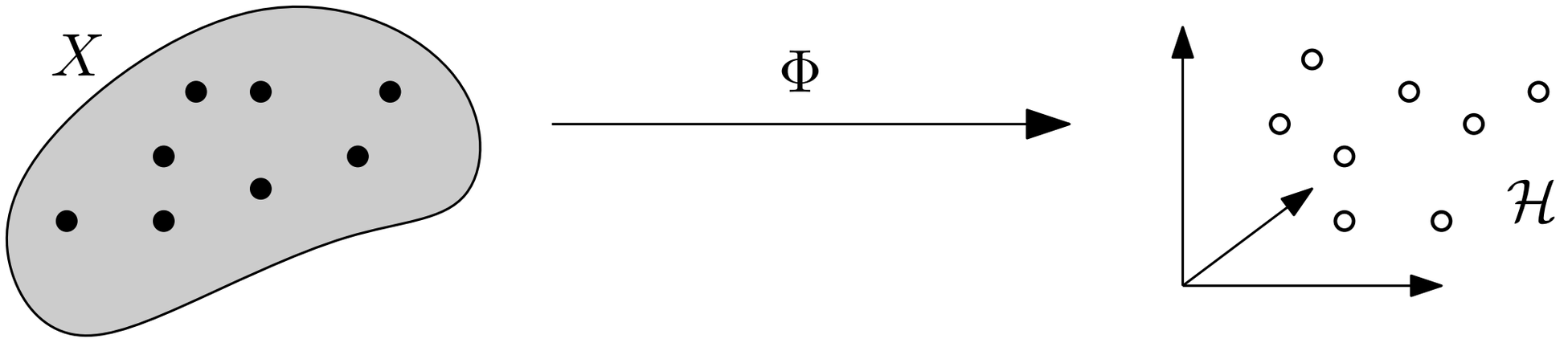
$$\forall x \in X,\, \forall f \in \mathcal{H},\; f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$$

reproducing property

Terminology:

- **feature space** $\mathcal{H}$, **feature map** Φ
- **feature vector** $\Phi(x)$
- **kernel** $k = \langle \Phi(\cdot),\, \Phi(\cdot) \rangle_{\mathcal{H}} : X \times X \to \mathbb{R}$

Case X Hilbert space:
$\mathcal{H} = X^*,\; \Phi(x) = \langle x, \cdot \rangle_X$
Φ isometric isomorphism [Riesz]
$\langle \cdot, \cdot \rangle_{\mathcal{H}} := \langle \Phi^{-1}(\cdot), \Phi^{-1}(\cdot) \rangle_X$



Reproducing Kernel Hilbert Space

Def: Let $\mathcal{H} \subset \mathbb{R}^X$ Hilbert, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
Then, $\mathcal{H}$ is a **RKHS** on X if $\exists \Phi : X \to \mathcal{H}$ s.t.:

$$\forall x \in X, \forall f \in \mathcal{H}, \; f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$$

reproducing property

Prop: Given X, the kernel of a RKHS on X is unique.
Conversely, k is the kernel of at most one RKHS on X.

$$\rightarrow \Phi(x) = k(x, \cdot)$$

Reproducing Kernel Hilbert Space

[*Theory of Reproducing Kernels*, Aronszajn, Trans. Amer. Math. Soc., 1950]

Def: Let $\mathcal{H} \subset \mathbb{R}^X$ Hilbert, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
Then, $\mathcal{H}$ is a **RKHS** on X if $\exists \Phi : X \to \mathcal{H}$ s.t.:

$$\forall x \in X, \forall f \in \mathcal{H}, \; f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$$

reproducing property

Prop: Given X, the kernel of a RKHS on X is unique. Conversely, k is the kernel of at most one RKHS on X.

Thm: The function $k : X \times X \to \mathbb{R}$ is a kernel iff it is *positive (semi-)definite*, i.e. $\forall n \in \mathbb{N}$, $\forall x_1, \ldots, x_n \in X$, the Gram matrix $(k(x_i, x_j))_{i,j}$ is positive semi-definite.

Reproducing Kernel Hilbert Space

[*Theory of Reproducing Kernels*, Aronszajn, Trans. Amer. Math. Soc., 1950]

Def: Let $\mathcal{H} \subset \mathbb{R}^X$ Hilbert, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
Then, $\mathcal{H}$ is a **RKHS** on X if $\exists \Phi : X \to \mathcal{H}$ s.t.:

$$\forall x \in X, \forall f \in \mathcal{H}, \; f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$$

reproducing property

Prop: Given X, the kernel of a RKHS on X is unique. Conversely, k is the kernel of at most one RKHS on X.

Thm: The function $k : X \times X \to \mathbb{R}$ is a kernel iff it is *positive (semi-)definite*, i.e. $\forall n \in \mathbb{N}$, $\forall x_1, \ldots, x_n \in X$, the Gram matrix $(k(x_i, x_j))_{i,j}$ is positive semi-definite.

Examples in $X = (\mathbb{R}^d, \langle \cdot, \cdot \rangle)$:

- linear: $k(x, y) = \langle x, y \rangle$
- polynomial: $k(x, y) = (1 + \langle x, y \rangle)^N = \sum_{n_1 + \cdots + n_d = N} \binom{N}{n_1, \ldots, n_d} x_1^{n_1} \ldots x_d^{n_d} \, y_1^{n_1} \ldots y_d^{n_d}$
- Gaussian: $k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2} \right)$, $\sigma > 0$.

Reproducing Kernel Hilbert Space

[*A correspondence between Bayesian estimation on stochastic processes and smoothing by splines*, Kimeldorf, Wahba, The Annals Math. Stat., 1970]

Def: Let $\mathcal{H} \subset \mathbb{R}^X$ Hilbert, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
Then, $\mathcal{H}$ is a **RKHS** on X if $\exists \Phi : X \to \mathcal{H}$ s.t.:

$$\forall x \in X, \forall f \in \mathcal{H}, \; f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$$

reproducing property

Thm: (Representer)
Given RKHS $\mathcal{H}$ with kernel k, any function $f^* \in \mathcal{H}$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(\|f\|_{\mathcal{H}})$$

is of the form $f^*(\cdot) = \sum_{j=1}^{n} \alpha_j k(x_j, \cdot)$, where $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$.

Reproducing Kernel Hilbert Space

[*A correspondence between Bayesian estimation on stochastic processes and smoothing by splines*, Kimeldorf, Wahba, The Annals Math. Stat., 1970]

Def: Let $\mathcal{H} \subset \mathbb{R}^X$ Hilbert, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
Then, $\mathcal{H}$ is a **RKHS** on X if $\exists \Phi : X \to \mathcal{H}$ s.t.:

$$\forall x \in X, \forall f \in \mathcal{H}, \; f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$$

reproducing property

Thm: (Representer)
Given RKHS $\mathcal{H}$ with kernel k, any function $f^* \in \mathcal{H}$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(\|f\|_{\mathcal{H}})$$

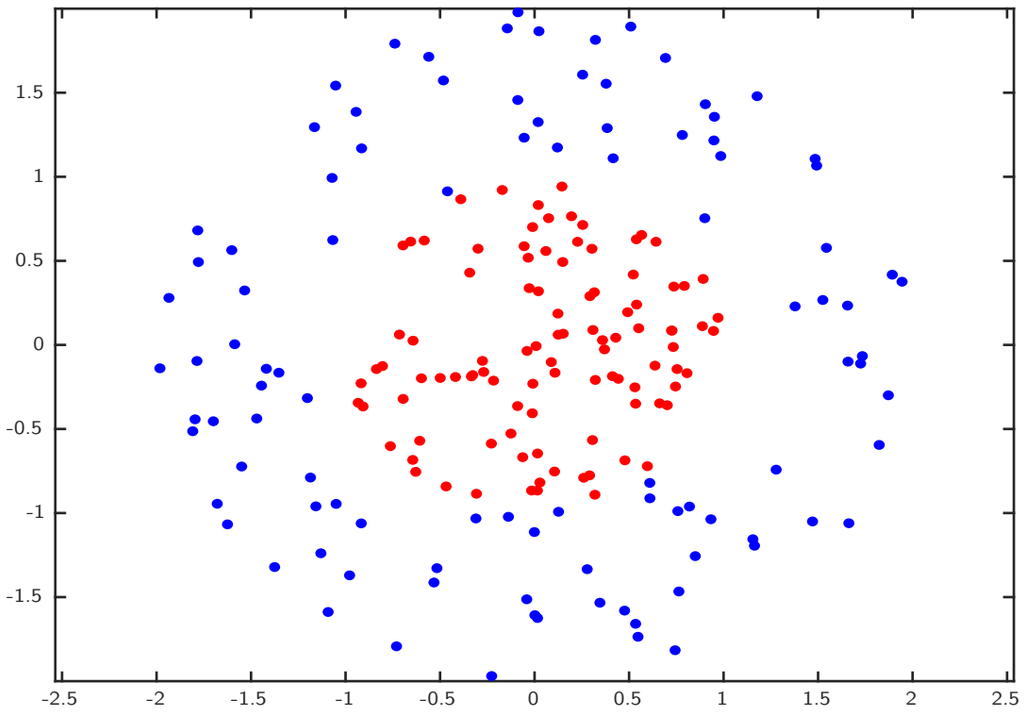
is of the form $f^*(\cdot) = \sum_{j=1}^{n} \alpha_j k(x_j, \cdot)$, where $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$.

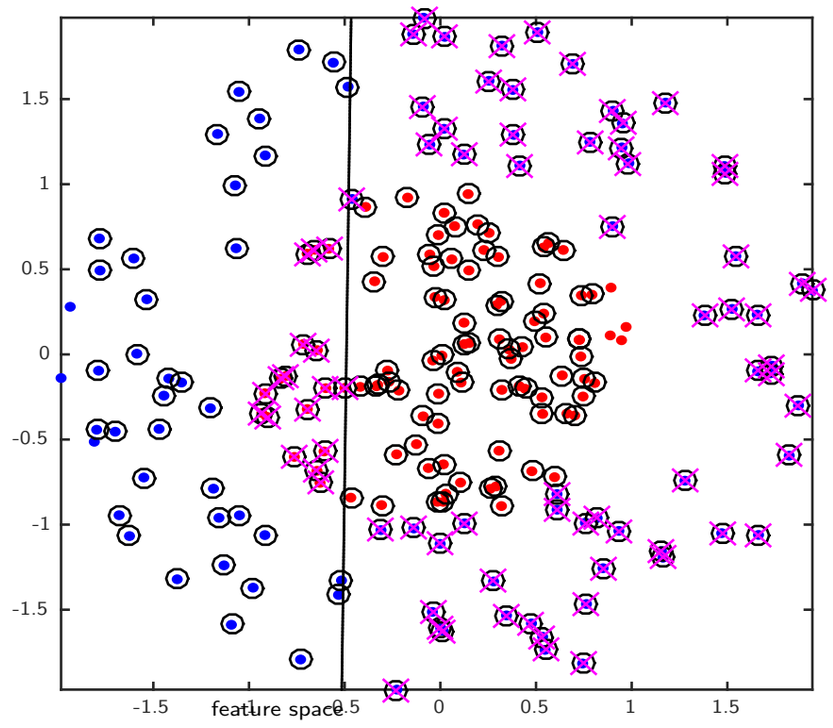
$$\rightsquigarrow \quad \operatorname*{argmin}_{\alpha} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{n} \alpha_j k(x_j, x_i)\right) + \Omega\left(\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)\right)$$

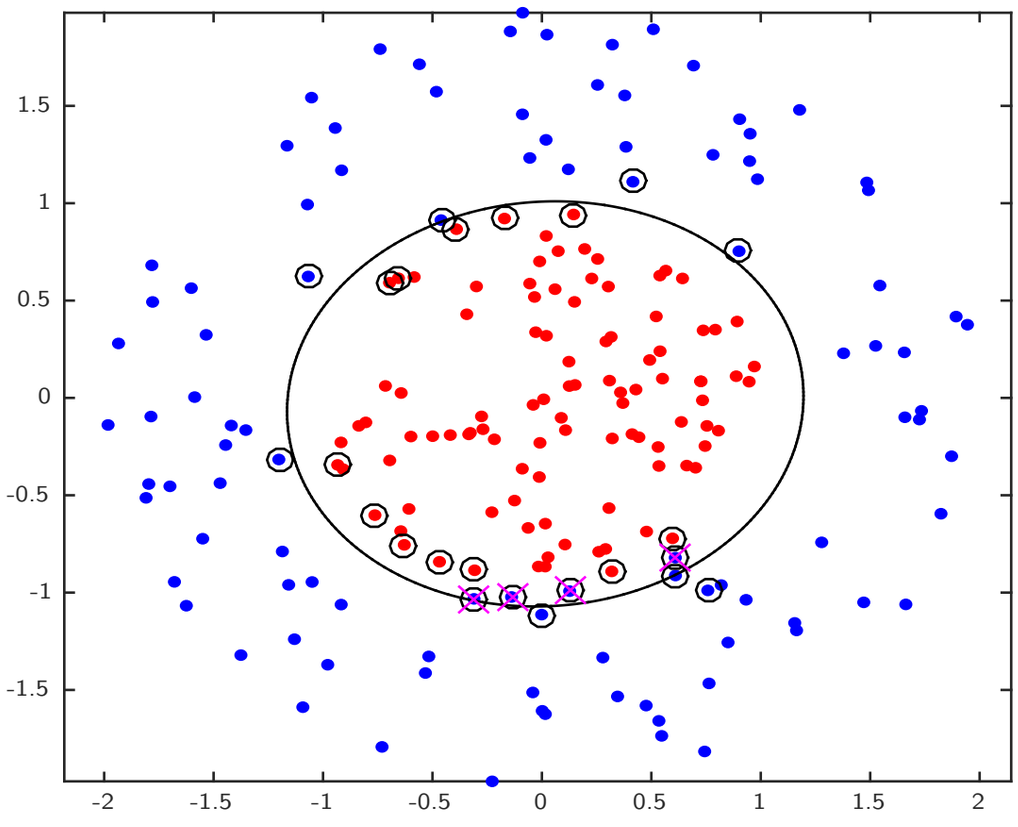
where $\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$ and $K = (k(x_i, x_j))_{ij}$

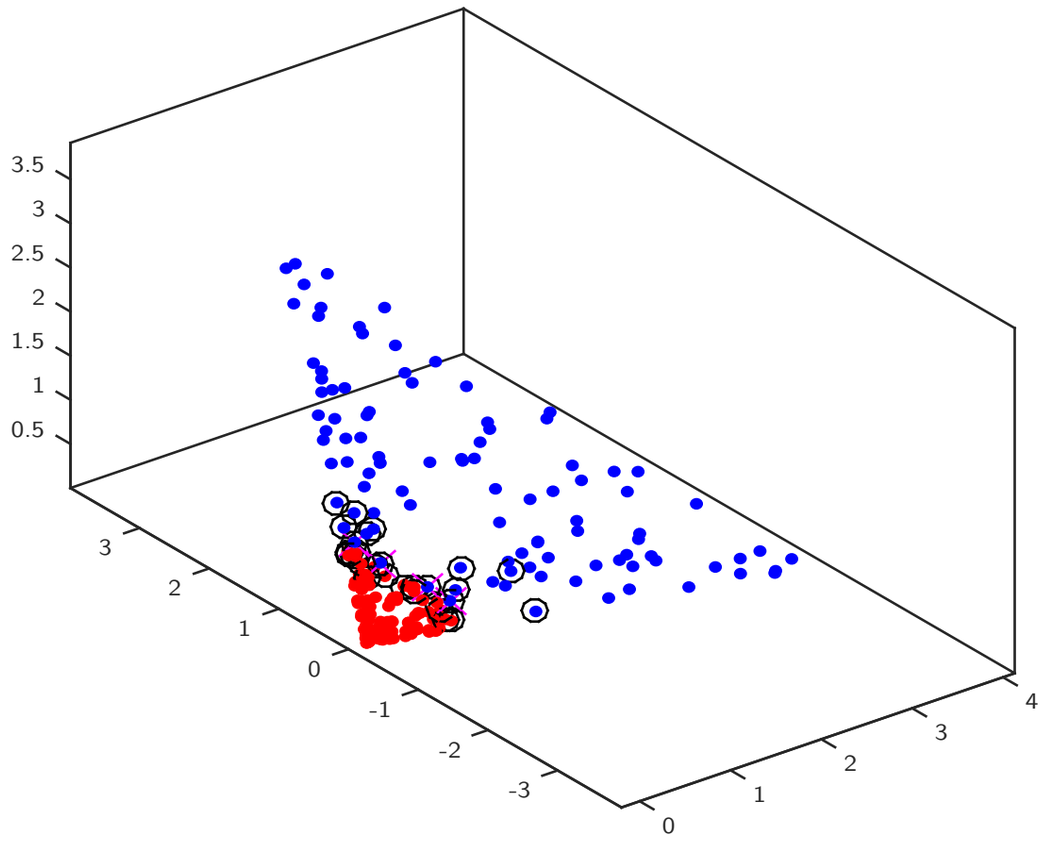
only the $k(x_i, x_j)$ are required to minimize (**kernel trick**)

Kernel Trick









Kernels for persistence diagrams

Three potential approaches:

- build kernel from kernels (algebraic operations)

  - **sum of kernels $\longleftrightarrow$ concatenation of feature spaces**

$$k_1(x,y) + k_2(x,y) = \left\langle \begin{pmatrix} \Phi_1(x) \\ \Phi_2(x) \end{pmatrix}, \begin{pmatrix} \Phi_1(y) \\ \Phi_2(y) \end{pmatrix} \right\rangle$$

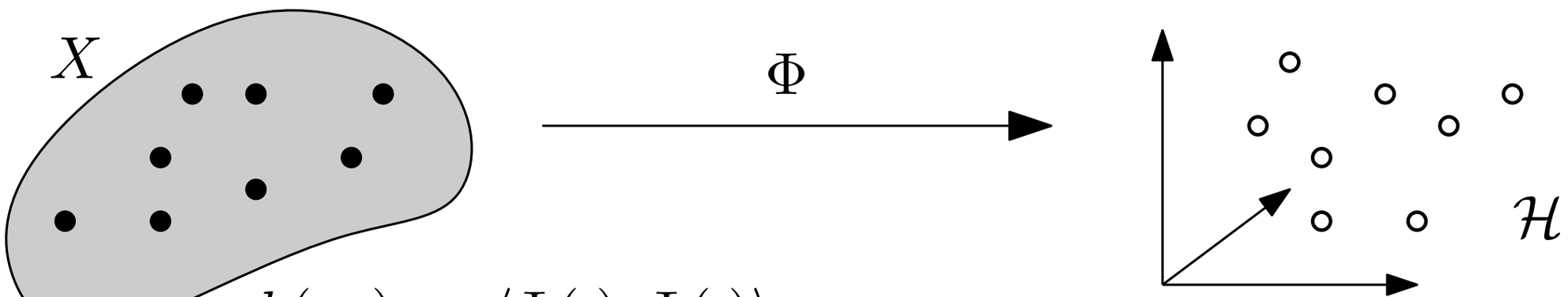
- product of kernels $\longleftrightarrow$ tensor product of feature spaces

$$k_1(x,y)k_2(x,y) = \left\langle \Phi_1(x)\Phi_2(x)^T, \Phi_1(y)\Phi_2(y)^T \right\rangle$$

Kernels for persistence diagrams

Three potential approaches:

- build kernel from kernels (algebraic operations)
- 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

[*A correspondence between Bayesian estimation on stochastic processes and smoothing by splines*, Kimeldorf, Wahba, The Annals Math. Stat., 1970]

Three potential approaches:

- build kernel from kernels (algebraic operations)
- define explicit feature map $\Phi : X \to \mathcal{H}$ (vectorization)
- define kernel from metric via radial basis function

Thm: If $d : X \times X \to \mathbb{R}_+$ is symmetric and *conditionally negative semidefinite* (cnsd), 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 $\sigma > 0$.

Kernels for persistence diagrams

[*A correspondence between Bayesian estimation on stochastic processes and smoothing by splines*, Kimeldorf, Wahba, The Annals Math. Stat., 1970]

Three potential approaches:

- build kernel from kernels (algebraic operations)
- define explicit feature map $\Phi : X \to \mathcal{H}$ (vectorization)
- define kernel from metric via radial basis function

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

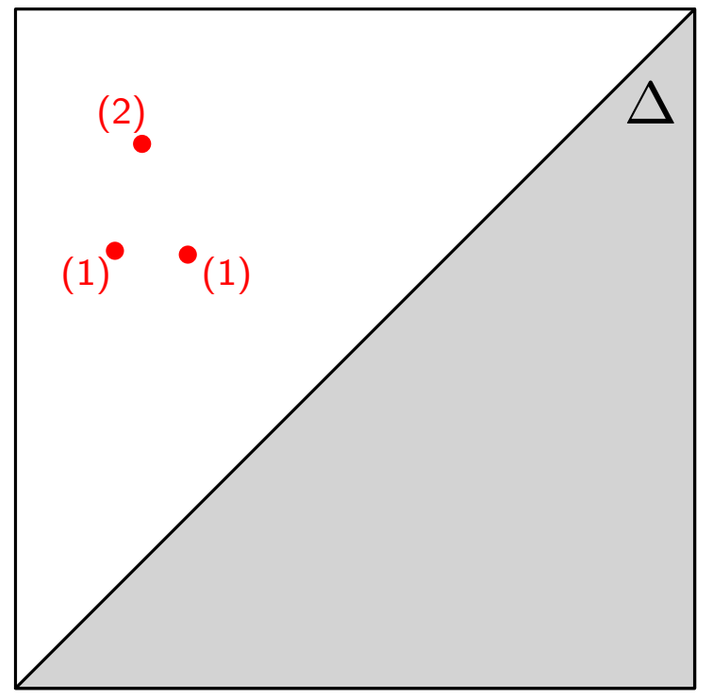
$$\forall n \in \mathbb{N},\ \forall x_1, \ldots, x_n \in X,\ \sum_{i=1}^{n} \alpha_i = 0 \Longrightarrow \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, d(x_i, x_j) \leq 0,$$

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

Q: does this apply to persistence diagrams?

Space of persistence diagrams

Persistence diagram $\equiv$ **finite** multiset in the open half-plane $\Delta \times \mathbb{R}_{>0}$



Space of persistence diagrams

Persistence diagram $\equiv$ **finite** multiset in the open half-plane $\Delta \times \mathbb{R}_{>0}$

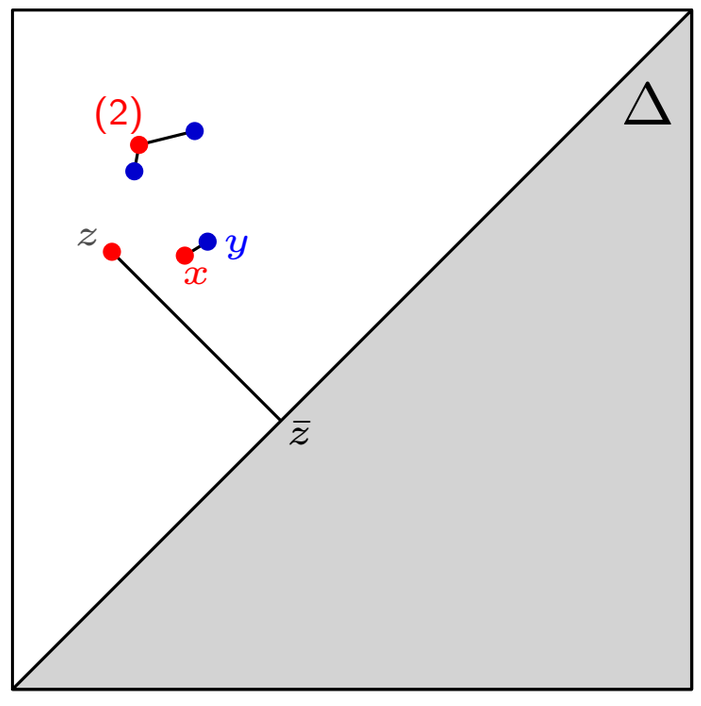
Given a **partial matching** $M : X \leftrightarrow Y$:

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

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

The **cost of** M is:

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



Space of persistence diagrams

Persistence diagram $\equiv$ **finite** multiset in the open half-plane $\Delta \times \mathbb{R}_{>0}$

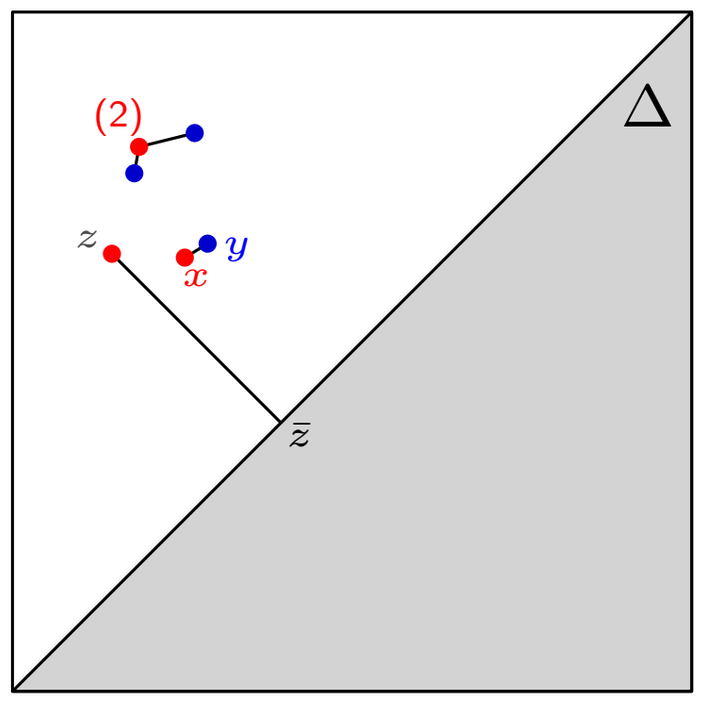
Given a **partial matching** $M : X \leftrightarrow Y$:

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

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

The **cost of** M is:

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



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

$$d_p(X, Y) := \inf_{M : X \leftrightarrow Y} c_p(M)$$

Def: bottleneck distance:

$$d_b(X, Y) := \lim_{p \to \infty} d_p(X, Y)$$

Space of persistence diagrams

Persistence diagram $\equiv$ **finite** multiset in the open half-plane $\Delta \times \mathbb{R}_{>0}$

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

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

cost of an unmatched point $z \in$

d_p is **NOT** cnsd

$\Rightarrow$ previous theorem is not applicable

The **cost of** M is:

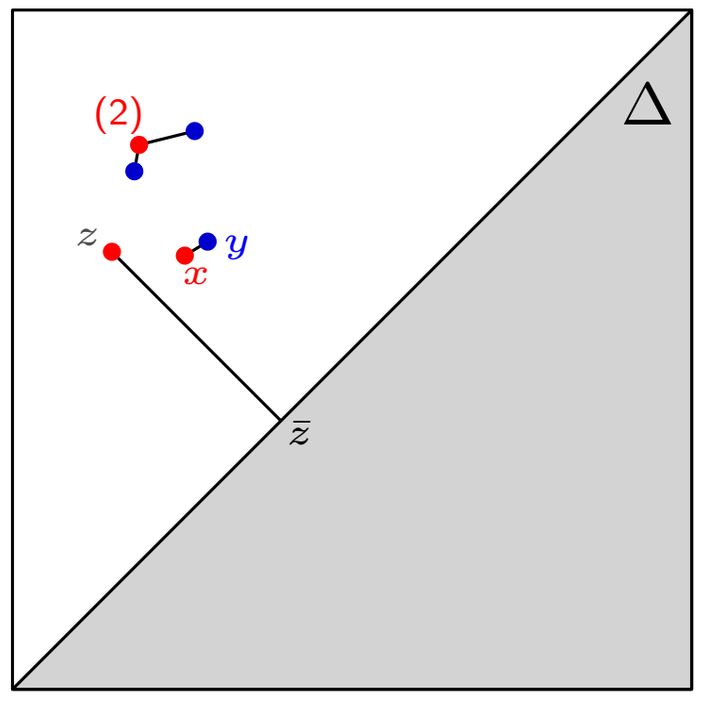
$$c_p(M) := \left(\sum_{(x,\, y) \text{ matched}} c_p(x, y) \; + \sum_{z \text{ unmatched}} c_p(z) \right)^{1/p}$$

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

$$d_p(X, Y) := \inf_{M : X \leftrightarrow Y} c_p(M)$$

Def: bottleneck distance:

$$d_b(X, Y) := \lim_{p \to \infty} d_p(X, Y)$$

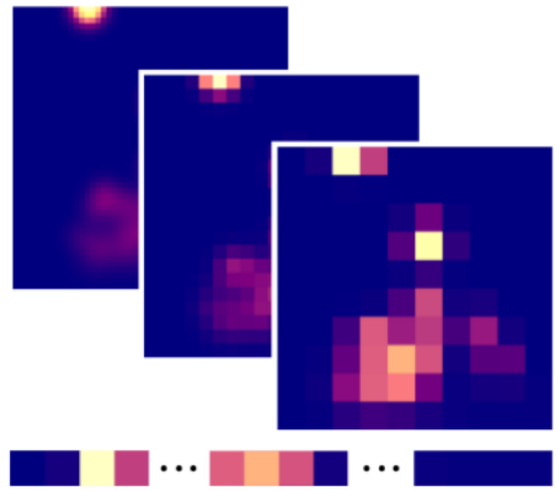


Kernels for persistence diagrams

State of the Art: define ϕ explicitly (vectorization) via:

- images

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]



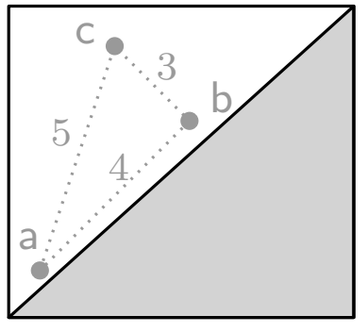
Kernels for persistence diagrams

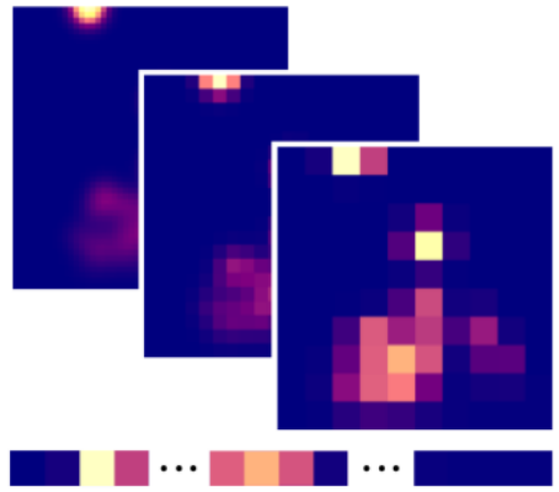
State of the Art: define ϕ explicitly (vectorization) via:

- images

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]

$$\begin{array}{c} \\ a \\ b \\ c \end{array}\begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0 & 4 & 5 \\ 4 & 0 & 3 \\ 5 & 3 & 0 \end{bmatrix} \end{array}$$





- finite metric spaces

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]

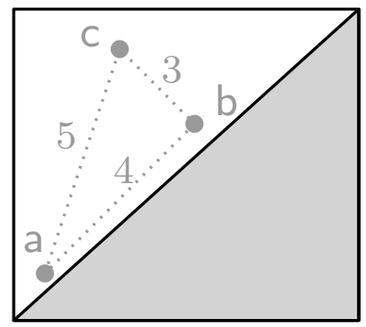
Kernels for persistence diagrams

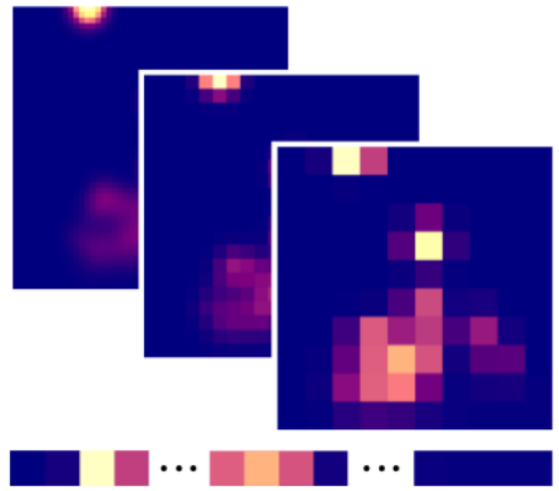
State of the Art: define ϕ explicitly (vectorization) via:

- images

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]

	a	b	c
a	0	4	5
b	4	0	3
c	5	3	0





- finite metric spaces

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]

- polynomial roots or evaluations

[*Tropical coordinates on the space of persistence barcodes*, Kalisnik, FoCM, 2018]

$$\{p_1, \ldots, p_n\} \mapsto (P_1(p_1, \ldots, p_n), \ldots, P_r(p_1, \ldots, p_n), \ldots)$$

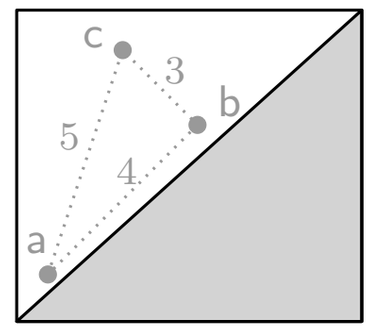
Kernels for persistence diagrams

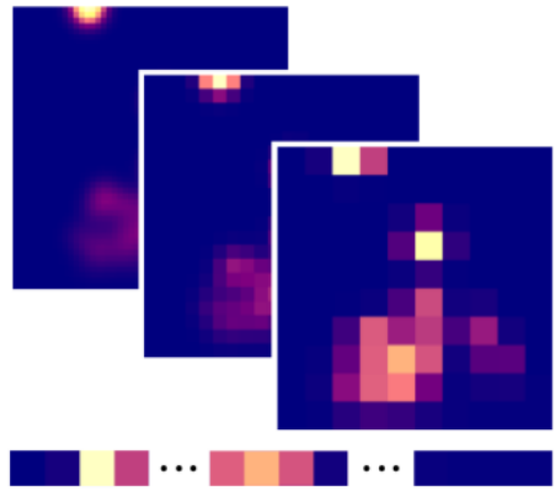
State of the Art: define ϕ explicitly (vectorization) via:

- images

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]

	a	b	c
a	0	4	5
b	4	0	3
c	5	3	0





- finite metric spaces

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]

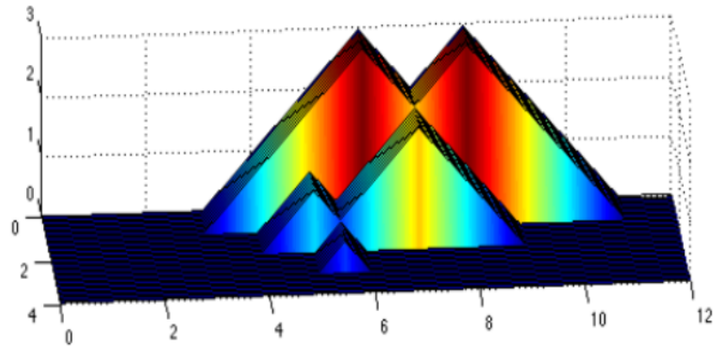
- polynomial roots or evaluations

[*Tropical coordinates on the space of persistence barcodes*, Kalisnik, FoCM, 2018]

$$\{p_1, \ldots, p_n\} \mapsto (P_1(p_1, \ldots, p_n), \ldots, P_r(p_1, \ldots, p_n), \ldots)$$

- landscapes

[*Statistical Topological Data Analysis using Persistence Landscapes*, Bubenik, JMLR, 2015]



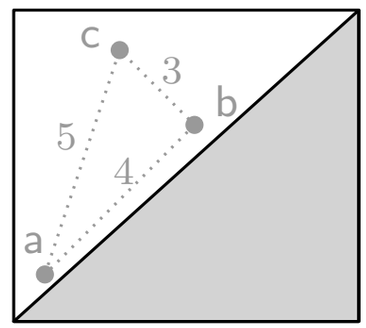
Kernels for persistence diagrams

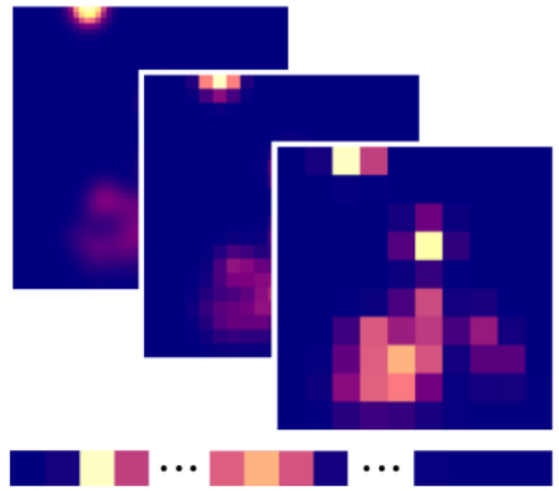
State of the Art: define ϕ explicitly (vectorization) via:

- images

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]

	a	b	c
a	0	4	5
b	4	0	3
c	5	3	0





- finite metric spaces

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]

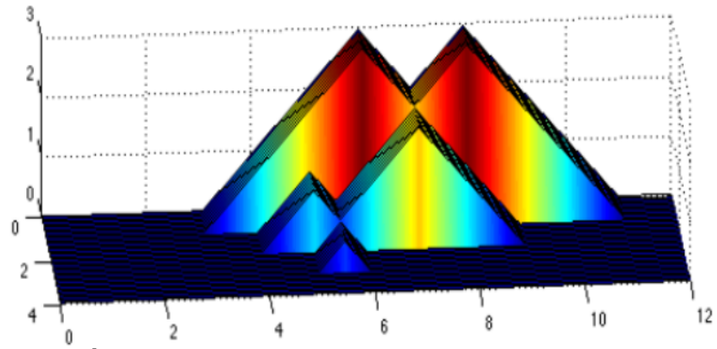
- polynomial roots or evaluations

[*Tropical coordinates on the space of persistence barcodes*, Kalisnik, FoCM, 2018]

$$\{p_1, \ldots, p_n\} \mapsto (P_1(p_1, \ldots, p_n), \ldots, P_r(p_1, \ldots, p_n), \ldots)$$

- landscapes

[*Statistical Topological Data Analysis using Persistence Landscapes*, Bubenik, JMLR, 2015]



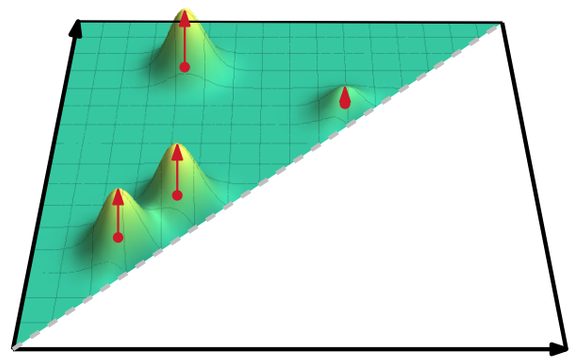
- discrete measures:

→ Fisher information

[*Persistence Fisher kernel: a Riemannian manifold kernel for persistence diagrams*, Le, Yamada, NeurIPS, 2018]

→ convolution with weighted kernel

[*Persistence weighted Gaussian kernel for topological data analysis*, Kusano, Hiraoka, Fukumizu, ICML, 2016]



→ heat diffusion

[*A stable multi-scale kernel for topological machine learning*, Reininghaus et al., CVPR, 2015]

Kernels for persistence diagrams

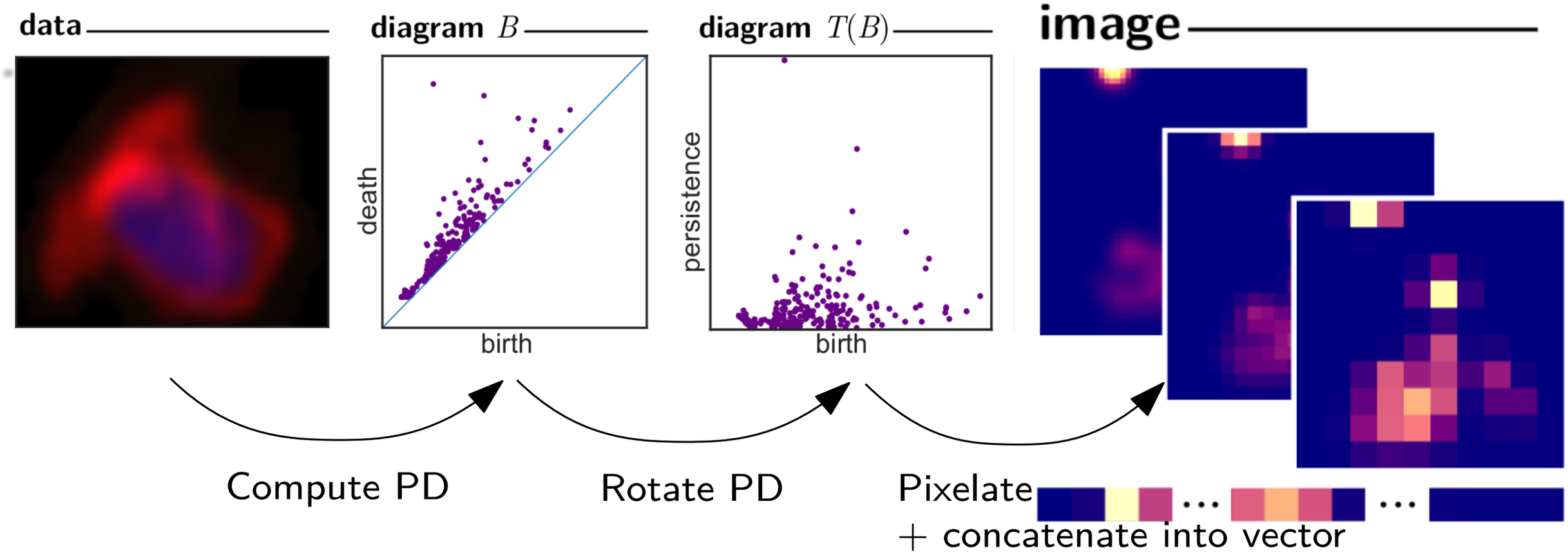
	images	metric spaces	polynomials	landscapes	discrete measures
ambient Hilbert space	$(\mathbb{R}^d, \|.\|_2)$	$(\mathbb{R}^d, \|.\|_2)$	$\ell_2(\mathbb{R})$	$L_2(\mathbb{N}\times\mathbb{R})$	$L_2(\mathbb{R}^2)$
positive (semi-)definiteness	✓	✓	✓	✓	✓
$\|\phi(\cdot)-\phi(\cdot)\|_{\mathcal{H}} \leq \phi(d_p)$	✓	✓	✓	✓	✓
$\|\phi(\cdot)-\phi(\cdot)\|_{\mathcal{H}} \geq \psi(d_p)$	✗	✗	✗	✗	✗
injectivity	✗	✗	✓	✓	✓
universality	✗	✗	✗	✗	✓
algorithmic cost	f. map: $O(n^2)$ kernel: $O(d)$	f. map: $O(n^2)$ kernel: $O(d)$	f. map: $O(nd)$ kernel: $O(d)$	$O(n^2)$	$O(n^2)$

Kernels for persistence diagrams

	images	metric spaces	polynomials	landscapes	discrete measures
ambient Hilbert space	$(\mathbb{R}^d, \|.\|_2)$	$(\mathbb{R}^d, \|.\|_2)$	$\ell_2(\mathbb{R})$	$L_2(\mathbb{N}\times\mathbb{R})$	$L_2(\mathbb{R}^2)$
positive (semi-)definiteness	✓	✓	✓	✓	✓
$\|\phi(\cdot)-\phi(\cdot)\|_{\mathcal{H}} \leq \phi(d_p)$	✓	✓	✓	✓	✓
$\|\phi(\cdot)-\phi(\cdot)\|_{\mathcal{H}} \geq \psi(d_p)$	✗	✗	✗	✗	✗
injectivity	✗	✗	✓	✓	✓
universality	✗	✗	✗	✗	✓
algorithmic cost	f. map: $O(n^2)$ kernel: $O(d)$	f. map: $O(n^2)$ kernel: $O(d)$	f. map: $O(nd)$ kernel: $O(d)$	$O(n^2)$	$O(n^2)$

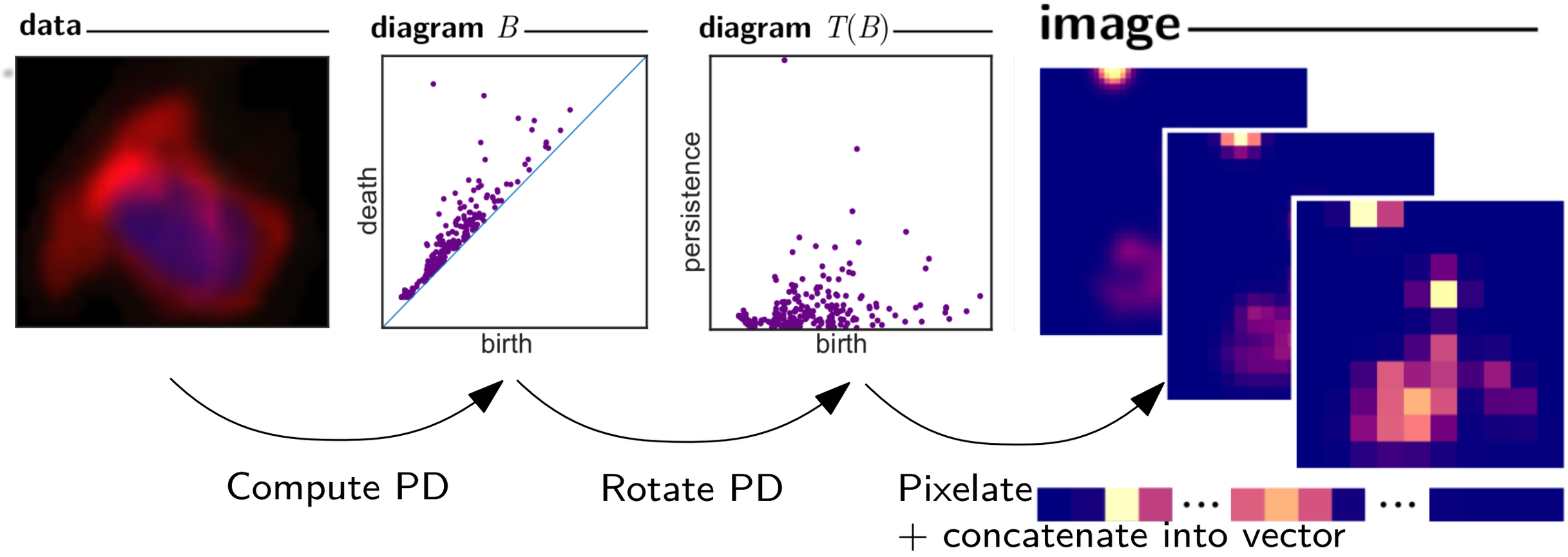
Explicit Feature Map in $\mathbb{R}^d$

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]

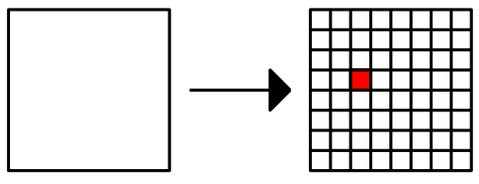


Explicit Feature Map in $\mathbb{R}^d$

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]



Discretize plane into a grid:



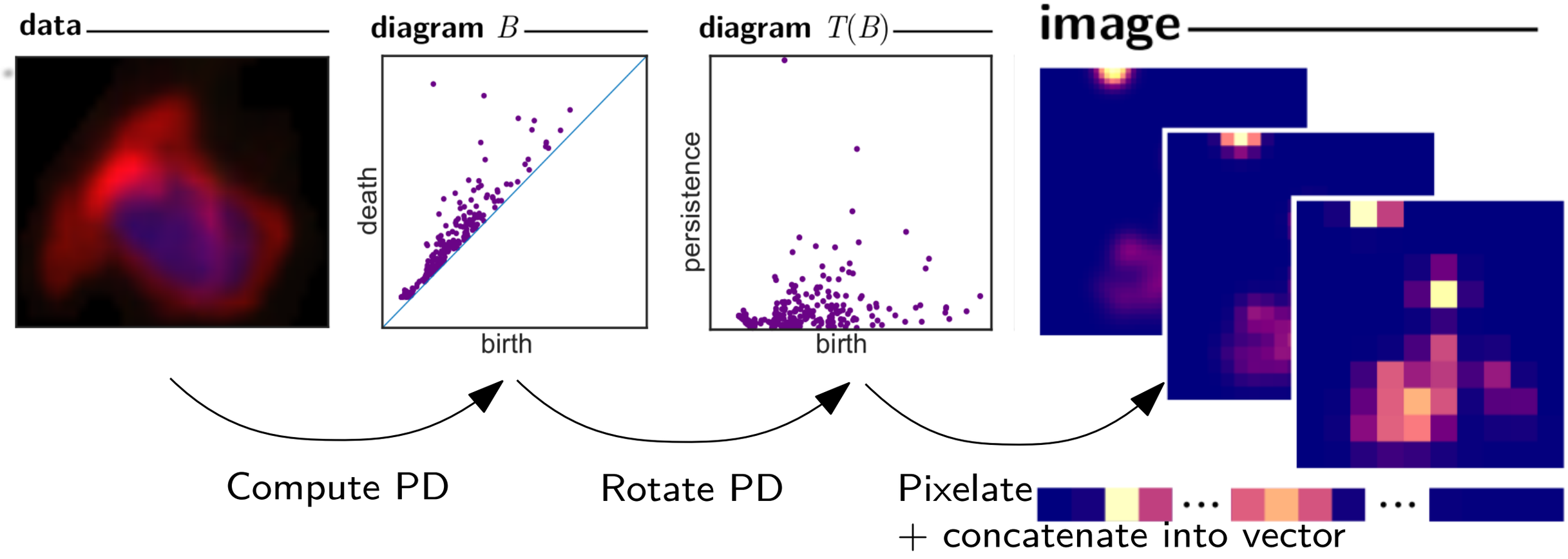
For each grid pixel P, compute $I(P) = \sum_{p \in D} \int \int_P w(p) \cdot \phi_p$.

Concatenate all $I(P)$ into a single vector $\mathrm{PI}(D)$.

Ex: $\phi_p = \mathcal{N}(p, \sigma)$.

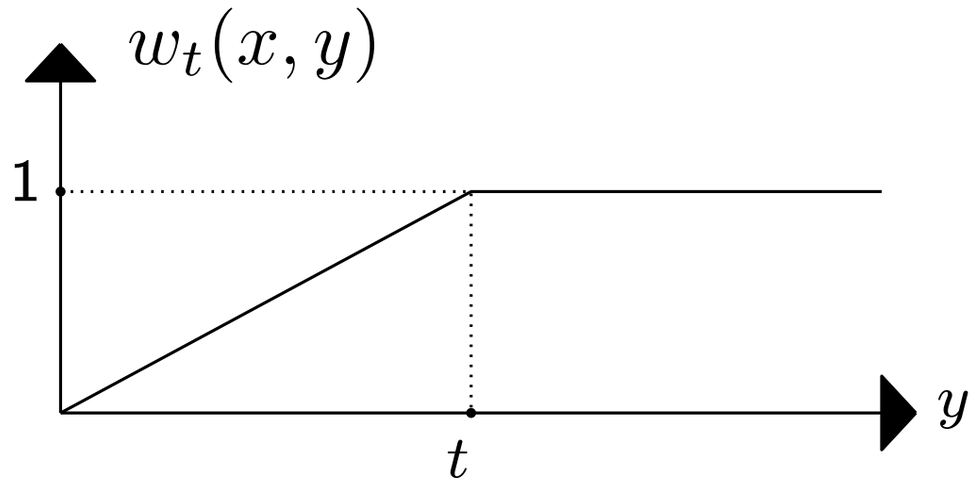
Explicit Feature Map in $\mathbb{R}^d$

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]



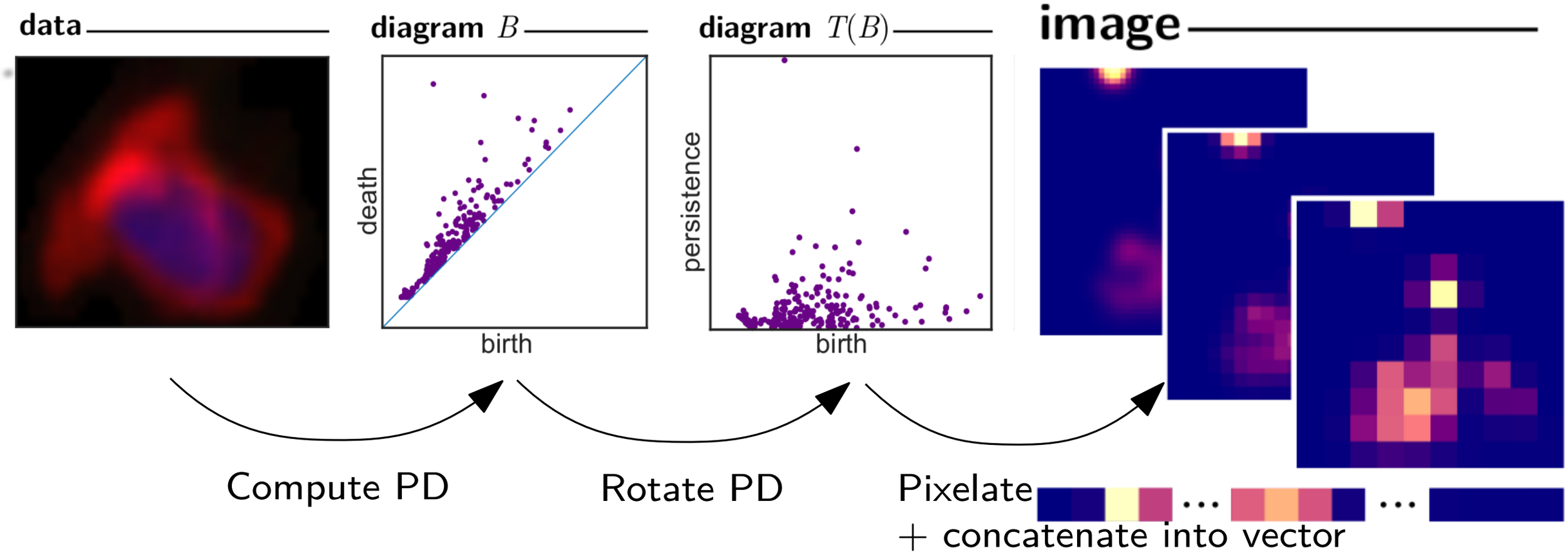
Weight functions that preserve stability must satisfy $w(p) \to 0$ when $d(p, \Delta) \to 0$.

[*Understanding the topology and the geometry of the persistence diagram space via optimal partial transport*, Divol, Lacombe, JACT, 2020]



Explicit Feature Map in $\mathbb{R}^d$

[*Persistence Images: A Stable Vector Representation of Persistent Homology*, Adams et al., JMLR, 2017]



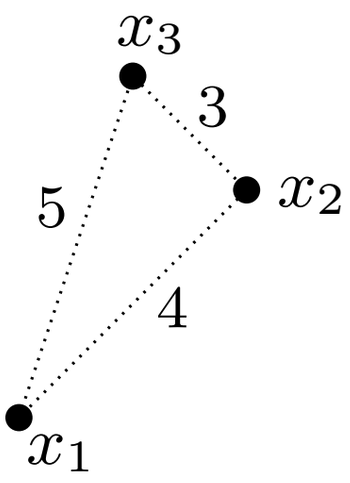
Prop: The following inequalities hold:

- $\|\mathrm{PI}(D) - \mathrm{PI}(D')\|_\infty \leq C(w, \phi_p)\, d_1(D, D')$.
- $\|\mathrm{PI}(D) - \mathrm{PI}(D')\|_2 \leq \sqrt{d} C(w, \phi_p)\, d_1(D, D')$.

Explicit Feature Map in $\mathbb{R}^d$

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]

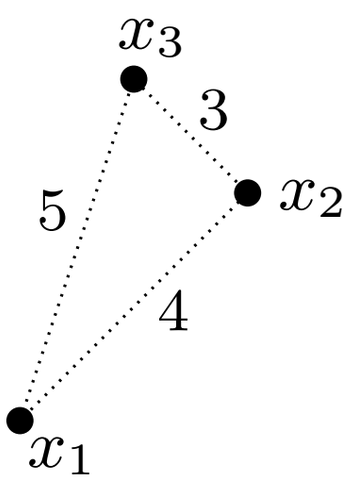
finite metric space



Explicit Feature Map in $\mathbb{R}^d$

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]

finite metric space



ϕ_1

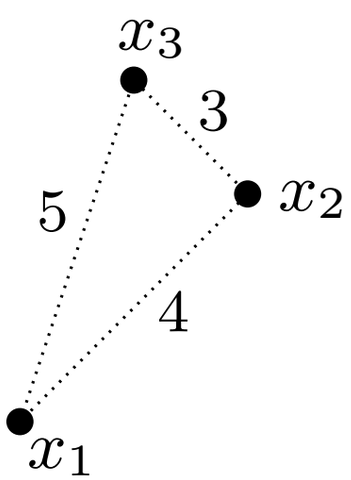
distance matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 0 & 4 & 5 \\ 4 & 0 & 3 \\ 5 & 3 & 0 \end{bmatrix} \end{array}$$

Explicit Feature Map in $\mathbb{R}^d$

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]

finite metric space



ϕ_1

distance matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 0 & 4 & 5 \\ 4 & 0 & 3 \\ 5 & 3 & 0 \end{bmatrix} \end{array}$$

ϕ_2

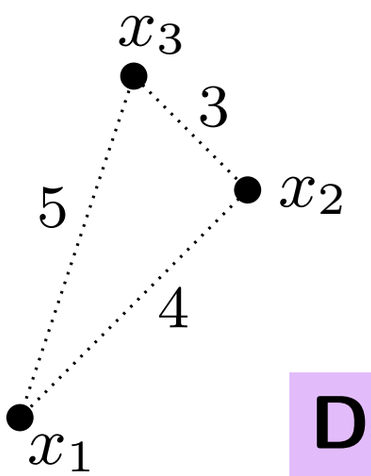
sorted sequence
with finite support

$(5, 4, 3, 0, \cdots)$

Explicit Feature Map in $\mathbb{R}^d$

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]

finite metric space



ϕ_1

distance matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 0 & 4 & 5 \\ 4 & 0 & 3 \\ 5 & 3 & 0 \end{bmatrix} \end{array}$$

Def: $\Phi = \Phi_3 \circ \Phi_2 \circ \Phi_1$

$(5, 4, 3, 0, \cdots, 0)$

finite-dimensional vector

ϕ_3

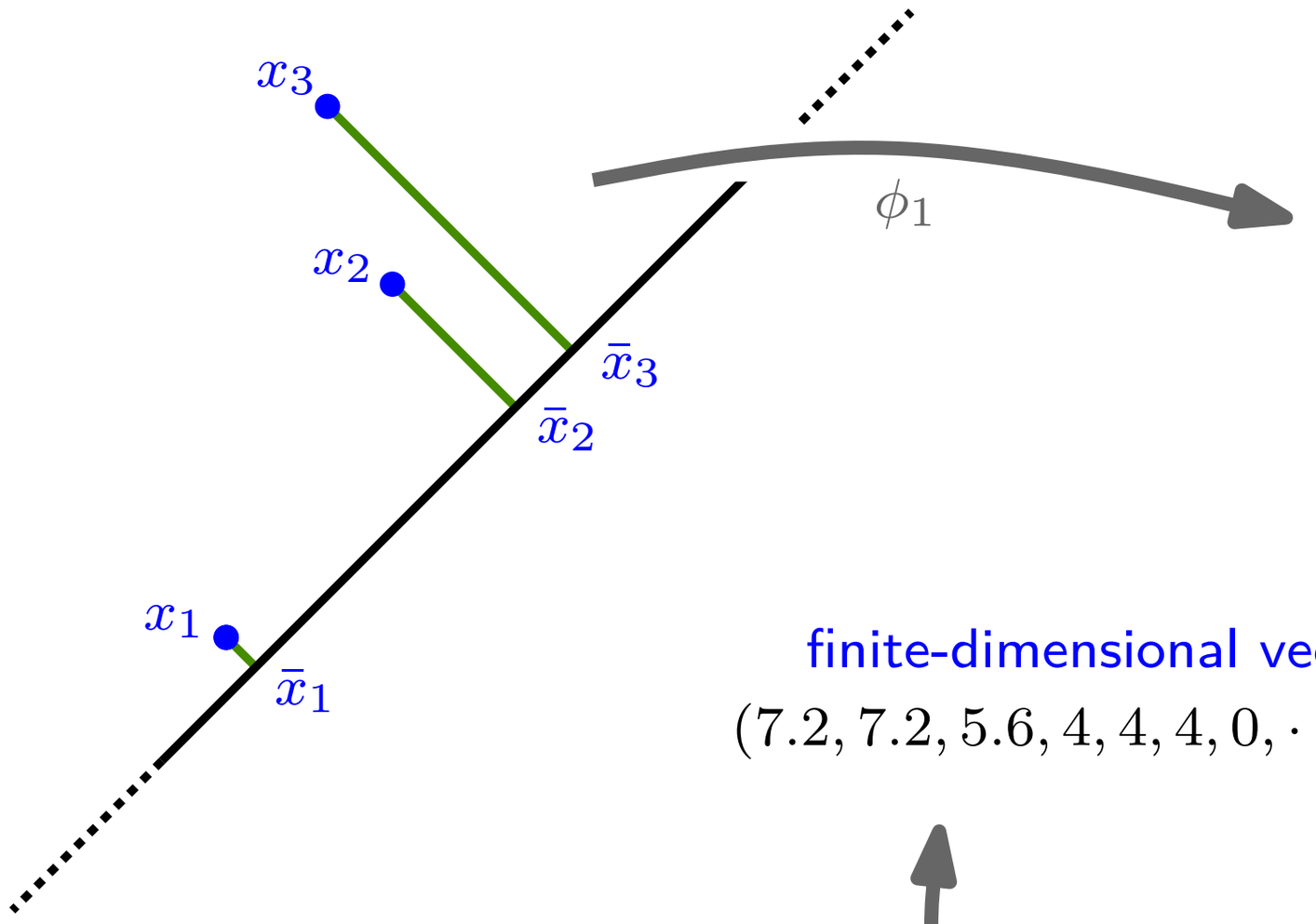
sorted sequence
with finite support

ϕ_2

$(5, 4, 3, 0, \cdots)$

Explicit Feature Map in $\mathbb{R}^d$

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]



distance matrix

$$\begin{bmatrix} D_{ij} = \min\{ & \|x_i - x_j\|_\infty, \\ & \|x_i - \bar{x}_i\|_\infty, \\ & \|x_j - \bar{x}_j\|_\infty \} \end{bmatrix}$$

ϕ_2

finite-dimensional vector

$(7.2, 7.2, 5.6, 4, 4, 4, 0, \cdots, 0)$

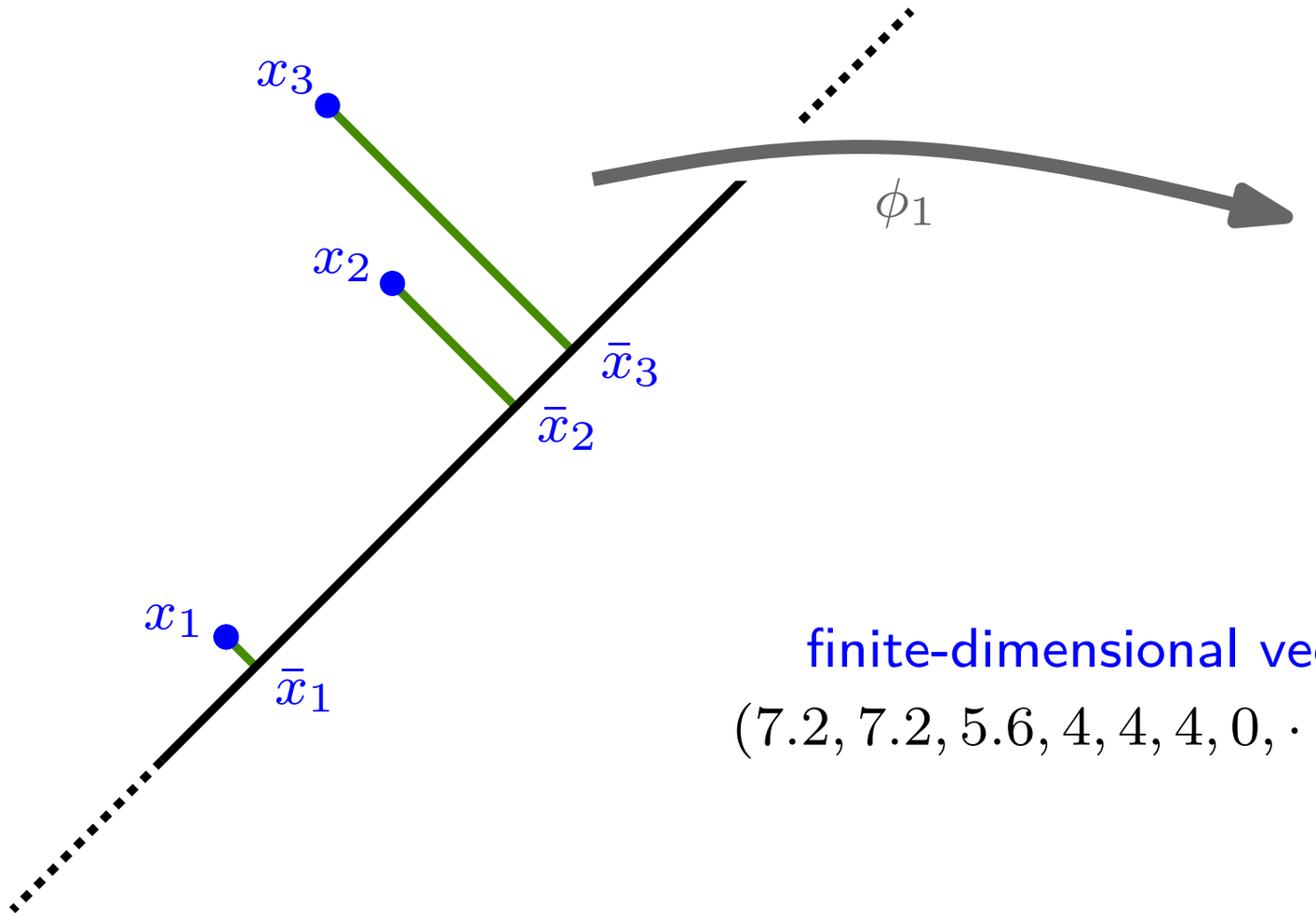
ϕ_3

$(7.2, 7.2, 5.6, 4, 4, 4, 0, \cdots)$

sorted sequence
with finite support

Explicit Feature Map in $\mathbb{R}^d$

[*Stable topological signatures for points on 3D shapes*, C., Oudot, Ovsjanikov, SGP, 2015]



distance matrix

$$\begin{bmatrix} D_{ij} = \min\{ & \|x_i - x_j\|_\infty, \\ & \|x_i - \bar{x}_i\|_\infty, \\ & \|x_j - \bar{x}_j\|_\infty \} \end{bmatrix}$$

finite-dimensional vector

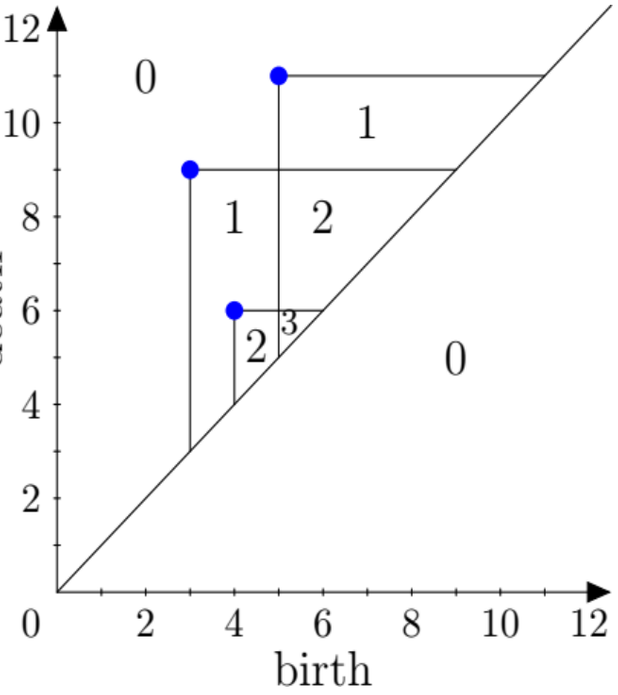
$$(7.2, 7.2, 5.6, 4, 4, 4, 0, \cdots, 0)$$

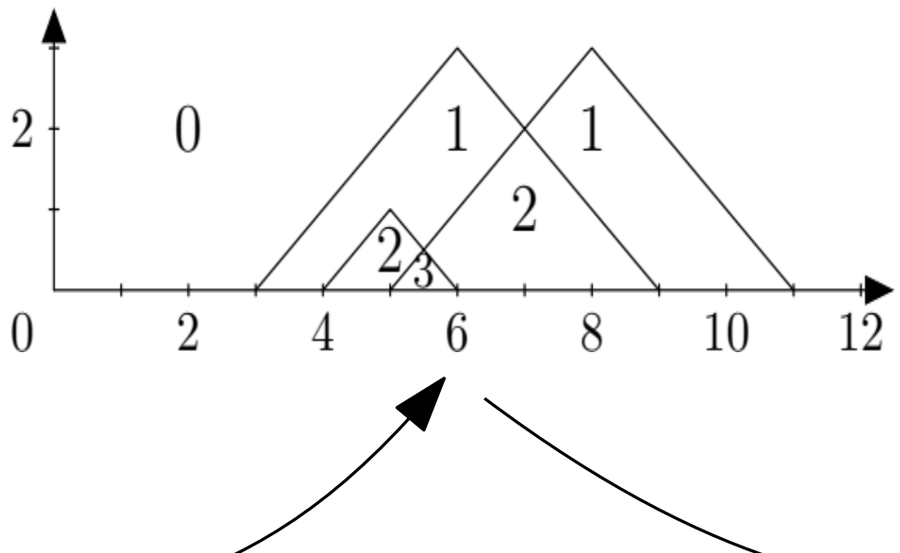
Prop: The following inequalities hold:

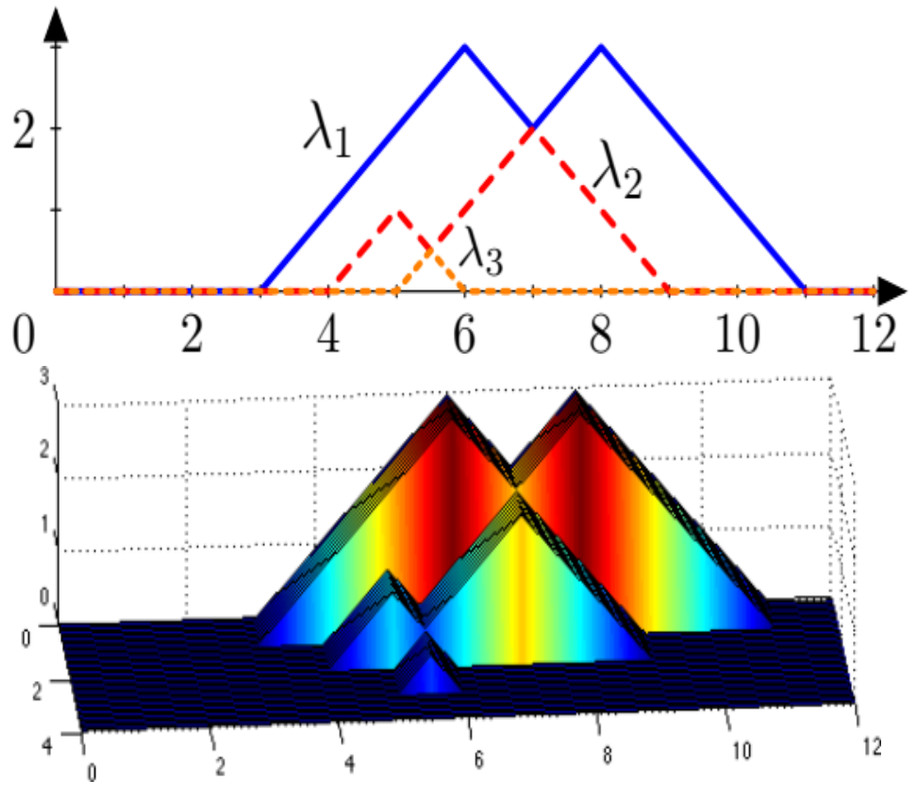
- $\|\Phi(D) - \Phi(D')\|_\infty \leq 2\, d_b(D, D')$.
- $\|\Phi(D) - \Phi(D')\|_2 \leq 2\sqrt{d}\, d_b(D, D')$.

Explicit Feature Map in Function Space

[*Statistical Topological Data Analysis using Persistence Landscapes*, Bubenik, JMLR, 2015]





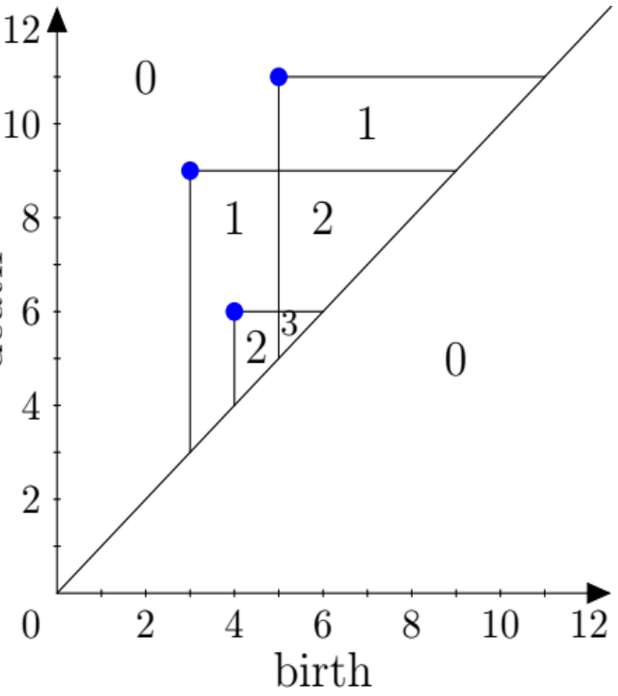


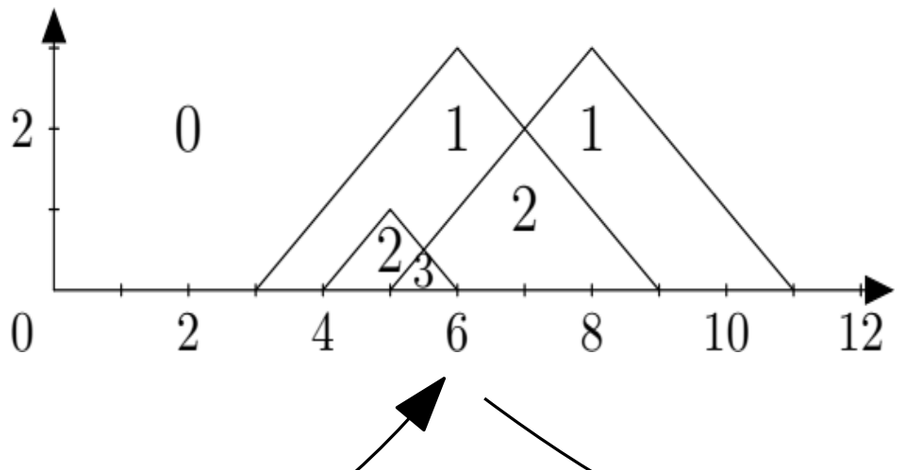
Rotate PD
Compute rank function

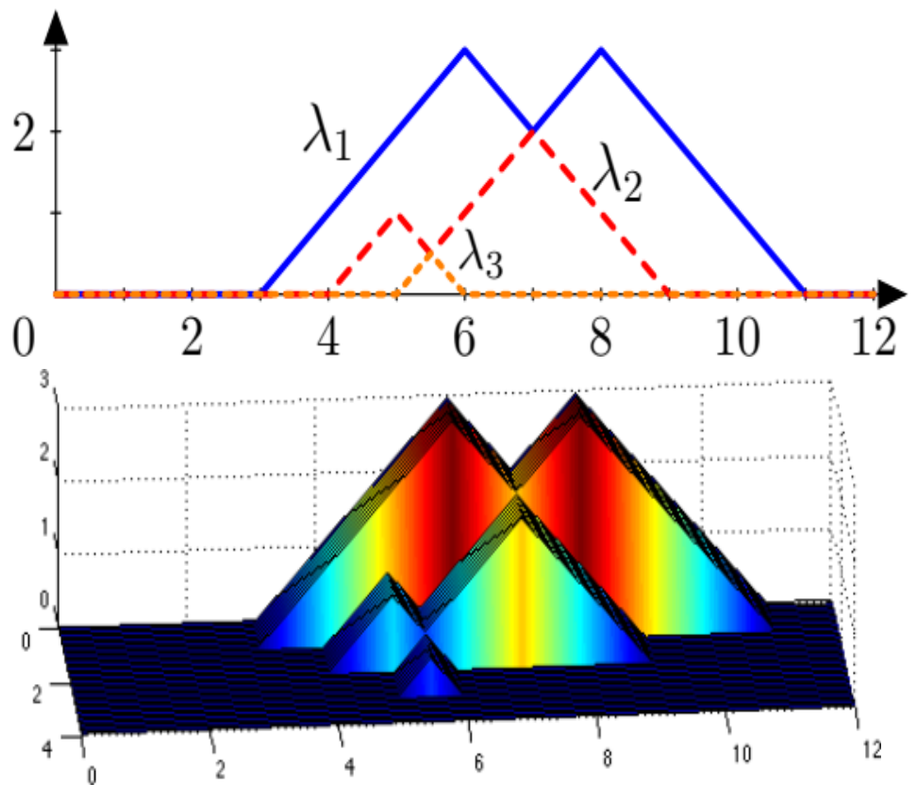
Use boundaries of rank function

Explicit Feature Map in Function Space

[*Statistical Topological Data Analysis using Persistence Landscapes*, Bubenik, JMLR, 2015]







Rotate PD
Compute rank function

Use boundaries of rank function

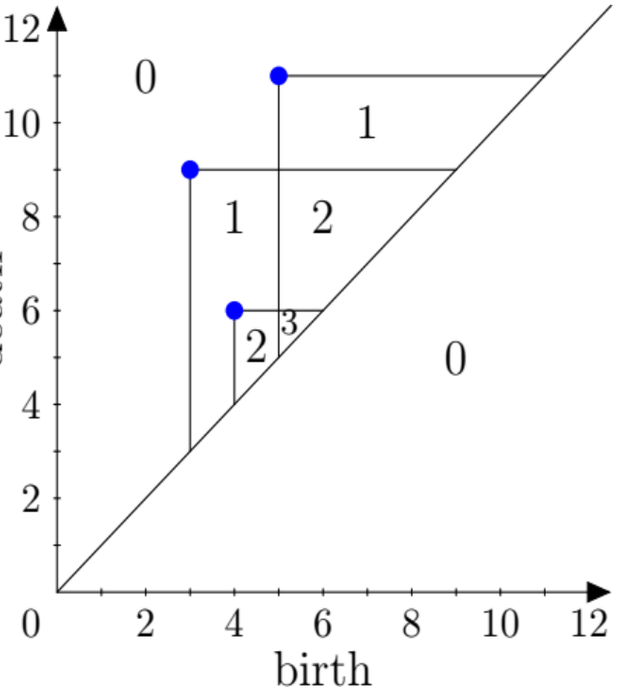
$x \leq y \Longrightarrow f^{-1}(-\infty, x) \subseteq f^{-1}(-\infty, y)$

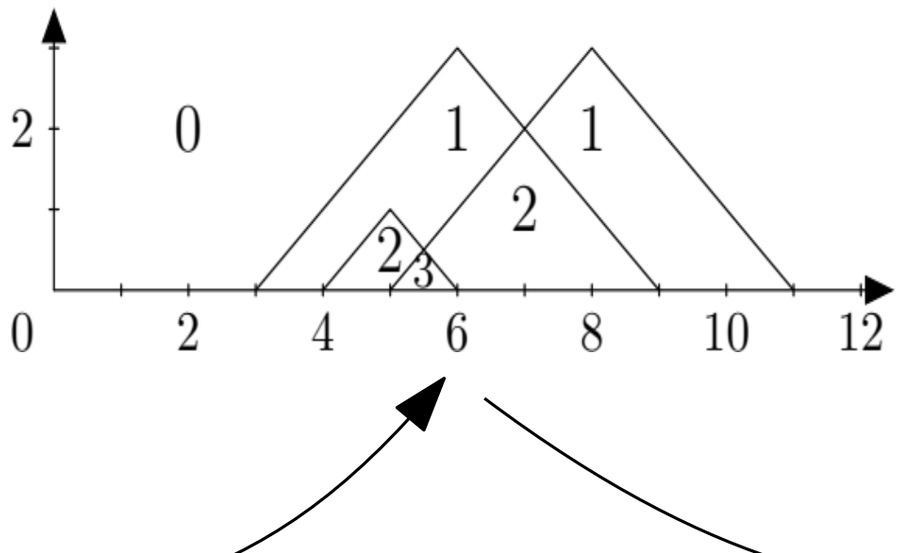
$\iota_x^y : H(f^{-1}(-\infty, x)) \to H(f^{-1}(-\infty, y))$ induced linear map

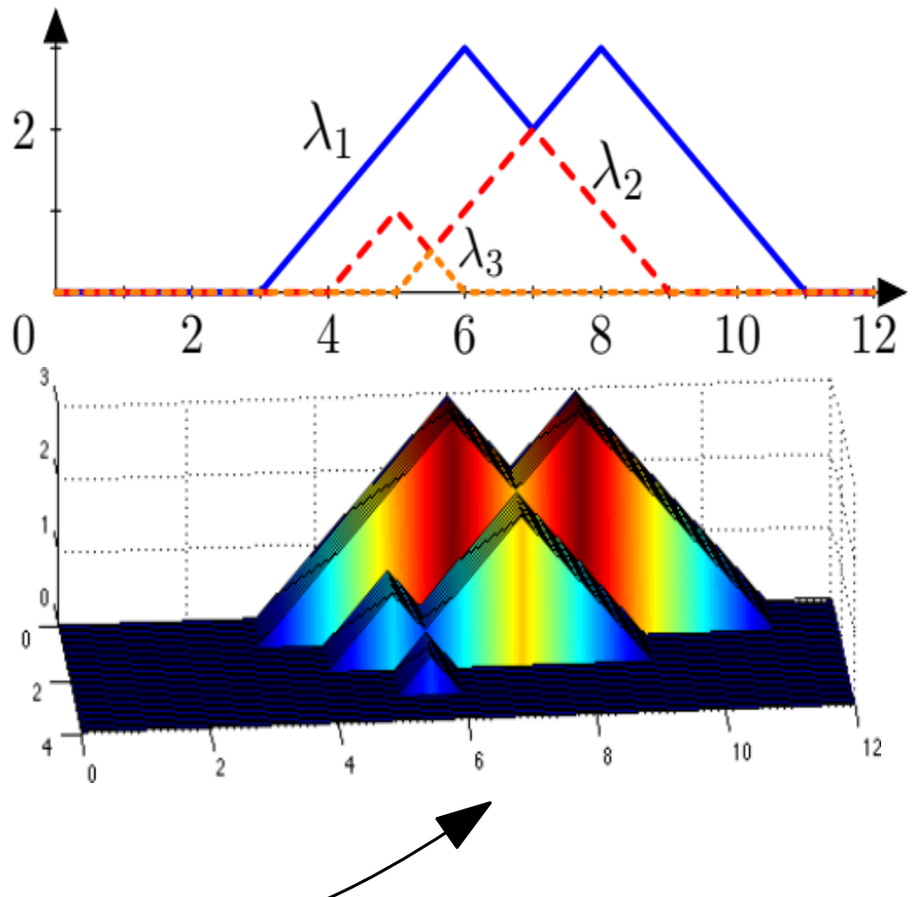
Rank function is defined as $\lambda(x, y) = \operatorname{rank} \iota_x^y$

Explicit Feature Map in Function Space

[*Statistical Topological Data Analysis using Persistence Landscapes*, Bubenik, JMLR, 2015]







Rotate PD
Compute rank function

Use boundaries of rank function

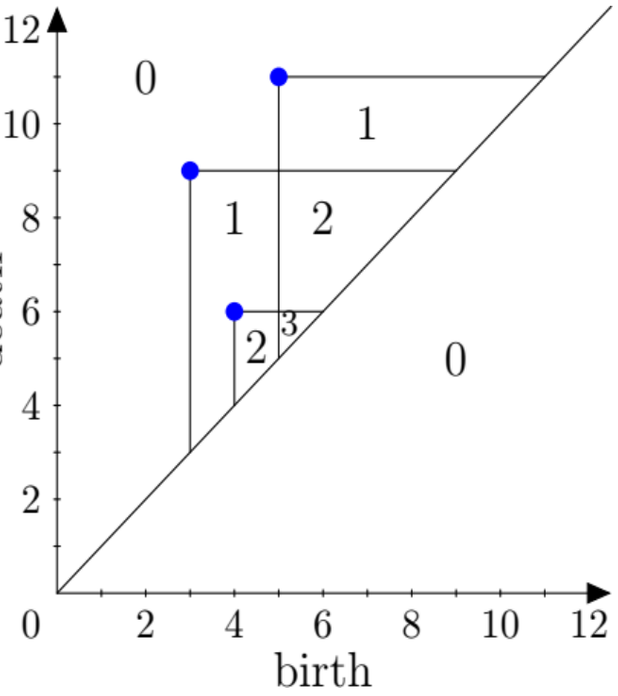
Boundaries of rank function: $\lambda_i(t) = \sup\{s \geq 0 : \lambda(t-s, t+s) \geq i\}$

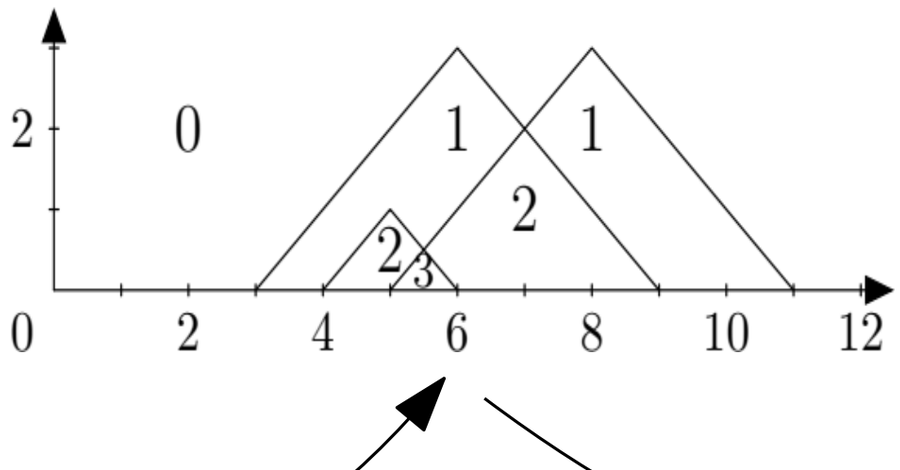
Landscape $\Lambda : \mathbb{R}^2 \to \mathbb{R}$ is defined as: $\Lambda(i,t) = \lambda_{\lfloor i \rfloor}(t)$

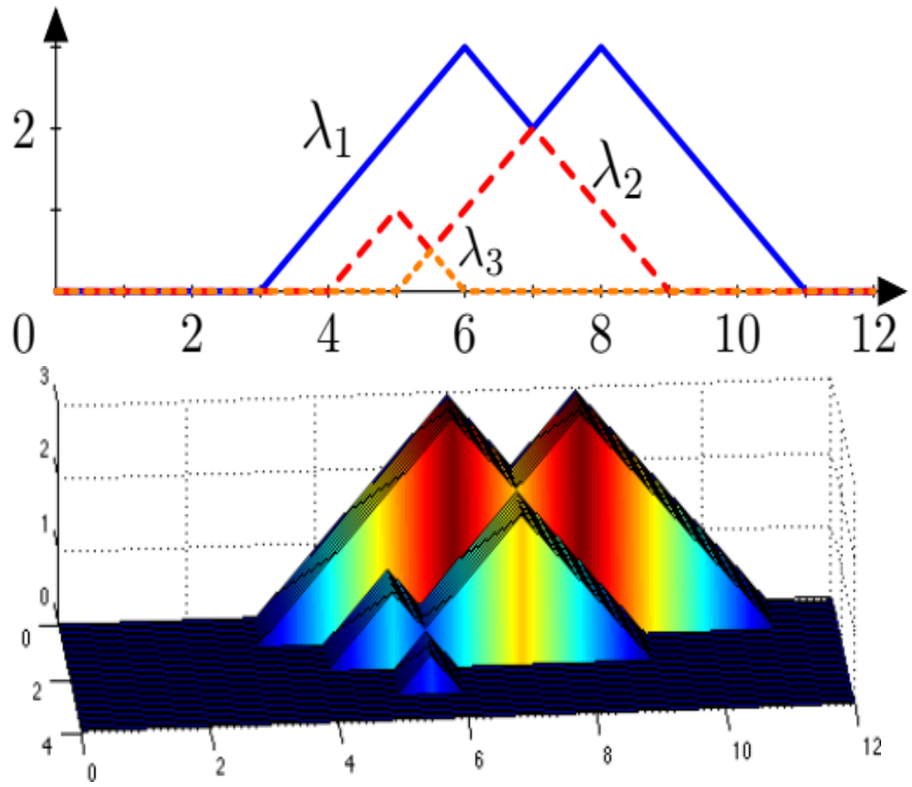
They can equivalently be defined as: $\Lambda(i,t) = i$-th $\max\{\lambda_j(t)\}$

Explicit Feature Map in Function Space

[*Statistical Topological Data Analysis using Persistence Landscapes*, Bubenik, JMLR, 2015]







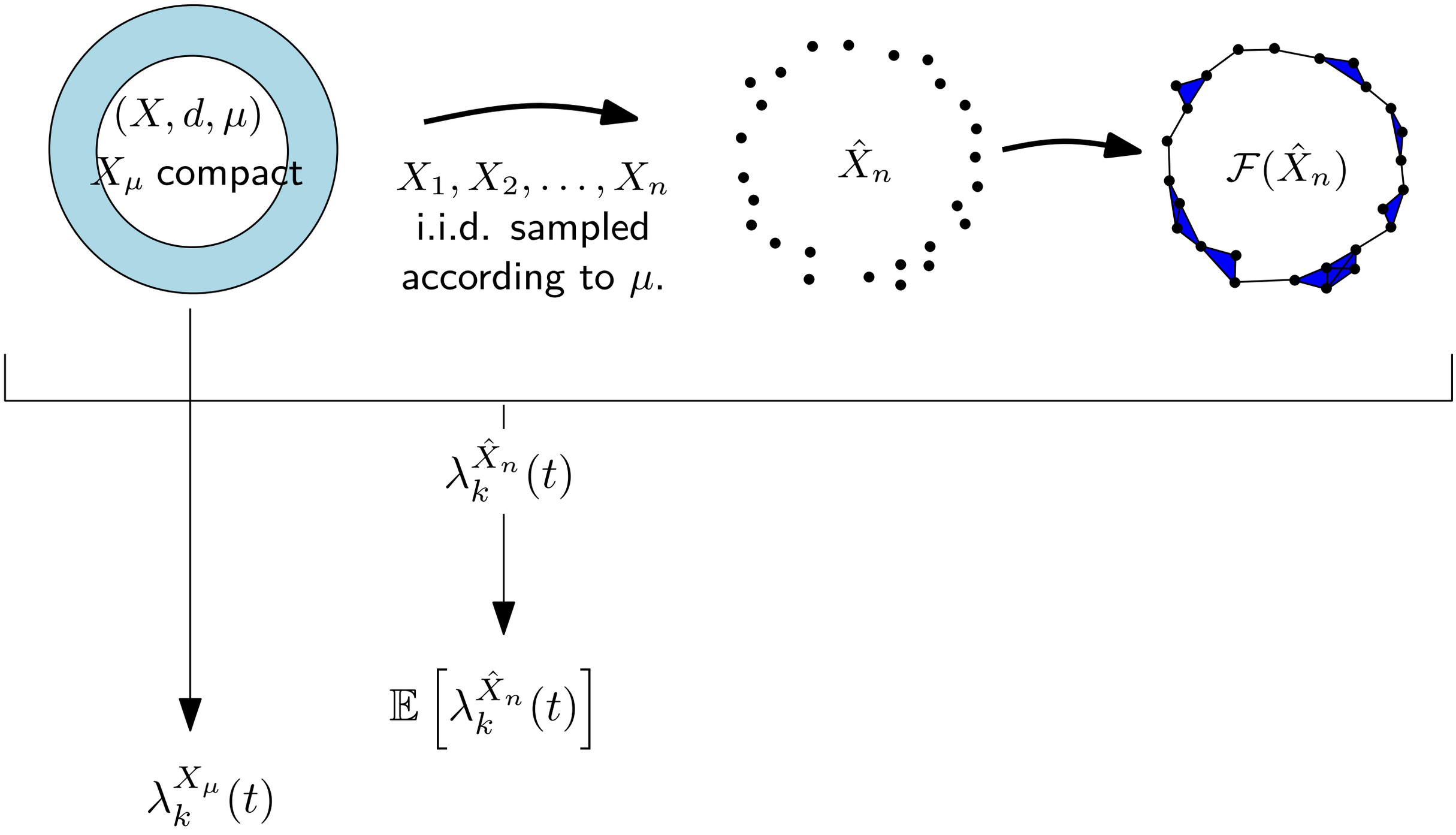
Rotate PD
Compute rank function

Use boundaries of rank function

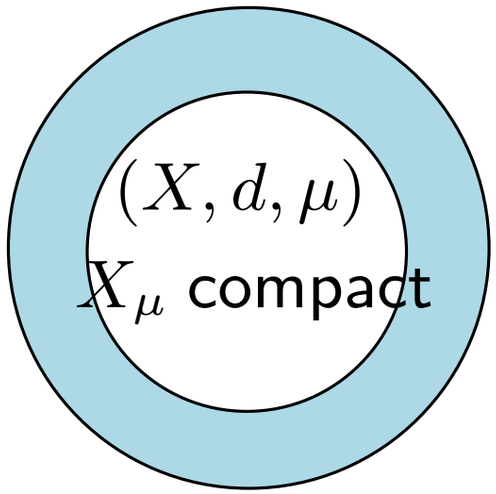
Prop: The following inequalities hold:

- $\|\Lambda(D) - \Lambda(D')\|_\infty \leq d_b(D, D')$.
- $\min\{1, C(D, D')\|\Lambda(D) - \Lambda(D')\|_2\} \leq d_2(D, D')$.

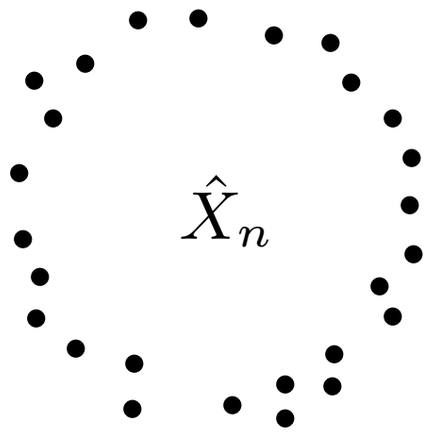
Statistics with landscapes

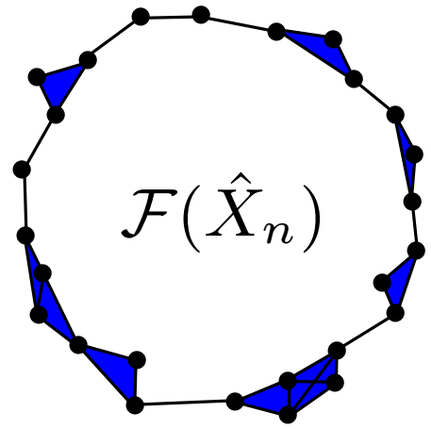


Statistics with landscapes



$X_1, X_2, \ldots, X_n$
i.i.d. sampled
according to μ.





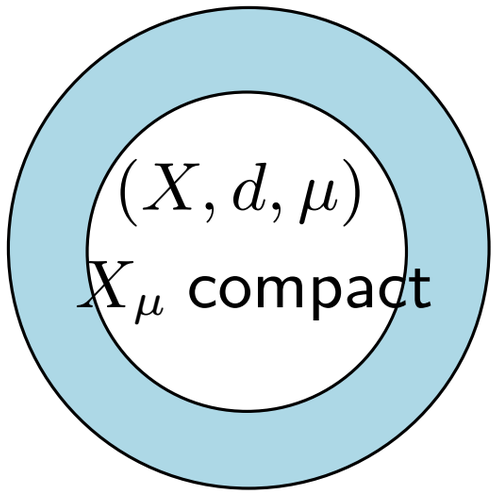
$\lambda_k^{\hat{X}_n}(t)$

$\lambda_k^{X_\mu}(t)$

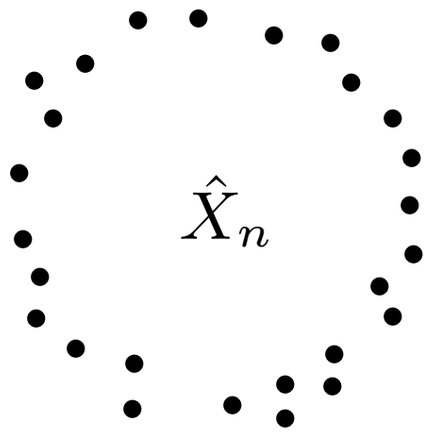
$\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t)\right]$

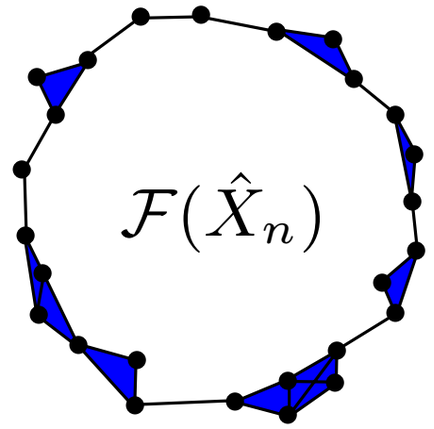
$\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t) - \lambda_k^{X_\mu}(t)\right] \to 0$ as $n \to +\infty$

Statistics with landscapes



$X_1, X_2, \ldots, X_n$
i.i.d. sampled
according to μ.





$\lambda_k^{\hat{X}_n}(t)$

$\lambda_k^{\hat{X}_n^1}(t)$ $\lambda_k^{\hat{X}_n^2}(t)$ $\lambda_k^{\hat{X}_n^3}(t)$ $\ldots$ $\lambda_k^{\hat{X}_n^m}(t)$

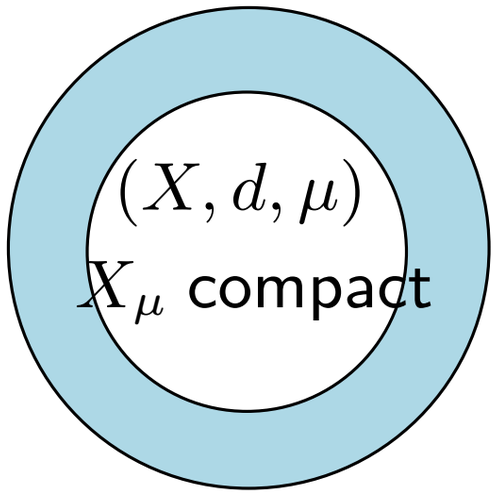
$\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t)\right]$

$\bar{\lambda}_k^m(t) := \frac{1}{m}\sum_{i=1}^m \lambda_k^{\hat{X}_n^i}(t)$

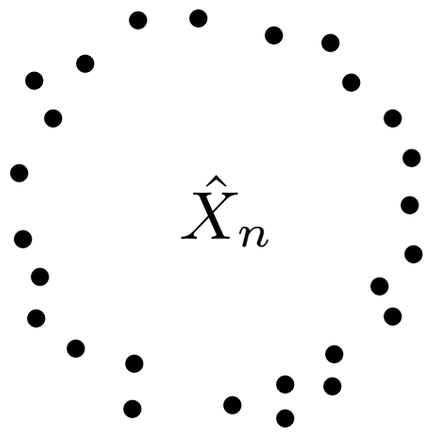
$\lambda_k^{X_\mu}(t)$

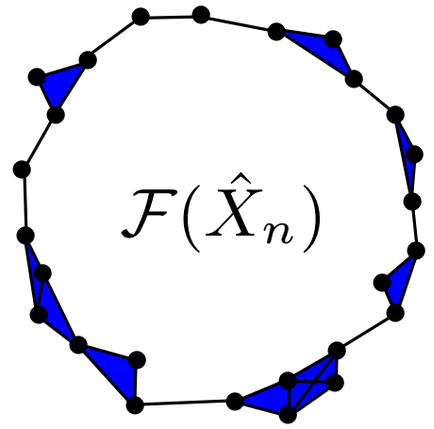
$\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t) - \lambda_k^{X_\mu}(t)\right] \to 0$ as $n \to +\infty$

Statistics with landscapes



$X_1, X_2, \ldots, X_n$
i.i.d. sampled
according to μ.





$\lambda_k^{\hat{X}_n}(t)$ $\quad$ $\lambda_k^{\hat{X}_n^1}(t)$ $\lambda_k^{\hat{X}_n^2}(t)$ $\lambda_k^{\hat{X}_n^3}(t)$ $\ldots$ $\lambda_k^{\hat{X}_n^m}(t)$

$\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t)\right]$? $\bar{\lambda}_k^m(t) := \frac{1}{m}\sum_{i=1}^m \lambda_k^{\hat{X}_n^i}(t)$

$\lambda_k^{X_\mu}(t)$

$\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t) - \lambda_k^{X_\mu}(t)\right] \to 0$ as $n \to +\infty$

Bootstrapping landscapes

[*Stochastic convergence of persistence landscapes and silhouettes*, Chazal et al., JoCG, 2015]

Thm: Suppose that $\mathrm{var}(\bar{\lambda}_k^m(t)) > c > 0$ in an interval $[t_*, t^*] \subset [0, T]$, for some constant c. Then, given a confidence level $1-\alpha$, one has:

$$\mathbb{P}\left(\left|\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t)\right] - \bar{\lambda}_k^m(t)\right| \leq \frac{Z_{B,\alpha}}{\sqrt{m}} \ \forall t \in [t_*, t^*]\right) \geq 1 - \alpha - O\left(\frac{(\log m)^{7/8}}{m^{1/8}}\right),$$

where $Z_{B,\alpha}$ is a quantile of a multiplier bootstrap distribution.

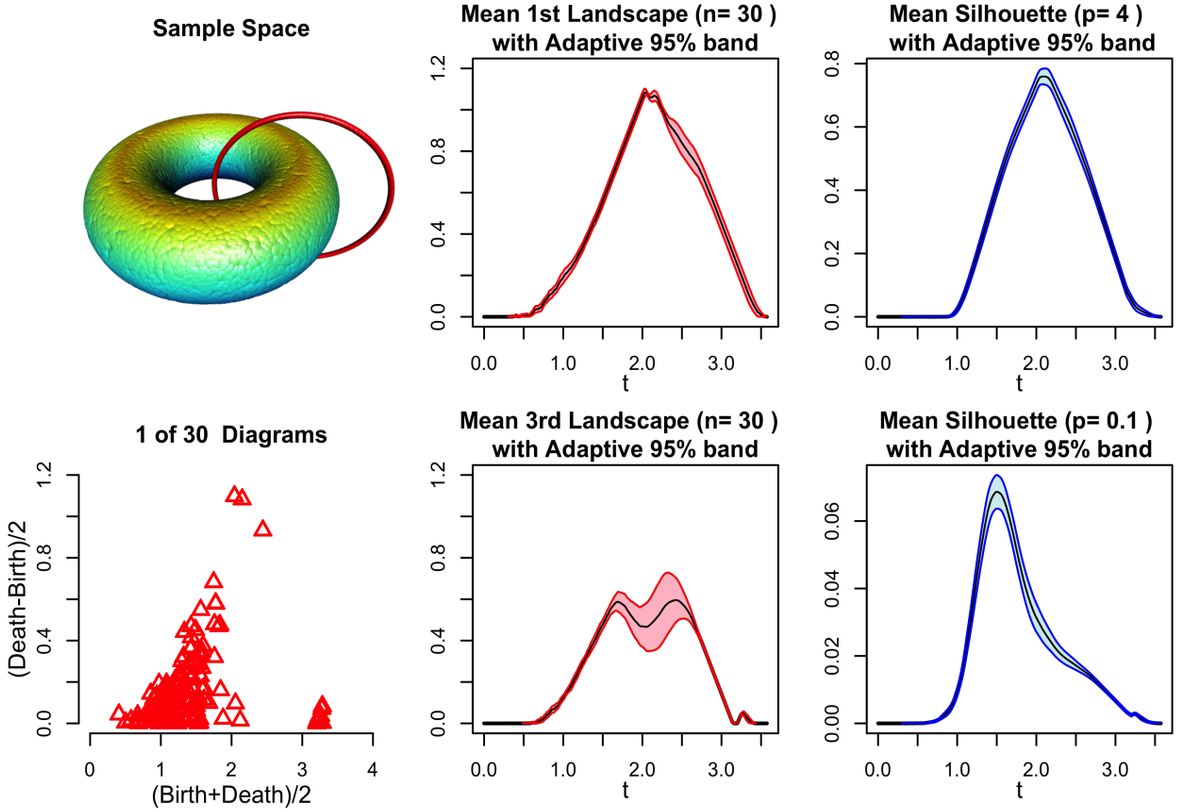
Bootstrapping landscapes

[*Stochastic convergence of persistence landscapes and silhouettes*, Chazal et al., JoCG, 2015]

Thm: Suppose that $\mathrm{var}(\bar{\lambda}_k^m(t)) > c > 0$ in an interval $[t_*, t^*] \subset [0, T]$, for some constant c. Then, given a confidence level $1-\alpha$, one has:

$$\mathbb{P}\left(\left|\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t)\right] - \bar{\lambda}_k^m(t)\right| \leq \frac{Z_{B,\alpha}}{\sqrt{m}} \ \forall t \in [t_*, t^*]\right) \geq 1 - \alpha - O\left(\frac{(\log m)^{7/8}}{m^{1/8}}\right),$$

where $Z_{B,\alpha}$ is a quantile of a multiplier bootstrap distribution.



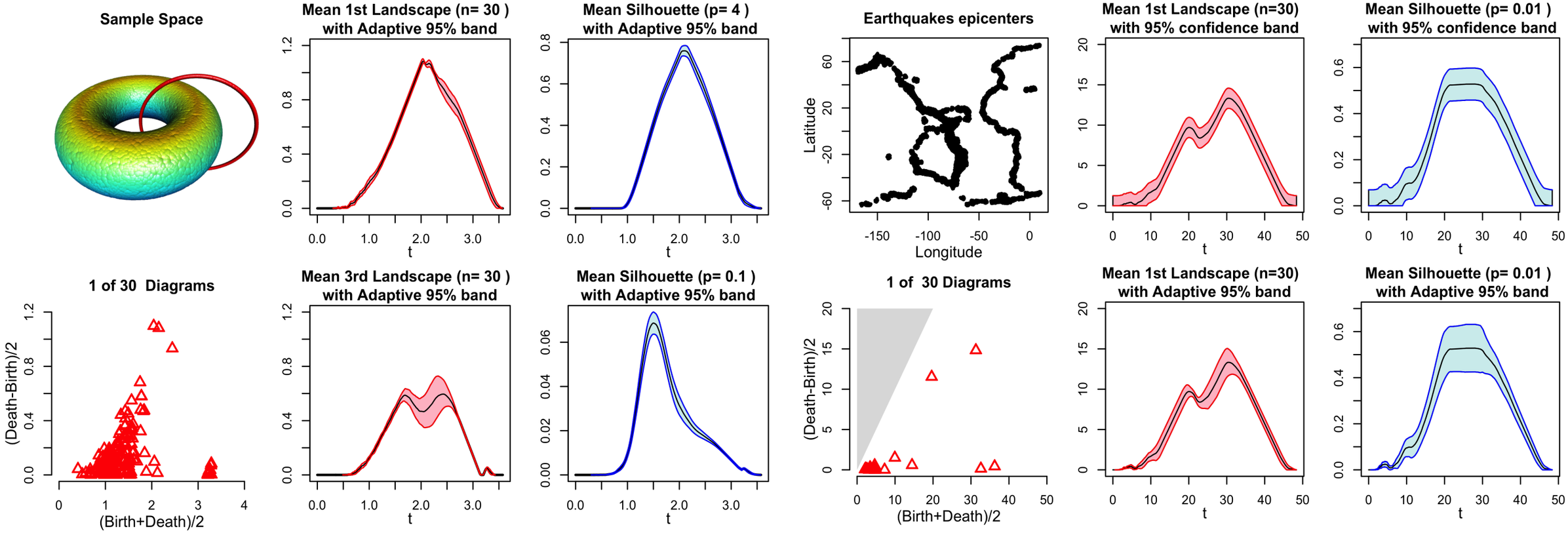
Bootstrapping landscapes

[*Stochastic convergence of persistence landscapes and silhouettes*, Chazal et al., JoCG, 2015]

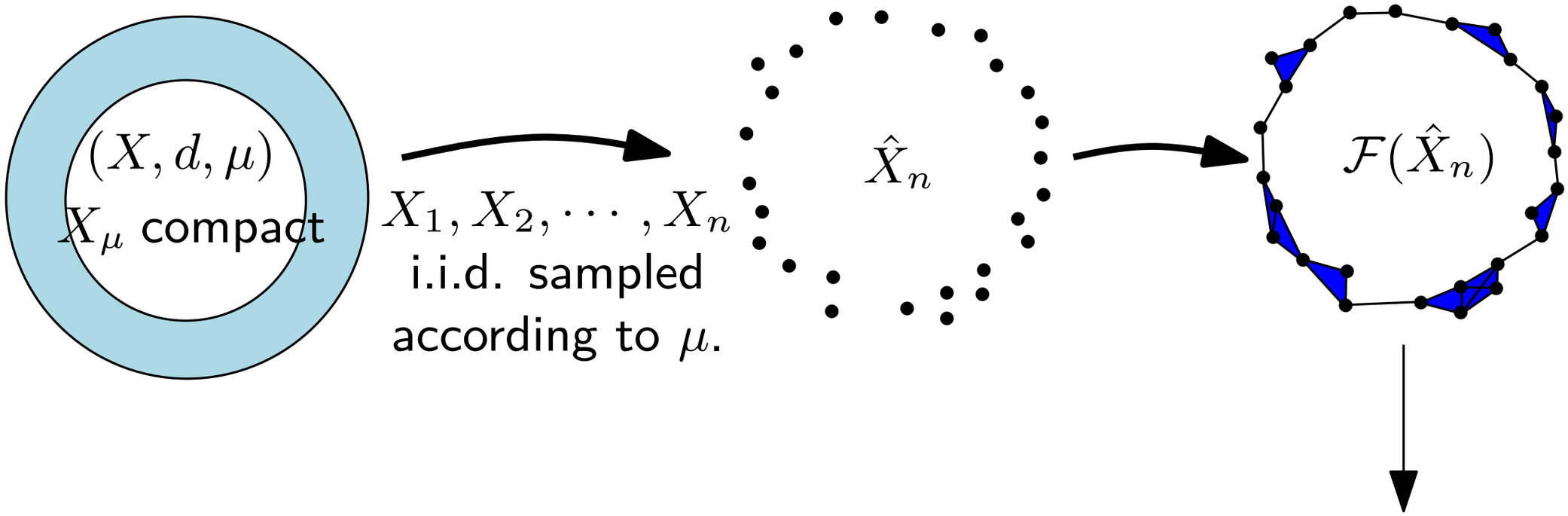
Thm: Suppose that $\mathrm{var}(\bar{\lambda}_k^m(t)) > c > 0$ in an interval $[t_* , t^*] \subset [0, T]$, for some constant c. Then, given a confidence level $1 - \alpha$, one has:

$$\mathbb{P}\left(\left|\mathbb{E}\left[\lambda_k^{\hat{X}_n}(t)\right] - \bar{\lambda}_k^m(t)\right| \leq \frac{Z_{B,\alpha}}{\sqrt{m}} \ \forall t \in [t_* , t^*]\right) \geq 1 - \alpha - O\left(\frac{(\log m)^{7/8}}{m^{1/8}}\right),$$

where $Z_{B,\alpha}$ is a quantile of a multiplier bootstrap distribution.

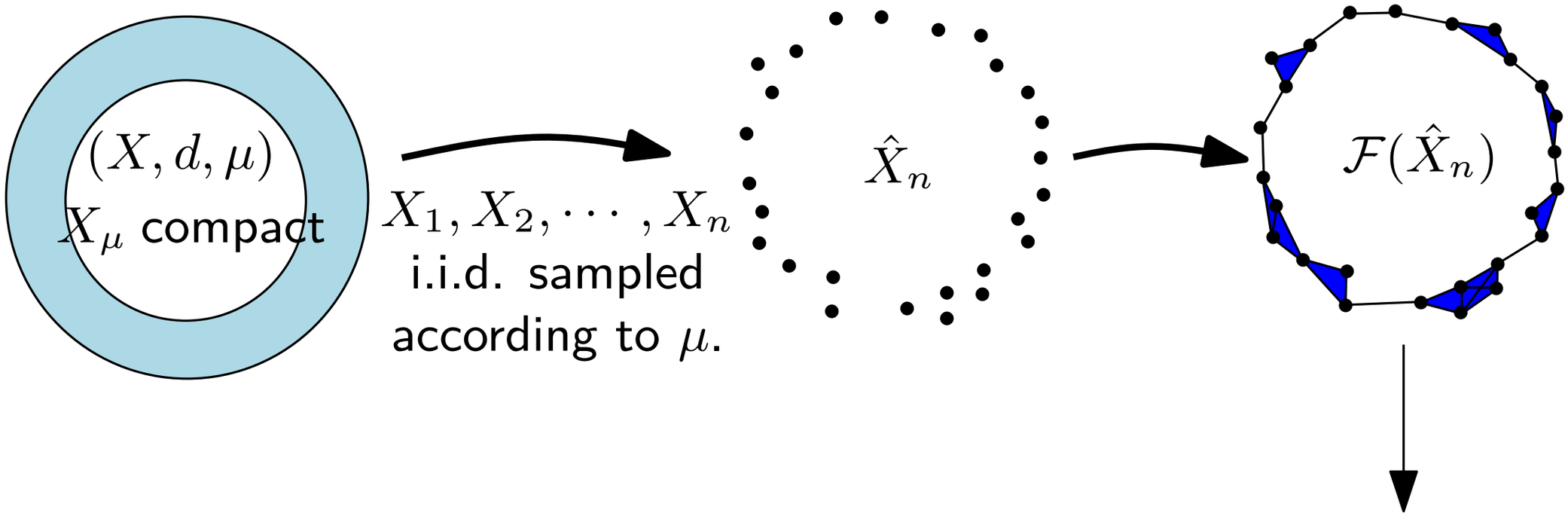


Stability of the mean landscape



$$\Lambda_{\mu,n}(t) = \mathbb{E}_{P_\mu}\left[\lambda_k^{\hat{X}_n}(t)\right]$$

Stability of the mean landscape

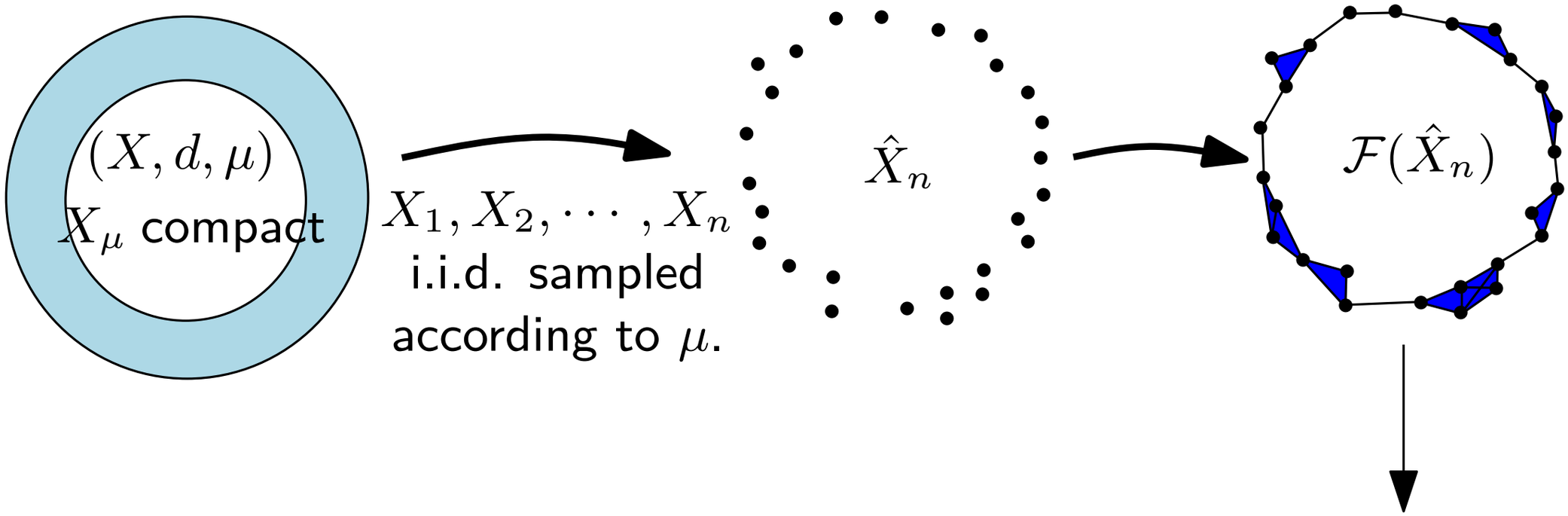


$$\Lambda_{\mu,n}(t) = \mathbb{E}_{P_\mu}\left[\lambda_k^{\hat{X}_n}(t)\right]$$

How bad is the dependence in μ?

Stability of the mean landscape

[*Subsampling methods for persistent homology*, Chazal et al., ICML, 2015]



$$\Lambda_{\mu,n}(t) = \mathbb{E}_{P_\mu}\left[\lambda_k^{\hat{X}_n}(t)\right]$$

How bad is the dependence in μ?

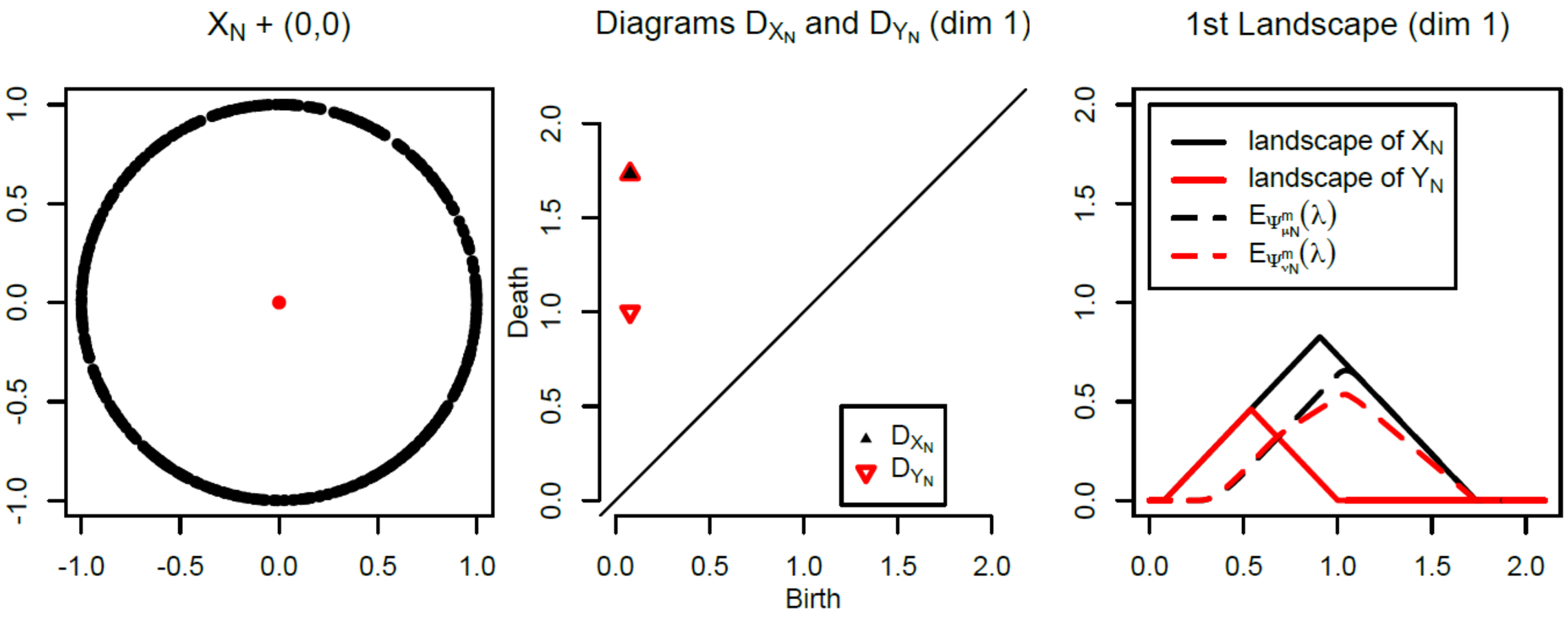
Thm: Let (X, d) be a metric space and let μ, ν be probability measures on X with compact supports. Then one has:

$$\|\Lambda_{\mu,n} - \Lambda_{\nu,n}\|_\infty \leq n^{\frac{1}{p}} W_p(\mu, \nu),$$

where W_p denotes the Wasserstein distance with cost function $d(\cdot, \cdot)^p$.

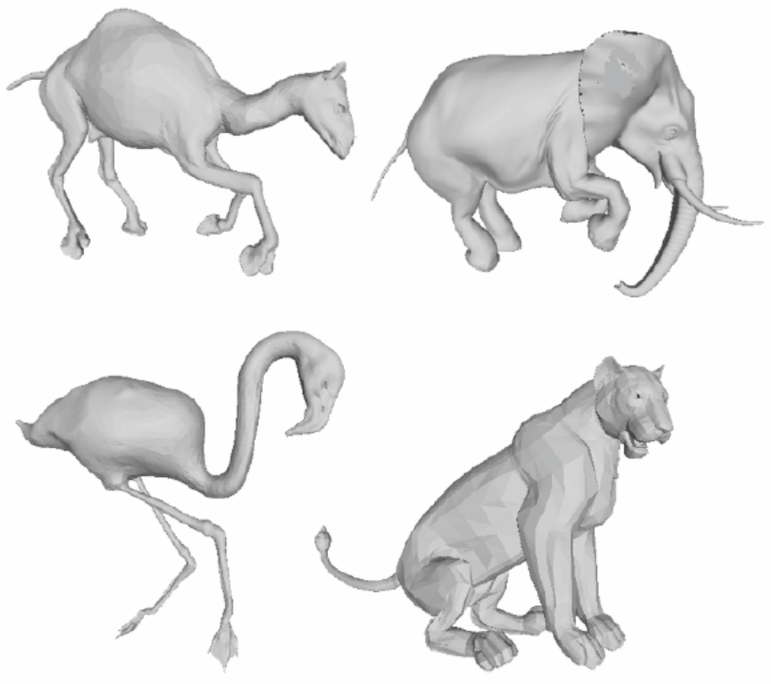
Numerical illustrations: confidence for landscapes

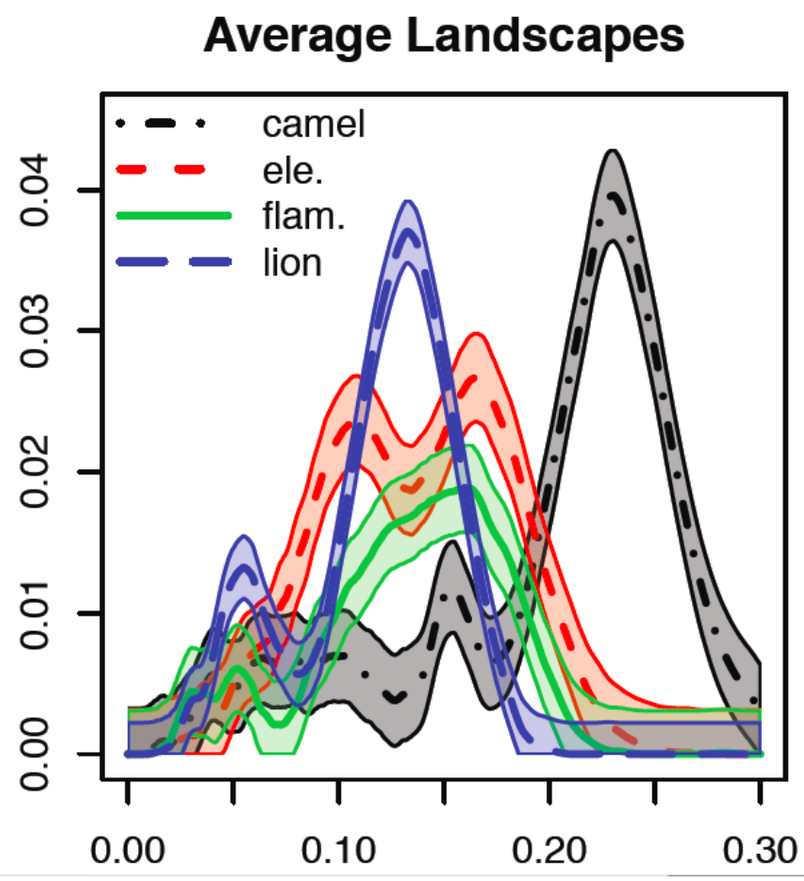
Example: Circle with one outlier.

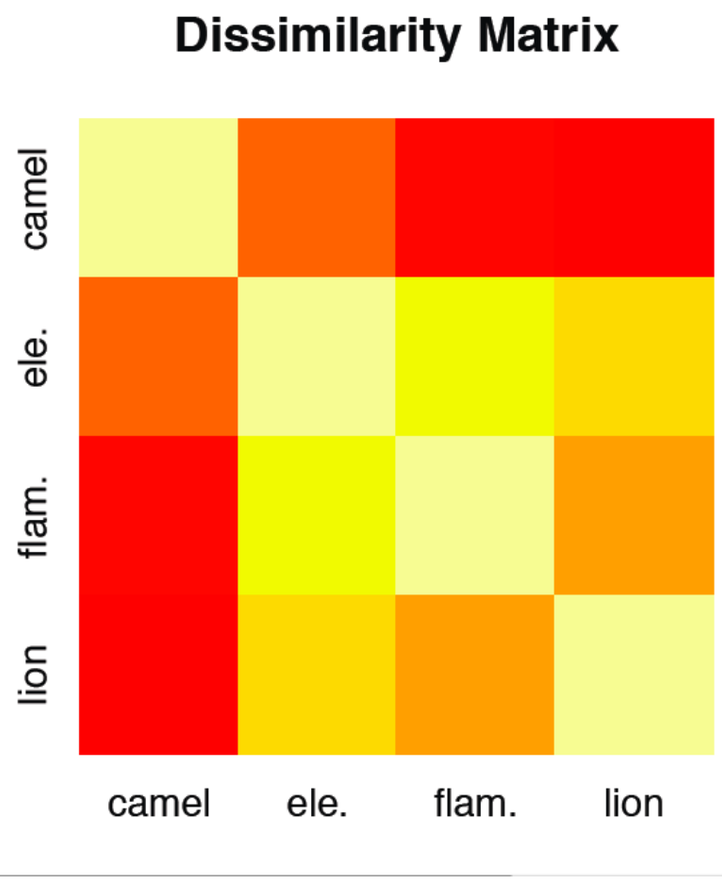


Numerical illustrations: confidence for landscapes

Example: 3D shapes

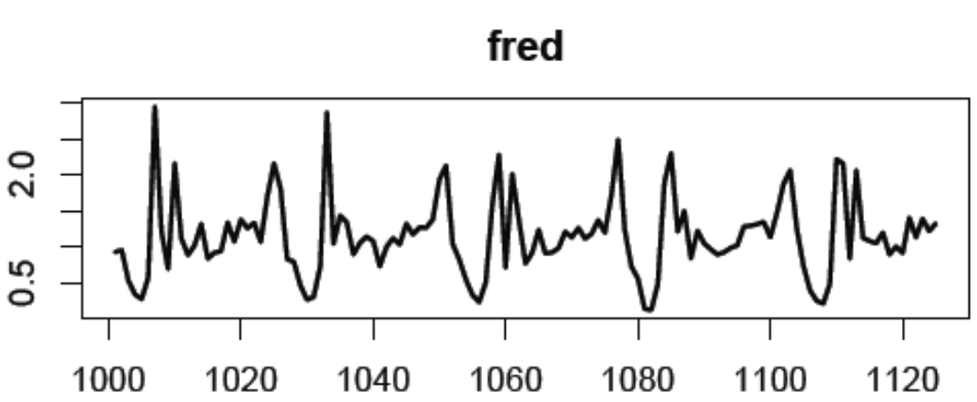


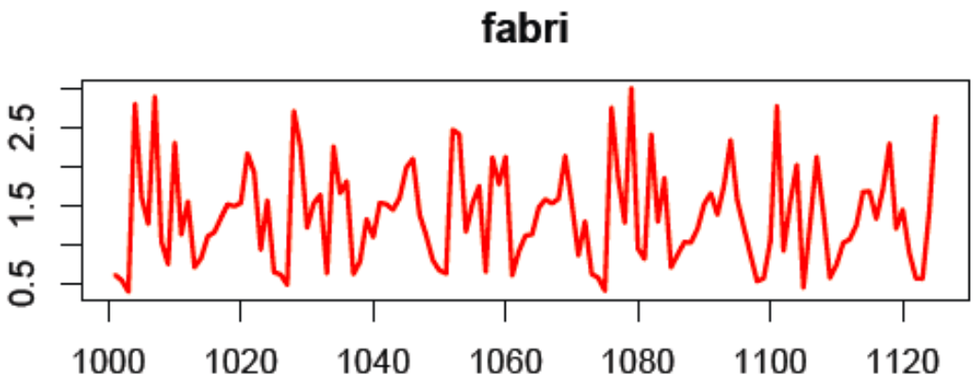


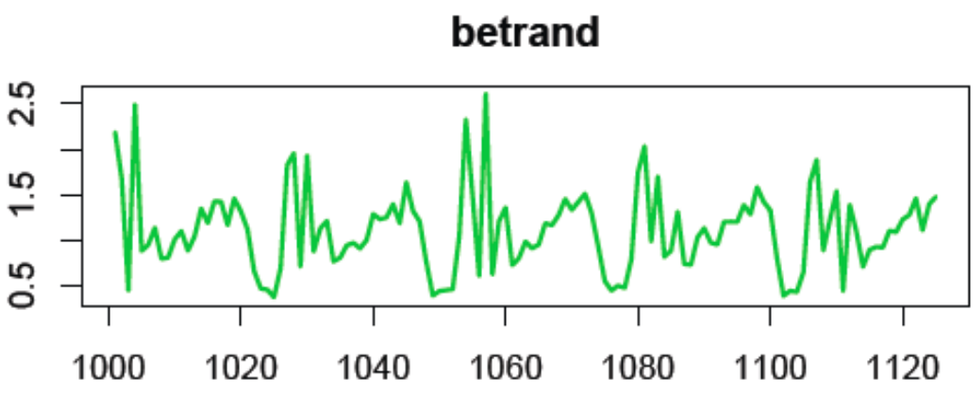


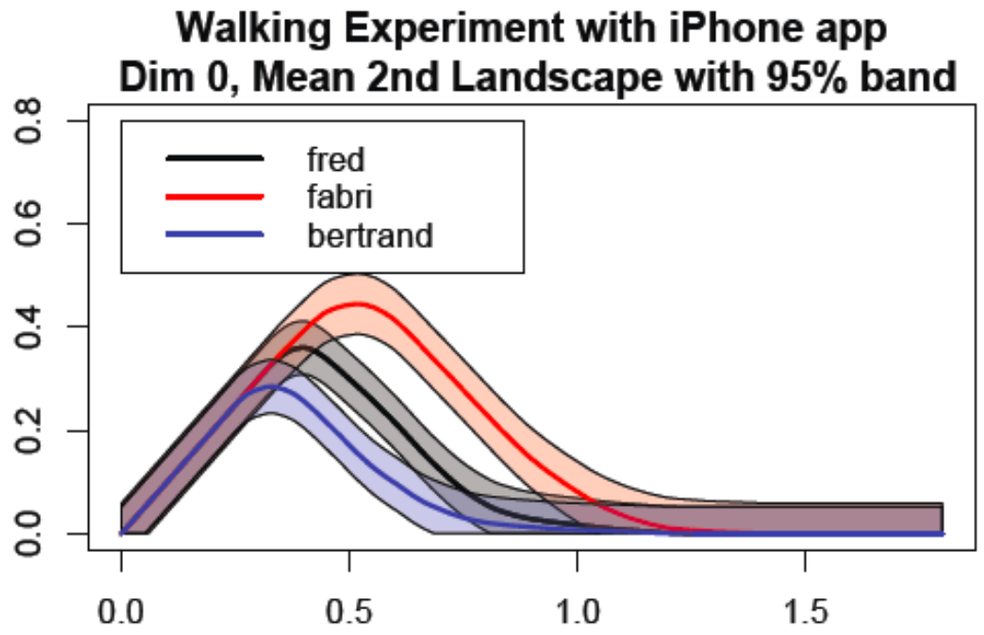
Numerical illustrations: confidence for landscapes

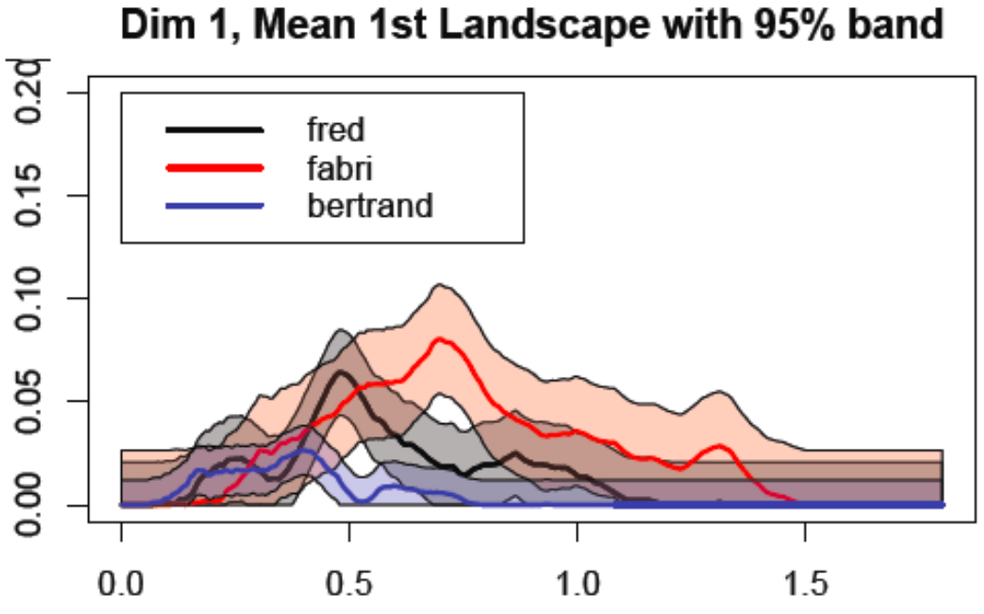
Example: Accelerometer data from smartphone.











I. Topological Inference on Persistence Diagrams for Point Clouds

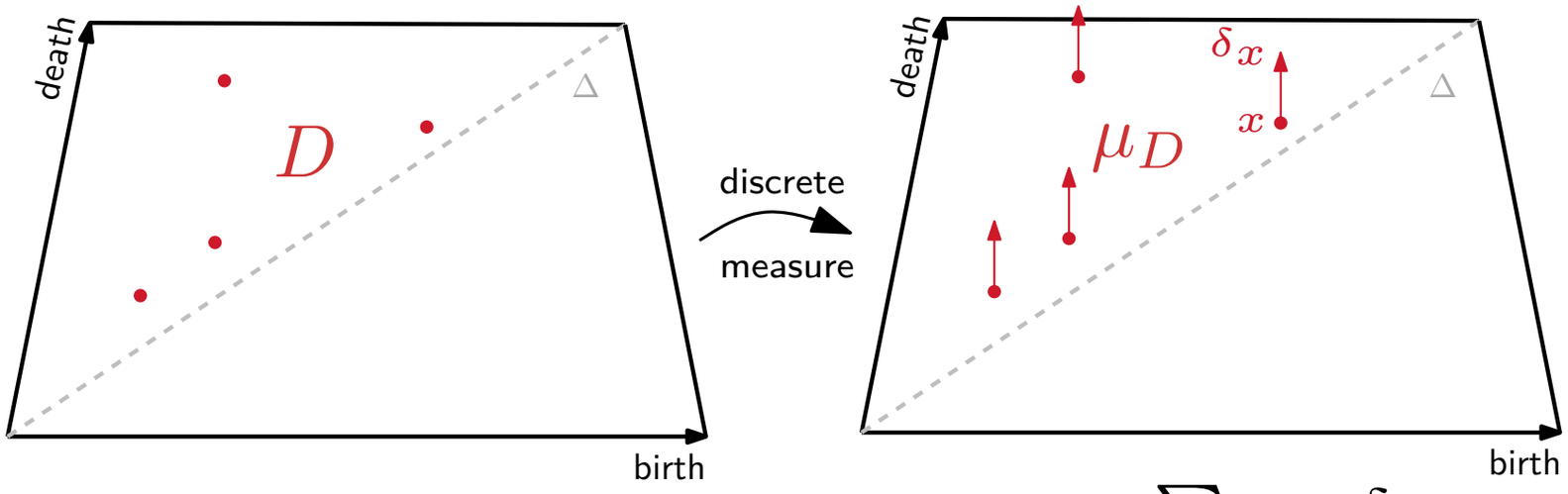
II. Topological Vectorizations for Inference and Machine Learning

(III. A Gaussian Kernel for Persistence Diagrams)

IV. Deep Learning for Persistence Diagrams

Explicit Feature Map in Function Space

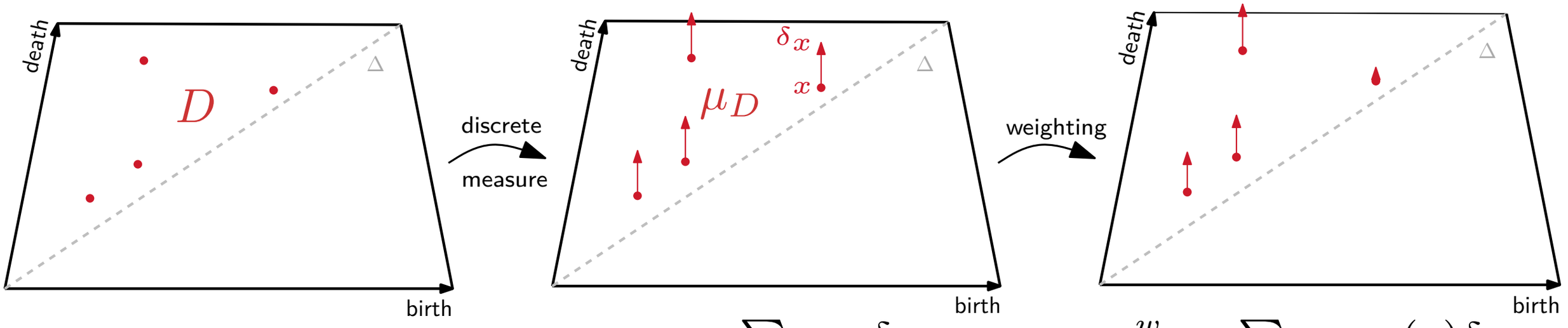
[*Persistence weighted Gaussian kernel for topological data analysis*, Kusano, Hiraoka, Fukumizu, ICML, 2016]



$$\mu_D := \sum_{x \in D} \delta_x$$

Explicit Feature Map in Function Space

[*Persistence weighted Gaussian kernel for topological data analysis*, Kusano, Hiraoka, Fukumizu, ICML, 2016]

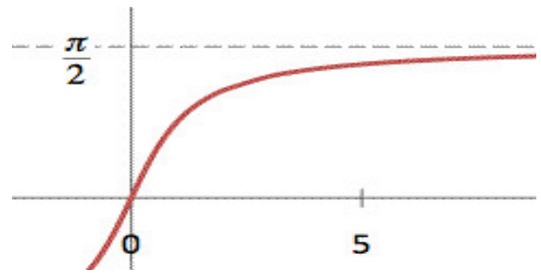


$$\mu_D := \sum_{x \in D} \delta_x$$

$$\mu_D^w := \sum_{x \in D} w(x)\delta_x$$

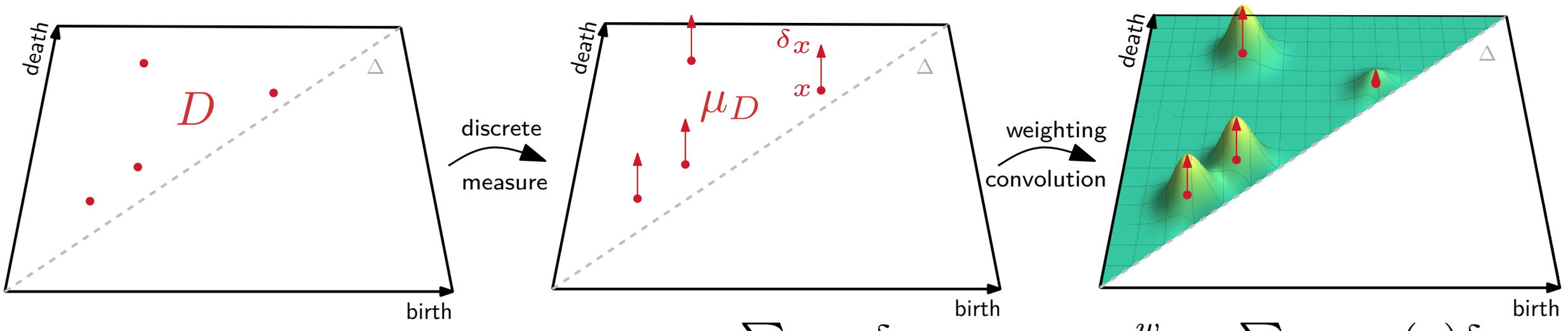
Pb: μ_D is unstable (points on diagonal disappear)

$w(x) := \arctan\left(c\, d(x, \Delta)^r\right),\ c, r > 0$



Explicit Feature Map in Function Space

[*Persistence weighted Gaussian kernel for topological data analysis*, Kusano, Hiraoka, Fukumizu, ICML, 2016]



$$\mu_D := \sum_{x \in D} \delta_x$$

$$\mu_D^w := \sum_{x \in D} w(x) \delta_x$$

$$\tilde{\mu}_D^w := \mu_D^w * \mathcal{N}(0, \sigma)$$

Pb: μ_D is unstable (points on diagonal disappear)

$$w(x) := \arctan\left(c\, d(x, \Delta)^r\right),\ c, r > 0$$

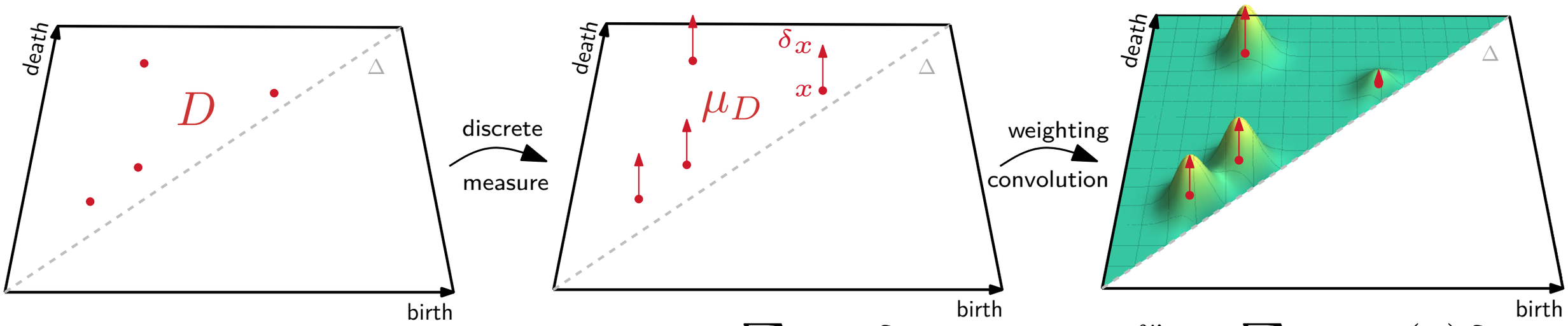
Def: $\phi(D)$ is the density function of $\mu_D^w * \mathcal{N}(0, \sigma)$ (w.r.t. Lebesgue measure):

$$\phi(D) := \frac{1}{\sqrt{2\pi}\sigma} \sum_{x \in D} \arctan(c\, d(x, \Delta)^r)\, \exp\left(-\frac{\| \cdot - x\|^2}{2\sigma^2}\right)$$

$$k(D, D') := \langle \phi(D),\, \phi(D') \rangle_{L_2(\Delta \times \mathbb{R}_+)}$$

Explicit Feature Map in Function Space

[*Persistence weighted Gaussian kernel for topological data analysis*, Kusano, Hiraoka, Fukumizu, ICML, 2016]



$$\mu_D := \sum_{x \in D} \delta_x$$

$$\mu_D^w := \sum_{x \in D} w(x) \delta_x$$

$$\tilde{\mu}_D^w := \mu_D^w * \mathcal{N}(0, \sigma)$$

Prop: The following hold:

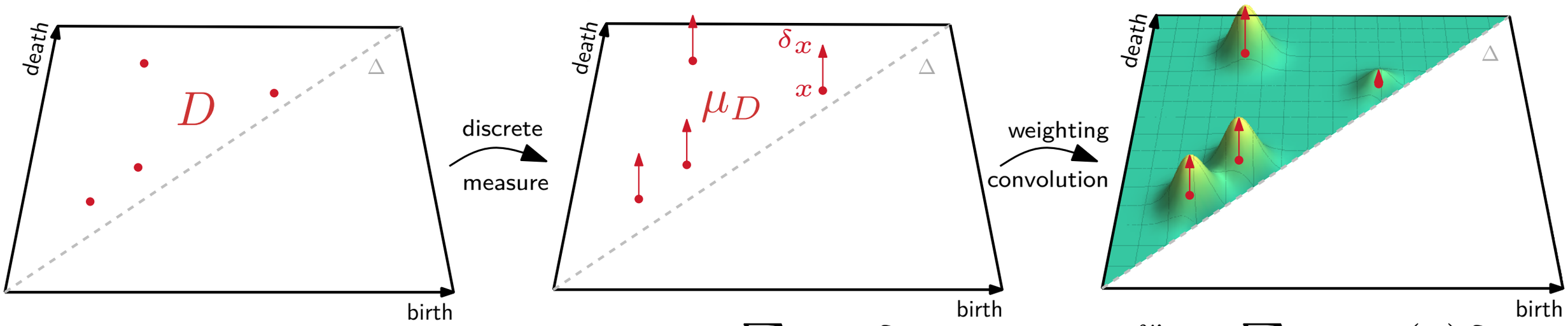
- $\|\phi(D) - \phi(D')\|_{\mathcal{H}} \leq C\, d_p(D, D')$.
- ϕ is injective and $\exp(k)$ is universal

$$\phi(D) := \frac{1}{\sqrt{2\pi}\sigma} \sum_{x \in D} \arctan(c\, d(x, \Delta)^r)\, \exp\left(-\frac{\|\cdot - x\|^2}{2\sigma^2}\right)$$

$$k(D, D') := \langle \phi(D),\, \phi(D') \rangle_{L_2(\Delta \times \mathbb{R}_+)}$$

Explicit Feature Map in Function Space

[*Persistence weighted Gaussian kernel for topological data analysis*, Kusano, Hiraoka, Fukumizu, ICML, 2016]



$$\mu_D := \sum_{x \in D} \delta_x$$

$$\mu_D^w := \sum_{x \in D} w(x) \delta_x$$

$$\tilde{\mu}_D^w := \mu_D^w * \mathcal{N}(0, \sigma)$$

Prop: The following hold:

- $\|\phi(D) - \phi(D')\|_{\mathcal{H}} \leq C\, d_p(D, D')$.
- ϕ is injective and $\exp(k)$ is universal

Pb: convolution reduces discriminativity $\rightarrow$ use discrete measure instead

$$\phi(D) := \frac{1}{\sqrt{2\pi}\sigma} \sum_{x \in D} \arctan(c\, d(x, \Delta)^r)\, \exp\left(-\frac{\|\cdot - x\|^2}{2\sigma^2}\right)$$

$$k(D, D') := \langle \phi(D), \phi(D') \rangle_{L_2(\Delta \times \mathbb{R}_+)}$$

Kernels for persistence diagrams

	images	metric spaces	polynomials	landscapes	discrete measures
ambient Hilbert space	$(\mathbb{R}^d, \|.\|_2)$	$(\mathbb{R}^d, \|.\|_2)$	$\ell_2(\mathbb{R})$	$L_2(\mathbb{N}\times\mathbb{R})$	$L_2(\mathbb{R}^2)$
positive (semi-)definiteness	✓	✓	✓	✓	✓
$\|\phi(\cdot)-\phi(\cdot)\|_{\mathcal{H}} \leq \phi(d_p)$	✓	✓	✓	✓	✓
$\|\phi(\cdot)-\phi(\cdot)\|_{\mathcal{H}} \geq \psi(d_p)$	✗	✗	✗	✗	✗
injectivity	✗	✗	✓	✓	✓
universality	✗	✗	✗	✗	✓
algorithmic cost	f. map: $O(n^2)$ kernel: $O(d)$	f. map: $O(n^2)$ kernel: $O(d)$	f. map: $O(nd)$ kernel: $O(d)$	$O(n^2)$	$O(n^2)$

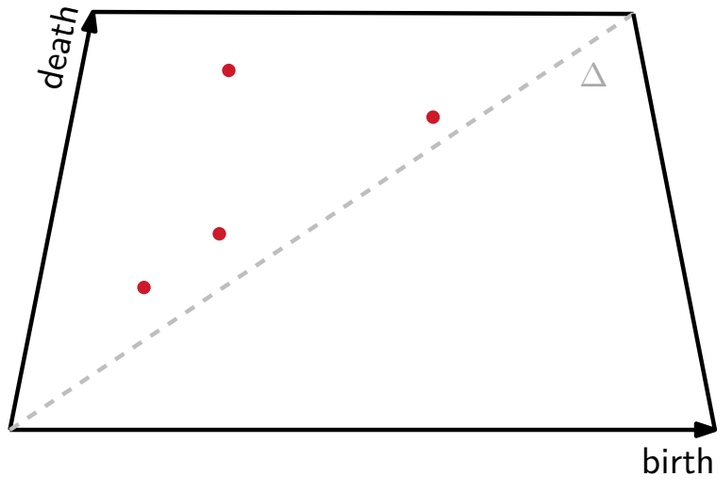
One kernel to rule them all...

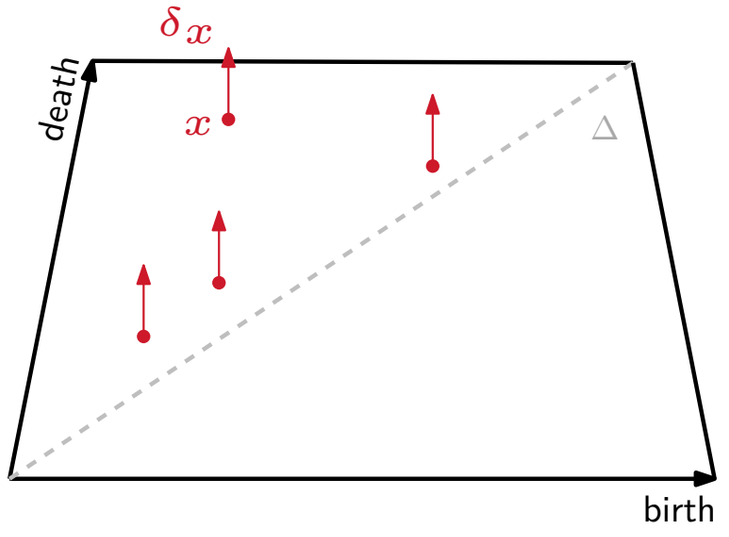
[*Sliced Wasserstein Kernel for persistence diagrams*, C., Cuturi, Oudot, ICML, 2017]

The Sliced Wasserstein Kernel is a kernel for PDs with:

- No feature map
- Provably stable
- Provably discriminative
- Mimicks the Gaussian kernel

Persistence diagrams as discrete measures

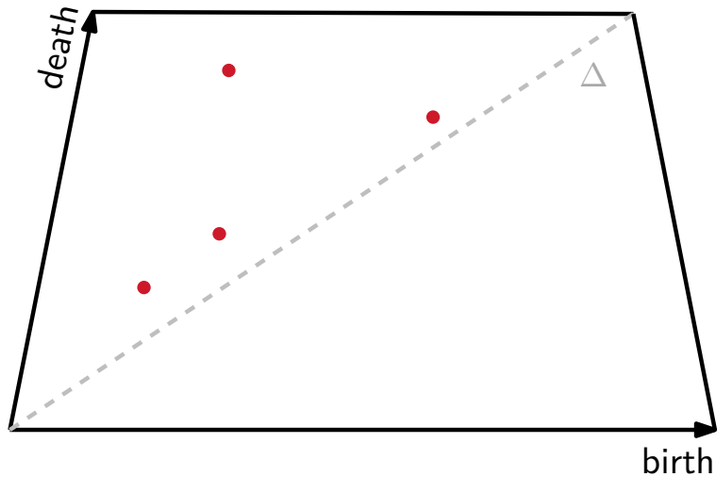


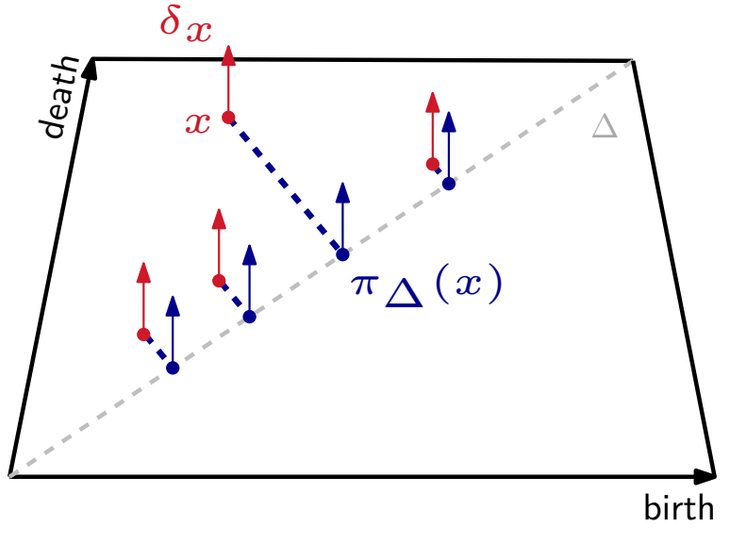


$$\mu_D := \sum_{x \in D} \delta_x$$

Pb: $d_p(D, D') \not\propto W_p(\mu_D, \mu_{D'})$ (W_p does not even make sense here)

Persistence diagrams as discrete measures





$$\mu_D := \sum_{x \in D} \delta_x$$

Pb: $d_p(D, D') \not\propto W_p(\mu_D, \mu_{D'})$ (W_p does not even make sense here)

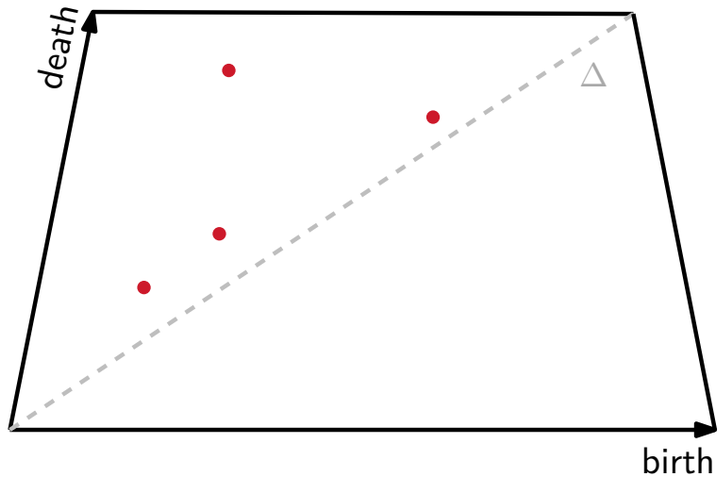
$\rightarrow$ given D, D', let

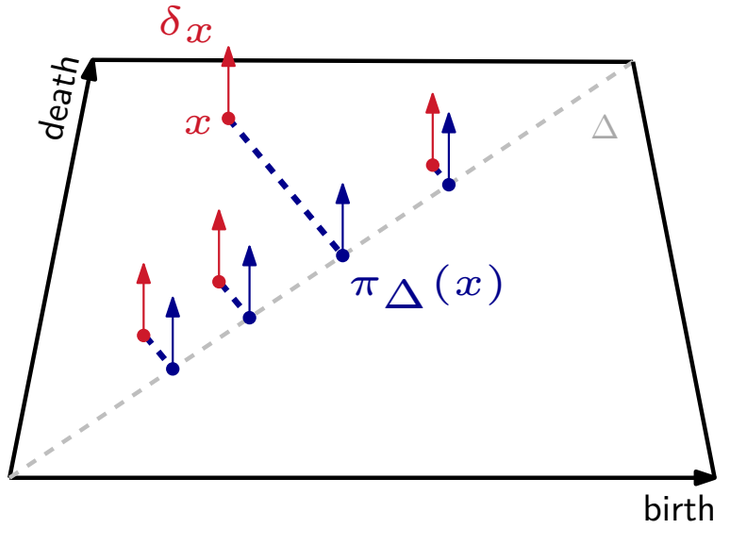
$$\bar{\mu}_D := \sum_{x \in D} \delta_x + \sum_{y \in D'} \delta_{\pi_\Delta(y)}$$

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

Then, $d_p(D, D') \leq W_p(\bar{\mu}_D, \bar{\mu}_{D'}) \leq 2\, d_p(D, D')$

Persistence diagrams as discrete measures





$$\mu_D := \sum_{x \in D} \delta_x$$

Pb: $d_p(D, D') \not\propto W_p(\mu_D, \mu_{D'})$ (W_p does not even make sense here)

$\rightarrow$ given D, D', let

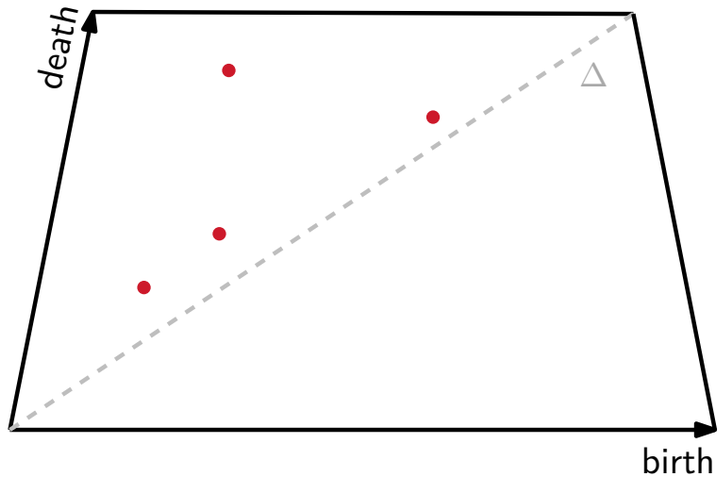
$$\bar{\mu}_D := \sum_{x \in D} \delta_x + \sum_{y \in D'} \delta_{\pi_\Delta(y)}$$

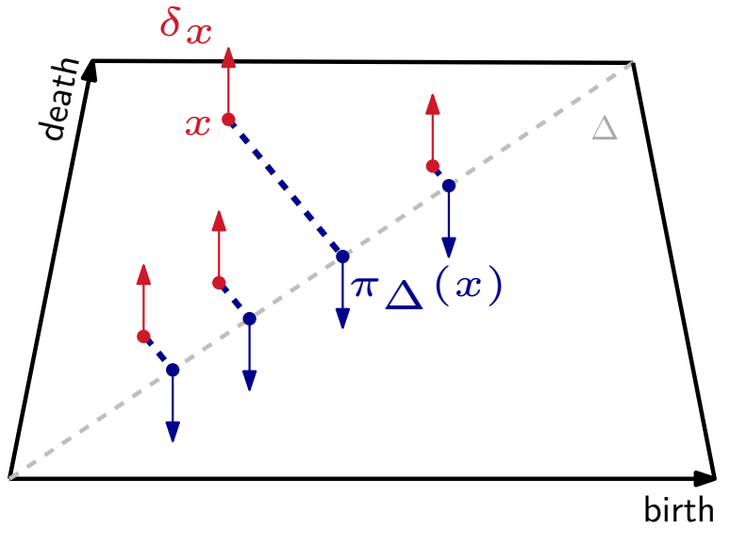
$$\bar{\mu}_{D'} := \sum_{y \in D'} \delta_y + \sum_{x \in D} \delta_{\pi_\Delta(x)}$$

Then, $d_p(D, D') \leq W_p(\bar{\mu}_D, \bar{\mu}_{D'}) \leq 2\, d_p(D, D')$

Pb: $\bar{\mu}_D$ depends on D'

Persistence diagrams as discrete measures





$$\mu_D := \sum_{x \in D} \delta_x$$

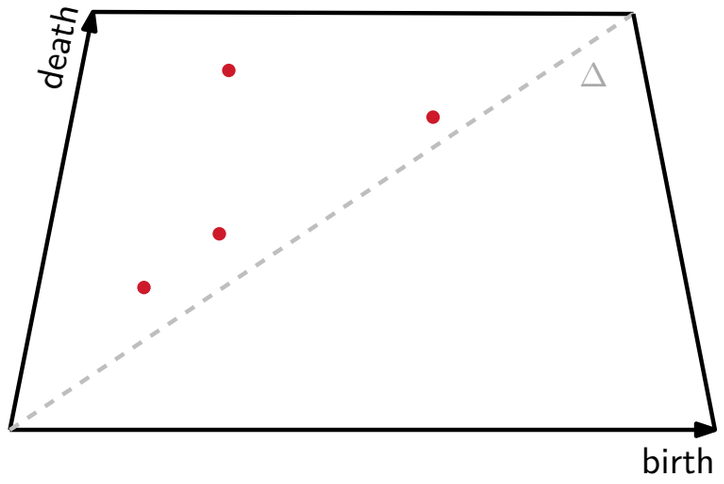
Pb: $d_p(D, D') \not\propto W_p(\mu_D, \mu_{D'})$ (W_p does not even make sense here)

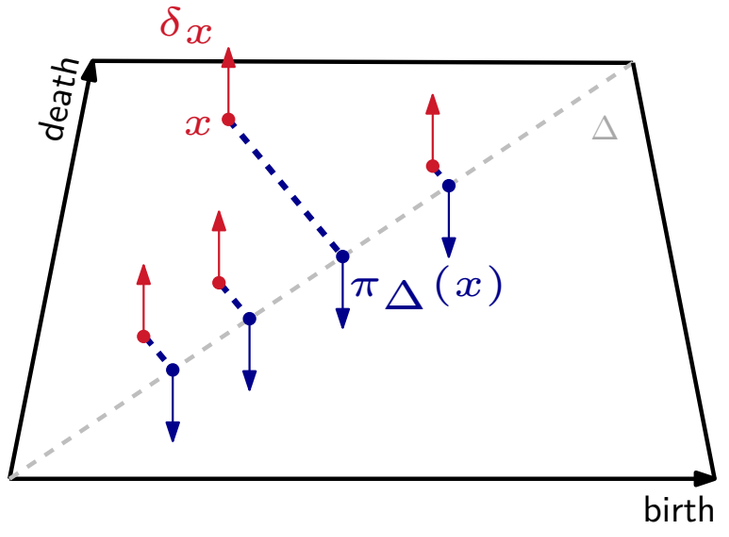
Solution: transfer mass negatively to μ_D:

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

$\rightarrow$ signed discrete measure of total mass zero

Persistence diagrams as discrete measures





$$\mu_D := \sum_{x \in D} \delta_x$$

Pb: $d_p(D, D') \not\propto W_p(\mu_D, \mu_{D'})$ (W_p does not even make sense here)

Solution: transfer mass negatively to μ_D:

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

$\rightarrow$ signed discrete measure of total mass zero with Kantorovich norm $\| \cdot \|_K$.

Persistence diagrams as discrete measures

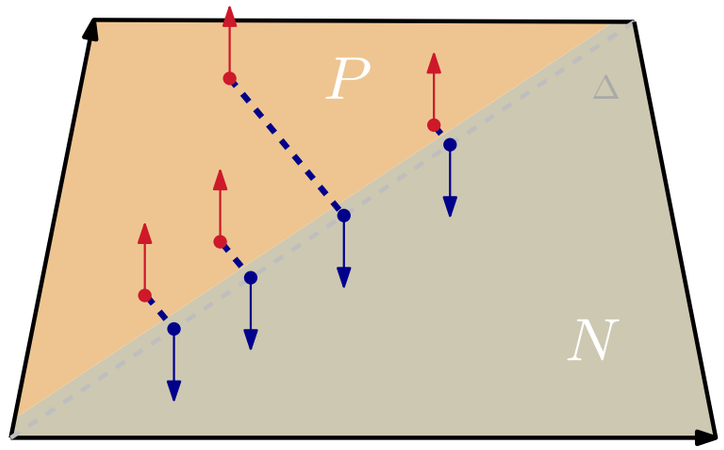
Hahn decomp. thm: For any $\mu \in \mathcal{M}_0(X, \Sigma)$ there exist measurable sets P, N such that:

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

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

Persistence diagrams as discrete measures

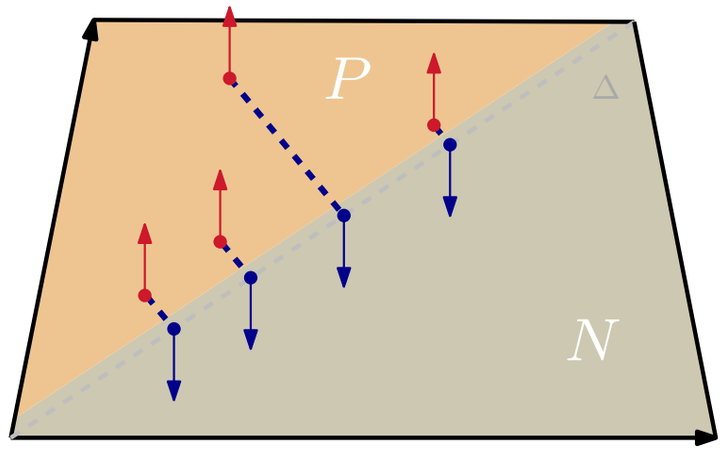
Hahn decomp. thm: For any $\mu \in \mathcal{M}_0(X, \Sigma)$ there exist measurable sets P, N such that:

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

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

for persistence diagrams: $\tilde{\mu}_D \quad \tilde{\mu}_{D'}$

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

A Wasserstein Gaussian kernel for PDs?

Thm: If $d : X \times X \to \mathbb{R}_+$ is symmetric and 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

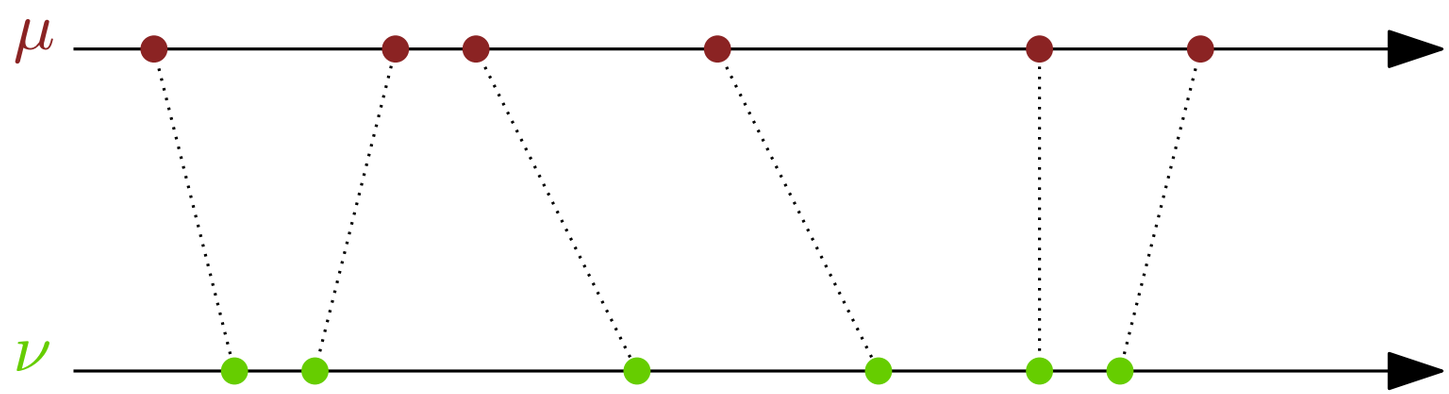
[*Sliced Wasserstein Kernel for persistence diagrams*, C., Cuturi, Oudot, ICML, 2017]

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*, C., Cuturi, Oudot, ICML, 2017]

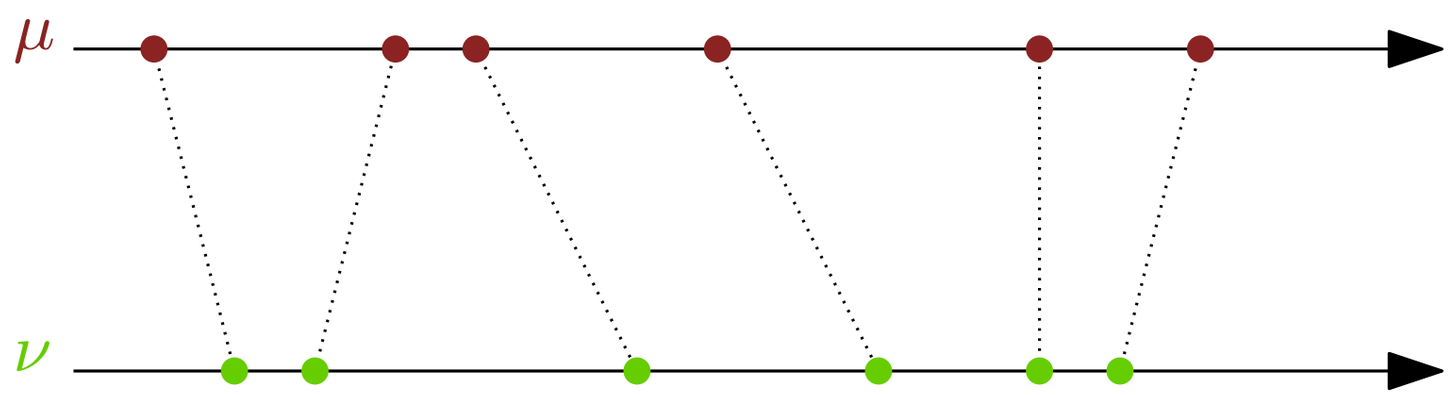
Special case: $X = \mathbb{R}$, μ, ν discrete measures of mass n

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

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)



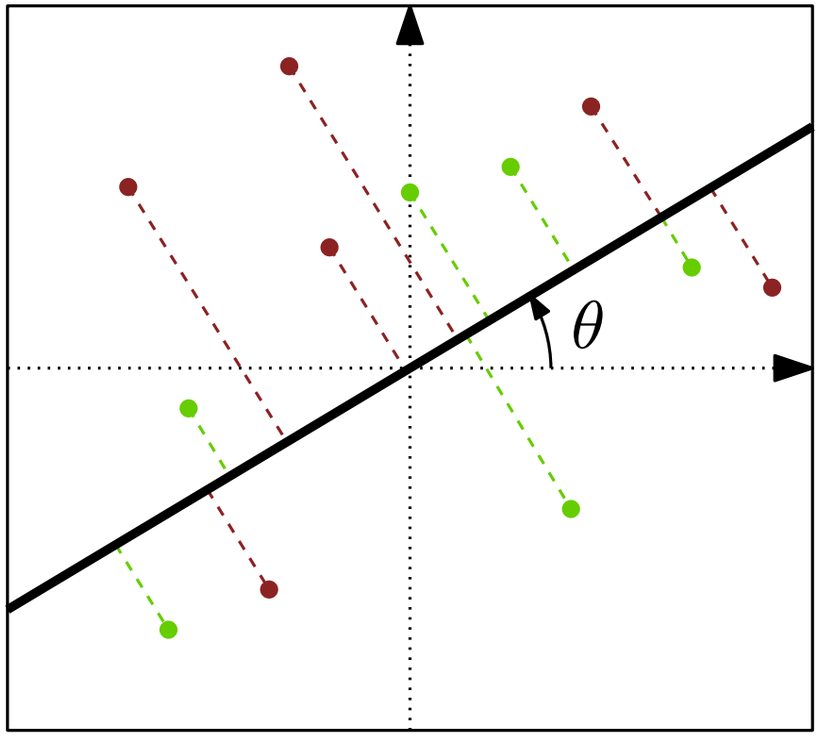
Sliced Wasserstein metric

[*Sliced Wasserstein Kernel for persistence diagrams*, C., 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 $\pi_\theta =$ orthogonal projection onto line passing through 0 with angle θ.



Sliced Wasserstein metric

[*Sliced Wasserstein Kernel for persistence diagrams*, C., 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 $\pi_\theta =$ orthogonal projection onto line passing through 0 with angle θ.

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

- satisfies the axioms of a metric
- well-defined barycenters, fast to compute via gradient descent, etc.
- conditionally negative semidefinite

Sliced Wasserstein kernel

[*Sliced Wasserstein Kernel for persistence diagrams*, C., 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: k_{SW} is positive semidefinite.

Sliced Wasserstein kernel

[*Sliced Wasserstein Kernel for persistence diagrams*, C., 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: k_{SW} is positive semidefinite.

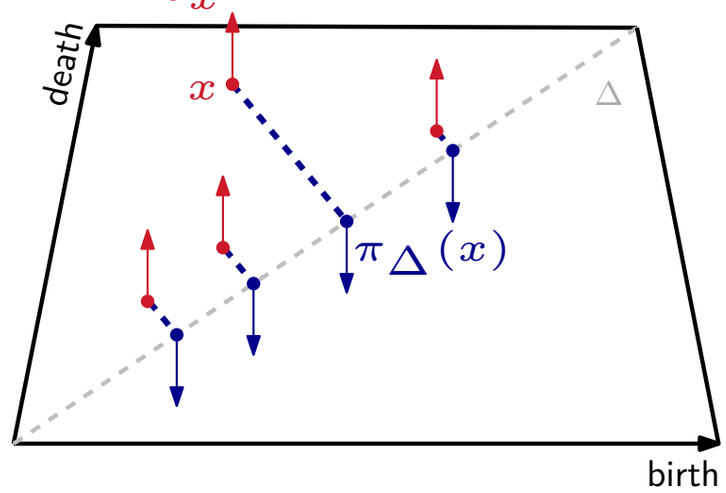
Application to persistence diagrams:

$$D \mapsto \mu_D := \sum_{x \in D} \delta_x$$

$$\mapsto \tilde{\mu}_D := \mu_D - \pi_\Delta \# \mu_D$$

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

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



Sliced Wasserstein kernel

[*Sliced Wasserstein Kernel for persistence diagrams*, C., 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 $D, D' \in \mathcal{D}_N$,

$$\frac{1}{2+4N(2N-1)}\, d_1(D,D') \;\leq\; SW_1(D,D') \;\leq\; 2\sqrt{2}\, d_1(D,D')$$

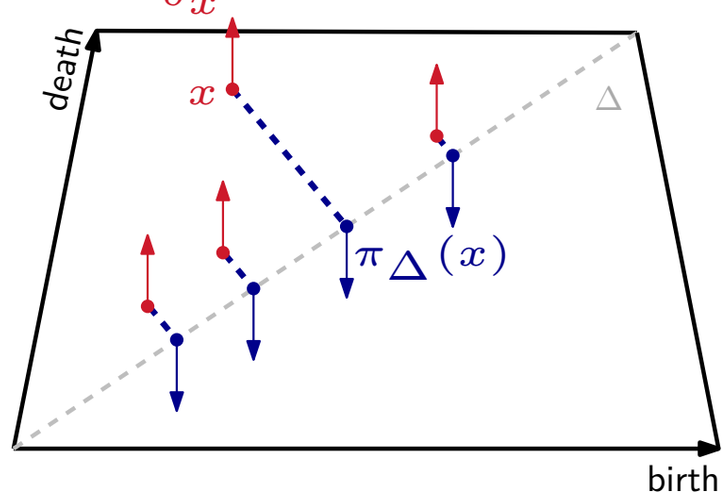
Application to persistence diagrams:

$$D \mapsto \mu_D := \sum\nolimits_{x\in D} \delta_x$$

$$\mapsto \tilde{\mu}_D := \mu_D - \pi_\Delta \# \mu_D$$

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

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



Sliced Wasserstein kernel

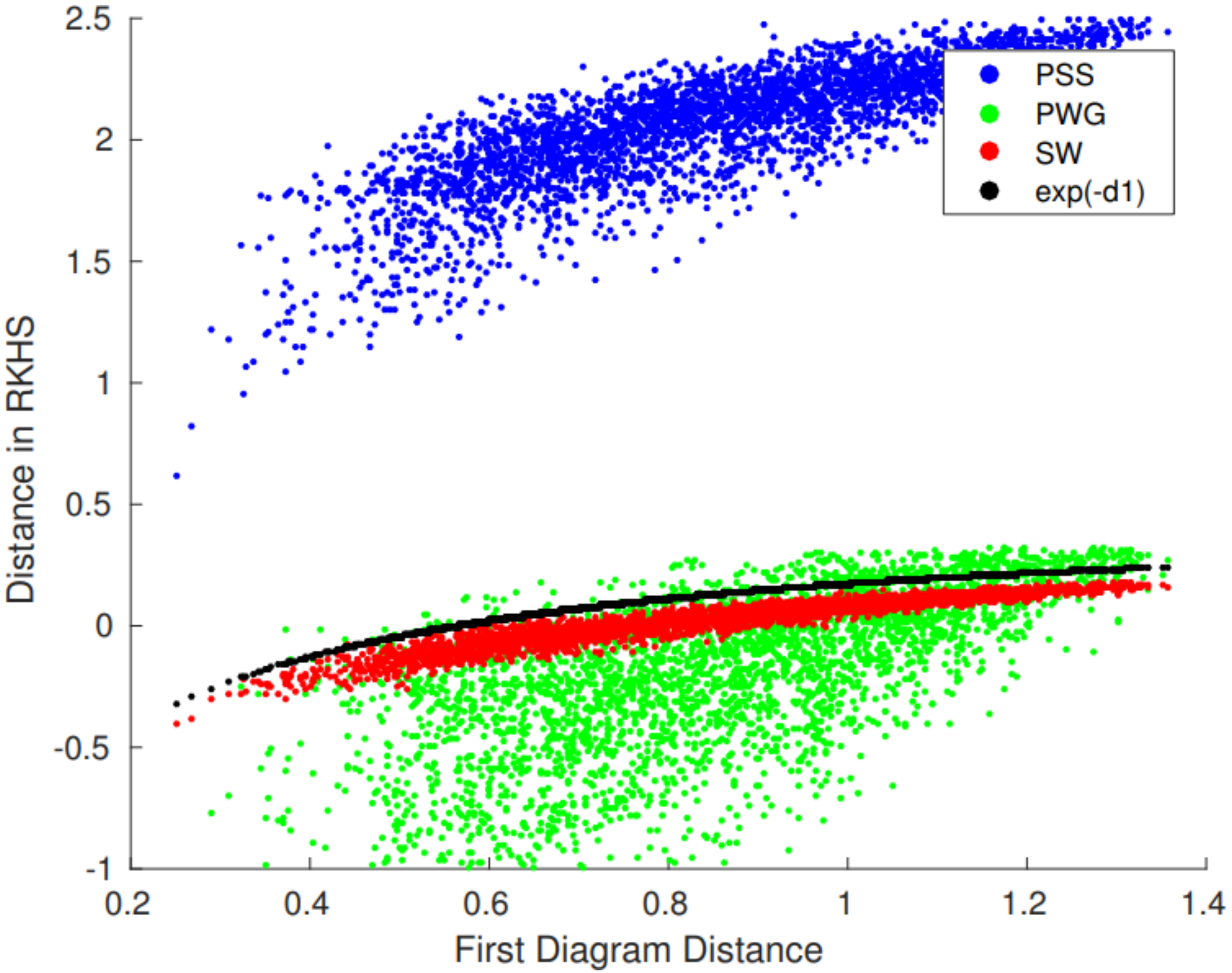
[*Sliced Wasserstein Kernel for persistence diagrams*, C., Cuturi, Oudot, ICML, 2017]

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 $D, D' \in \mathcal{D}_N$,

$$\frac{1}{2+4N(2N-1)}\, d_1(D,D') \;\leq\; SW_1(D,D') \;\leq\; 2\sqrt{2}\, d_1(D,D')$$

Cor: The feature map ϕ associated with k_{SW} is weakly metric-preserving: $\exists g, h$ nonzero except at 0 such that $g \circ d_1 \leq \|\phi(\cdot) - \phi(\cdot)\|_{\mathcal{H}} \leq h \circ d_1$.

Metric distortion in practice

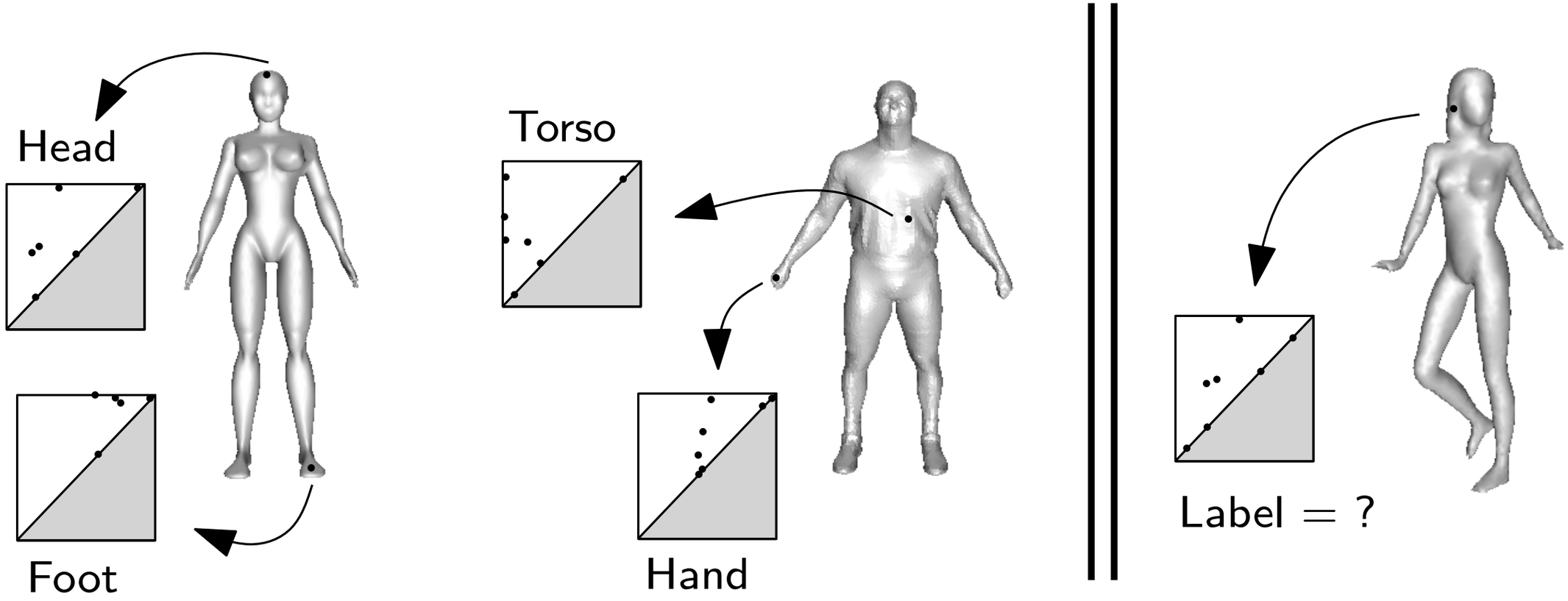


Application to supervised shape segmentation

Goal: segment 3D shapes based on examples.

Approach:

- train a (multiclass) classifier on PDs extracted from the training shapes
- apply classifier to PDs extracted from query shape



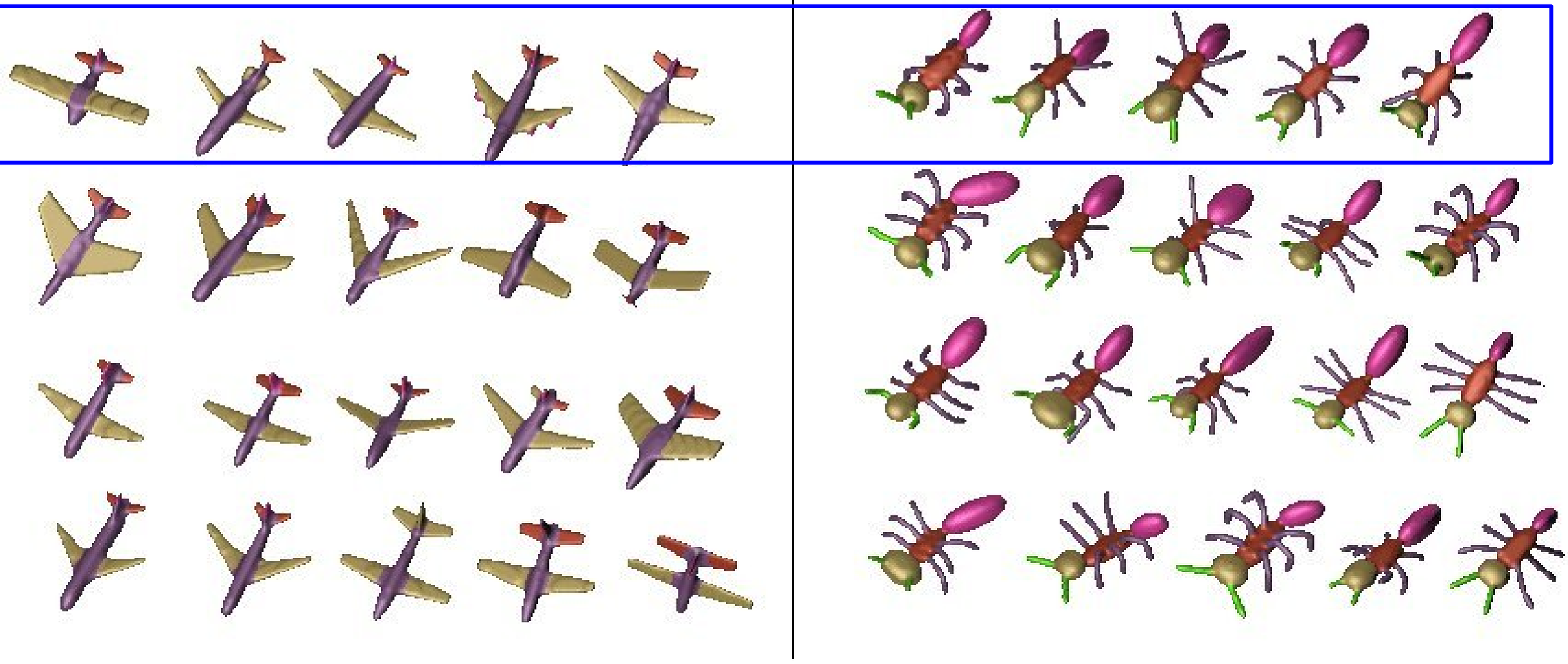
Application to supervised shape segmentation

Goal: segment 3D shapes based on examples.

Approach:

- train a (multiclass) classifier on PDs extracted from the training shapes
- apply classifier to PDs extracted from query shape

(training data)



Application to supervised shape segmentation

Goal: segment 3D shapes based on examples.

Approach:

- train a (multiclass) classifier on PDs extracted from the training shapes
- apply classifier to PDs extracted from query shape

Error rates (%) using TDA descriptors (kernels on barcodes):

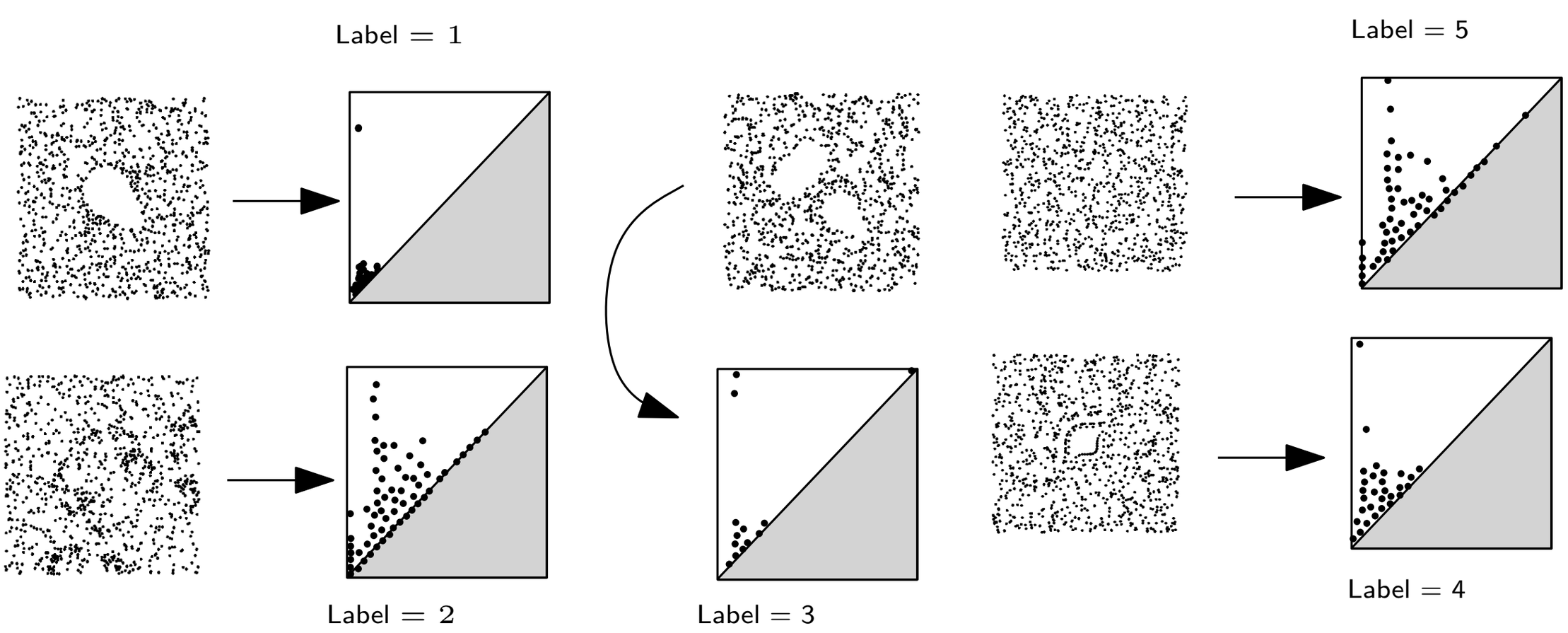
	TDA	geometry/stats	TDA + geometry/stats
Human	26.0	21.3	**11.3**
Airplane	27.4	18.7	**9.3**
Ant	7.7	9.7	**1.5**
FourLeg	27.0	25.6	**15.8**
Octopus	14.8	5.5	**3.4**
Bird	28.0	24.8	**13.5**
Fish	20.4	20.9	**7.7**

Application to supervised orbits classification

Goal: classify orbits of *linked twisted map*, modelling fluid flow dynamics.

Orbits described by (depending on parameter r):

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



Application to supervised orbits classification

Goal: classify orbits of *linked twisted map*, modelling fluid flow dynamics.

Orbits described by (depending on parameter r):

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

Accuracies (%) using only TDA descriptors (kernels on barcodes):

	k_{PSS}	k_{PWG}	k_{SW}
Orbit	64.0 ± 0.0	78.7 ± 0.0	$\mathbf{83.7} \pm 1.1$

(PDs as discrete measures)

Running times (in seconds on N-sized parameter space):

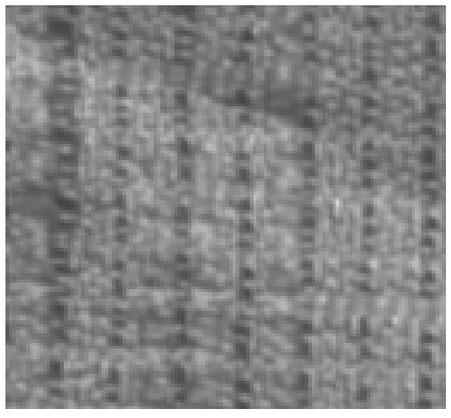
	k_{PSS}	k_{PWG}	k_{SW}
Orbit	$N \times 9183.4 \pm 65.6$	$N \times 69.2 \pm 0.9$	$385.8 \pm 0.2 + NC$

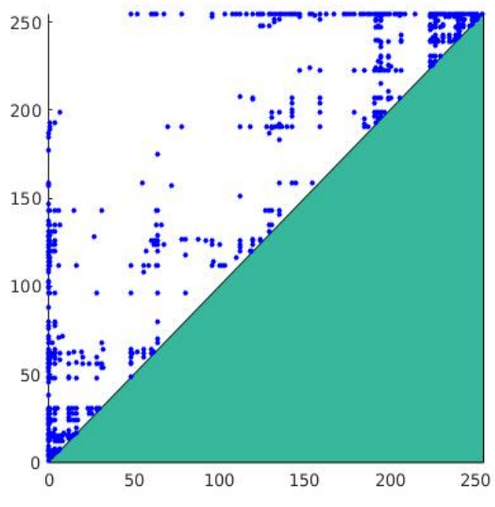
Application to supervised texture classification

Goal: classify textures from the OUTEX00000 database.

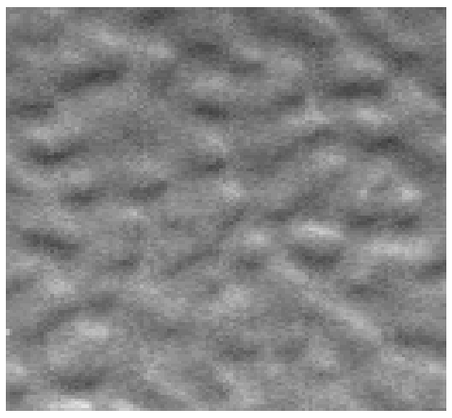
Textures described by CLBP (Compound Local Binary Pattern)

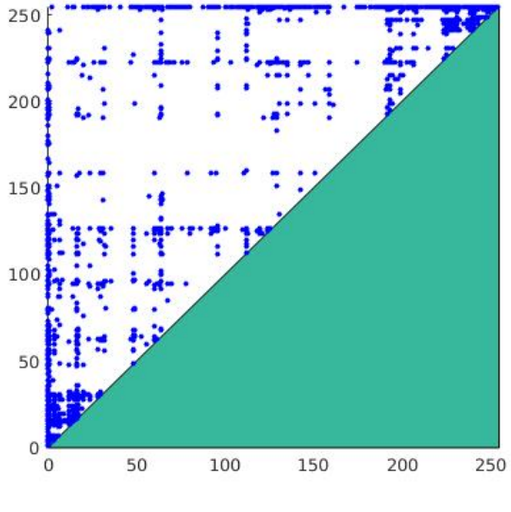
$\rightarrow$ apply degree-0 persistence on 1st sign component



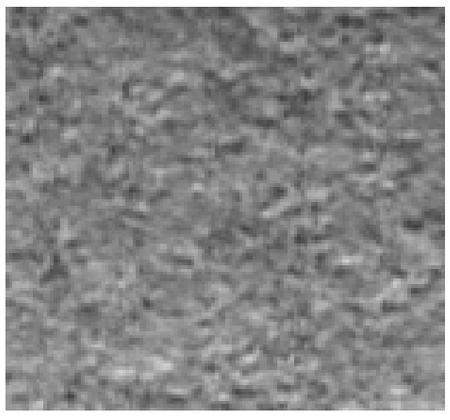


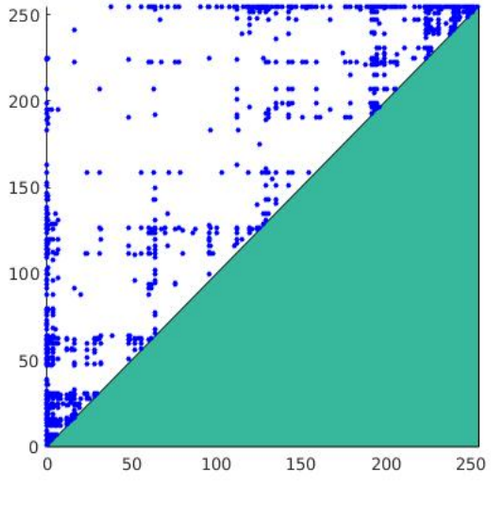
Label = Canvas





Label = Carpet





Label = Tile

Application to supervised texture classification

Goal: classify textures from the OUTEX00000 database.

Textures described by CLBP (Compound Local Binary Pattern)

$\rightarrow$ apply degree-0 persistence on 1st sign component

Accuracies (%) using only TDA descriptors (kernels on barcodes):

	k_{PSS}	k_{PWG}	k_{SW}
Orbit	$\mathbf{98.7} \pm 0.06$	96.7 ± 0.4	96.1 ± 0.1

(PDs as discrete measures)

Running times (in seconds on N-sized parameter space):

	k_{PSS}	k_{PWG}	k_{SW}
Orbit	$N \times 10337.4 \pm 140.5$	$N \times 45.9 \pm 0.6$	$126.4 \pm 0.2 + NC$

I. Topological Inference on Persistence Diagrams for Point Clouds

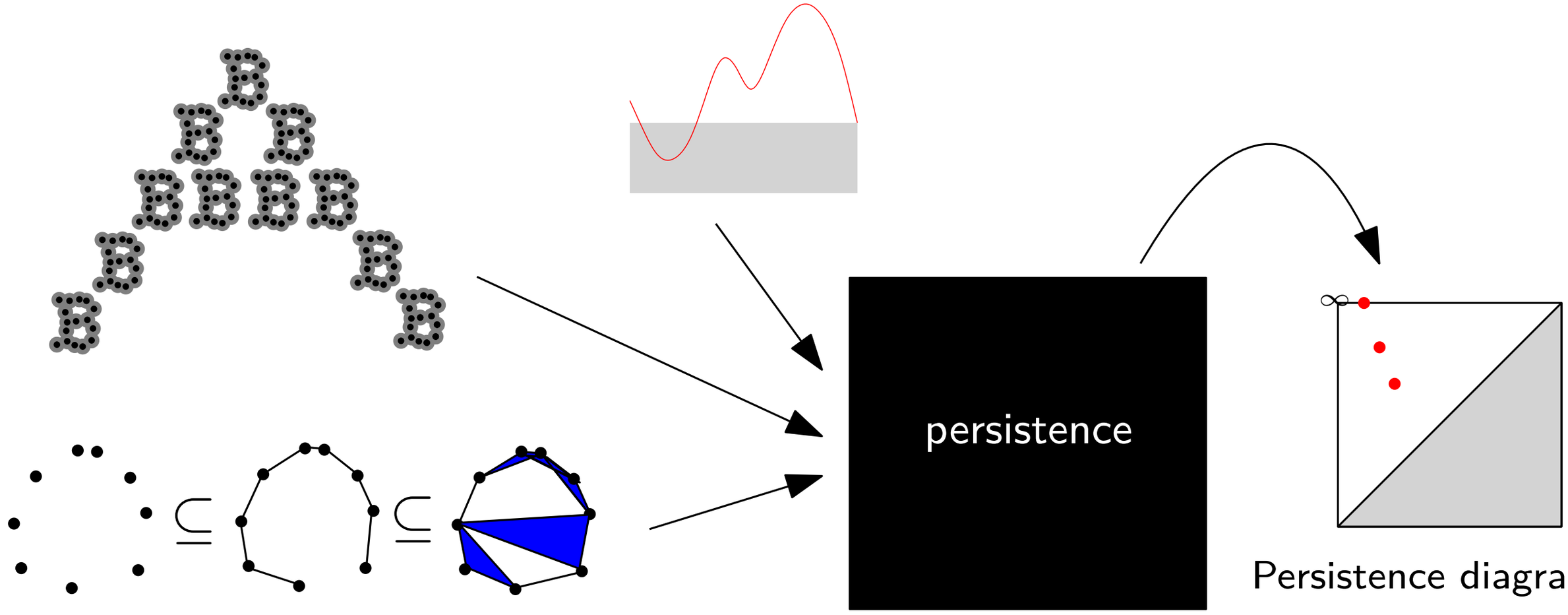
II. Topological Vectorizations for Inference and Machine Learning

(III. A Gaussian Kernel for Persistence Diagrams)

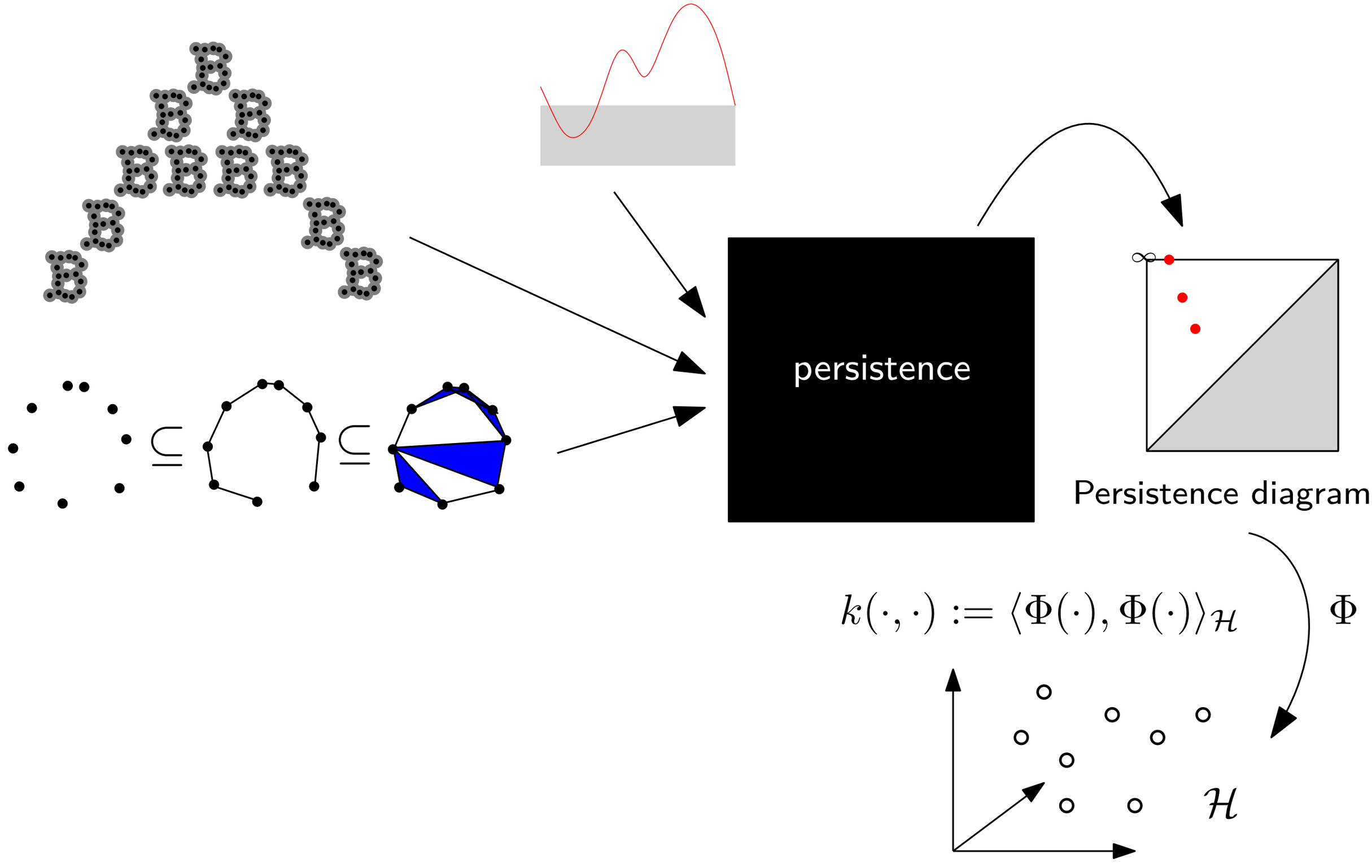
IV. Deep Learning for Persistence Diagrams

Persistence Diagrams and Machine Learning

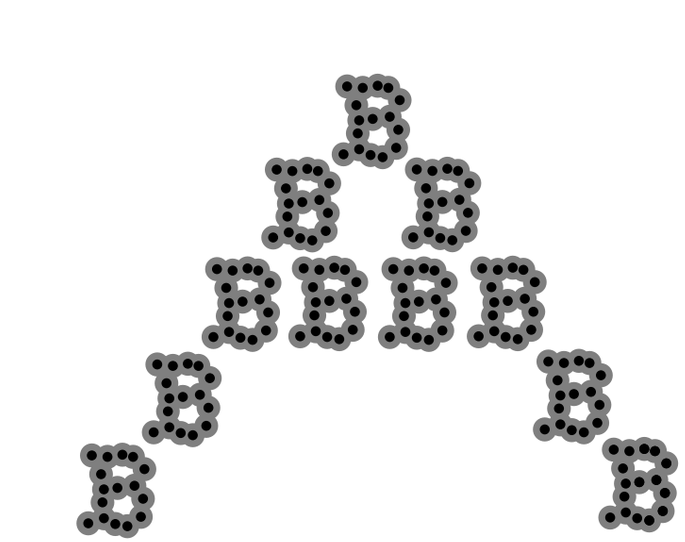
Persistence Diagrams and Machine Learning

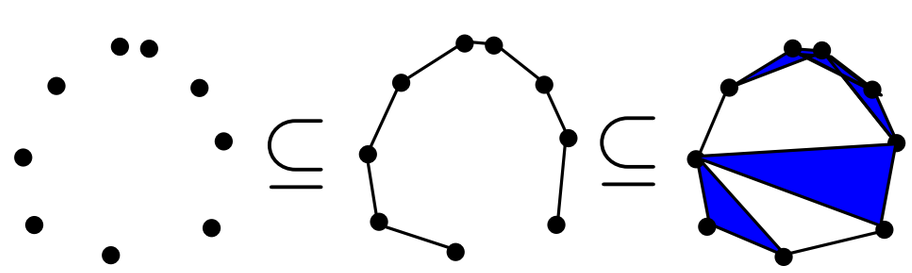


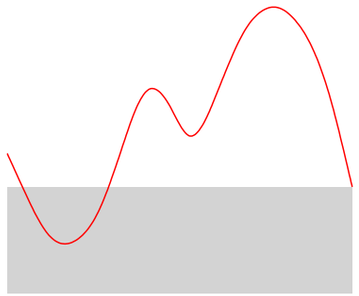
Persistence Diagrams and Machine Learning

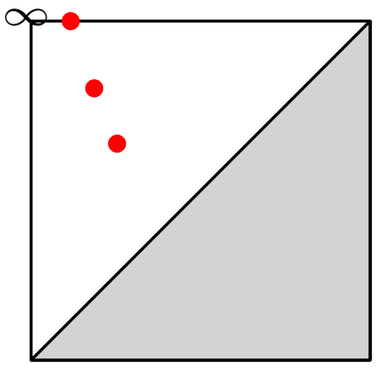


Persistence Diagrams and Machine Learning





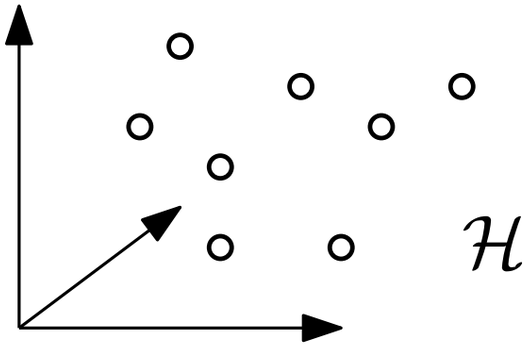




Persistence diagram

Φ

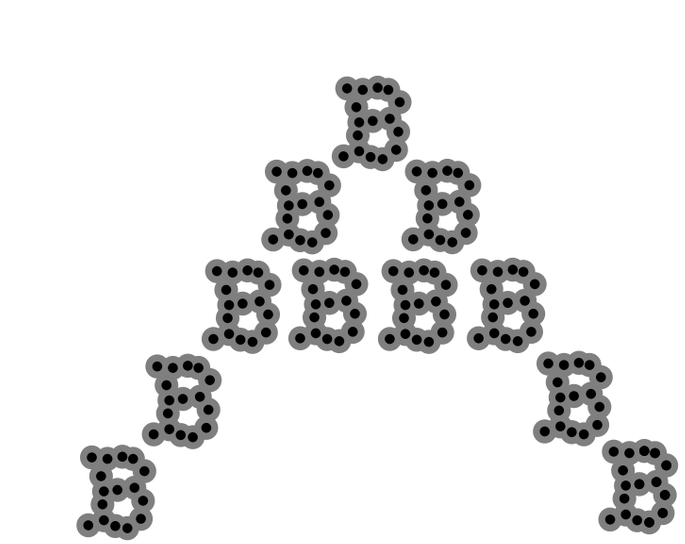
$k(\cdot,\cdot) := \langle \Phi(\cdot), \Phi(\cdot) \rangle_{\mathcal{H}}$

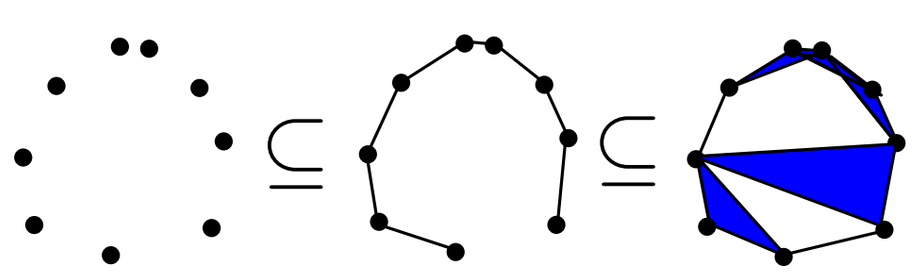


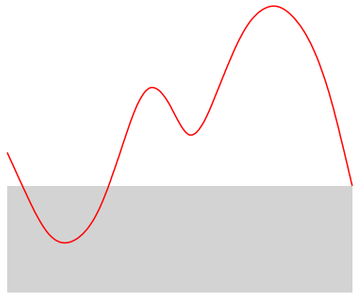
- Classifier (RF, SVM, NN etc.)
- Dim. red. (PCA, MDS, UMAP, t-SNE)
- Clustering (DBSCAN, K-means, etc.)

Etc.

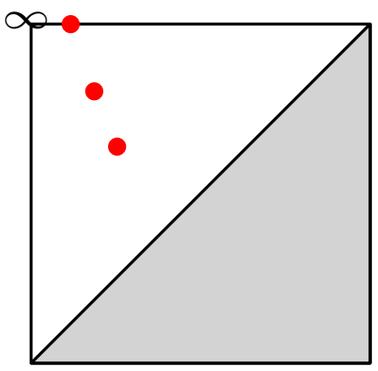
Persistence Diagrams and Machine Learning







persistence

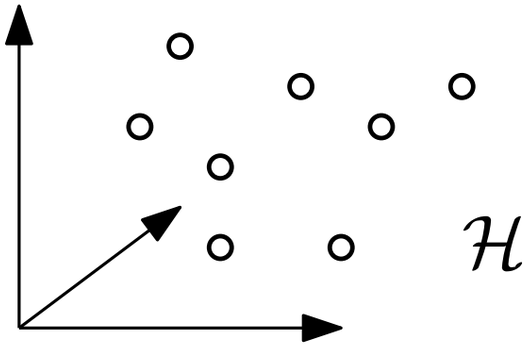


Persistence diagram

- Classifier (RF, SVM, NN etc.)
- Dim. red. (PCA, MDS, UMAP, t-SNE)
- Clustering (DBSCAN, K-means, etc.)

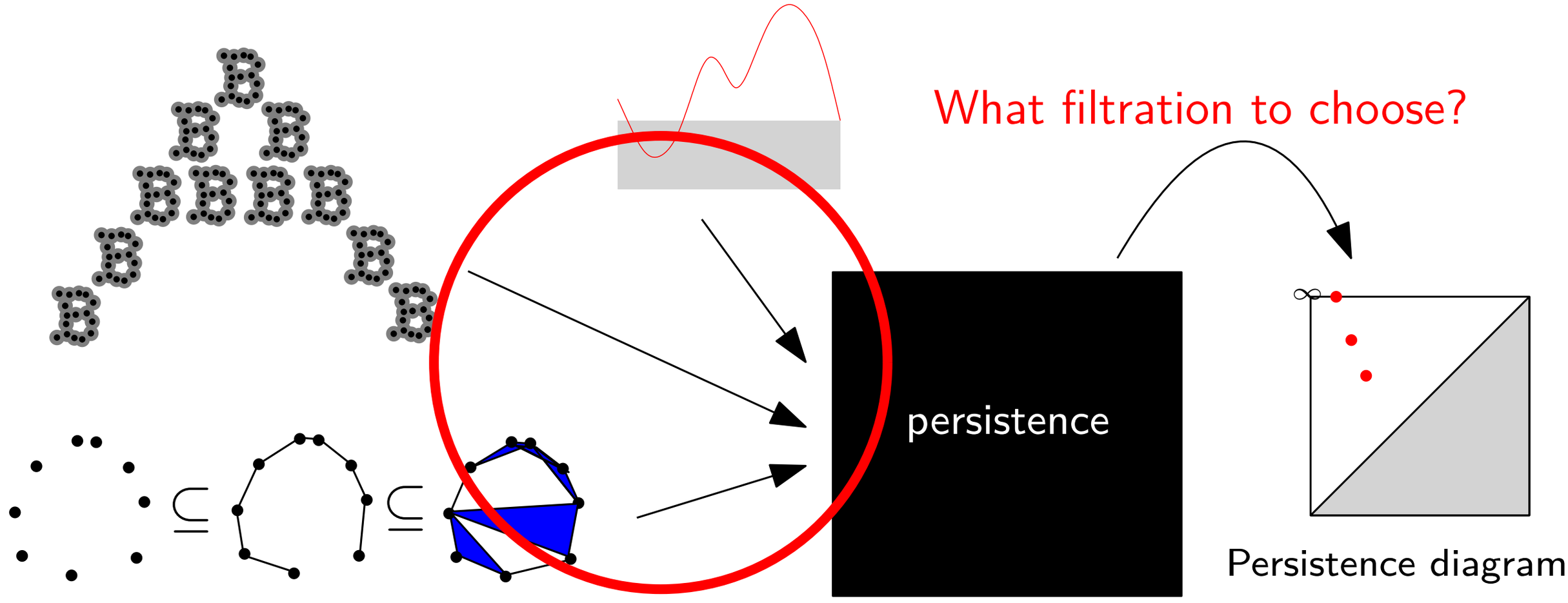
Etc.

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



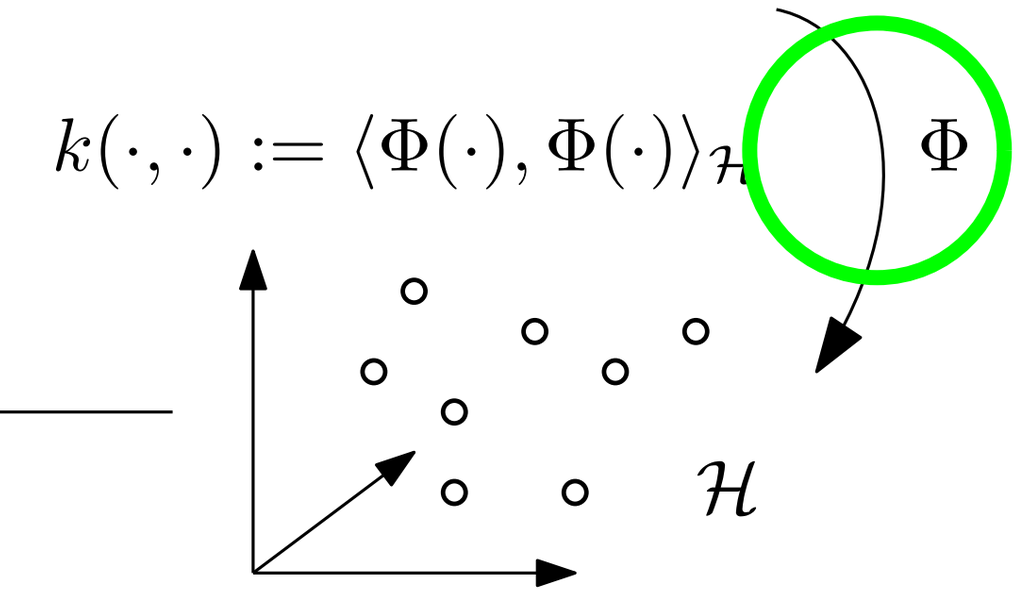
What linearization to choose?

Persistence Diagrams and Machine Learning



- Classifier (RF, SVM, NN etc.)
- Dim. red. (PCA, MDS, UMAP, t-SNE)
- Clustering (DBSCAN, K-means, etc.)

Etc.



What linearization to choose?

The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

Prop: $\mathcal{H}$ Hilbert with dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and distance $\| \cdot \|_{\mathcal{H}}$. Assume $d_{\mathcal{H}}$ and d_∞ or d_p are equivalent.

(i) $\mathcal{H} = \mathbb{R}^d \Rightarrow$ **Impossible**

even if the PDs are included in $[-L, L]^2$ and have less than N points

(ii) $\mathcal{H}$ separable, $p = 1 \Rightarrow$ either $A \to 0$ or $B \to +\infty$

when $L, N \to +\infty$

The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

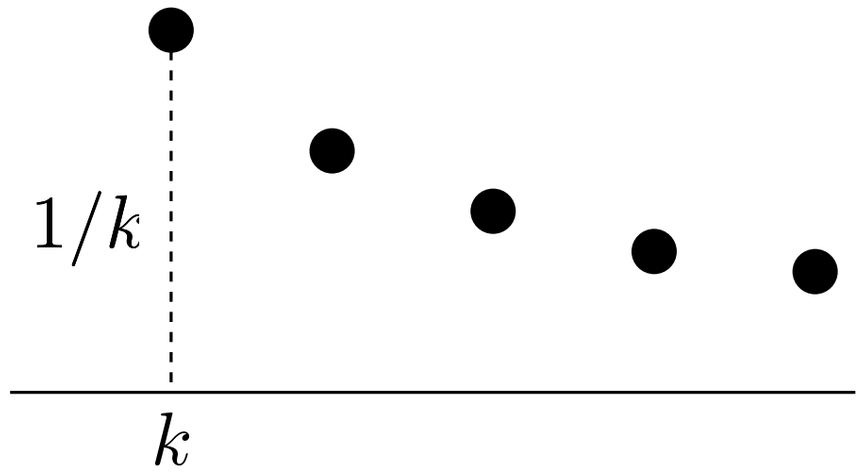
Proof:

(ii) The space of PDs with possibly infinite number of points is not separable with respect to d_1

Consider $S = \{D_u\}_{u \in \{0,1\}^{\mathbb{N}}}$

where $D_u = \left\{ \left(k, k + \frac{1}{k}\right) \,:\, u_k = 1 \right\}$

S is not countable with d_1



The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

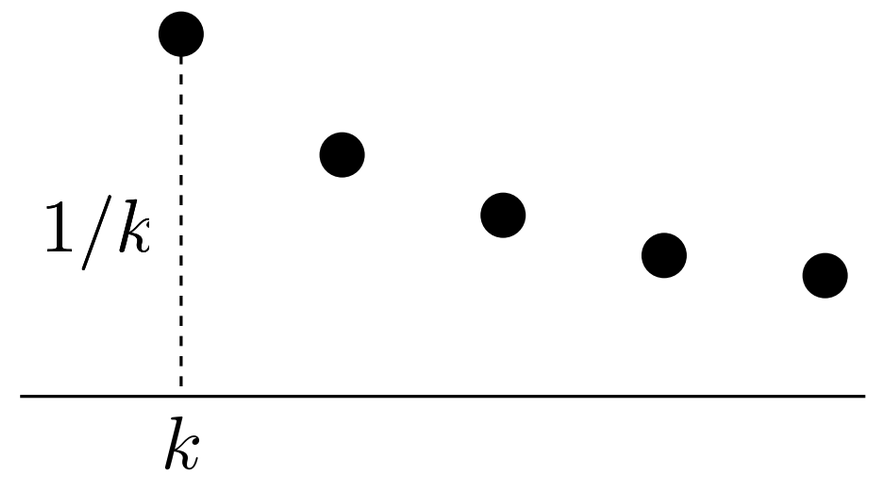
Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

Proof:

$$S = \{D_u\}_{u \in \{0,1\}^{\mathbb{N}}}$$

Indeed, let $S' \subseteq S$ be a dense set and $\epsilon > 0$



The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

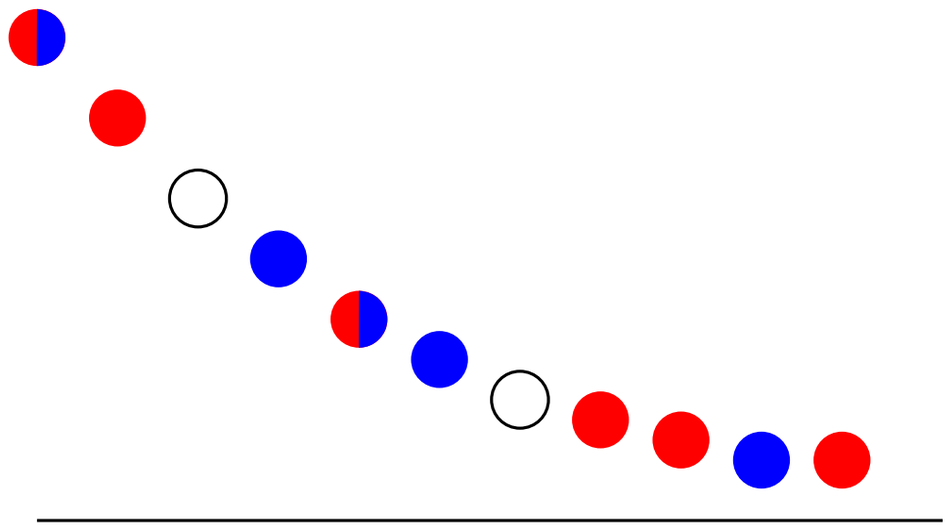
$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

Proof:

$$S = \{D_u\}_{u \in \{0,1\}^{\mathbb{N}}}$$

Indeed, let $S' \subseteq S$ be a dense set and $\epsilon > 0$

$$\forall D_u \in S,\, \exists D_{u'} \in S' : \mathrm{d}_1(D_u, D_{u'}) \leq \epsilon$$



The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

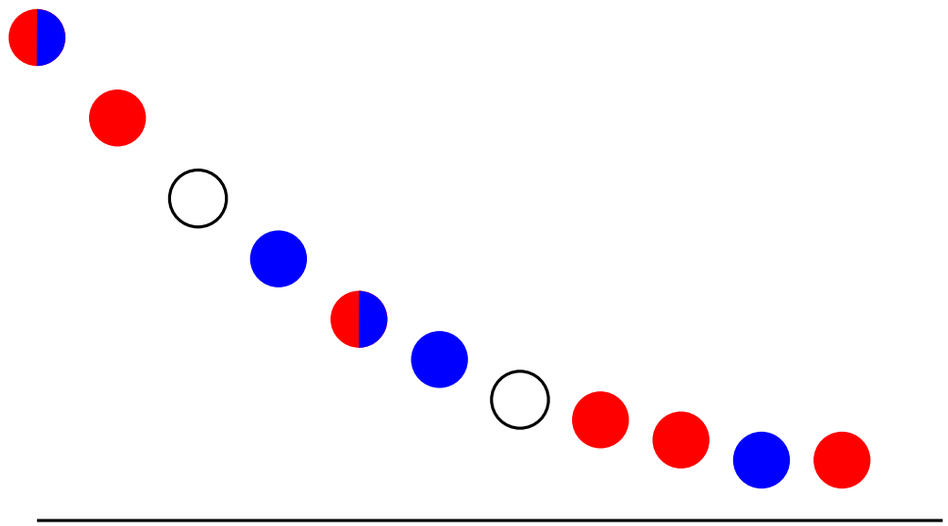
Proof:

$$S = \{D_u\}_{u \in \{0,1\}^{\mathbb{N}}}$$

Indeed, let $S' \subseteq S$ be a dense set and $\epsilon > 0$

$$\forall D_u \in S,\, \exists D_{u'} \in S' : \mathrm{d}_1(D_u, D_{u'}) \leq \epsilon$$

Supports of u' and u must differ on a finite number of terms only



The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

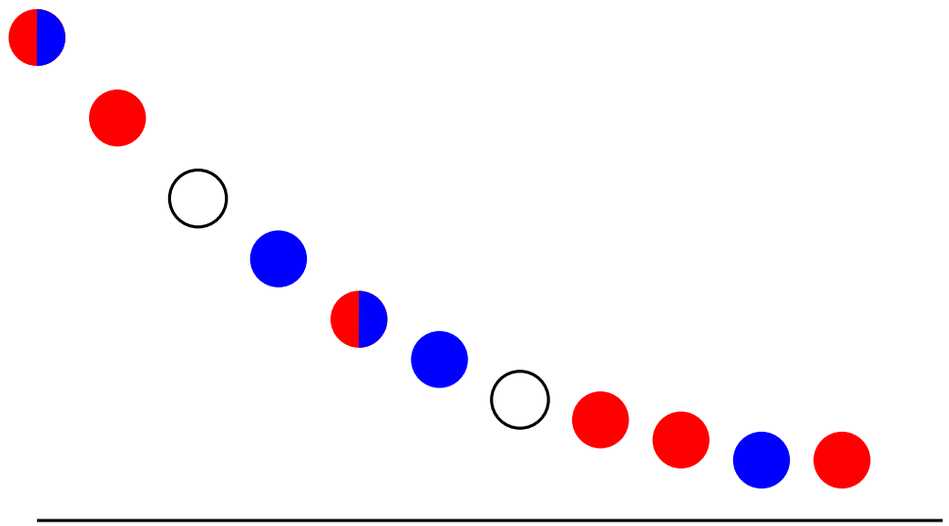
Proof:

$$S = \{D_u\}_{u \in \{0,1\}^{\mathbb{N}}}$$

Indeed, let $S' \subseteq S$ be a dense set and $\epsilon > 0$

$$\forall D_u \in S,\, \exists D_{u'} \in S' : \mathrm{d}_1(D_u, D_{u'}) \leq \epsilon$$

Supports of u' and u must differ on a finite number of terms only



$\Rightarrow$ card$(S') \geq$ card$(S/\sim)$

where $D_u \sim D_v \Leftrightarrow$ supp$(u) \triangle$ supp$(v) < \infty$

The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

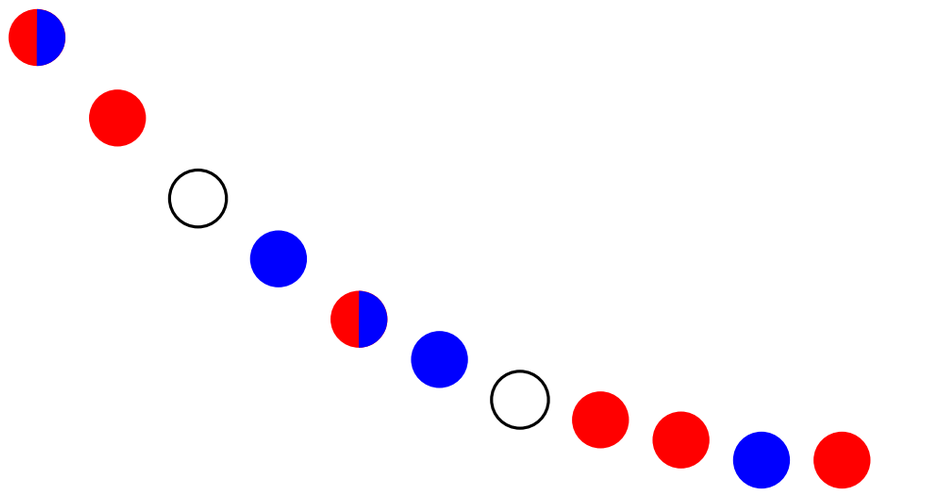
Proof:

$$S = \{D_u\}_{u \in \{0,1\}^{\mathbb{N}}}$$

Indeed, let $S' \subseteq S$ be a dense set and $\epsilon > 0$

$$\forall D_u \in S,\ \exists D_{u'} \in S' : \mathrm{d}_1(D_u, D_{u'}) \leq \epsilon$$

Supports of u' and u must differ on a finite number of terms only



$\Rightarrow$ card$(S') \geq$ card$(S/\sim)$ uncountable!

where $D_u \sim D_v \Leftrightarrow$ supp$(u) \triangle$ supp$(v) < \infty$

The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

Ex: Persistence surface $\Phi(D) = \sum_{p \in D} w(p) \cdot \exp\left(-\frac{\|\cdot - p\|_2^2}{2\sigma^2}\right)$

where $w((x,y)) = \arctan\left(C|y-x|^\alpha\right)$ with $C, \alpha > 0$

The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

Ex: Persistence surface $\Phi(D) = \sum_{p \in D} w(p) \cdot \exp\left(-\frac{\|\cdot - p\|_2^2}{2\sigma^2}\right)$

where $w((x,y)) = \arctan\left(C|y-x|^\alpha\right)$ with $C, \alpha > 0$

If $\alpha \geq 2$, S is in the domain of Φ.

The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

Proof:

(i) is a little more tricky

Def: Let (X, d) be a metric space. Given a subset $E \subset X$ and $r > 0$, let $N_r(E)$ be the least number of open balls of radius $\leq r$ that can cover E. The *Assouad dimension* of (X, d) is:

$$\dim_A(X, d) = \inf\{\alpha : \exists C \text{ s.t. } \sup_x N_{\beta r}(B(x, r)) \leq C\beta^{-\alpha},\ 0 < \beta \leq 1\}$$

The space of persistence diagrams

[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

Proof:

(i) is a little more tricky

Def: Let (X, d) be a metric space. Given a subset $E \subset X$ and $r > 0$, let $N_r(E)$ be the least number of open balls of radius $\leq r$ that can cover E. The *Assouad dimension* of (X, d) is:

$$\dim_A(X, d) = \inf\{\alpha : \exists C \text{ s.t. } \sup_x N_{\beta r}(B(x, r)) \leq C\beta^{-\alpha},\ 0 < \beta \leq 1\}$$

$\dim_A$ is preserved for equivalent metrics

$\dim_A(\mathcal{D}, d_p) = +\infty$ whereas $\dim_A(\mathbb{R}^d) = d$

The space of persistence diagrams

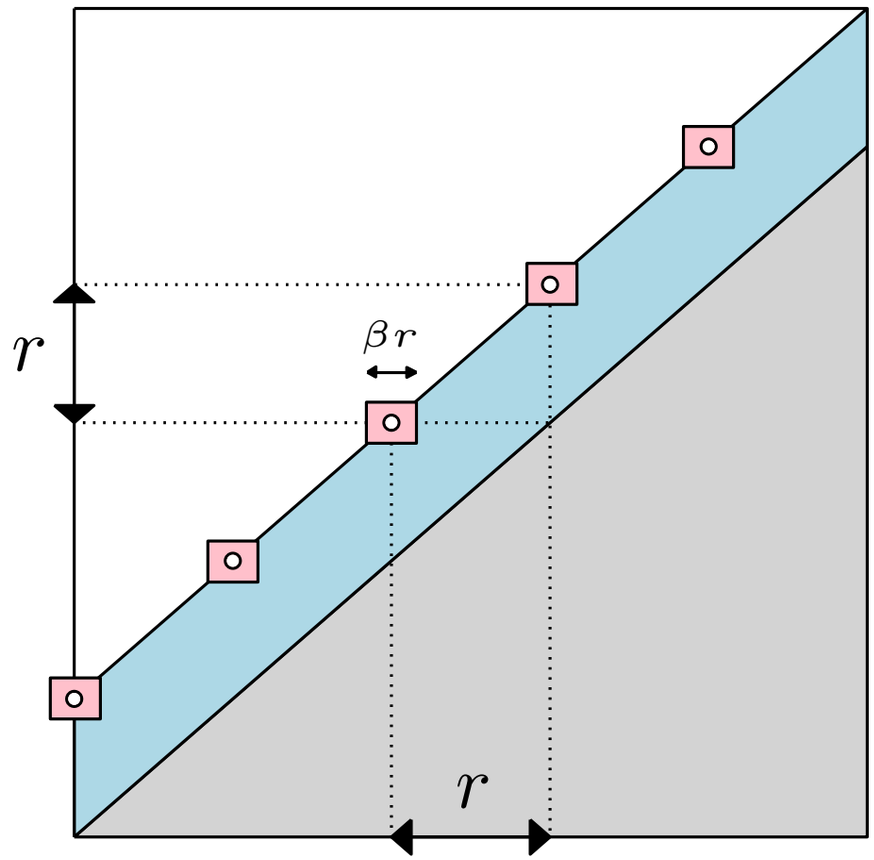
[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, C., SoCG, 2019]

Q: What happens in general when one embeds PDs in Hilbert?

Def: Two metrics d, d' are *equivalent* if

$$\exists 0 < A, B < +\infty \text{ s.t. } A\, d(\cdot,\cdot) \leq d'(\cdot,\cdot) \leq B\, d(\cdot,\cdot)$$

Proof:



Idea: Consider the ball of radius r around the empty diagram and diagrams with single points at distance r from Δ and from each other

The number of such diagrams increases to $+\infty$ as β goes to 0

$\dim_A$ is preserved for equivalent metrics

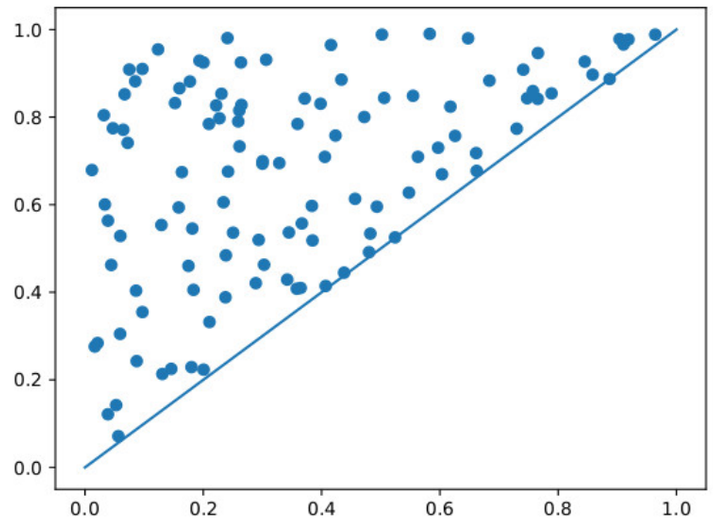
$\dim_A(\mathcal{D}, d_p) = +\infty$ whereas $\dim_A(\mathbb{R}^d) = d$

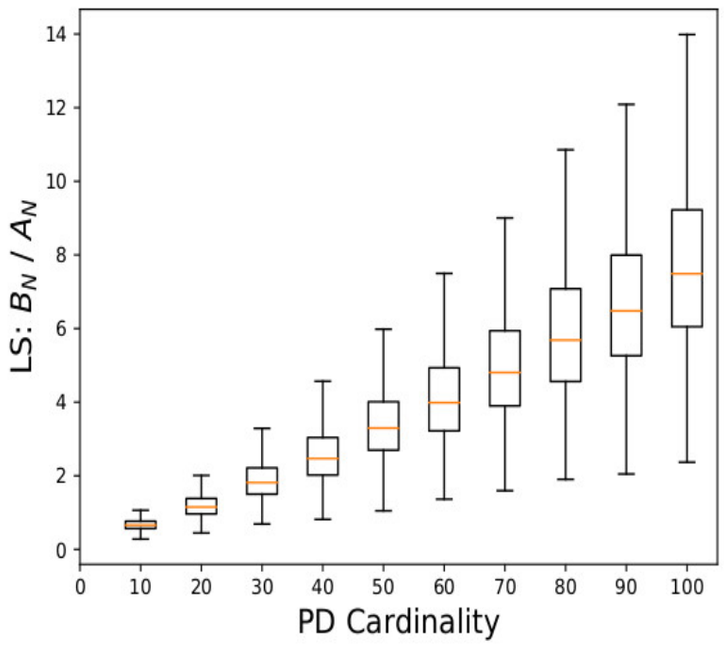
The space of persistence diagrams

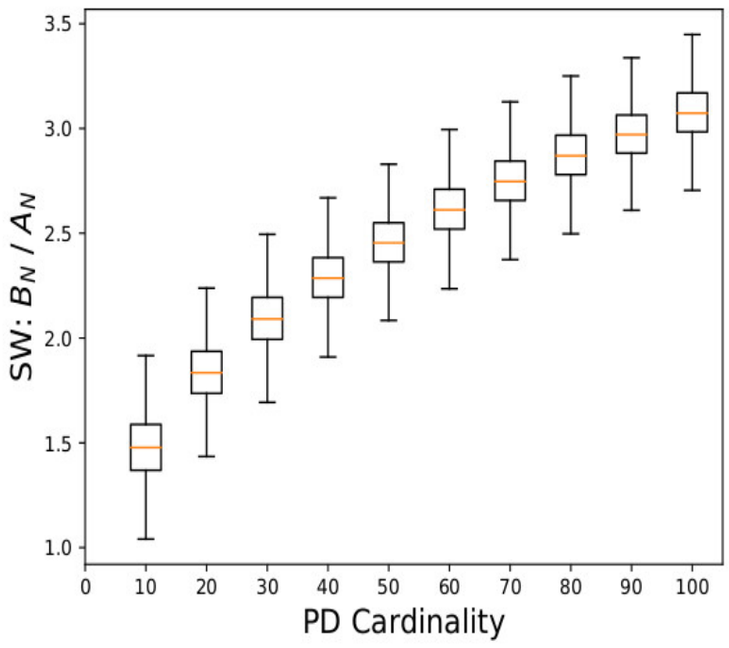
[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, Carrière, SoCG, 2019]

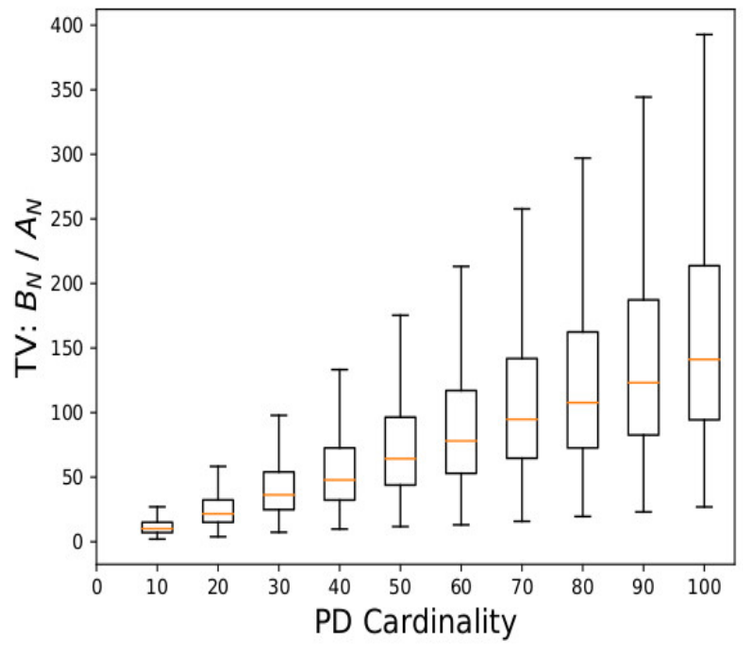
Illustrations:

We generate diagrams by uniformly sampling into the upper unit half-square







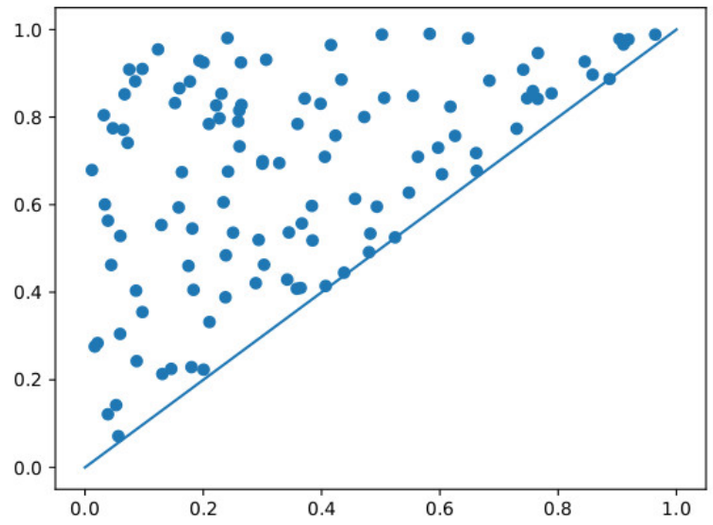


The space of persistence diagrams

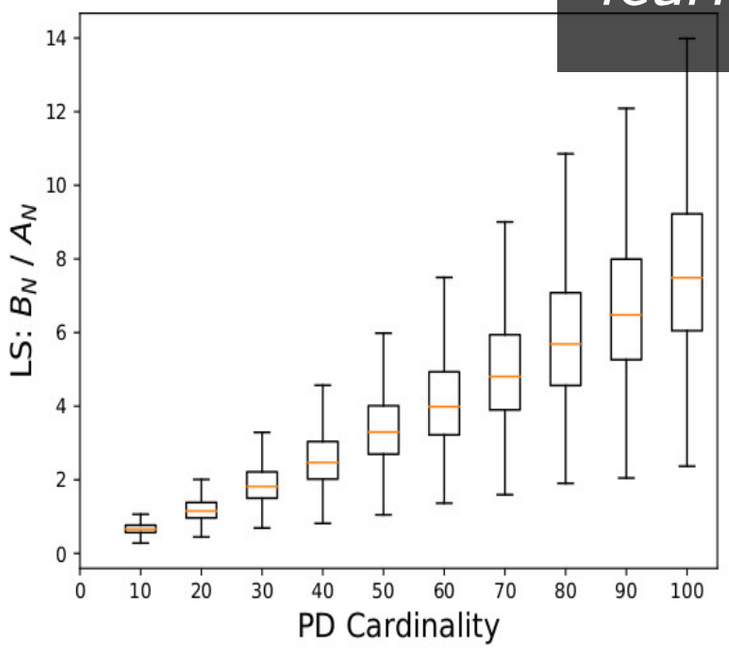
[*On the Metric Distortion of Embedding Persistence Diagrams into separable Hilbert Spaces*, Bauer, Carrière, SoCG, 2019]

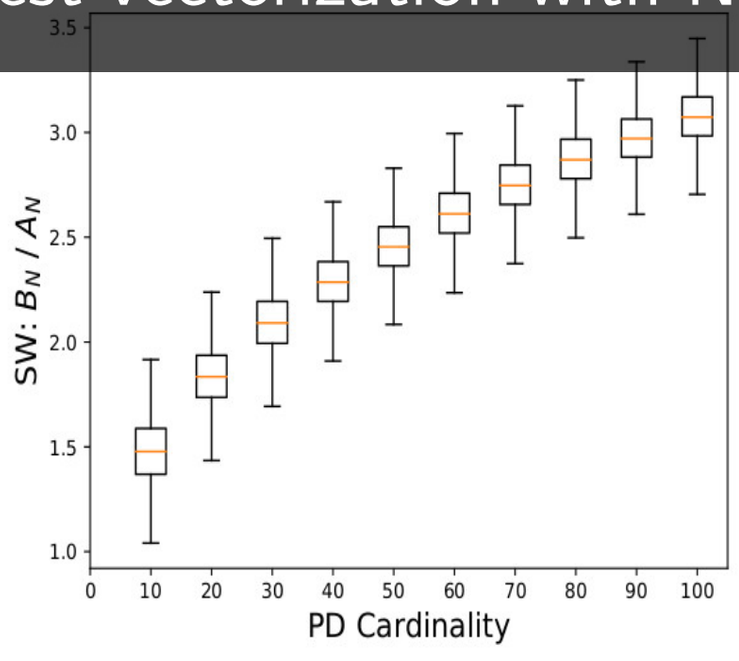
Illustrations:

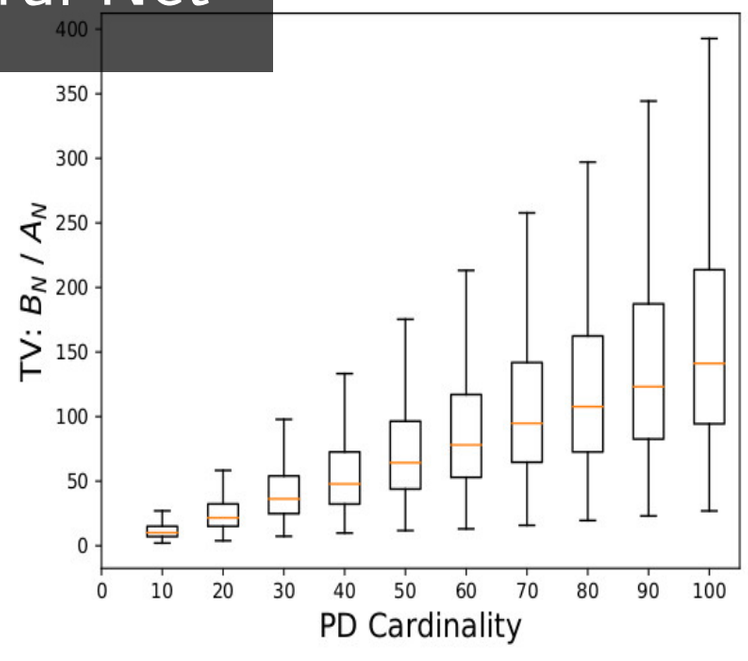
We generate diagrams by uniformly sampling into the upper unit half-square



Idea: Stay in Euclidean space $\mathbb{R}^d$ but *learn* best vectorization with Neural Net







The Deep Set architecture

[*Deep Sets*, Zaheer, Kottur, Ravanbakhsh, Poczos, Salakhutdinov, Smola, NeurIPS, 2017]

Deep Set is a novel neural net architecture that is able to handle sets instead of finite dimensional vectors

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

The Deep Set architecture

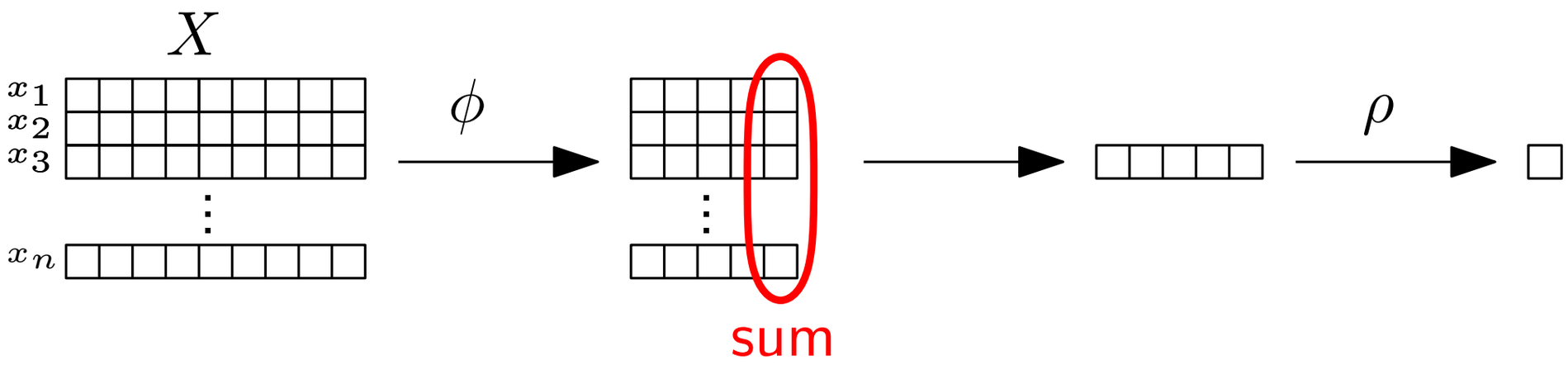
[*Deep Sets*, Zaheer, Kottur, Ravanbakhsh, Poczos, Salakhutdinov, Smola, NeurIPS, 2017]

Deep Set is a novel neural net architecture that is able to handle sets instead of finite dimensional vectors

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

Network is *permutation invariant*: $F(X) = \rho\left(\sum_i \phi(x_i)\right)$

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



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

The Deep Set architecture

[*Deep Sets*, Zaheer, Kottur, Ravanbakhsh, Poczos, Salakhutdinov, Smola, NeurIPS, 2017]

Deep Set is a novel neural net architecture that is able to handle sets instead of finite dimensional vectors

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

Network is *permutation invariant*: $F(X) = \rho\left(\sum_i \phi(x_i)\right)$

Universality theorem

Thm: 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.

Application to PDs

Application to PDs

Permutation invariant layers generalize several TDA approaches

Application to PDs

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images

Application to PDs

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ landscapes

Application to PDs

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ landscapes $\rightarrow$ Betti curves [*Time Series Classification via Topological Data Analysis*, Umeda, Trans. Jap. Soc. for AI, 2017]

Application to PDs

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ landscapes $\rightarrow$ Betti curves [*Time Series Classification via Topological Data Analysis*, Umeda, Trans. Jap. Soc. for AI, 2017]

But not all of them since $\mathbb{R}^2$ is not countable

Application to PDs

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ landscapes $\rightarrow$ Betti curves [*Time Series Classification via Topological Data Analysis*, Umeda, Trans. Jap. Soc. for AI, 2017]

But not all of them since $\mathbb{R}^2$ is not countable

Using any permutation invariant operation (such as max, min, kth largest value) allows to generalize other TDA approaches

Application to PDs

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ landscapes $\rightarrow$ Betti curves [*Time Series Classification via Topological Data Analysis*, Umeda, Trans. Jap. Soc. for AI, 2017]

But not all of them since $\mathbb{R}^2$ is not countable

Using any permutation invariant operation (such as max, min, kth largest value) allows to generalize other TDA approaches

$$\mathrm{PersLay}(D) = \rho\left(\mathsf{op}\{w(p) \cdot \phi(p)\}_{p \in D}\right)$$

Application to PDs

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ landscapes $\rightarrow$ Betti curves [*Time Series Classification via Topological Data Analysis*, Umeda, Trans. Jap. Soc. for AI, 2017]

But not all of them since $\mathbb{R}^2$ is not countable

Using any permutation invariant operation (such as max, min, kth largest value) allows to generalize other TDA approaches

$$\mathrm{PersLay}(D) = \rho\left(\mathsf{op}\{w(p) \cdot \phi(p)\}_{p \in D}\right)$$

Permutation-invariant operation

Application to PDs

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ landscapes $\rightarrow$ Betti curves [*Time Series Classification via Topological Data Analysis*, Umeda, Trans. Jap. Soc. for AI, 2017]

But not all of them since $\mathbb{R}^2$ is not countable

Using any permutation invariant operation (such as max, min, kth largest value) allows to generalize other TDA approaches

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

Permutation-invariant operation

Weight function

Application to PDs

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

Permutation invariant layers generalize several TDA approaches

$\rightarrow$ persistence images $\rightarrow$ landscapes $\rightarrow$ Betti curves [*Time Series Classification via Topological Data Analysis*, Umeda, Trans. Jap. Soc. for AI, 2017]

But not all of them since $\mathbb{R}^2$ is not countable

Using any permutation invariant operation (such as max, min, kth largest value) allows to generalize other TDA approaches

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

Permutation-invariant operation

Weight function

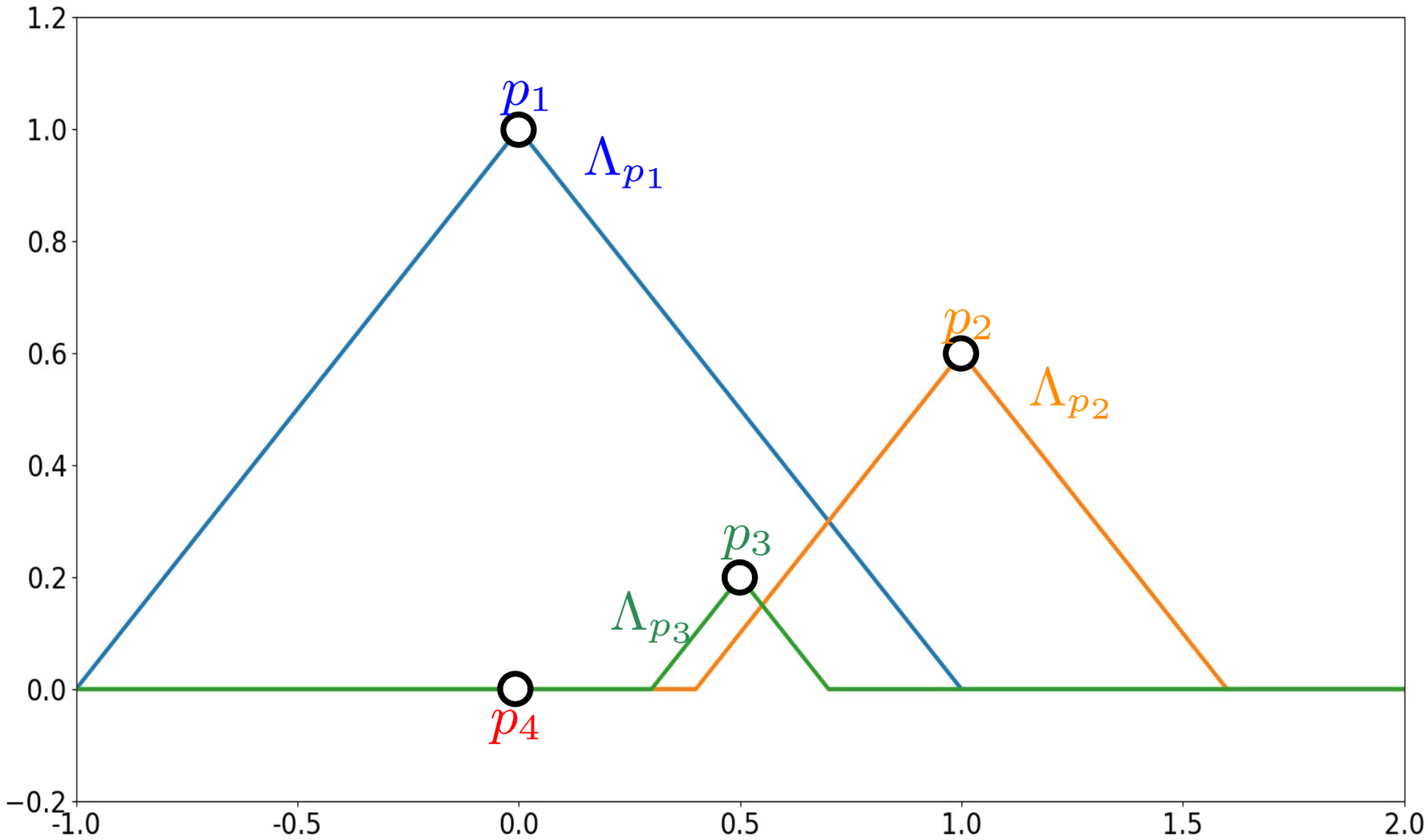
Point transformation

Application to PDs

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

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

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

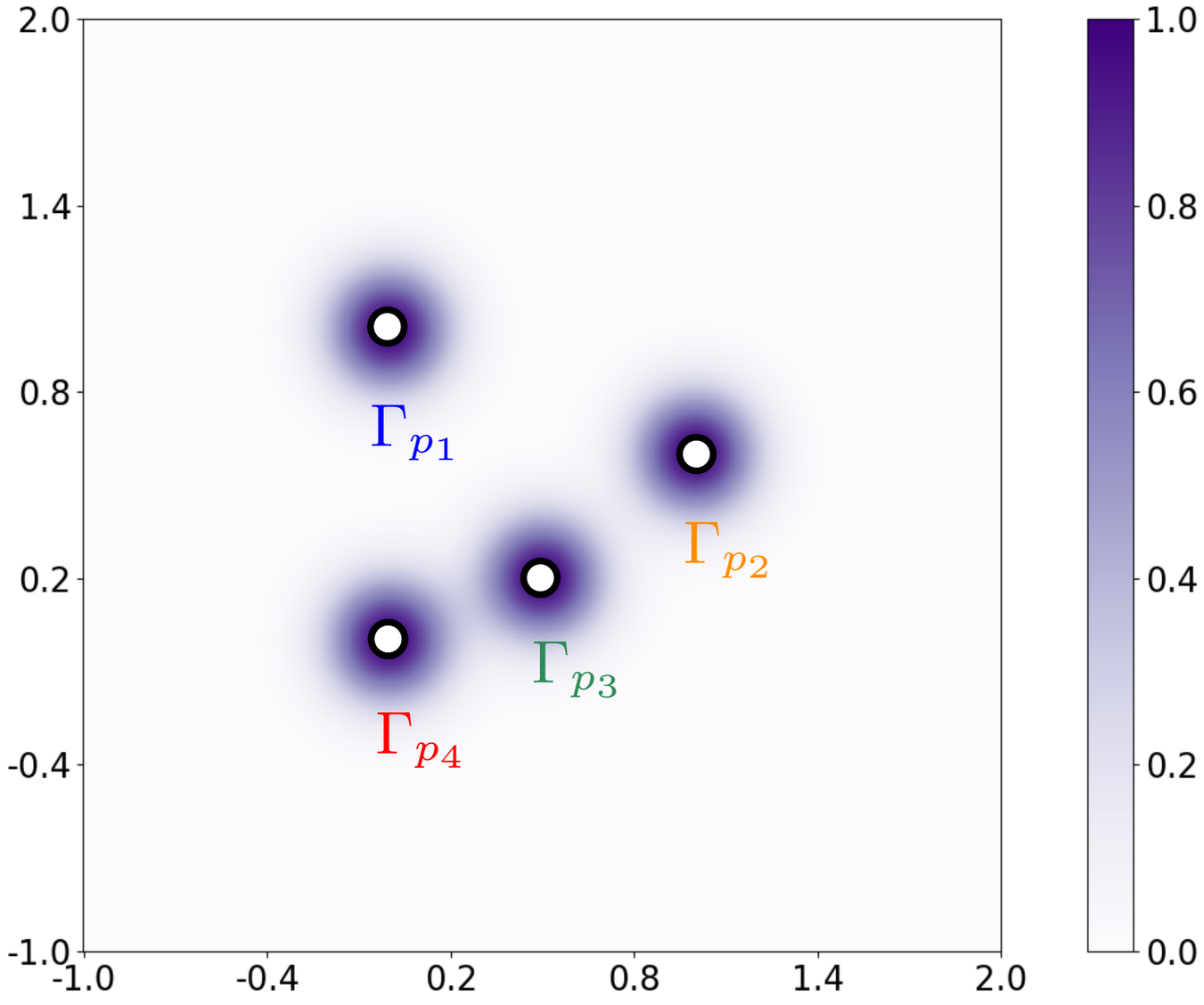


Application to PDs

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

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

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



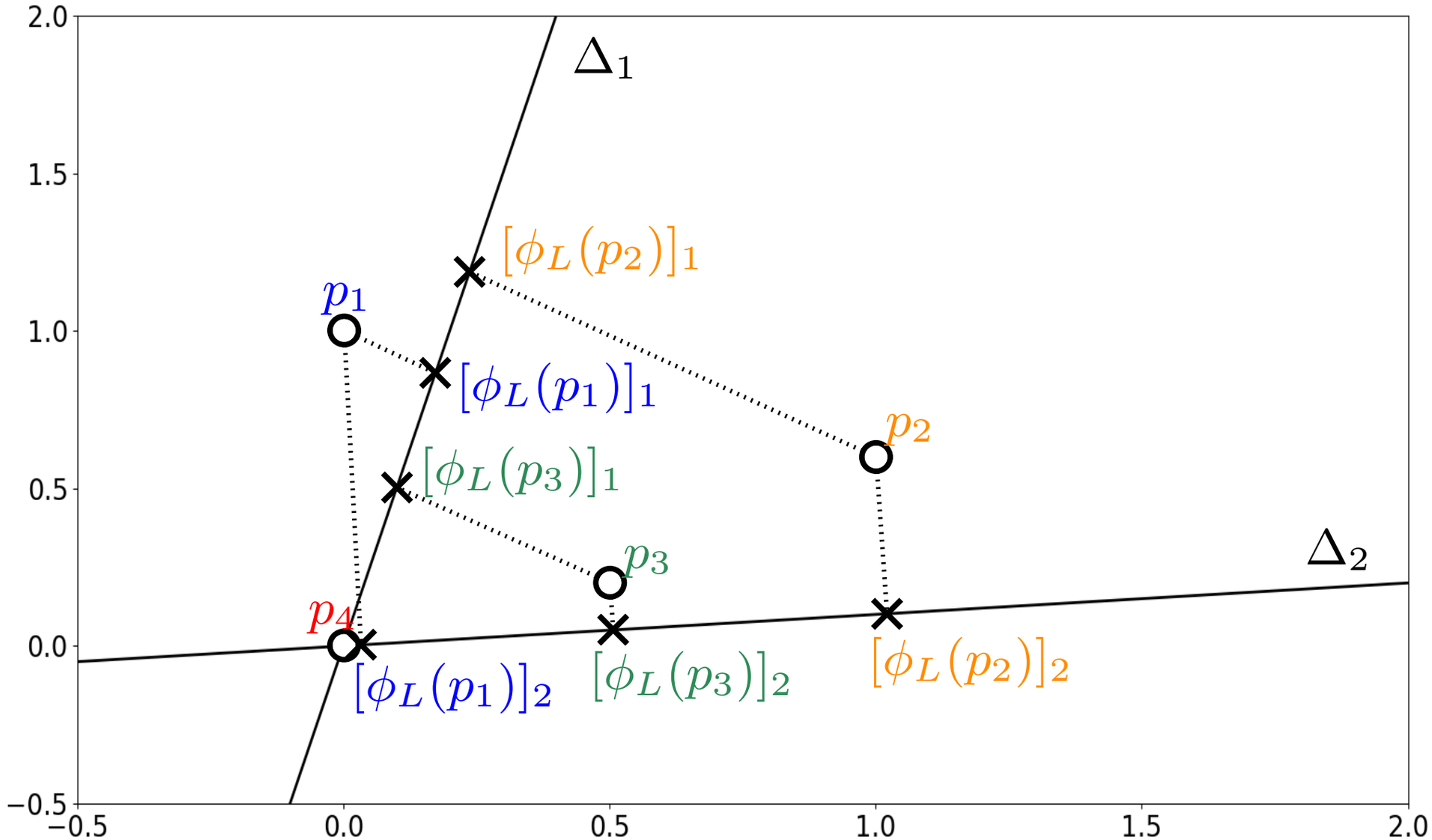
Application to PDs

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

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

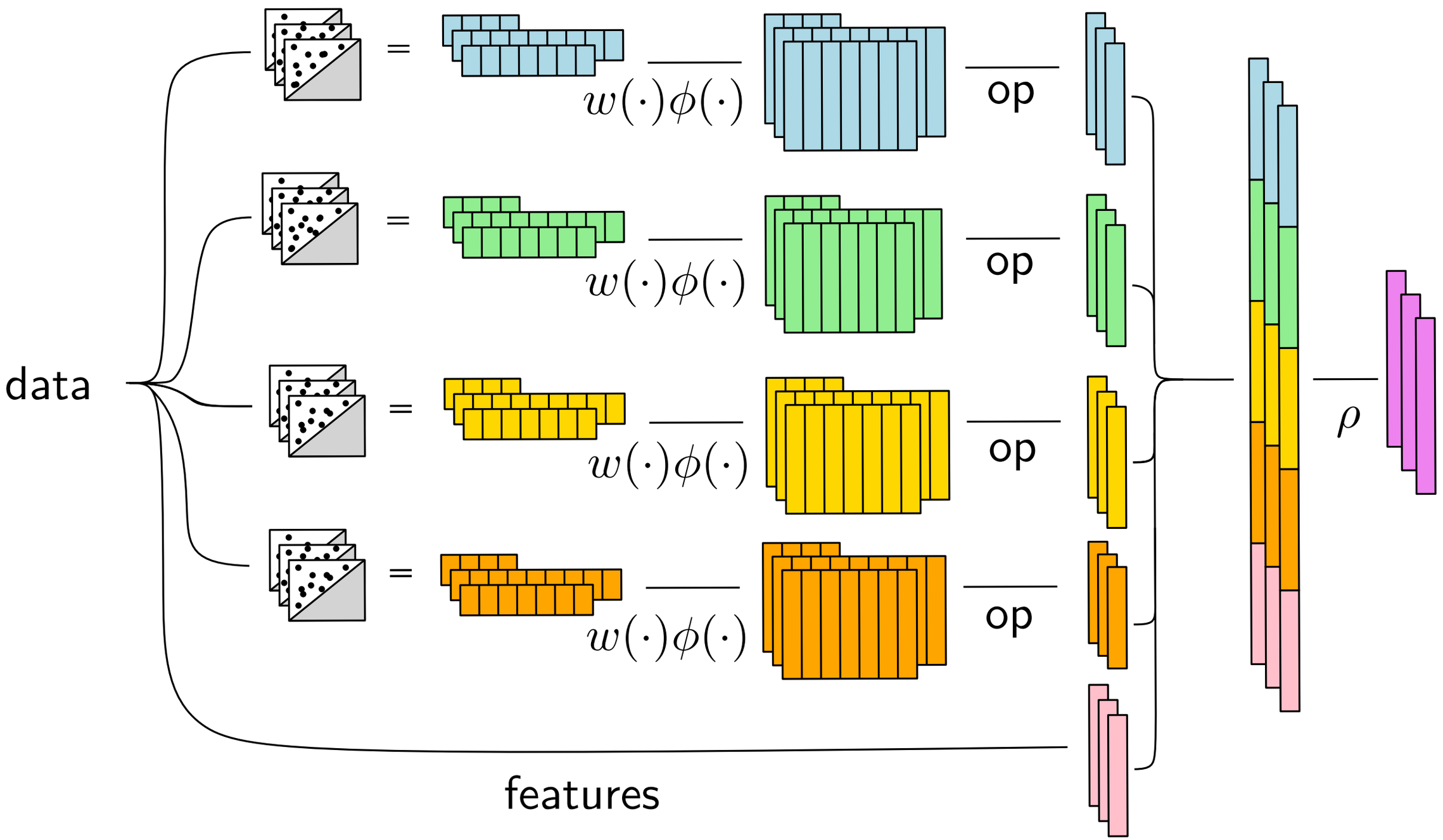
$b_{\Delta_1}, \cdots, b_{\Delta_q} \in \mathbb{R}$

$$\phi_L : p \mapsto \begin{bmatrix} \langle p, e_{\Delta_1} \rangle + b_{\Delta_1} \\ \langle p, e_{\Delta_2} \rangle + b_{\Delta_2} \\ \vdots \\ \langle p, e_{\Delta_q} \rangle + b_{\Delta_q} \end{bmatrix} \quad w(p) = 1 \quad \text{op} = \text{top-}k$$



Application to PDs

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]



Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

Let $G = (V, E)$ be a graph, A its adjacency matrix

D its degree matrix

and $L_w(G) = I - D^{-1/2} A D^{-1/2}$ its normalized Laplacian.

Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

Let $G = (V, E)$ be a graph, A its adjacency matrix

D its degree matrix

and $L_w(G) = I - D^{-1/2} A D^{-1/2}$ its normalized Laplacian.

$L_w(G)$ decomposes on a orthonormal basis $\phi_1 \dots \phi_n$

with eigenvalues $0 \leq \lambda_1 \leq \cdots \leq \lambda_n \leq 2$

Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

Let $G = (V, E)$ be a graph, A its adjacency matrix

D its degree matrix

and $L_w(G) = I - D^{-1/2}AD^{-1/2}$ its normalized Laplacian.

$L_w(G)$ decomposes on a orthonormal basis $\phi_1 \dots \phi_n$

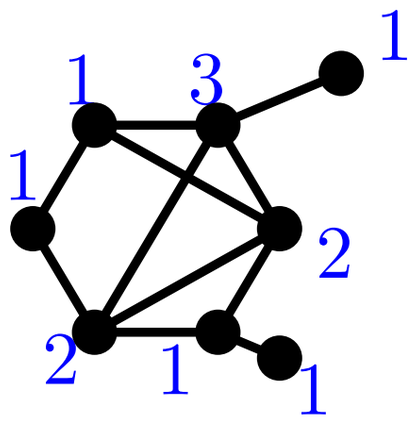
with eigenvalues $0 \leq \lambda_1 \leq \cdots \leq \lambda_n \leq 2$

Def: Let $t \geq 0$, and define the *Heat Kernel Signature* of param t:

$$\mathrm{hks}_{G,t} : v \mapsto \sum_{k=1}^{n} \exp(-\lambda_k t)\phi_k(v)^2$$

Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

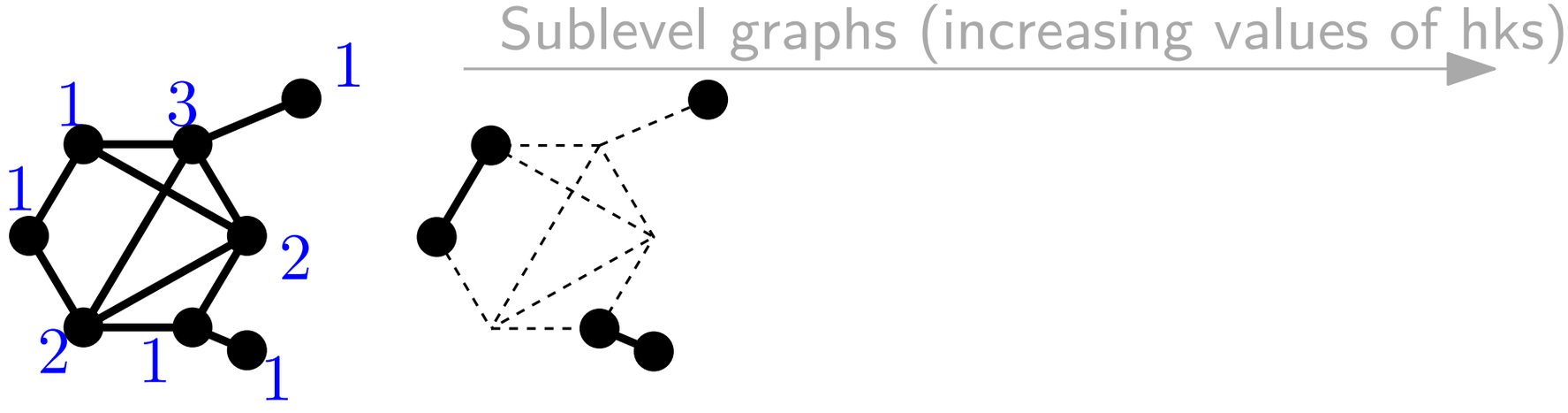


Def: Let $t \geq 0$, and define the *Heat Kernel Signature* of param t:

$$\mathrm{hks}_{G,t} : v \mapsto \sum_{k=1}^{n} \exp(-\lambda_k t)\phi_k(v)^2$$

Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

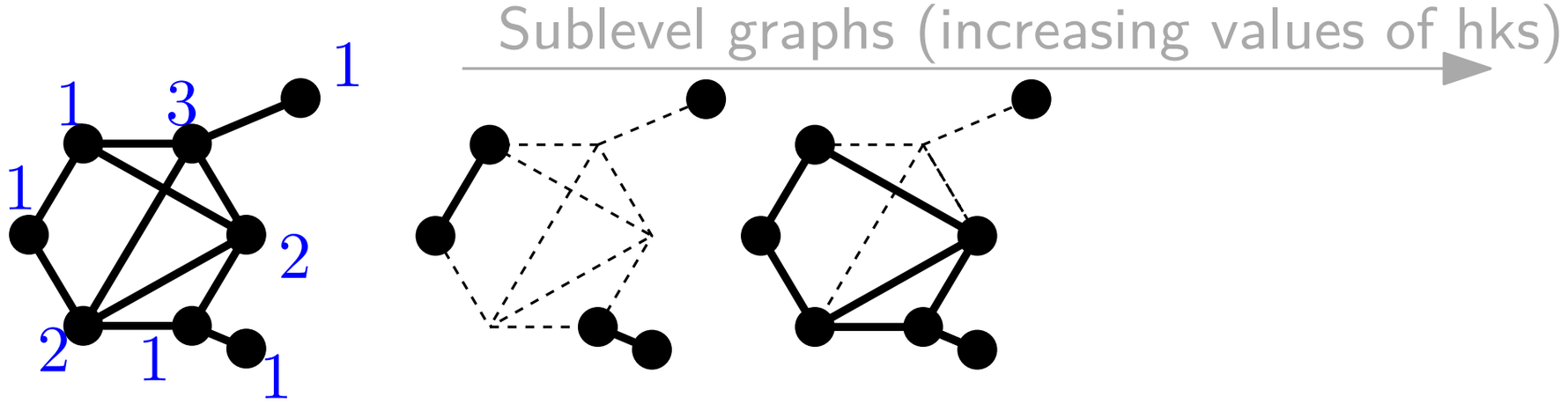


Def: Let $t \geq 0$, and define the *Heat Kernel Signature* of param t:

$$\mathrm{hks}_{G,t} : v \mapsto \sum_{k=1}^{n} \exp(-\lambda_k t)\phi_k(v)^2$$

Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

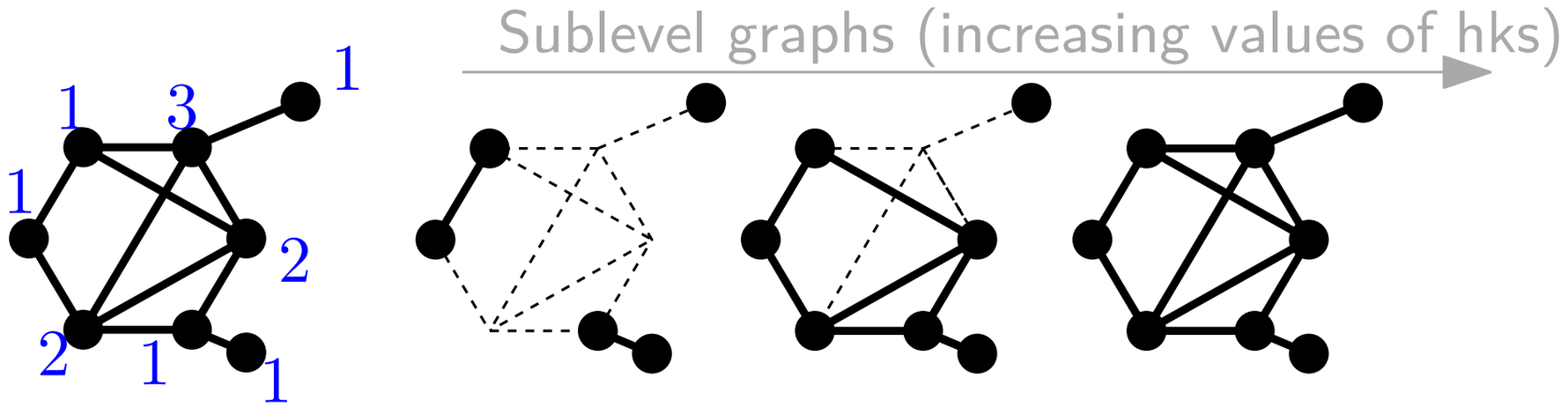


Def: Let $t \geq 0$, and define the *Heat Kernel Signature* of param t:

$$\mathrm{hks}_{G,t} : v \mapsto \sum_{k=1}^{n} \exp(-\lambda_k t)\phi_k(v)^2$$

Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

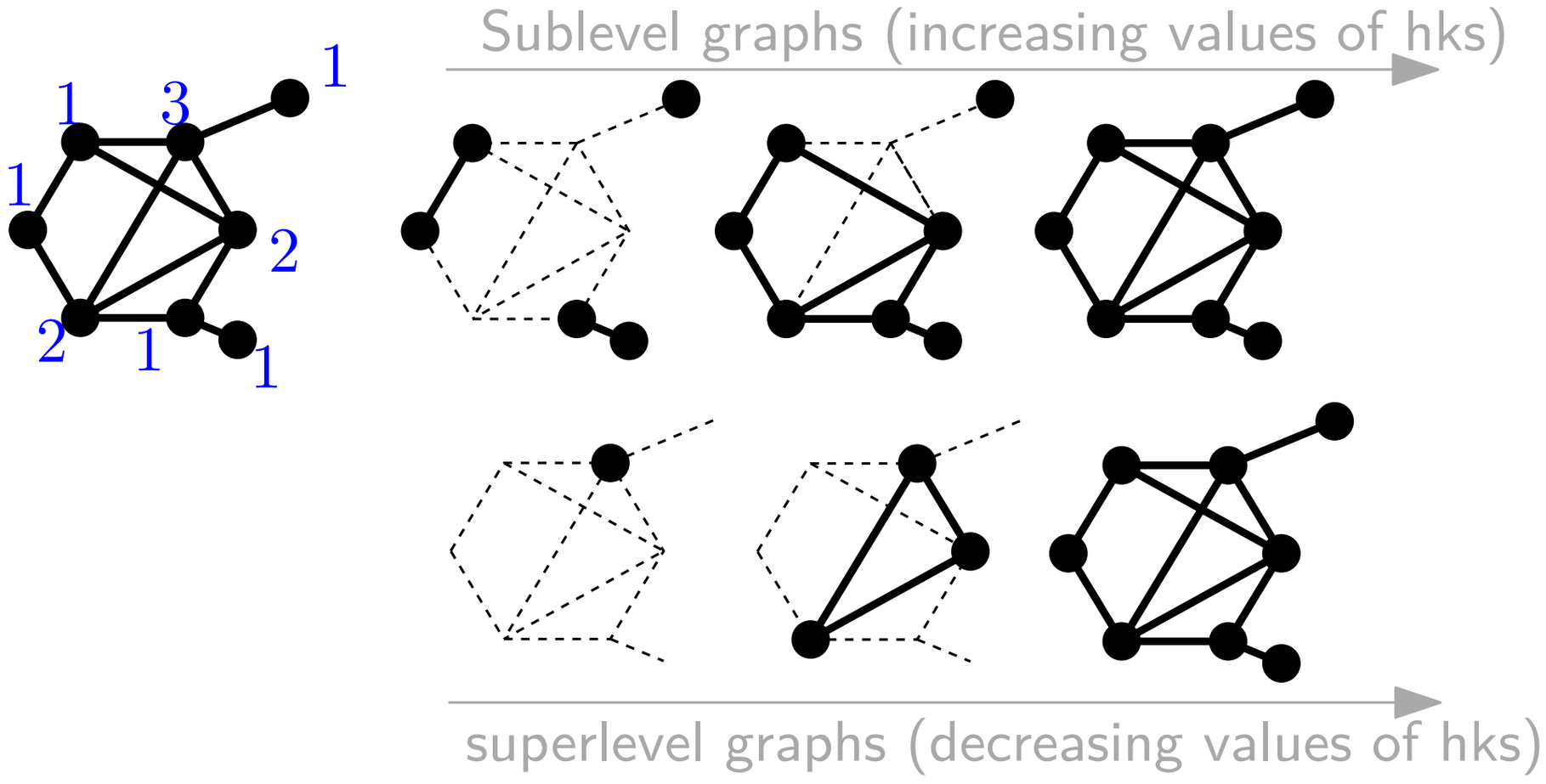


Def: Let $t \geq 0$, and define the *Heat Kernel Signature* of param t:

$$\mathrm{hks}_{G,t} : v \mapsto \sum_{k=1}^{n} \exp(-\lambda_k t)\phi_k(v)^2$$

Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]



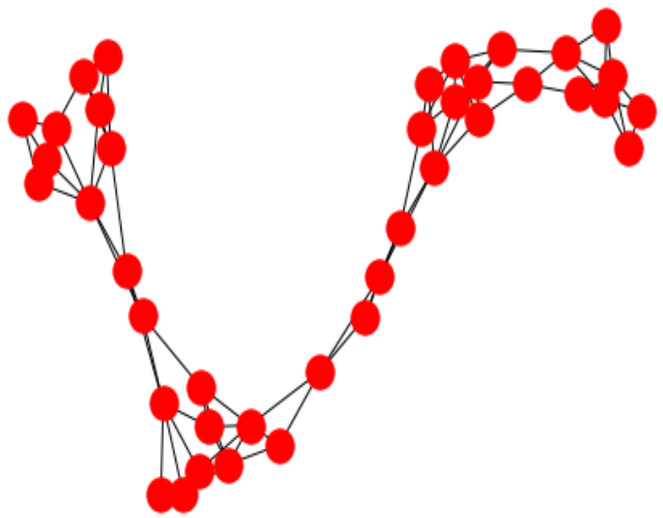
Def: Let $t \geq 0$, and define the *Heat Kernel Signature* of param t:

$$\mathrm{hks}_{G,t} : v \mapsto \sum_{k=1}^{n} \exp(-\lambda_k t)\phi_k(v)^2$$

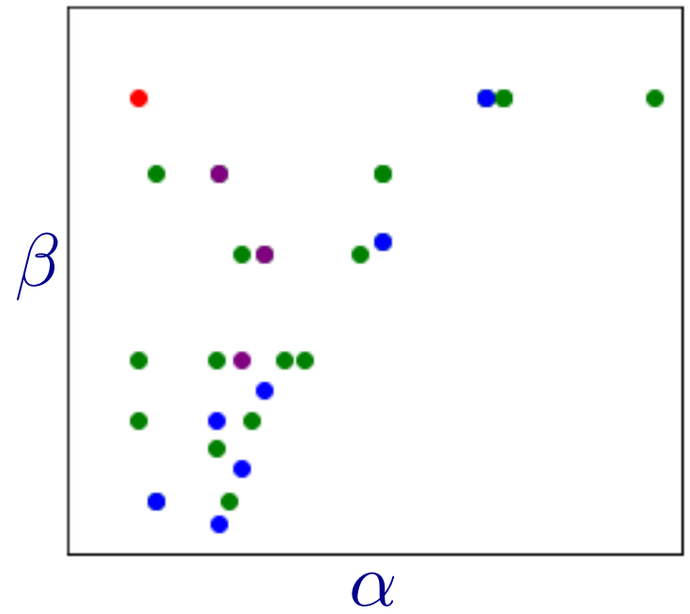
Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]

Graph from the `PROTEINS` dataset

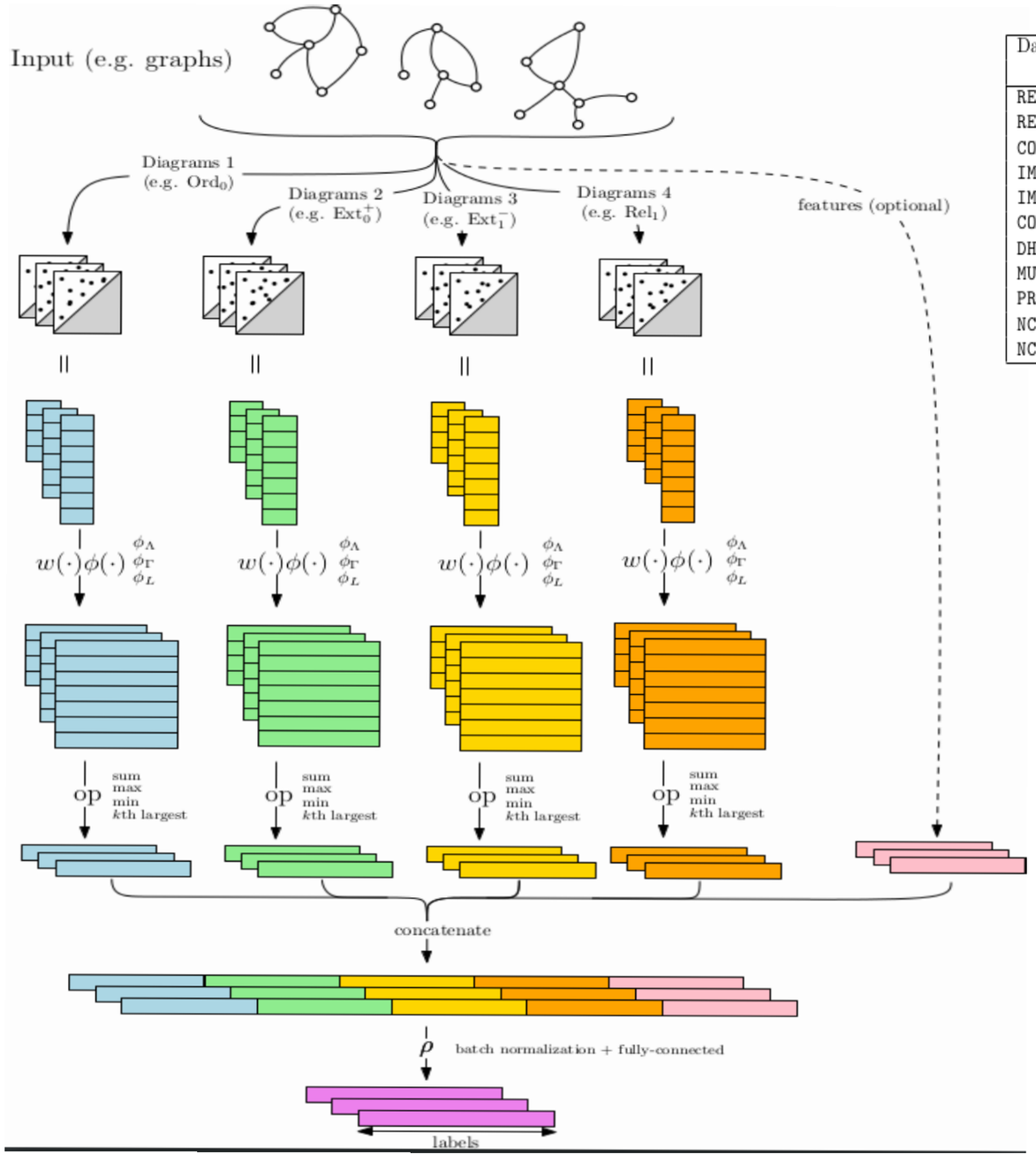


Corresponding extended persistence diagram



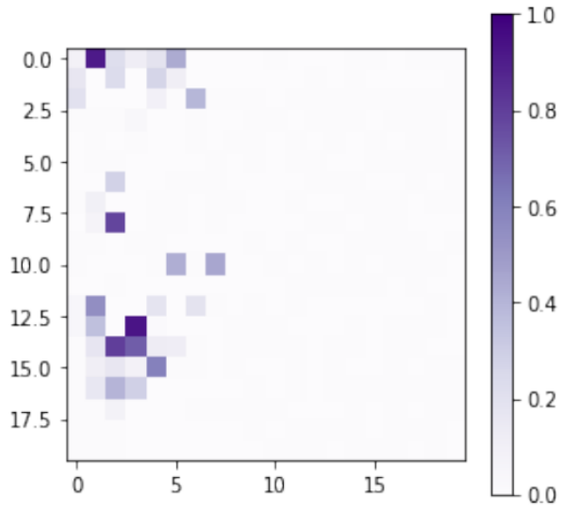
Application to graph classification

[*PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, C., Chazal, Ike, Lacombe, Royer, Umeda, AISTATS, 2019]



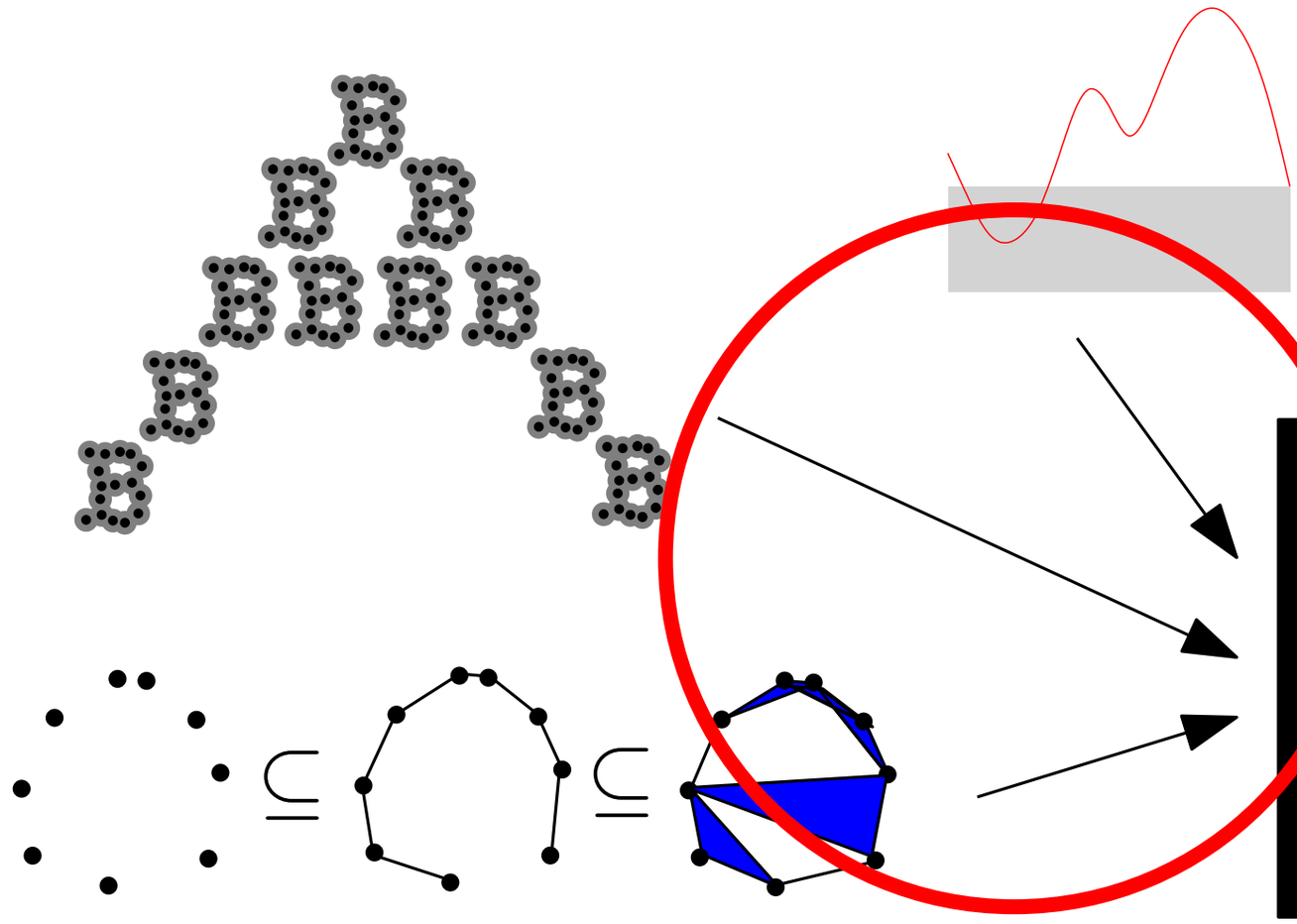
Dataset	SV[1]	RetGK* [2]	FGSD [3]	GCNN [4]	GIN [5]	PersLay	
						Mean	Max
REDDIT5K	—	56.1	47.8	52.9	57.0	55.6	56.5
REDDIT12K	—	48.7	—	46.6	—	47.7	49.1
COLLAB	—	81.0	80.0	79.6	80.1	76.4	78.0
IMDB-B	72.9	71.9	73.6	73.1	74.3	71.2	72.6
IMDB-M	50.3	47.7	52.4	50.3	52.1	48.8	52.2
COX2*	78.4	80.1	—	—	—	80.9	81.6
DHFR*	78.4	81.5	—	—	—	80.3	80.9
MUTAG*	88.3	90.3	92.1	86.7	89.0	89.8	91.5
PROTEINS*	72.6	75.8	73.4	76.3	75.9	74.8	75.9
NCI1*	71.6	84.5	79.8	78.4	82.7	73.5	74.0
NCI109*	70.5	—	78.8	—	—	69.5	70.1

Weight function learnt

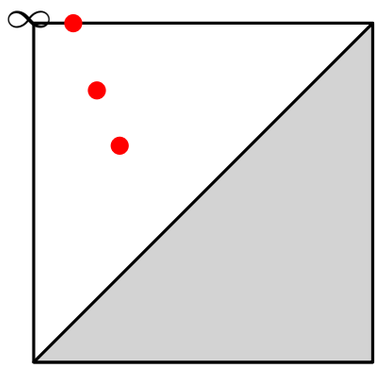


(after training on the MUTAG dataset)

Persistence diagrams and optimization



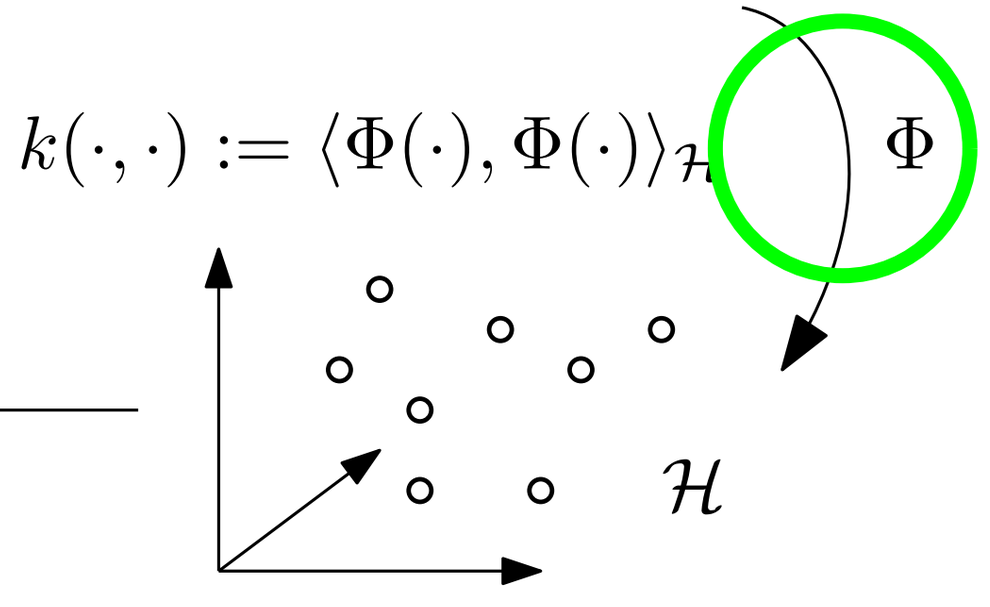
What filtration to choose?



Persistence diagram

- Classifier (RF, SVM, NN etc.)
- Dim. red. (PCA, MDS, UMAP, t-SNE)
- Clustering (DBSCAN, K-means, etc.)

Etc.



What linearization to choose?

Problem setting

Q: How to define ∇D?

Problem setting

Q: How to define ∇D?

Q: Given a parameterized family of functions $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, how to define $\nabla_\theta D_{f_\theta}$?

Q: Given a point cloud $X \subseteq \mathbb{R}^d$, how to define $\nabla_X D_{\mathrm{Rips}}(X)$?

Problem setting

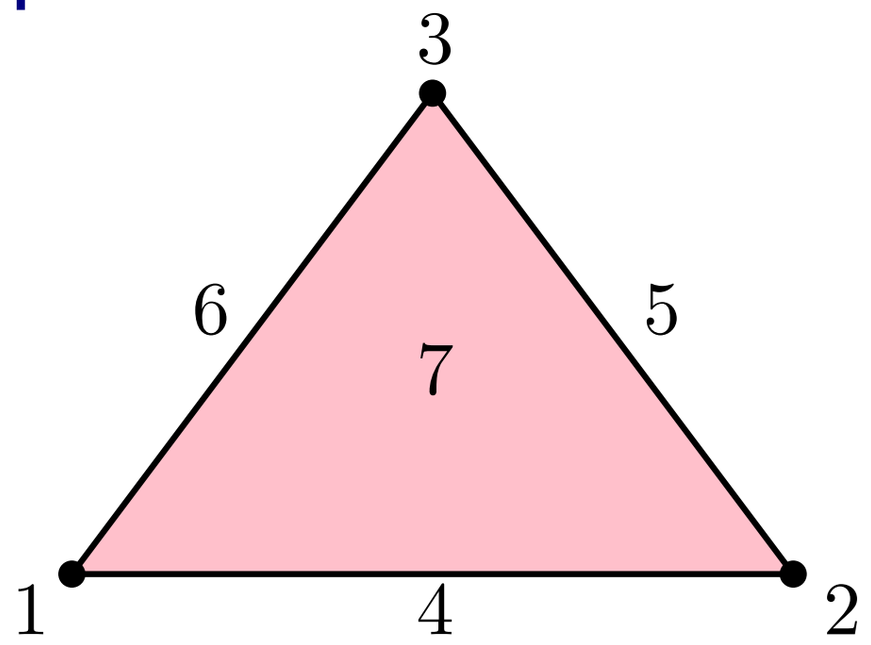
Q: How to define ∇D?

Q: Given a parameterized family of functions $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, how to define $\nabla_\theta D_{f_\theta}$?

Q: Given a point cloud $X \subseteq \mathbb{R}^d$, how to define $\nabla_X D_{\mathrm{Rips}}(X)$?

Idea: Let's go back to the PD construction...

Computation with matrix reduction



Input: simplicial filtration

(Persistent) homology can be computed by using the fact that each simplex is either:

positive, i.e., it *creates a new homology class*

negative, i.e., it *destroys an homology class*

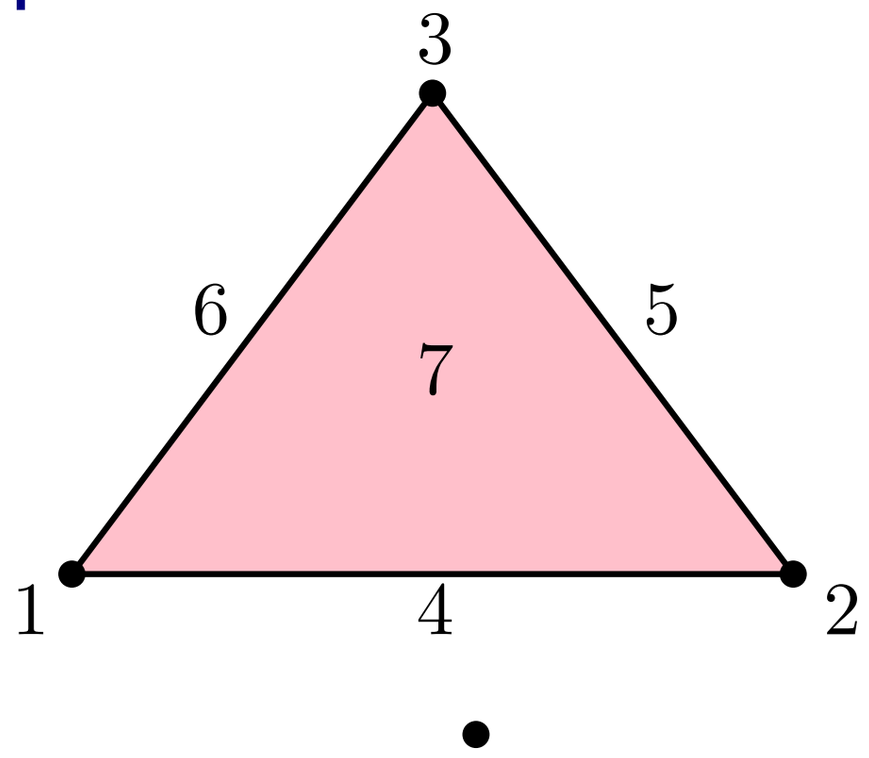
Computation with matrix reduction

Input: simplicial filtration

(Persistent) homology can be computed by using the fact that each simplex is either:

positive, i.e., it *creates a new homology class*

negative, i.e., it *destroys an homology class*



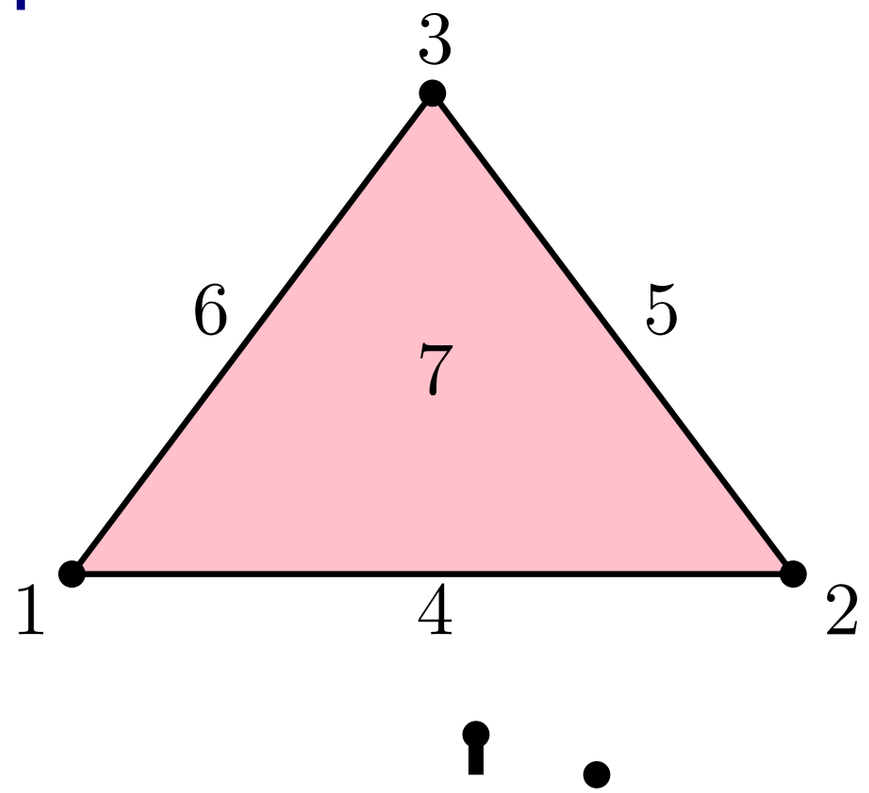
Computation with matrix reduction

Input: simplicial filtration

(Persistent) homology can be computed by using the fact that each simplex is either:

positive, i.e., it *creates a new homology class*

negative, i.e., it *destroys an homology class*



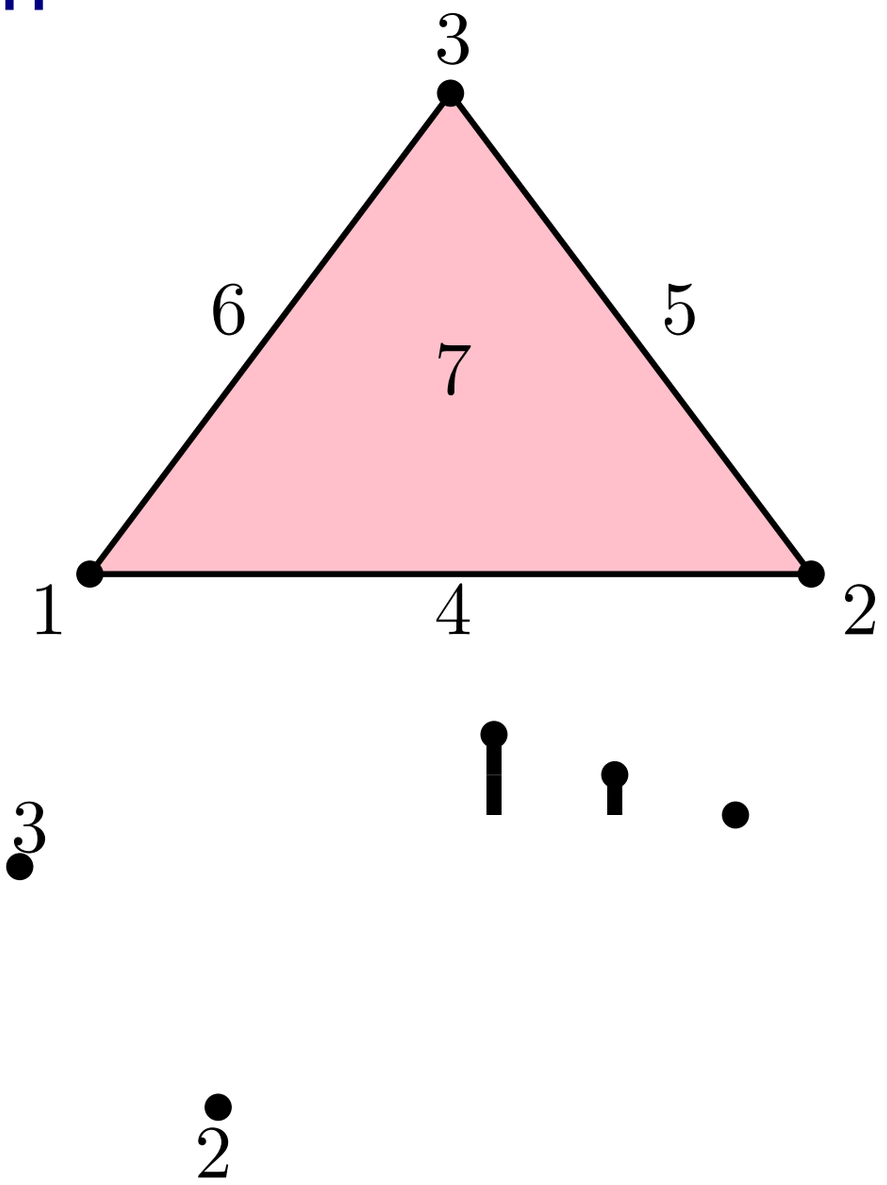
Computation with matrix reduction

Input: simplicial filtration

(Persistent) homology can be computed by using the fact that each simplex is either:

positive, i.e., it *creates a new homology class*

negative, i.e., it *destroys an homology class*



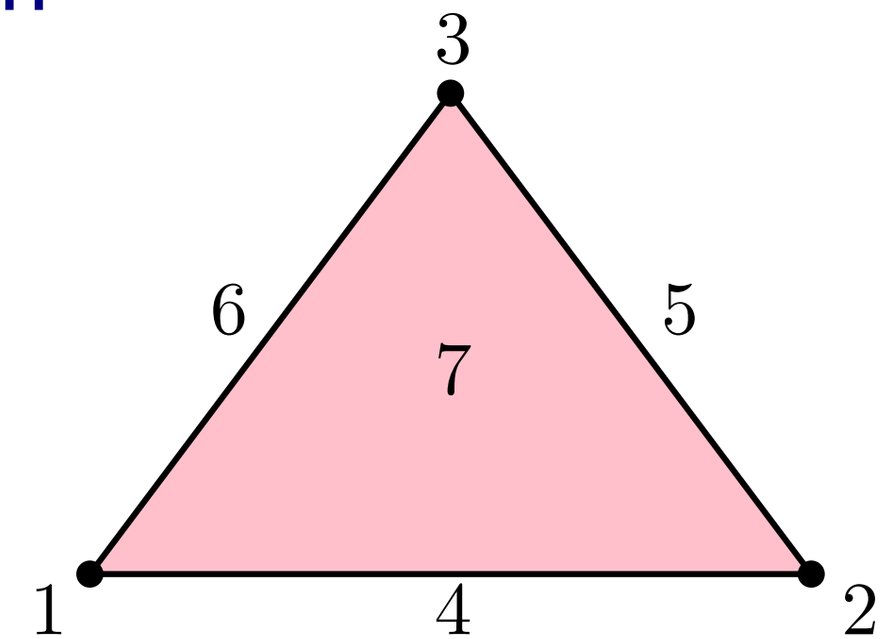
Computation with matrix reduction

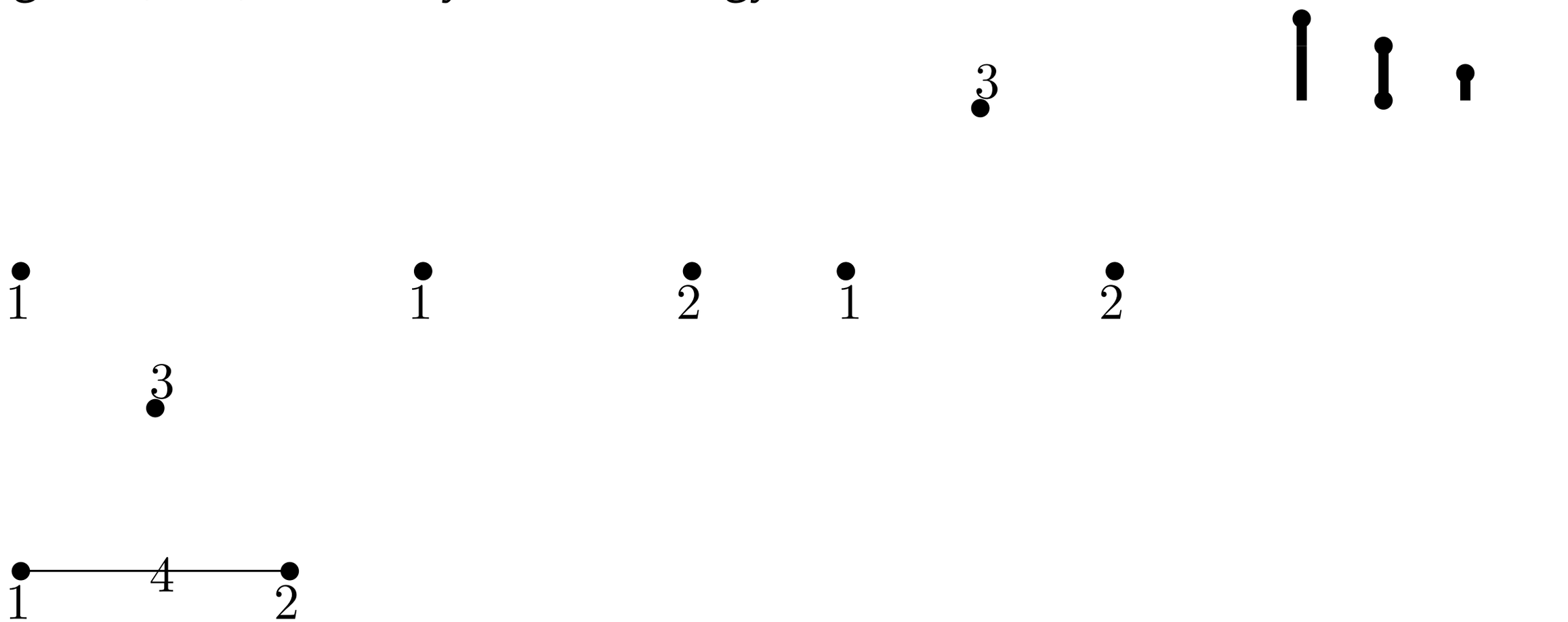
Input: simplicial filtration

(Persistent) homology can be computed by using the fact that each simplex is either:

positive, i.e., it *creates a new homology class*

negative, i.e., it *destroys an homology class*





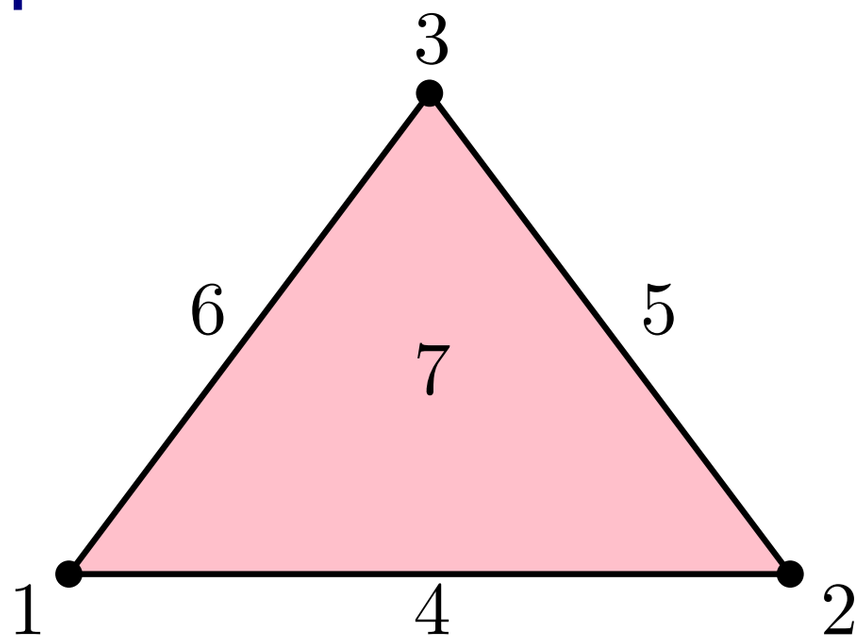
Computation with matrix reduction

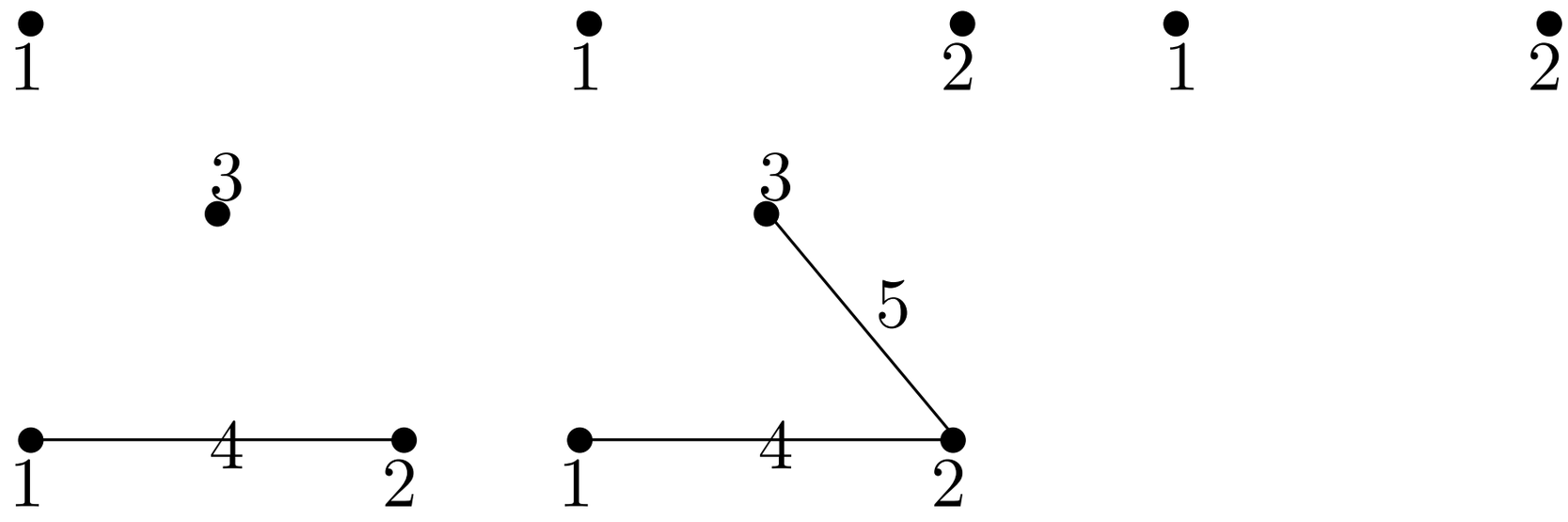
Input: simplicial filtration

(Persistent) homology can be computed by using the fact that each simplex is either:

positive, i.e., it *creates a new homology class*

negative, i.e., it *destroys an homology class*





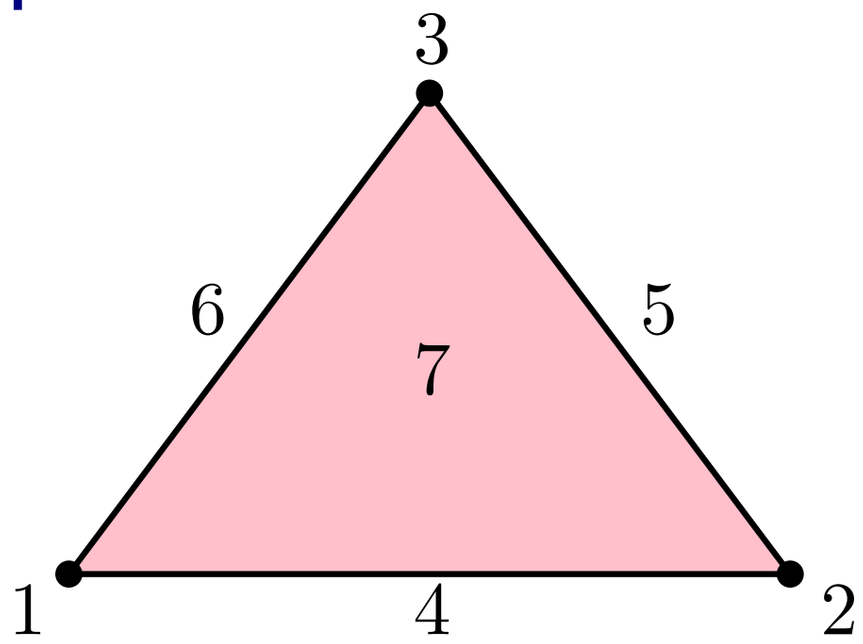
Computation with matrix reduction

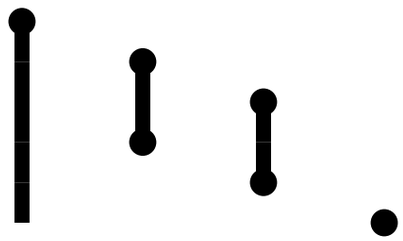
Input: simplicial filtration

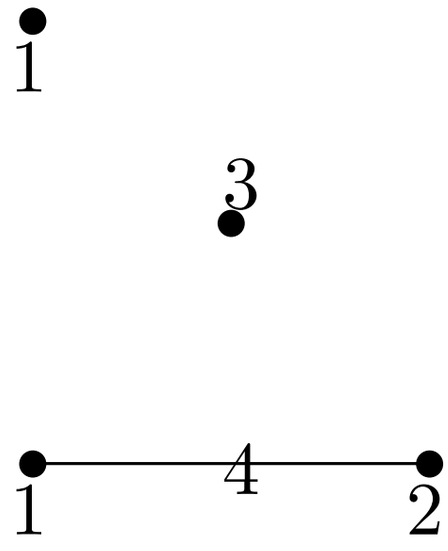
(Persistent) homology can be computed by using the fact that each simplex is either:

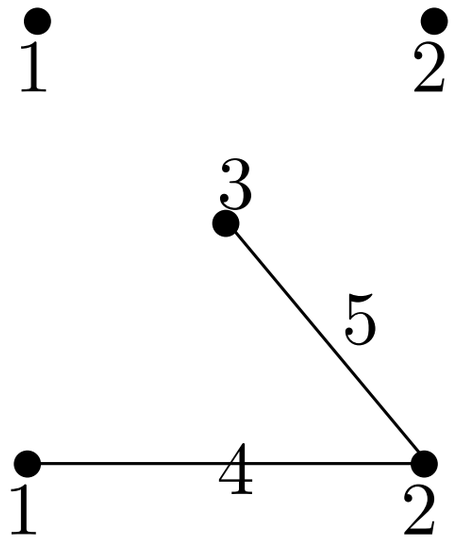
positive, i.e., it *creates a new homology class*

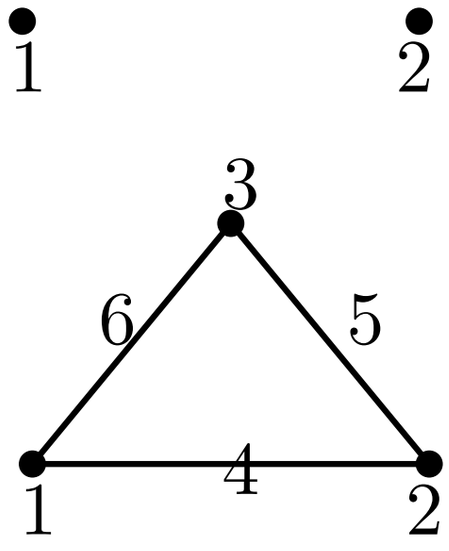
negative, i.e., it *destroys an homology class*











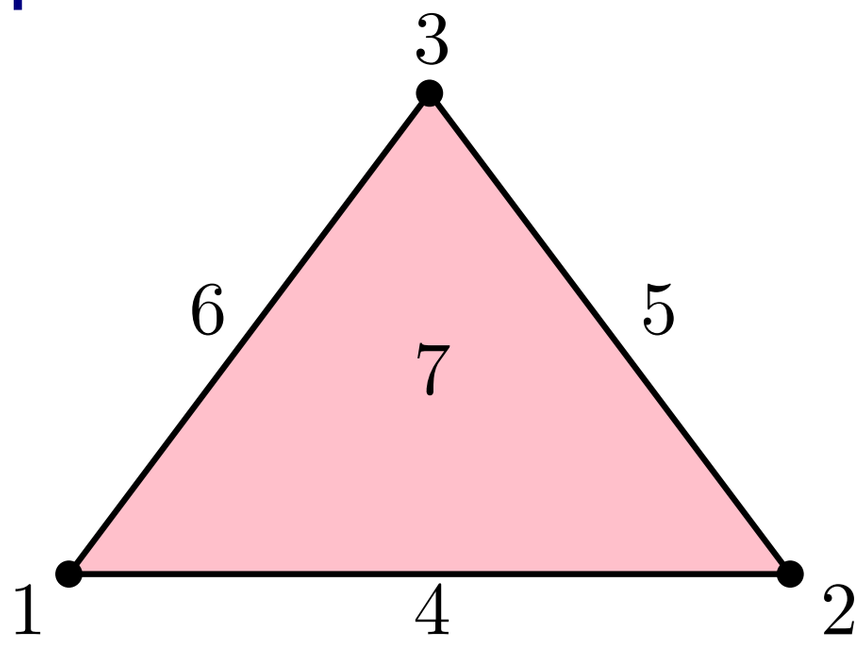
Computation with matrix reduction

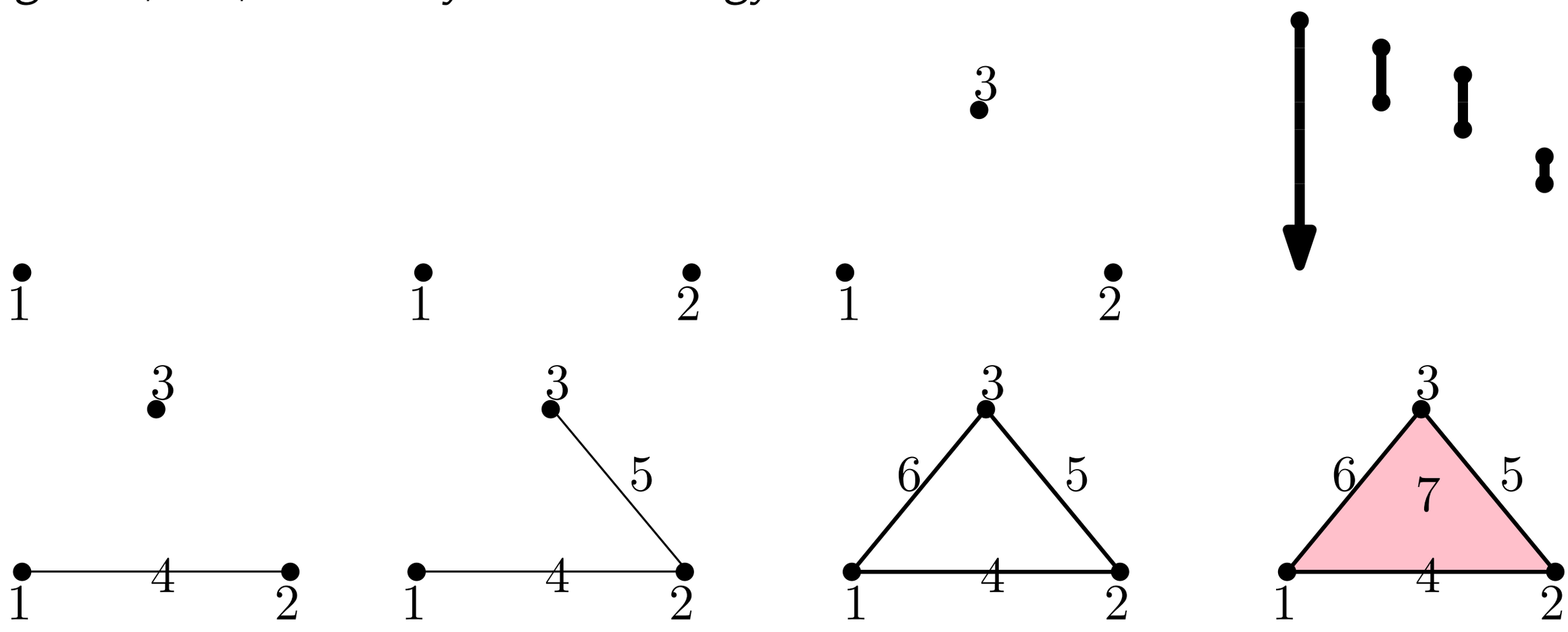
Input: simplicial filtration

(Persistent) homology can be computed by using the fact that each simplex is either:

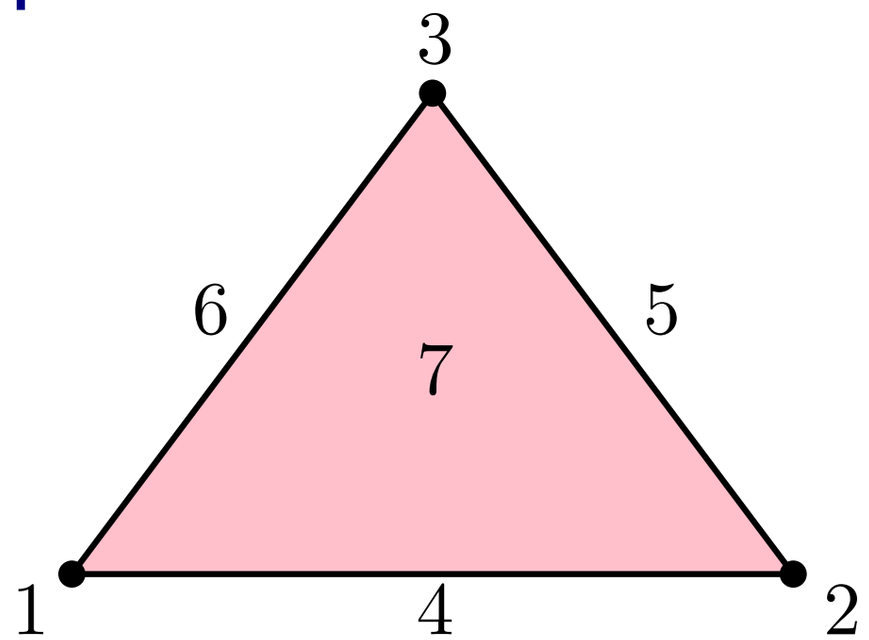
positive, i.e., it *creates a new homology class*

negative, i.e., it *destroys an homology class*





Computation with matrix reduction



Input: simplicial filtration

Output: boundary matrix
reduced to column-echelon form

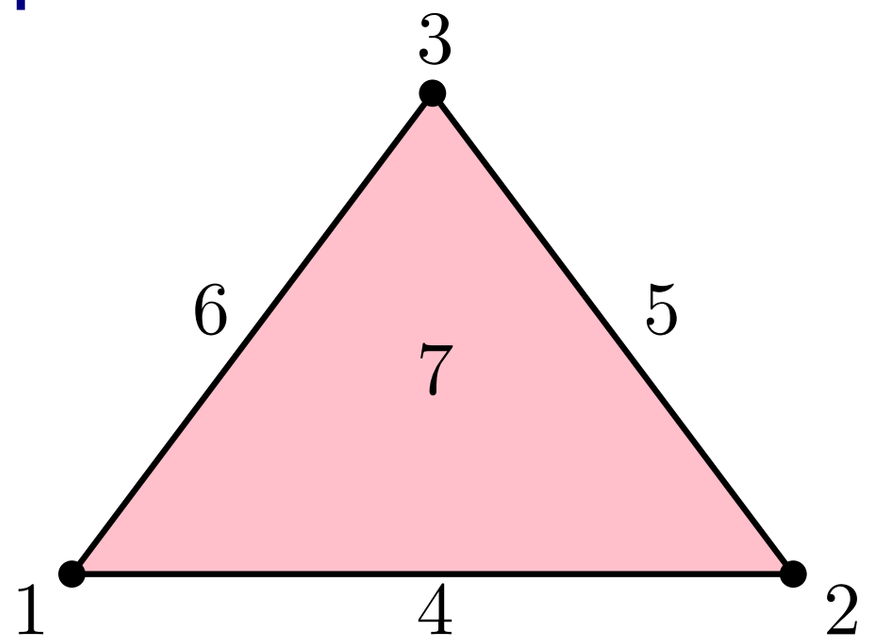
◯ simplex pairs give finite intervals:
$[2,4)$, $[3,5)$, $[6,7)$

□ unpaired simplices give infinite intervals: $[1,+\infty)$

	1	2	3	4	5	6	7
1				$*$		$*$	
2				$*$	$*$		
3					$*$	$*$	
4							$*$
5							$*$
6							$*$
7							

	1	2	3	4	5	6	7
1				$*$			
2				1	$*$		
3					1		
4							$*$
5							$*$
6							1
7							

Computation with matrix reduction



Input: simplicial filtration

Output: boundary matrix
reduced to column-echelon form

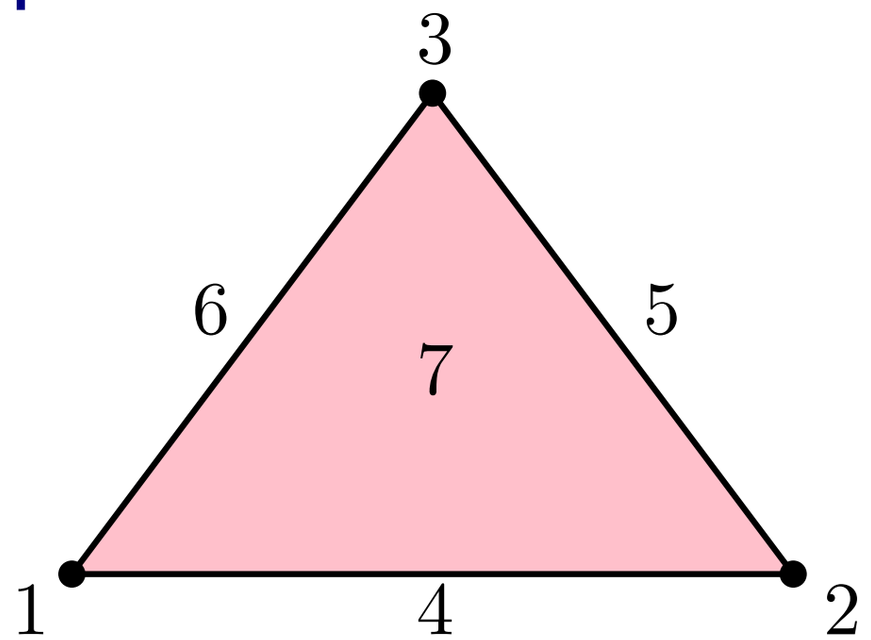
◯ simplex pairs give finite intervals:
$[2,4)$, $[3,5)$, $[6,7)$

□ unpaired simplices give infinite intervals: $[1,+\infty)$

A persistence diagram D is made of all $(\mathcal{F}(\sigma_+), \mathcal{F}(\sigma_-)) \in \mathbb{R}^2$ where σ_+ (resp. σ_-) is positive (resp. negative), and $\mathcal{F}$ is the filtration function.

	1	2	3	4	5	6	7
1				$*$			
2				(1)	$*$		
3					(1)		
4							$*$
5							$*$
6							(1)
7							

Computation with matrix reduction



Input: simplicial filtration

Output: boundary matrix
reduced to column-echelon form

◯ simplex pairs give finite intervals:
$[2,4)$, $[3,5)$, $[6,7)$

□ unpaired simplices give infinite intervals: $[1,+\infty)$

A persistence diagram D is made of all $(\mathcal{F}(\sigma_+),\mathcal{F}(\sigma_-)) \in \mathbb{R}^2$ where σ_+ (resp. σ_-) is positive (resp. negative), and $\mathcal{F}$ is the filtration function.

Thus we can define the gradient of a point $p=(\mathcal{F}(\sigma_+),\mathcal{F}(\sigma_-)) \in D$ as

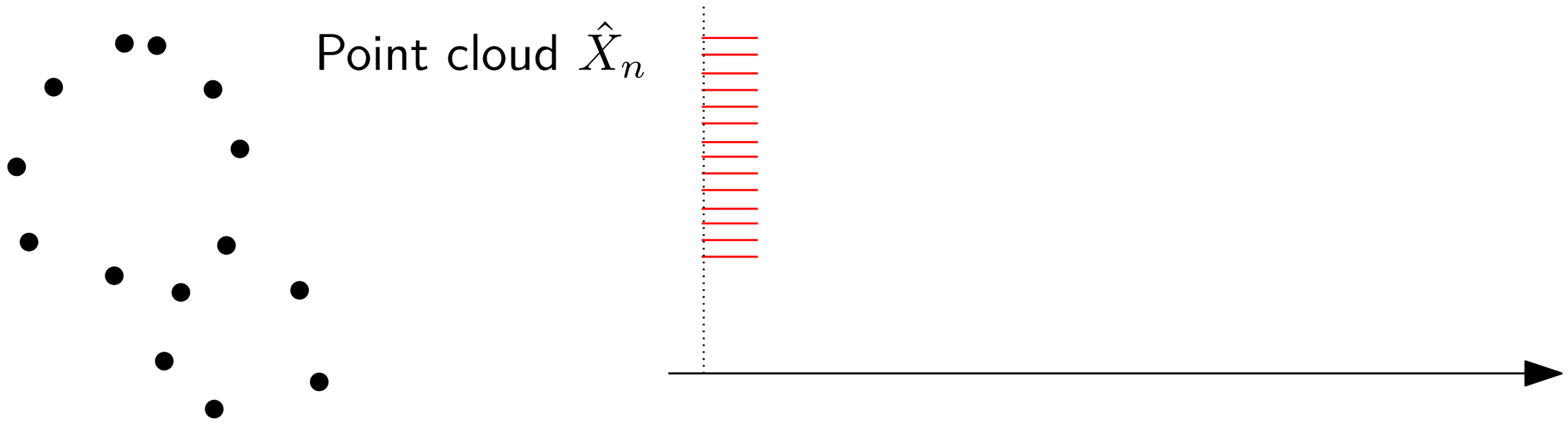
$$\nabla p = [\nabla\mathcal{F}(\sigma_+),\nabla\mathcal{F}(\sigma_-)]$$

	1	2	3	4	5	6	7
1				$*$			
2				1	$*$		
3					1		
4							$*$
5							$*$
6							1
7							

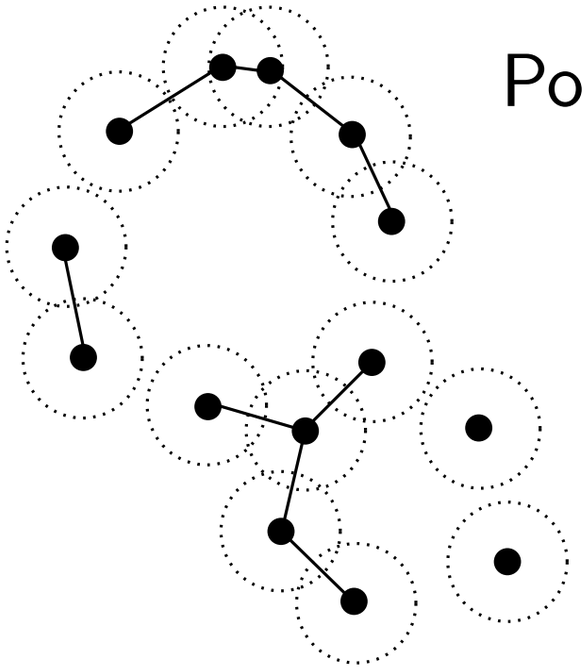
Example: Vietoris-Rips gradient

Q: Define and compute Vietoris-Rips gradient?

Example: Vietoris-Rips gradient

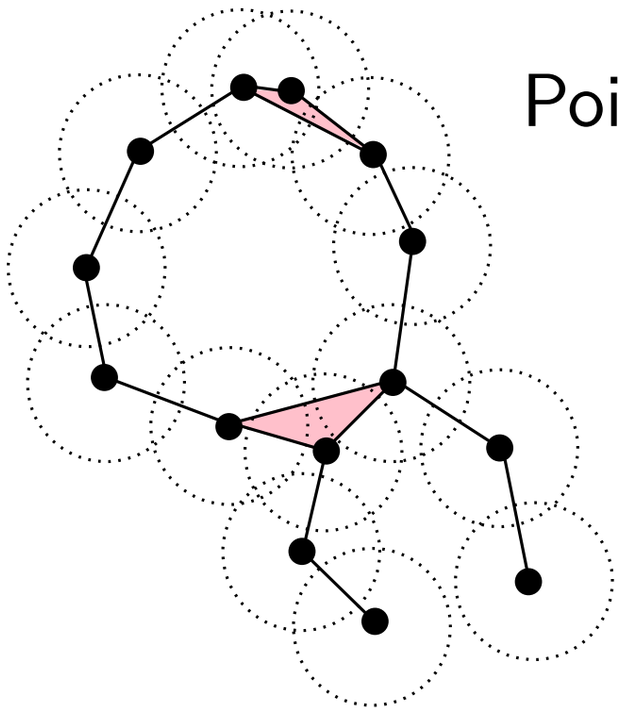


Example: Vietoris-Rips gradient

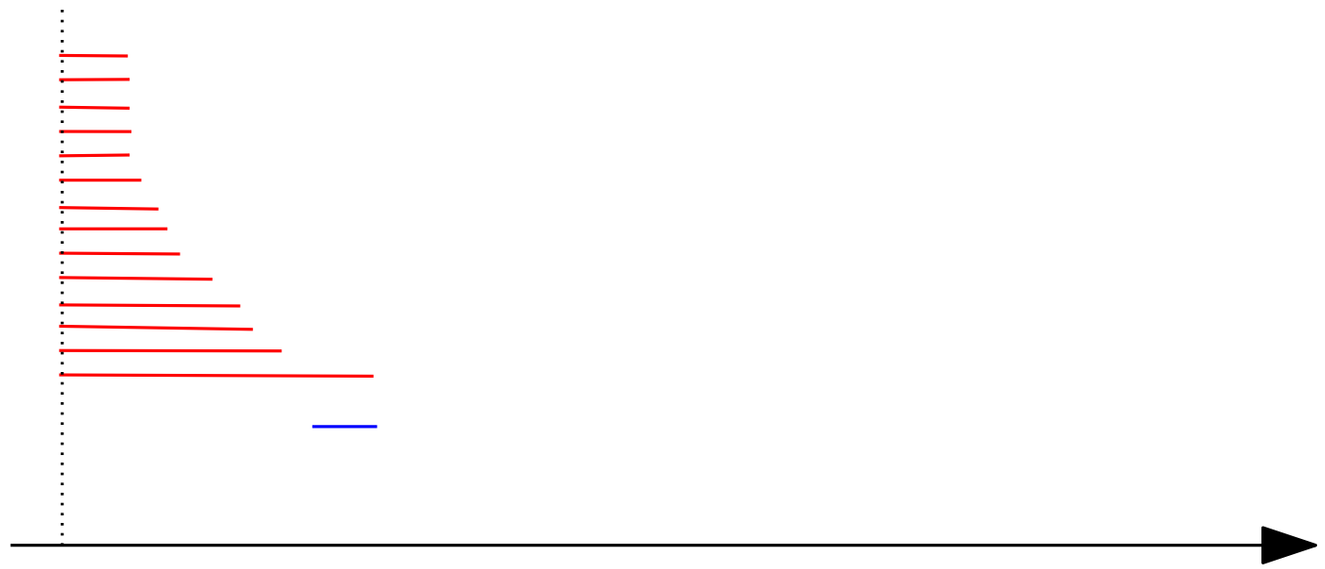


Point cloud $\hat{X}_n$

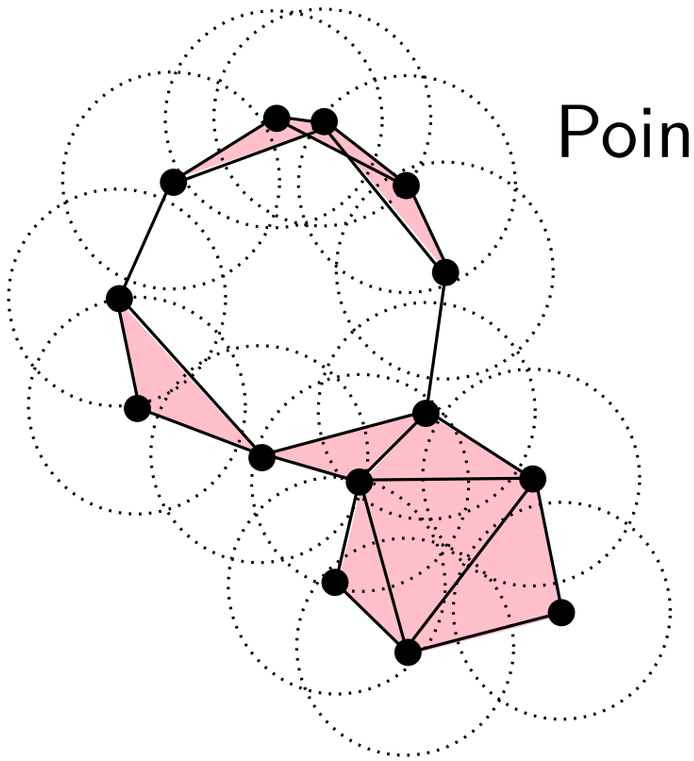
Example: Vietoris-Rips gradient



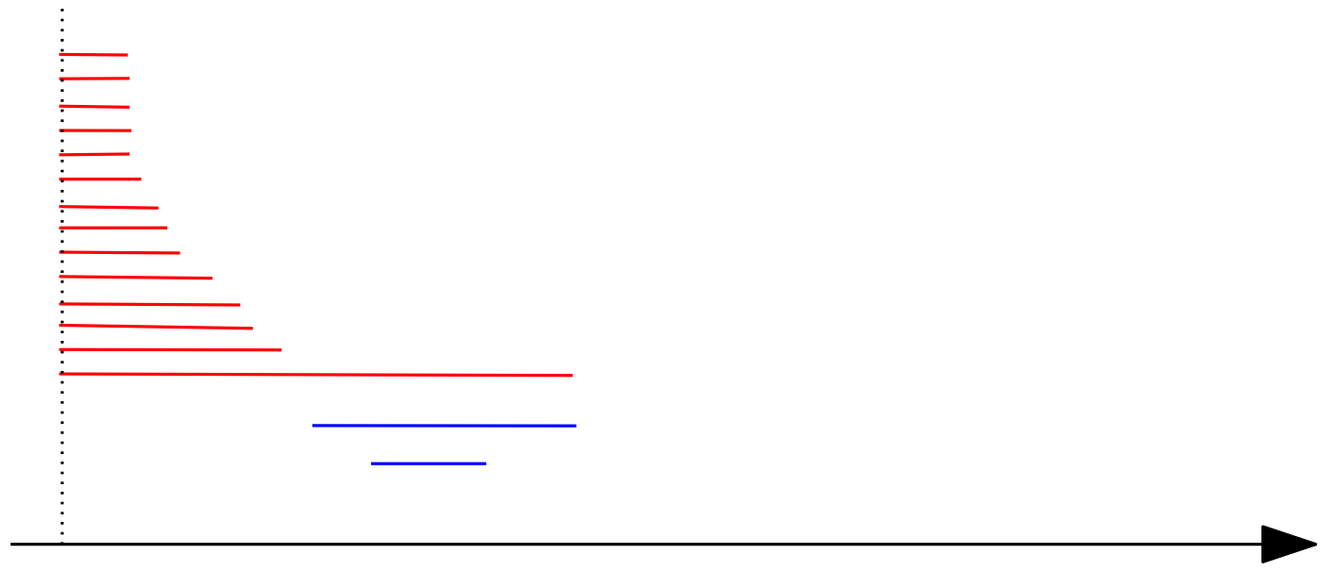
Point cloud $\hat{X}_n$



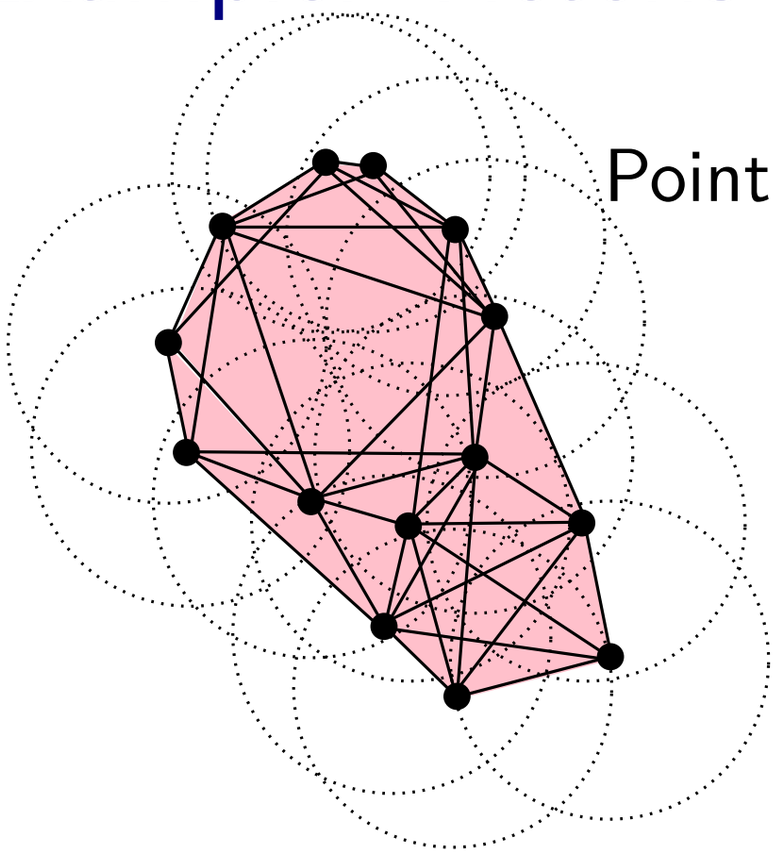
Example: Vietoris-Rips gradient



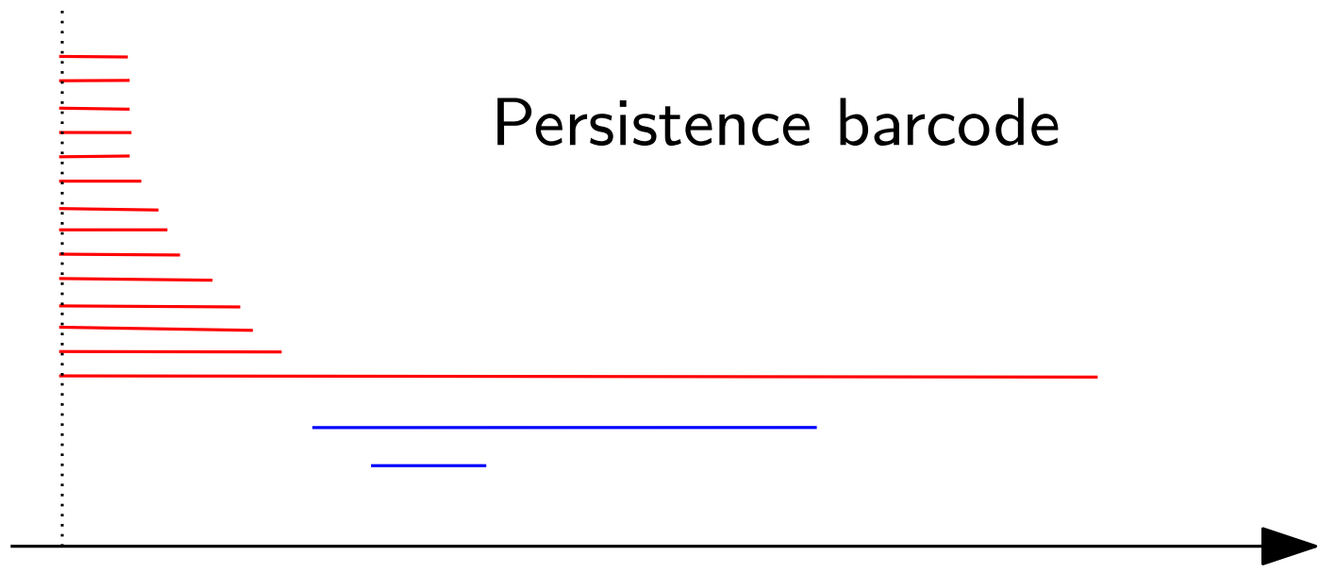
Point cloud $\hat{X}_n$



Example: Vietoris-Rips gradient

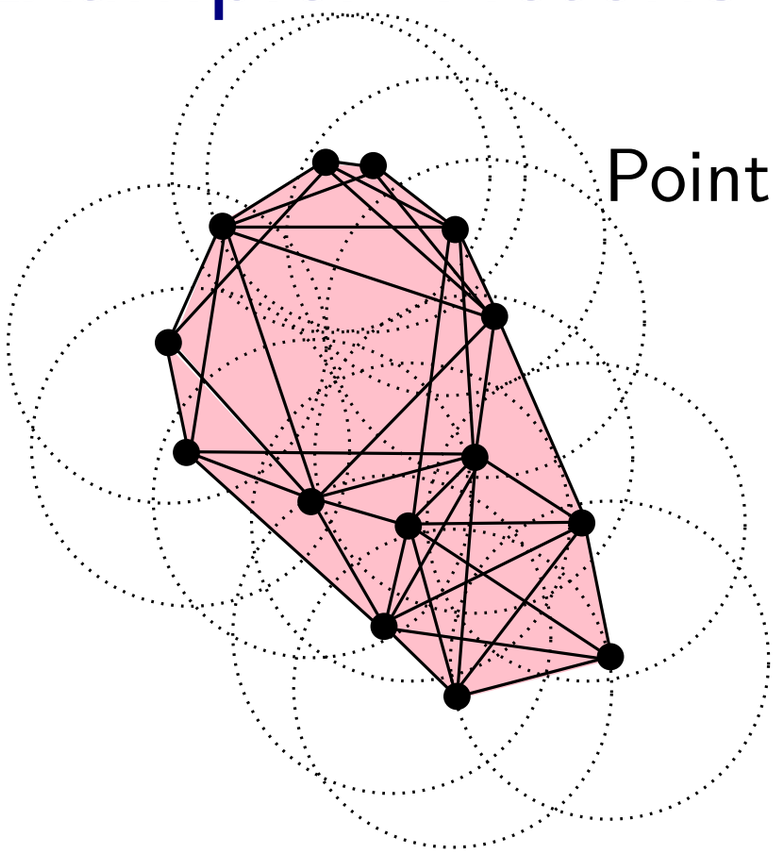


Point cloud $\hat{X}_n$

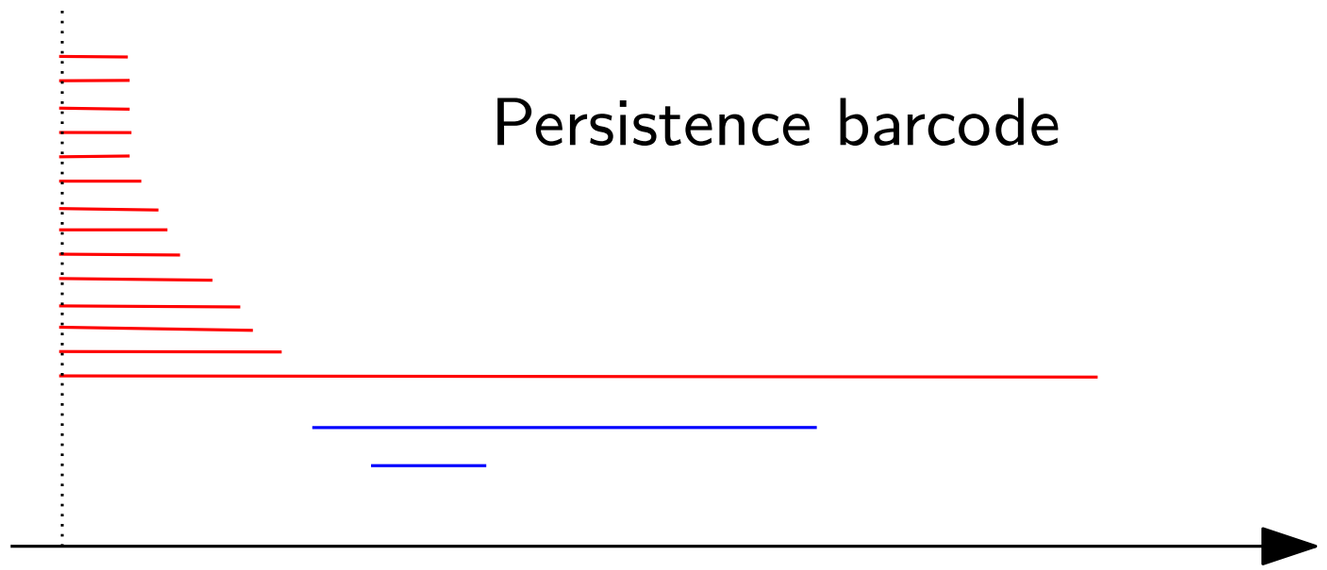


Persistence barcode

Example: Vietoris-Rips gradient



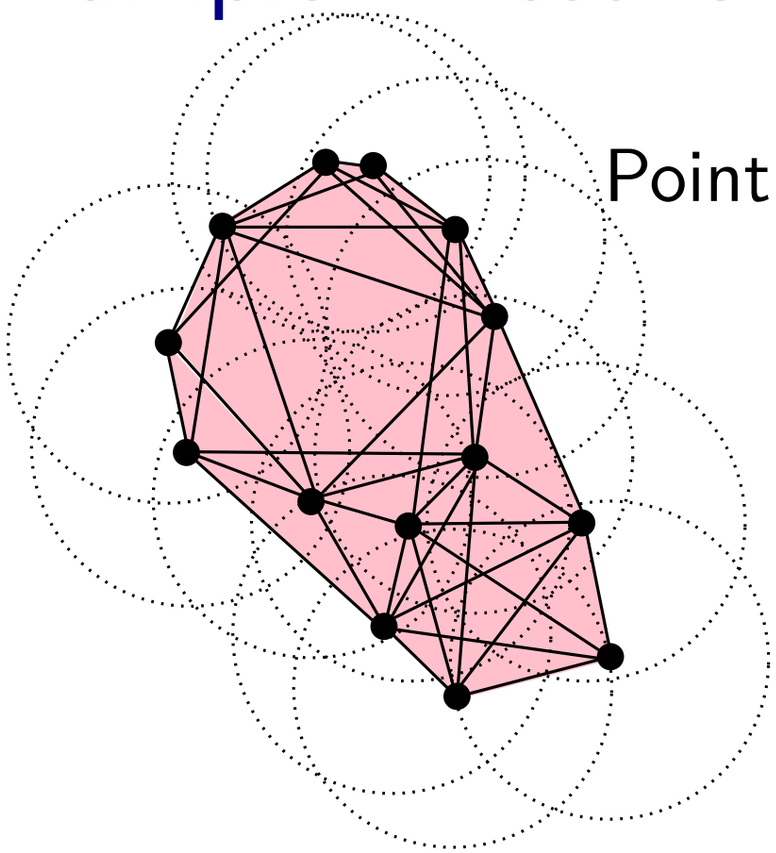
Point cloud $\hat{X}_n$



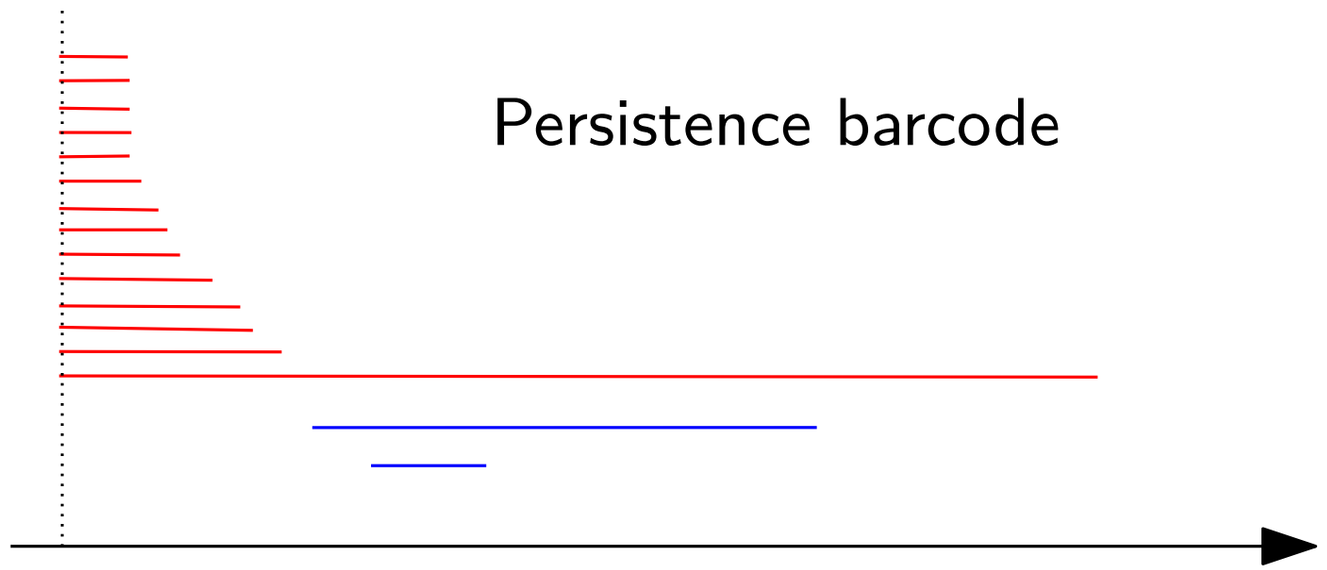
Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_{i,j} \|v_i - v_j\|$$

Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$

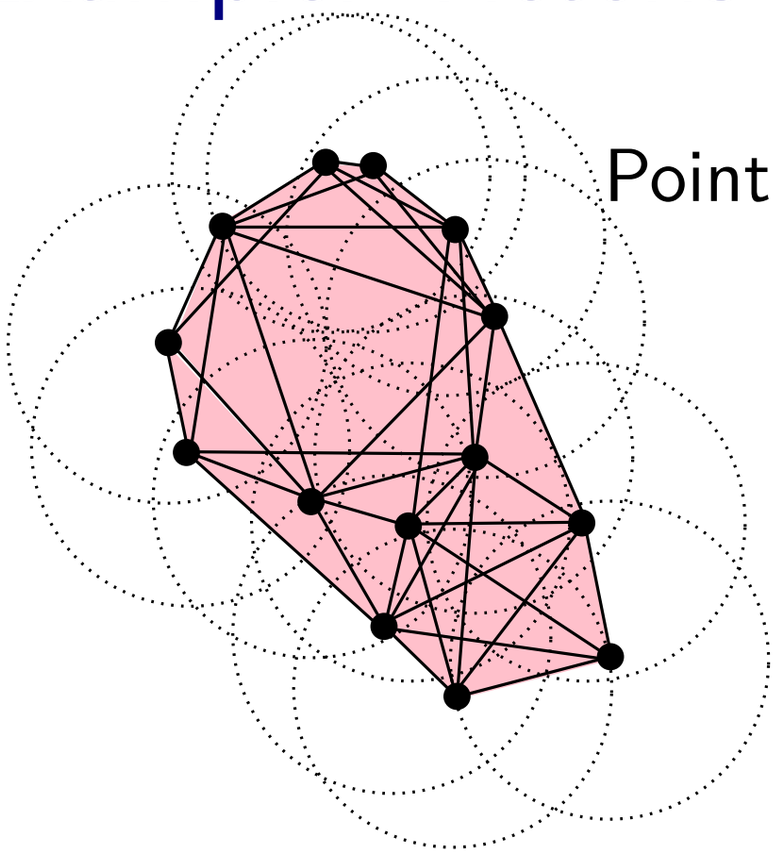


Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

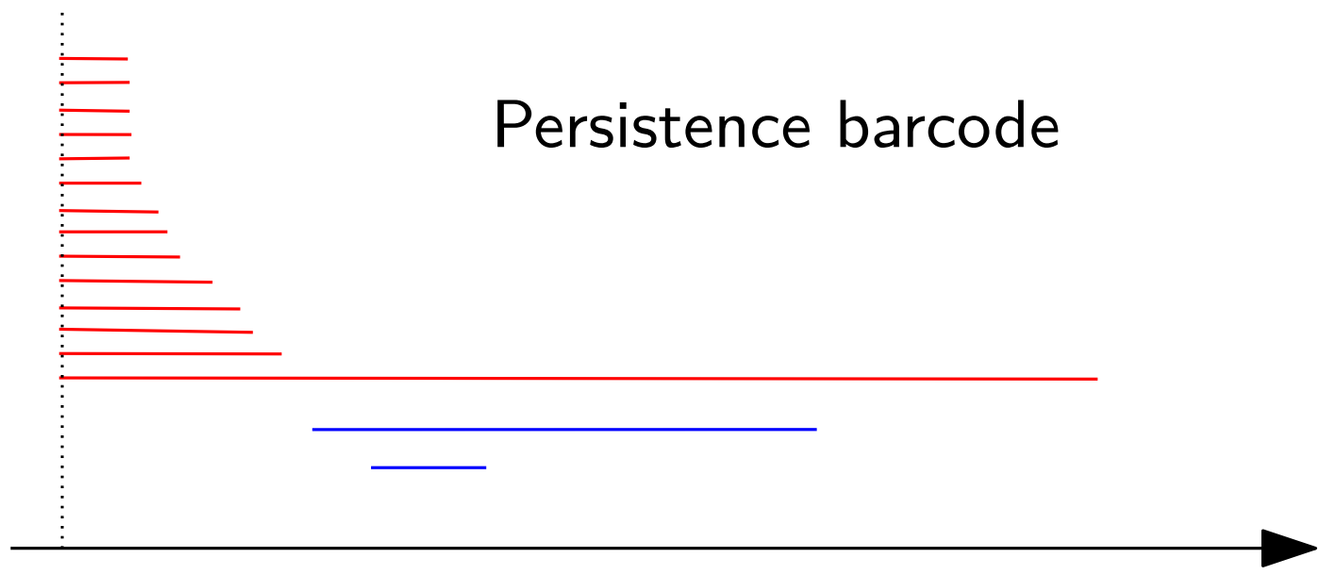
$$\mathcal{F}(\sigma) = \max_{i,j} \|v_i - v_j\|$$

Let $p = (\mathcal{F}(\sigma_+), \mathcal{F}(\sigma_-)) \in D_{\mathrm{Rips}}(X)$

Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$



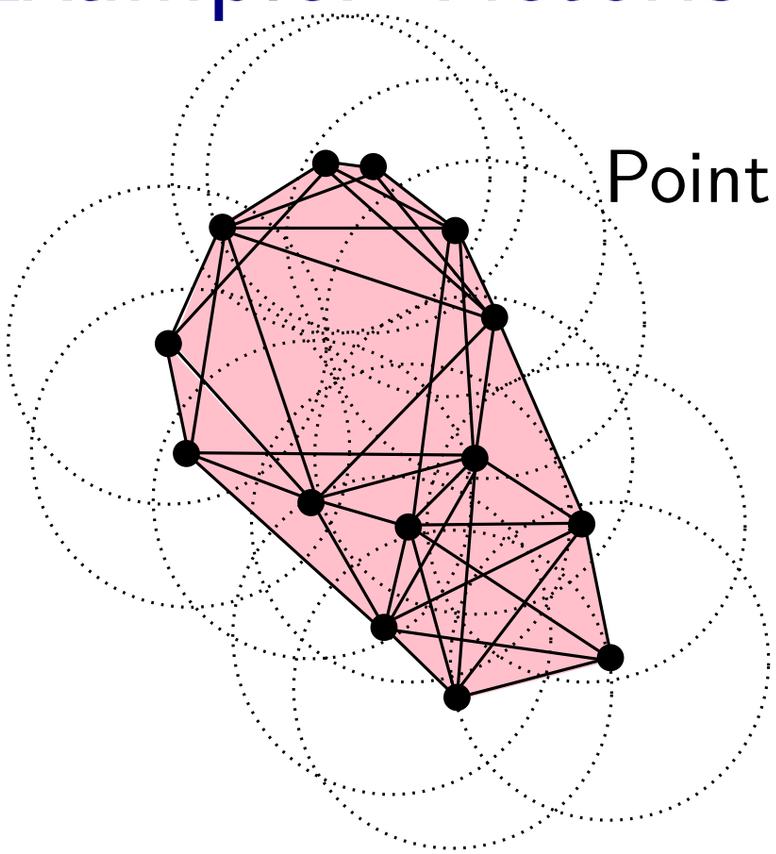
Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_{i,j} \|v_i - v_j\|$$

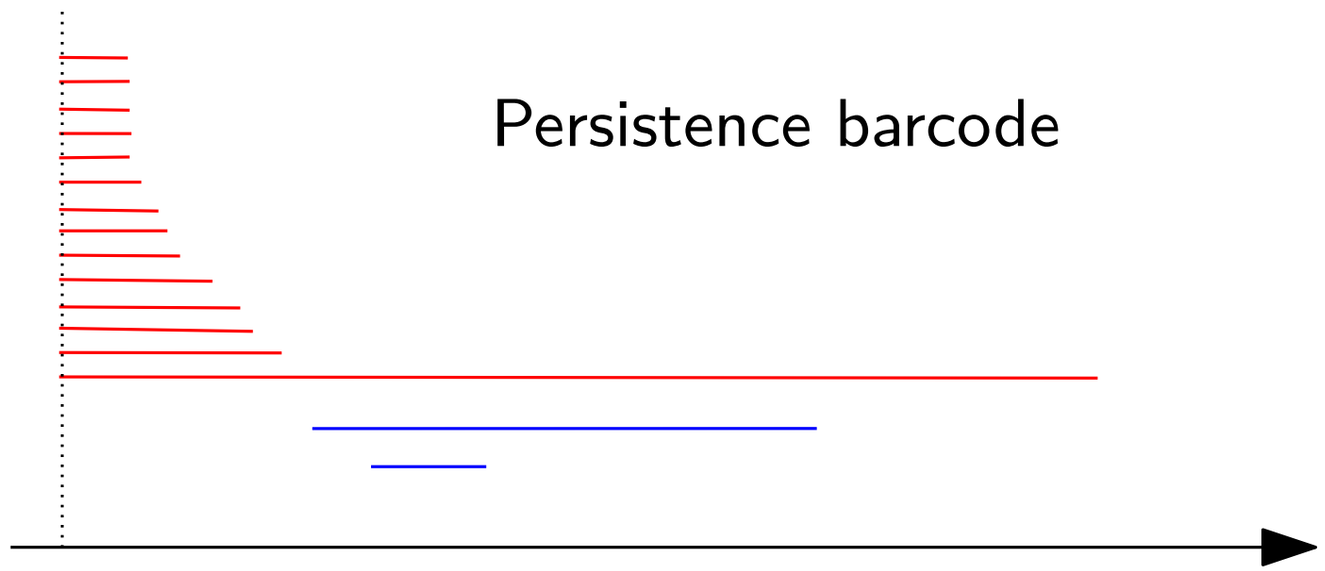
Let $p = (\mathcal{F}(\sigma_+), \mathcal{F}(\sigma_-)) \in D_{\mathrm{Rips}}(X)$

with $\sigma_+ = \{v_0, \ldots, v_k\}$ and $\sigma_- = \{w_0, \ldots, w_{k+1}\}$

Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$



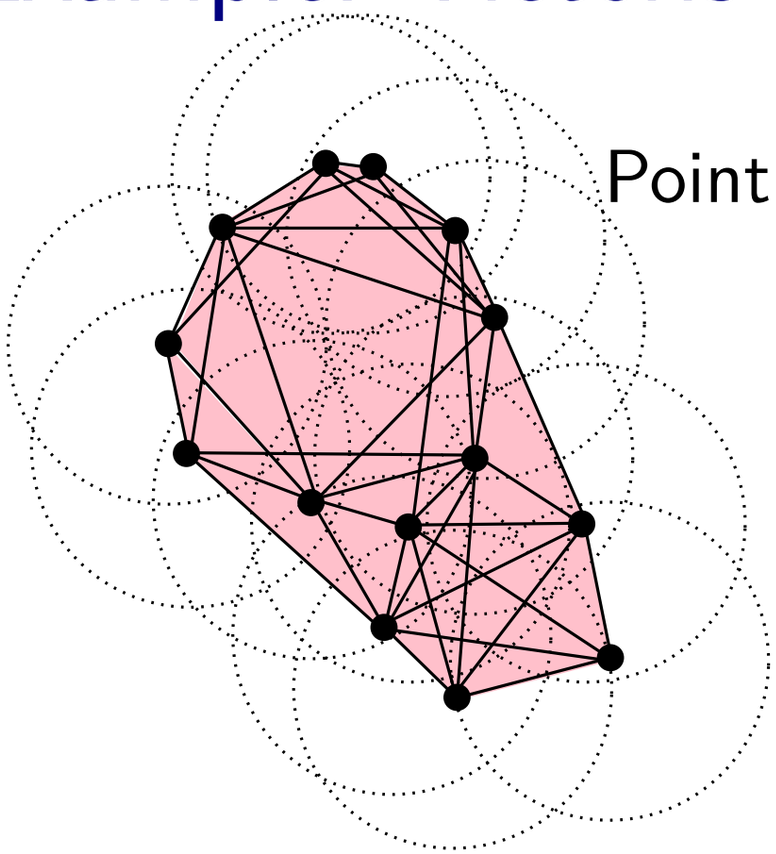
Persistence barcode

$$\nabla_X p = \left[\frac{\partial}{\partial X} \| v_{i^*} - v_{j^*} \|, \frac{\partial}{\partial X} \| w_{a^*} - w_{b^*} \| \right]$$

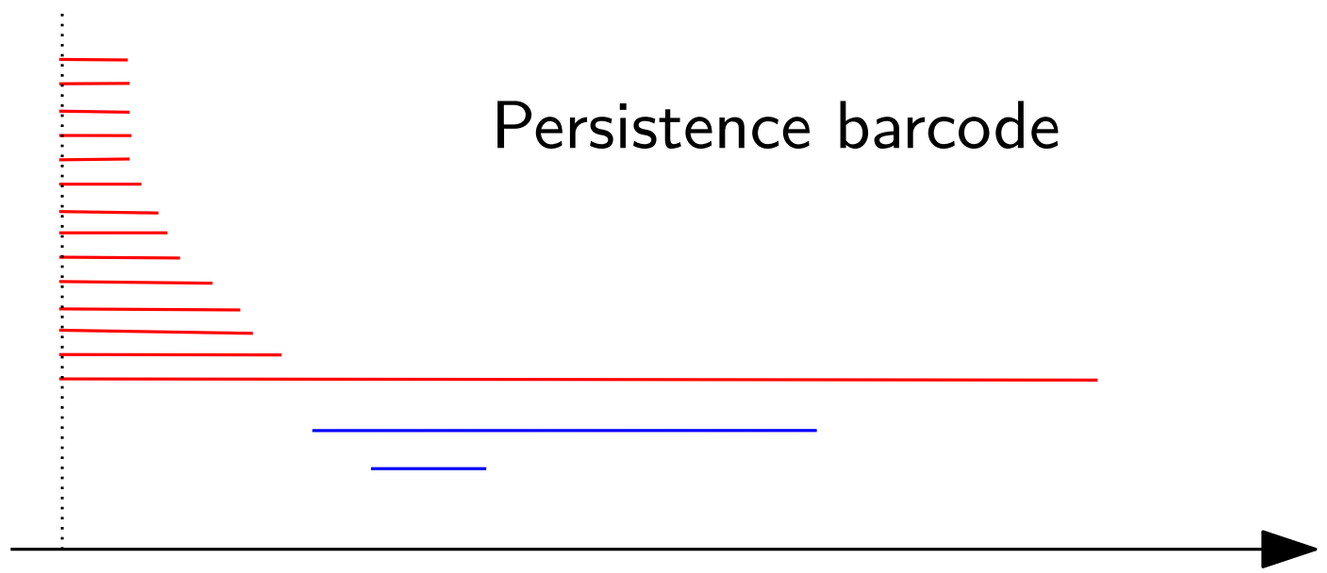
Let $p = (\mathcal{F}(\sigma_+), \mathcal{F}(\sigma_-)) \in D_{\mathrm{Rips}}(X)$

with $\sigma_+ = \{v_0, \ldots, v_k\}$ and $\sigma_- = \{w_0, \ldots, w_{k+1}\}$

Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$



Persistence barcode

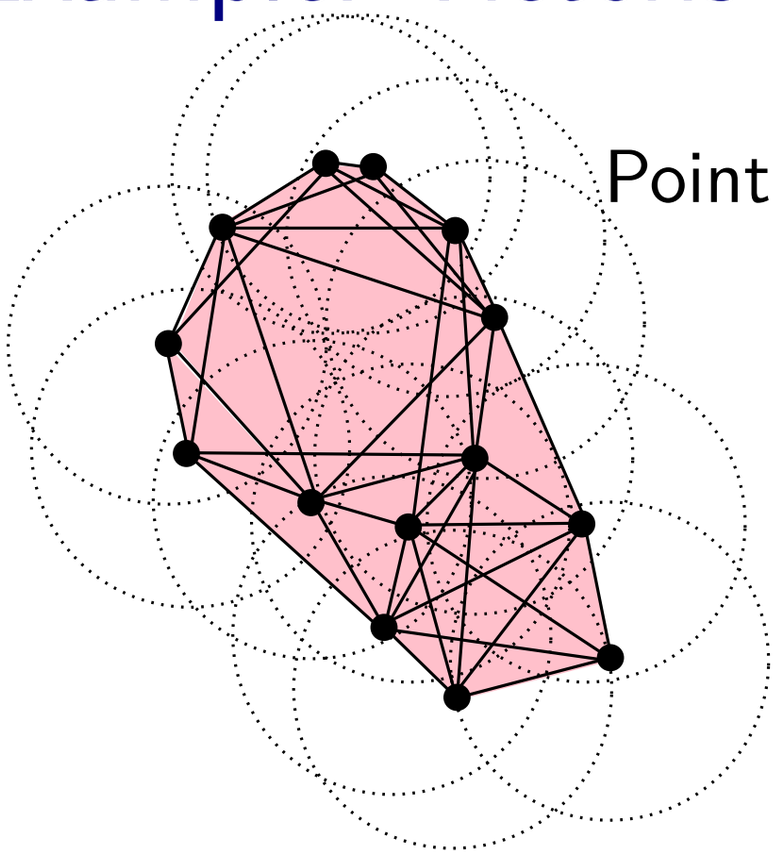
$$\nabla_X p = \left[\frac{\partial}{\partial X} \| v_{i^*} - v_{j^*} \|, \frac{\partial}{\partial X} \| w_{a^*} - w_{b^*} \| \right]$$

$$\frac{\partial}{\partial v_i^{(d)}} \| v_{i^*} - v_{j^*} \| = (-) \frac{1}{\| v_{i^*} - v_{j^*} \|} (v_{i^*}^{(d)} - v_{j^*}^{(d)}) \text{ if } i = i^* \ (j^*) \text{ and } 0 \text{ otherwise}$$

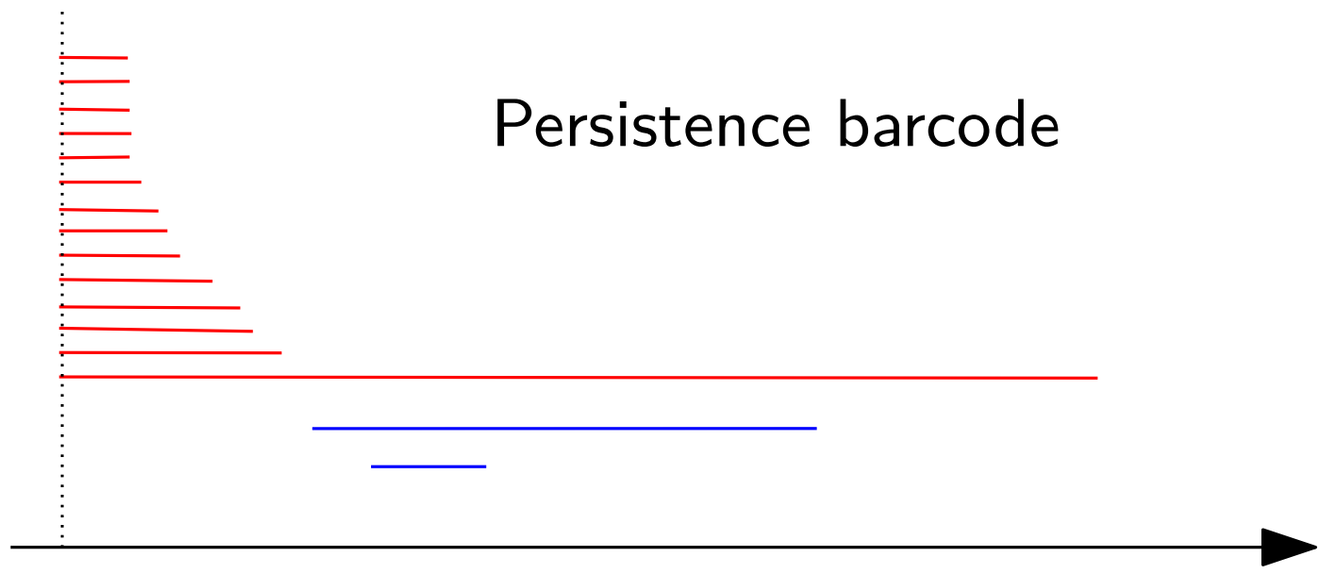
Let $p = (\mathcal{F}(\sigma_+), \mathcal{F}(\sigma_-)) \in D_{\mathrm{Rips}}(X)$

with $\sigma_+ = \{v_0, \ldots, v_k\}$ and $\sigma_- = \{w_0, \ldots, w_{k+1}\}$

Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$

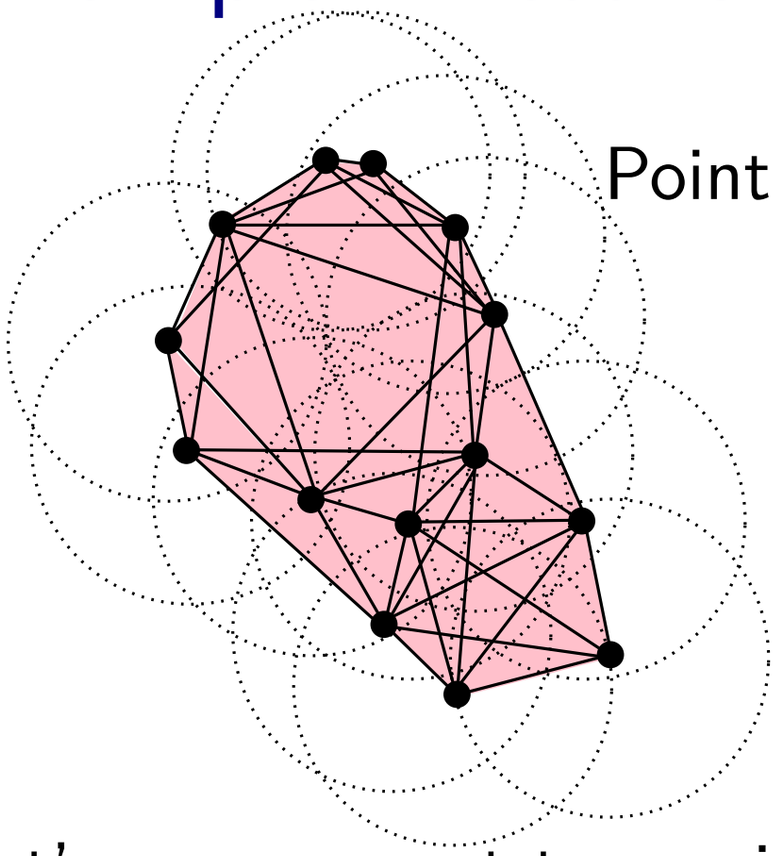


$$\nabla_X p = \left[\frac{\partial}{\partial X} \| v_{i^*} - v_{j^*} \|, \frac{\partial}{\partial X} \| w_{a^*} - w_{b^*} \| \right]$$

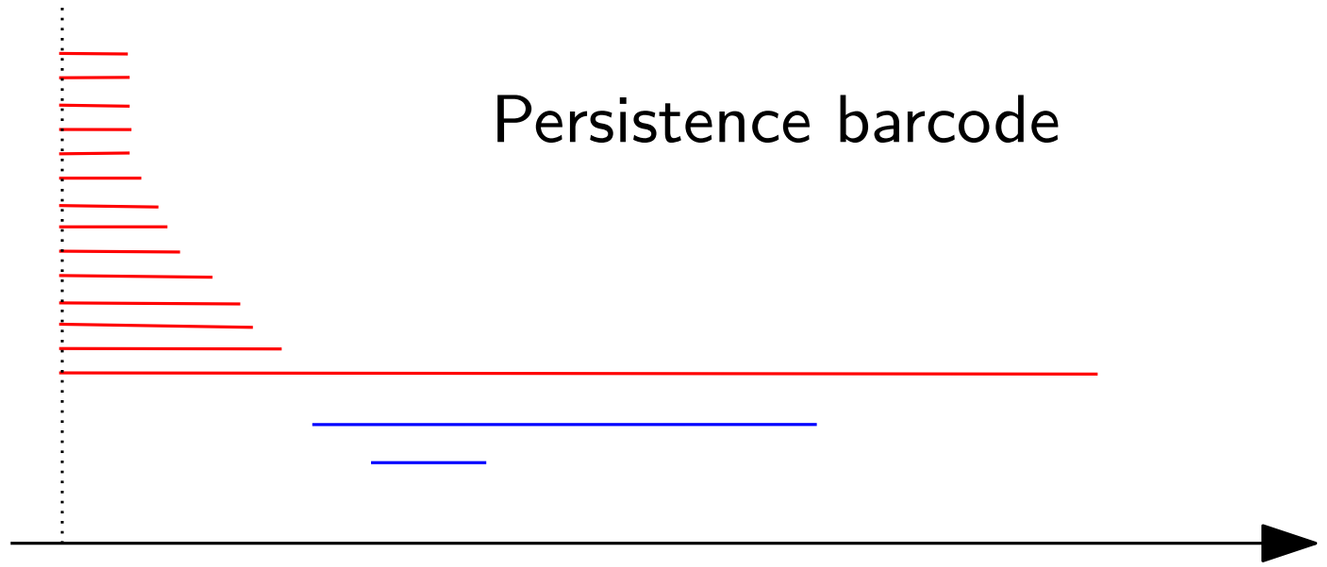
$\frac{\partial}{\partial v_i^{(d)}} \| v_{i^*} - v_{j^*} \| = (-) \frac{1}{\| v_{i^*} - v_{j^*} \|} (v_{i^*}^{(d)} - v_{j^*}^{(d)})$ if $i = i^*$ (j^*) and 0 otherwise

With this gradient rule, one can do gradient descent with any function of persistence!

Example: Vietoris-Rips gradient

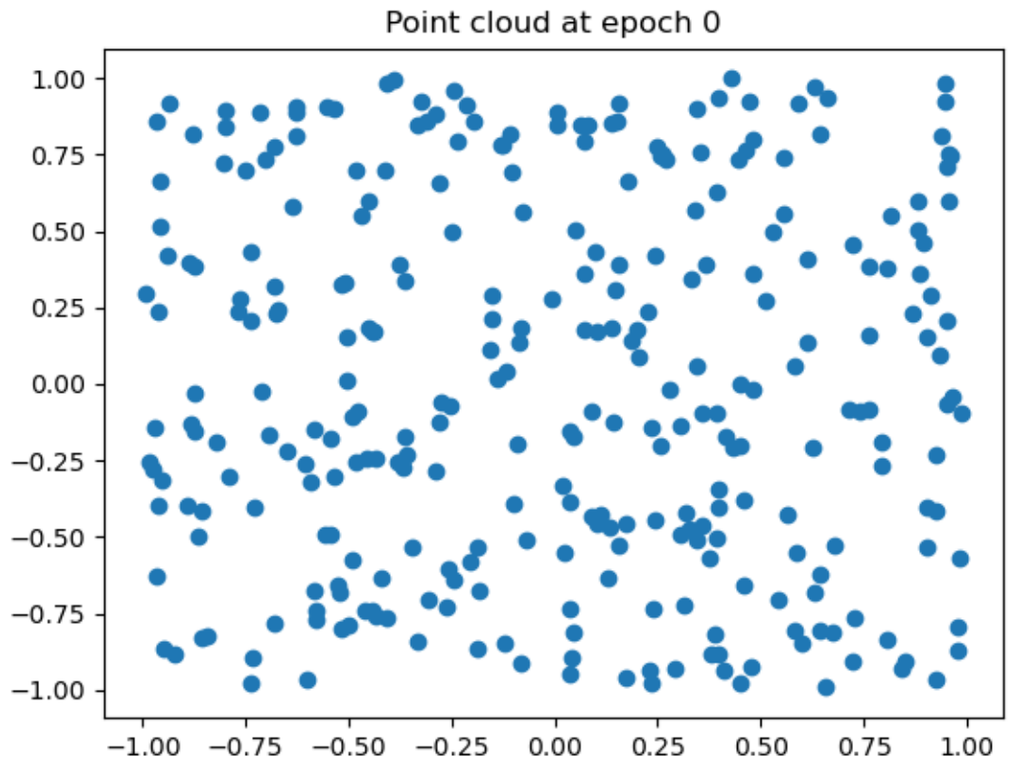


Point cloud $\hat{X}_n$

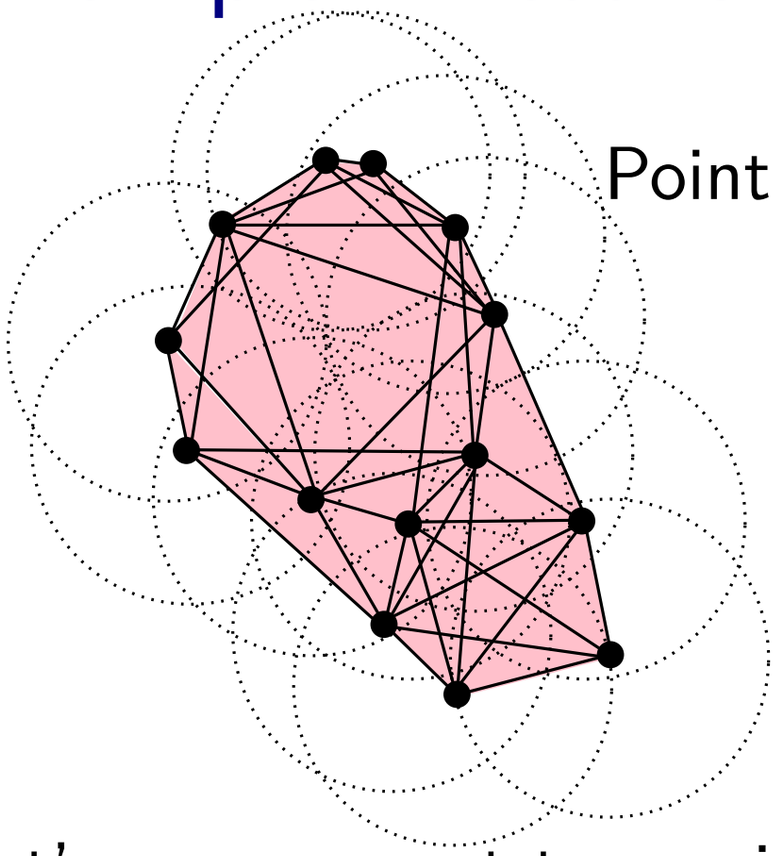


Persistence barcode

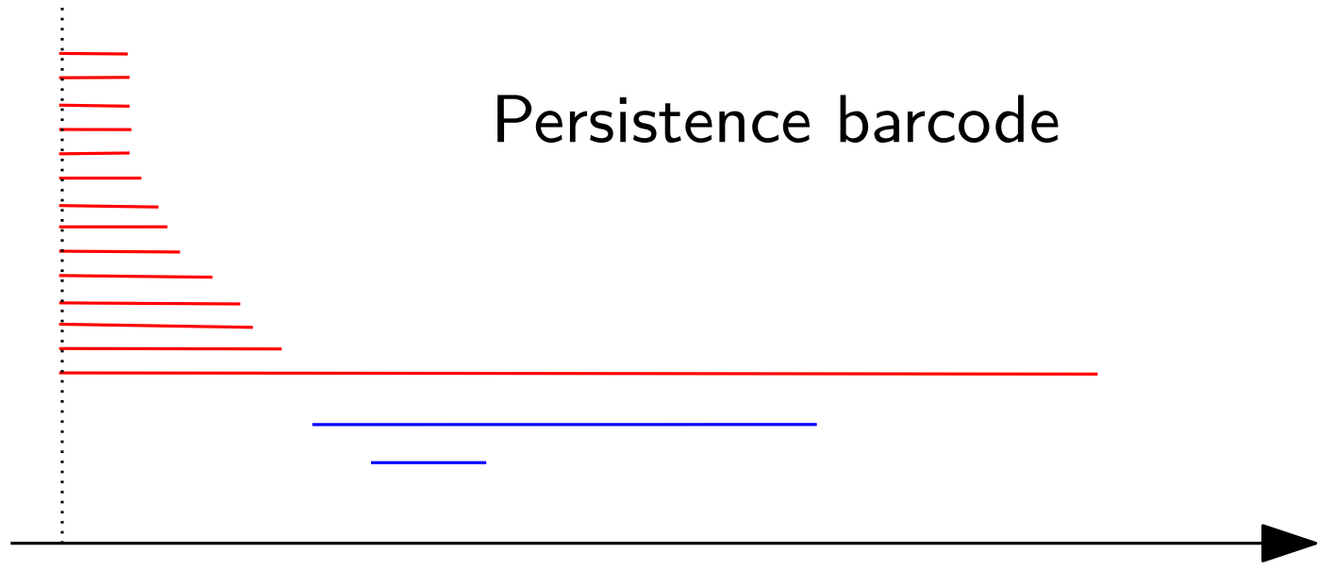
Let's say we want to maximize the number of holes in that point cloud.



Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$



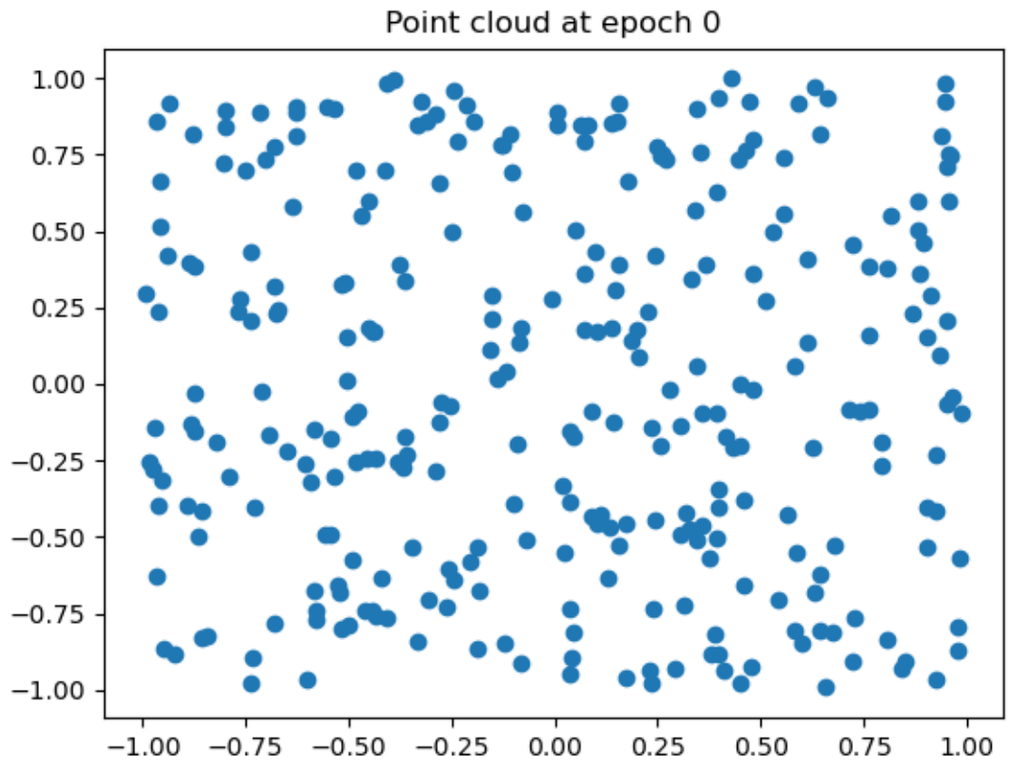
Persistence barcode

Let's say we want to maximize the number of holes in that point cloud.

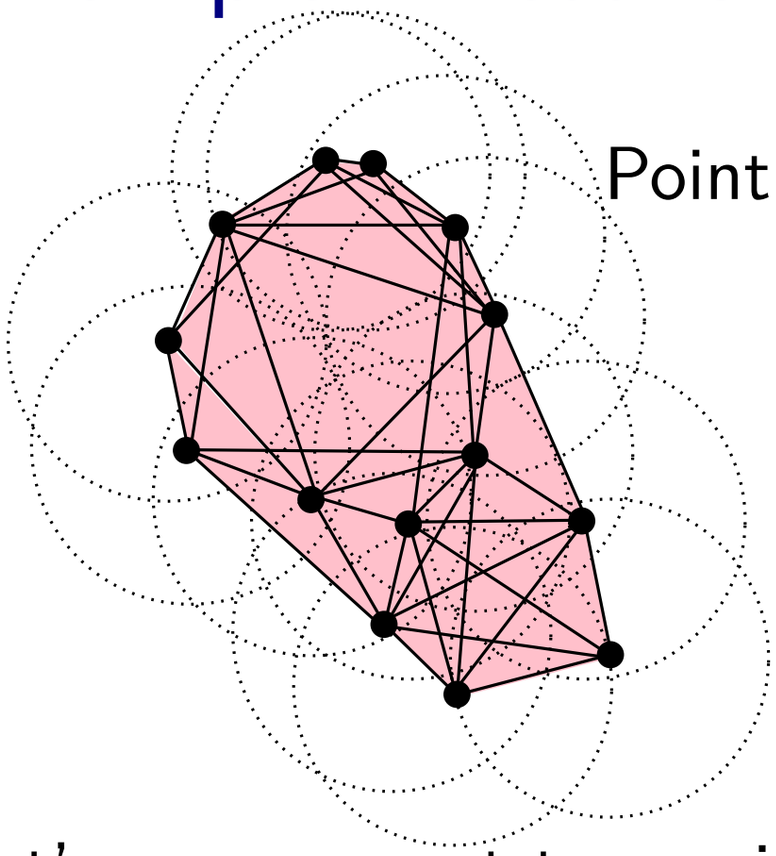
We can use gradient descent to minimize loss

$$\mathcal{L}(X) = -\sum_{p} \|p\|_2^2,$$

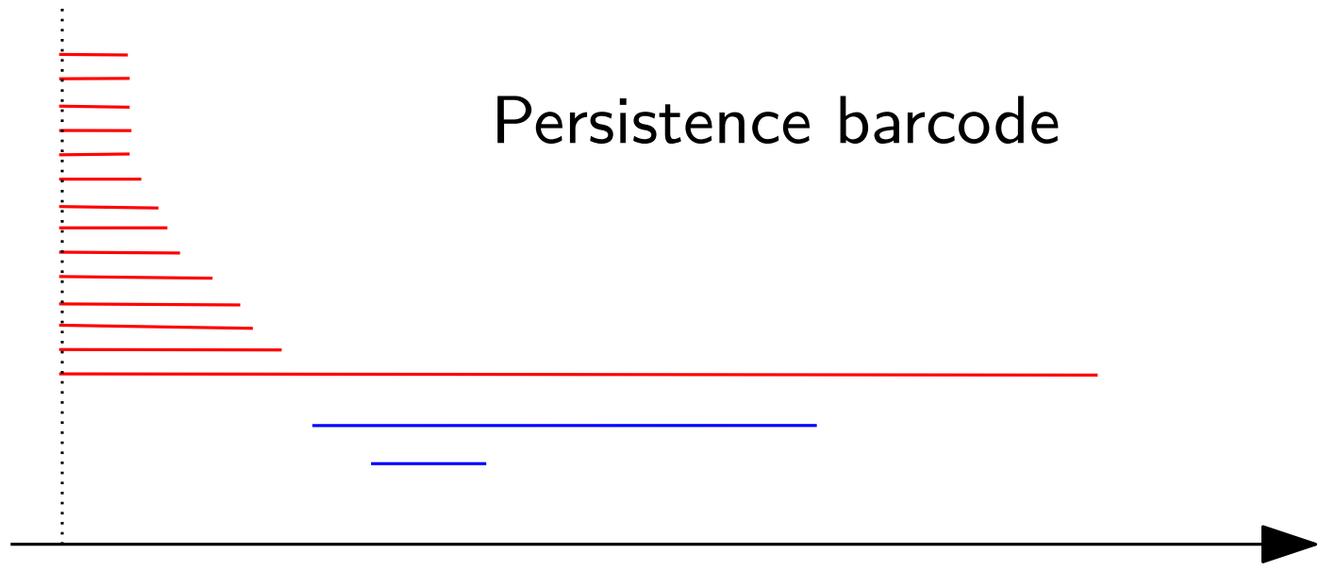
with $p \in D_{\mathrm{Rips}}(X)$ (in hom. 1).



Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$

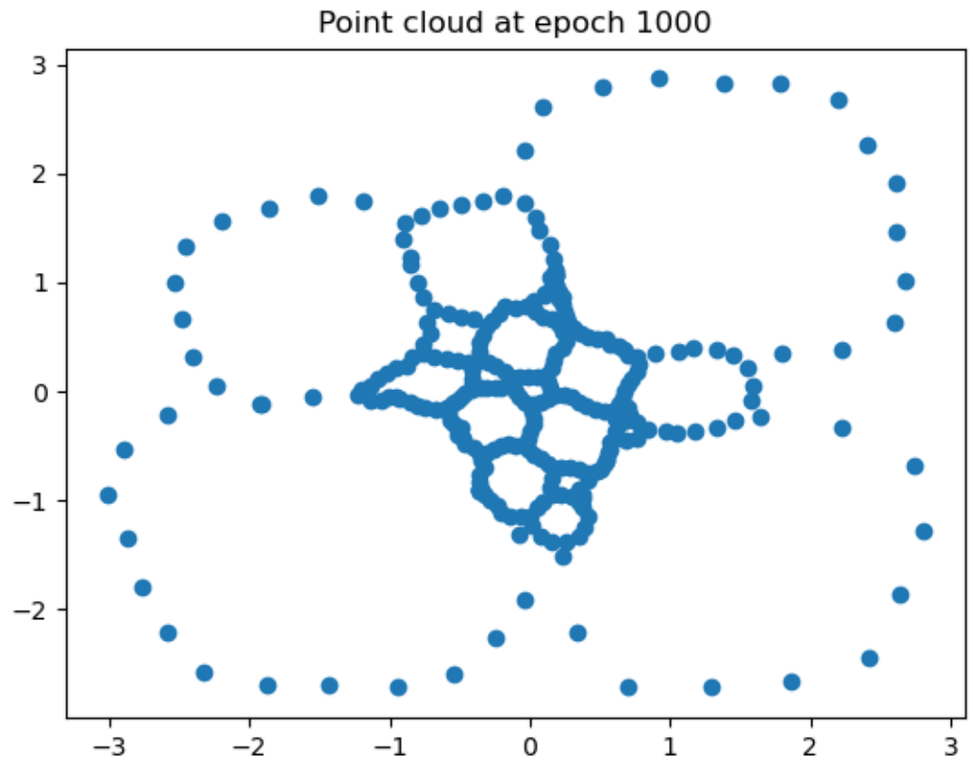


Let's say we want to maximize the number of holes in that point cloud.

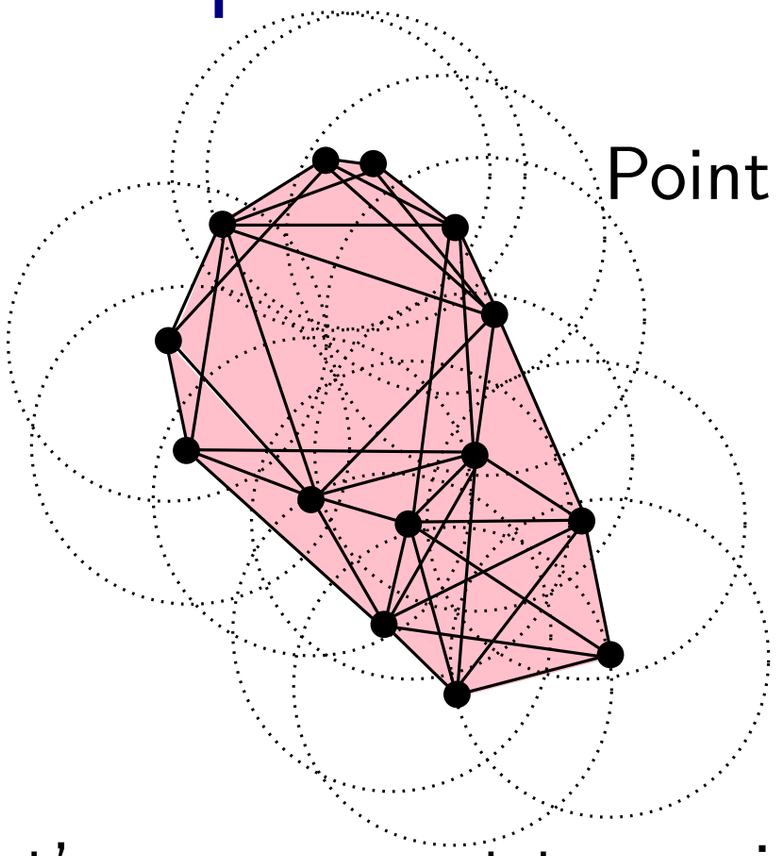
We can use gradient descent to minimize loss

$$\mathcal{L}(X) = -\sum_{p} \|p\|_2^2,$$

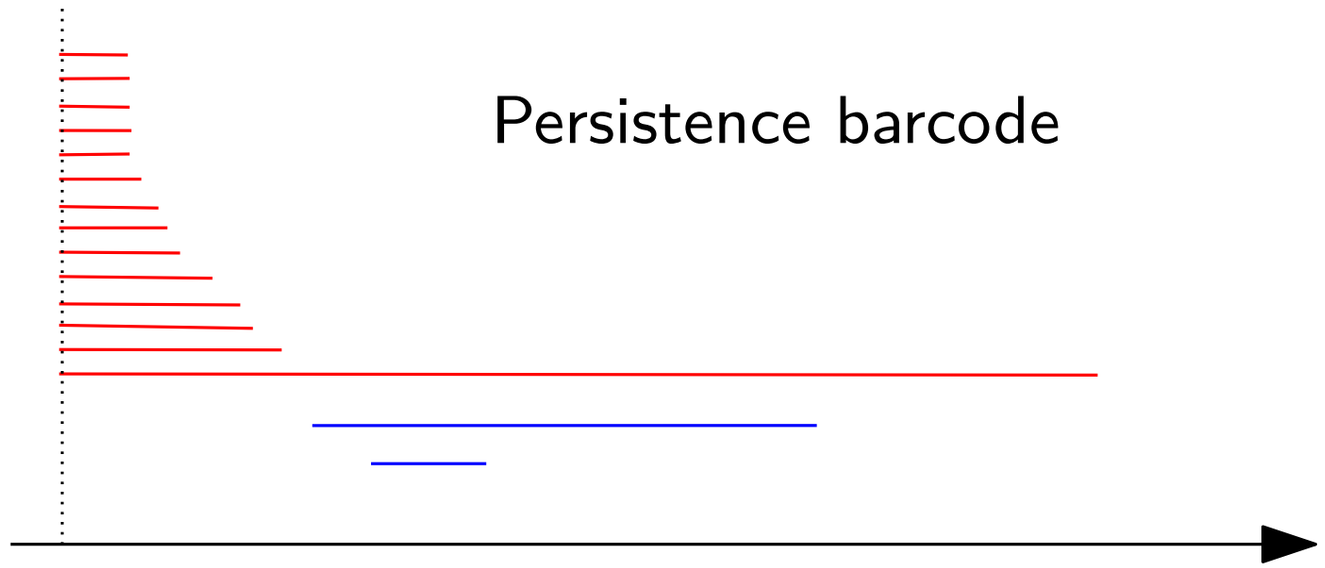
with $p \in D_{\mathrm{Rips}}(X)$ (in hom. 1).



Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$



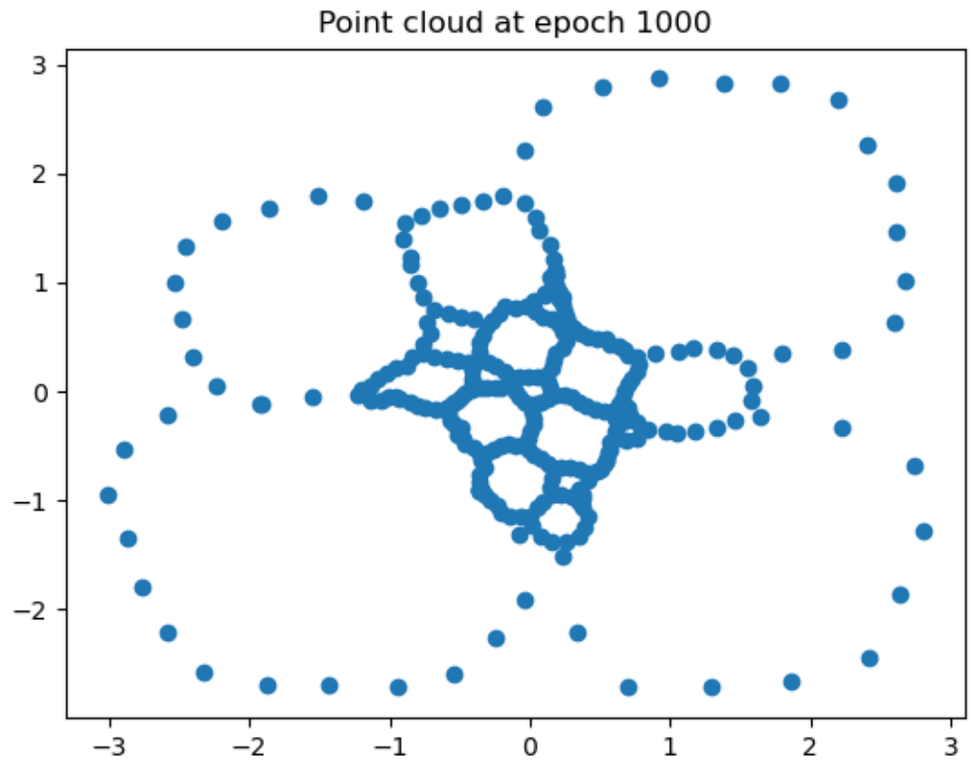
Persistence barcode

Let's say we want to maximize the number of holes in that point cloud.

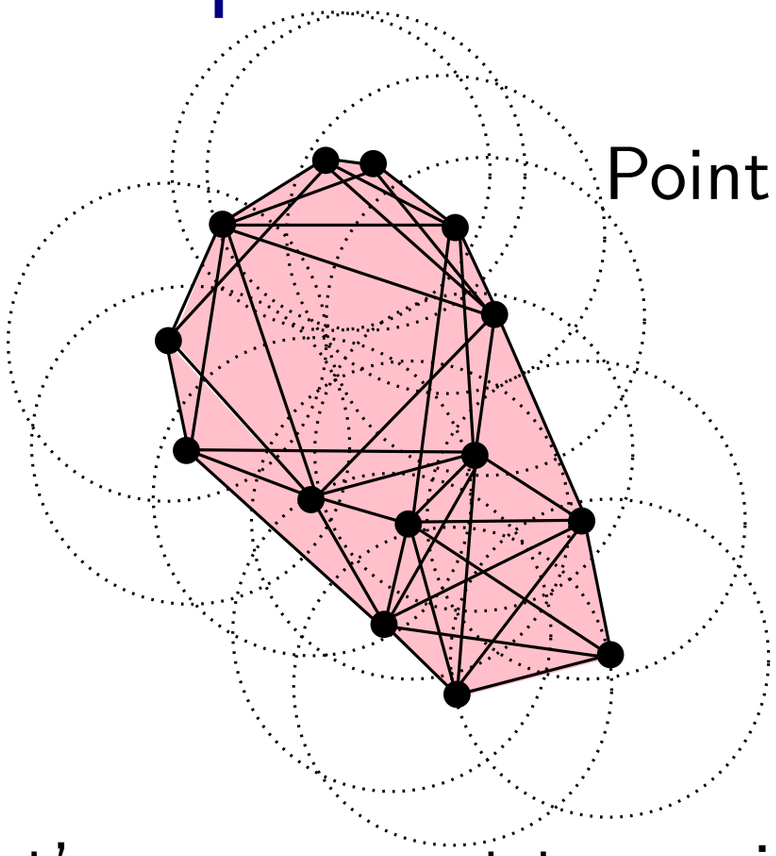
We can use gradient descent to minimize loss

$$\mathcal{L}(X) = -\sum_{p} \|p\|_2^2 + d(X, C),$$

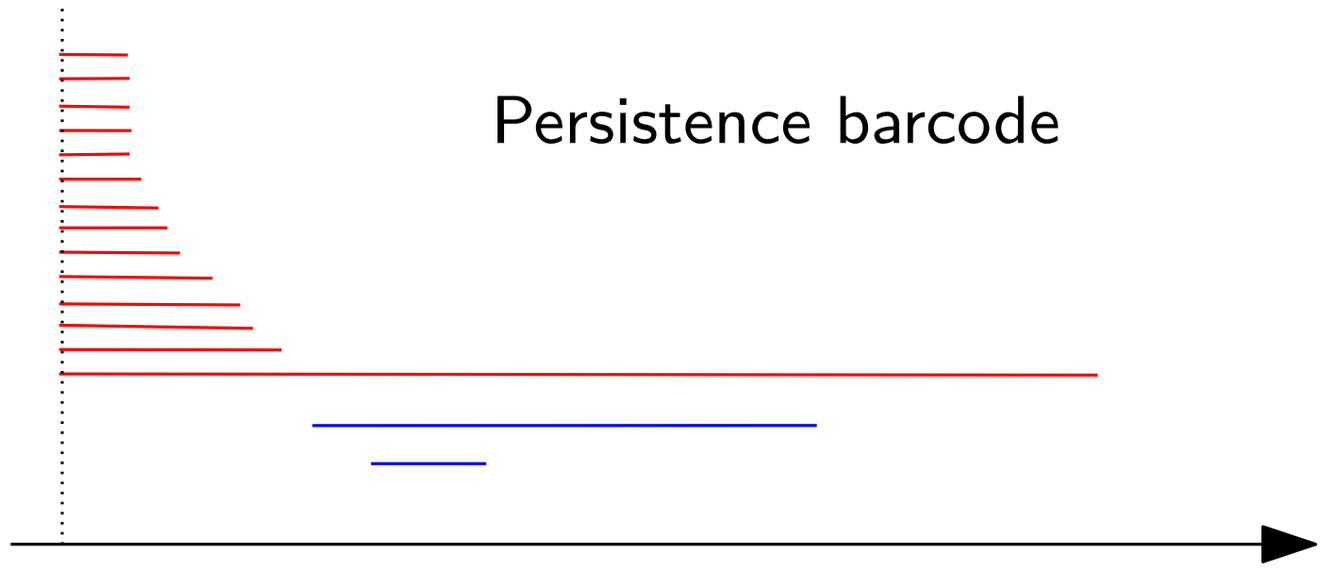
with $p \in D_{\mathrm{Rips}}(X)$ and C unit square.



Example: Vietoris-Rips gradient



Point cloud $\hat{X}_n$



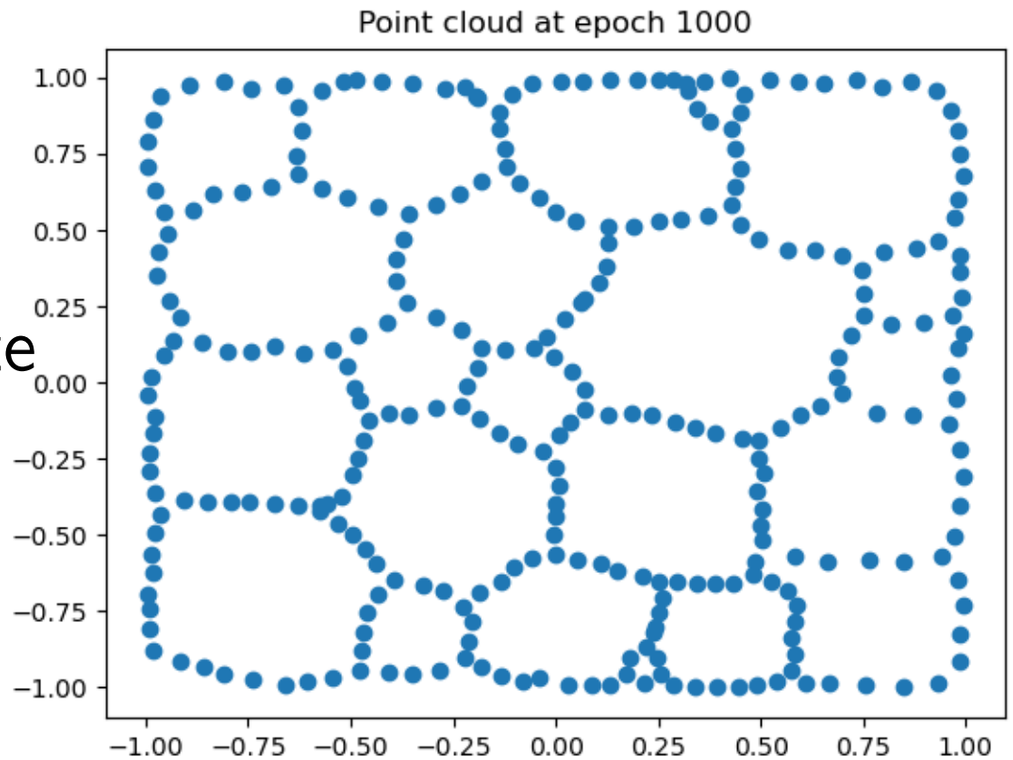
Persistence barcode

Let's say we want to maximize the number of holes in that point cloud.

We can use gradient descent to minimize loss

$$\mathcal{L}(X) = -\sum_{p} \|p\|_2^2 + d(X, C),$$

with $p \in D_{\mathrm{Rips}}(X)$ and C unit square.



Example: Sublevel sets

Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_i \ f_\theta(v_i)$$

Example: Sublevel sets

Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_i \ f_\theta(v_i)$$

$$\nabla_\theta p = \left[\frac{\partial}{\partial\theta} f_\theta(v_{i^*}), \frac{\partial}{\partial\theta} f_\theta(w_{a^*})\right]$$

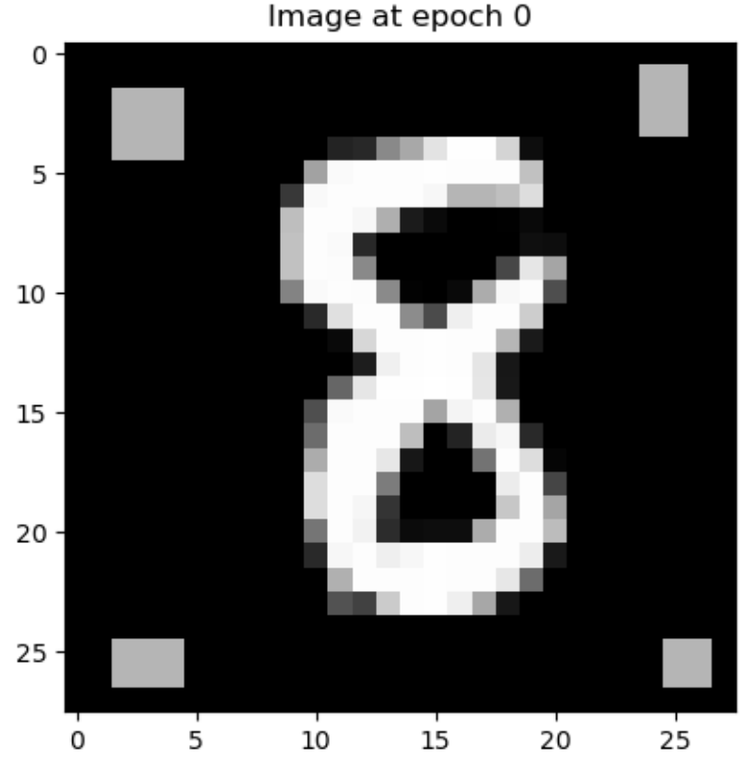
Example: Sublevel sets

Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_i \ f_\theta(v_i)$$

$$\nabla_\theta p = \left[\frac{\partial}{\partial\theta} f_\theta(v_{i^*}), \frac{\partial}{\partial\theta} f_\theta(w_{a^*})\right]$$

Let's say we want to remove the stains in that image.



Example: Sublevel sets

Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_i \ f_\theta(v_i)$$

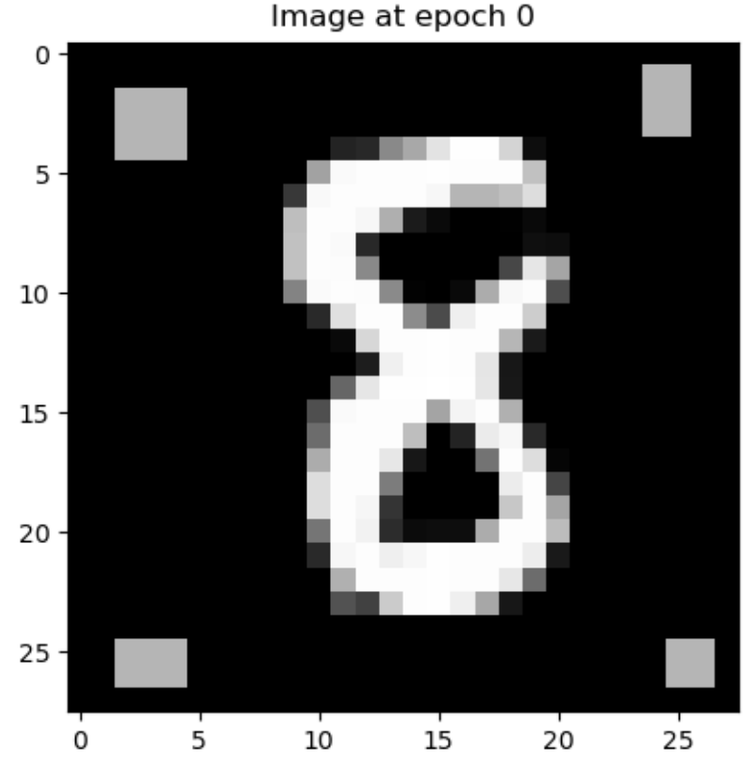
$$\nabla_\theta p = \left[\frac{\partial}{\partial\theta} f_\theta(v_{i^*}), \frac{\partial}{\partial\theta} f_\theta(w_{a^*})\right]$$

Let's say we want to remove the stains in that image.

We can use gradient descent to minimize loss

$$\mathcal{L}(X) = \sum_p \|p\|_2^2,$$

with $p \in D_{\text{Pixel}}(I)$ (in hom. 0).



Example: Sublevel sets

Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_i \ f_\theta(v_i)$$

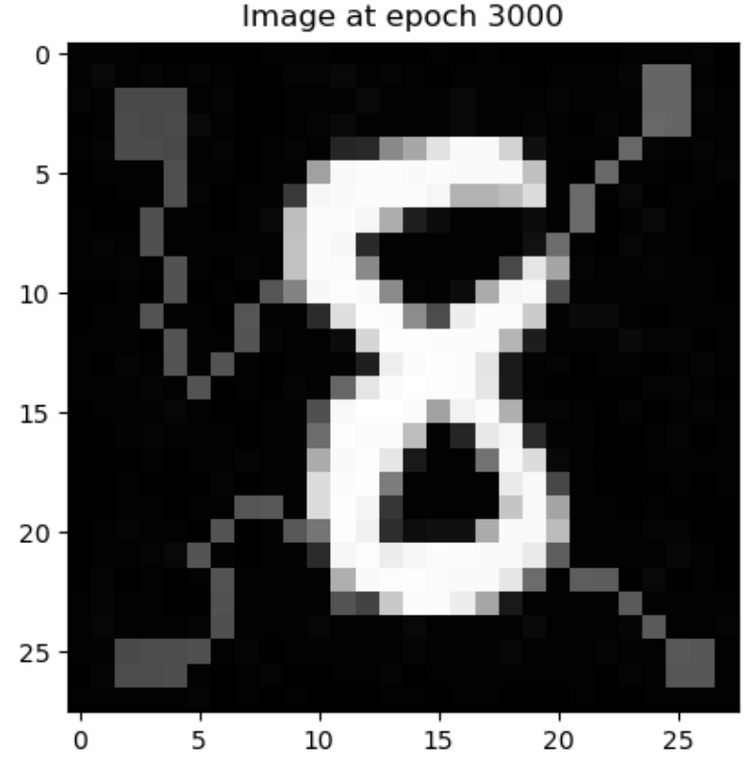
$$\nabla_\theta p = \left[\frac{\partial}{\partial\theta} f_\theta(v_{i^*}), \frac{\partial}{\partial\theta} f_\theta(w_{a^*})\right]$$

Let's say we want to remove the stains in that image.

We can use gradient descent to minimize loss

$$\mathcal{L}(X) = \sum_p \|p\|_2^2,$$

with $p \in D_{\text{Pixel}}(I)$ (in hom. 0).



Example: Sublevel sets

Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_i \ f_\theta(v_i)$$

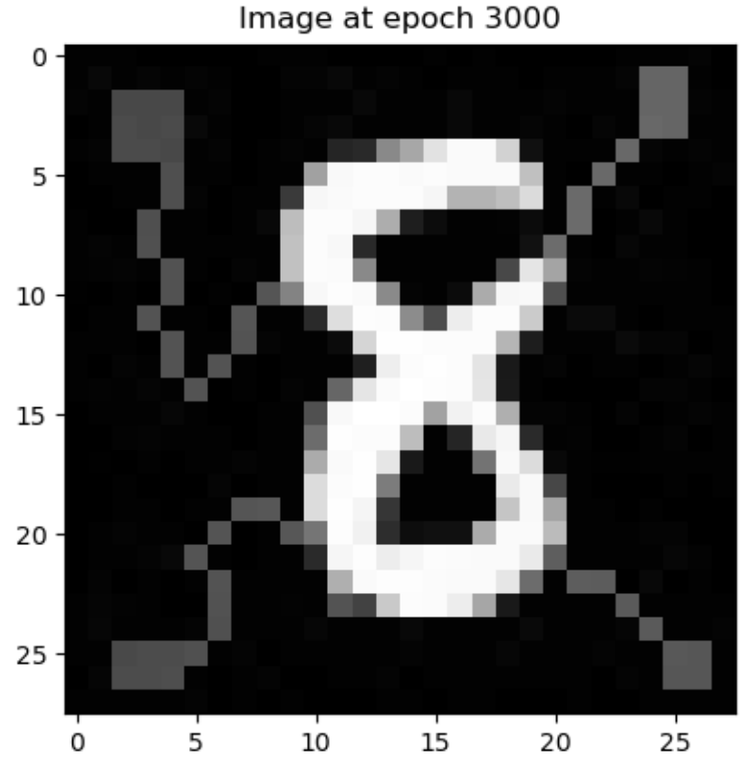
$$\nabla_\theta p = \left[\frac{\partial}{\partial\theta} f_\theta(v_{i^*}), \frac{\partial}{\partial\theta} f_\theta(w_{a^*})\right]$$

Let's say we want to remove the stains in that image.

We can use gradient descent to minimize loss

$$\mathcal{L}(X) = \sum_p \|p\|_2^2 + \sum_{P\in I} \max\{|P|, |1-P|\},$$

with $p \in D_{\text{Pixel}}(I)$.



Example: Sublevel sets

Given k-dim. simplex $\sigma = [v_0, \ldots, v_k]$, one has

$$\mathcal{F}(\sigma) = \max_i \ f_\theta(v_i)$$

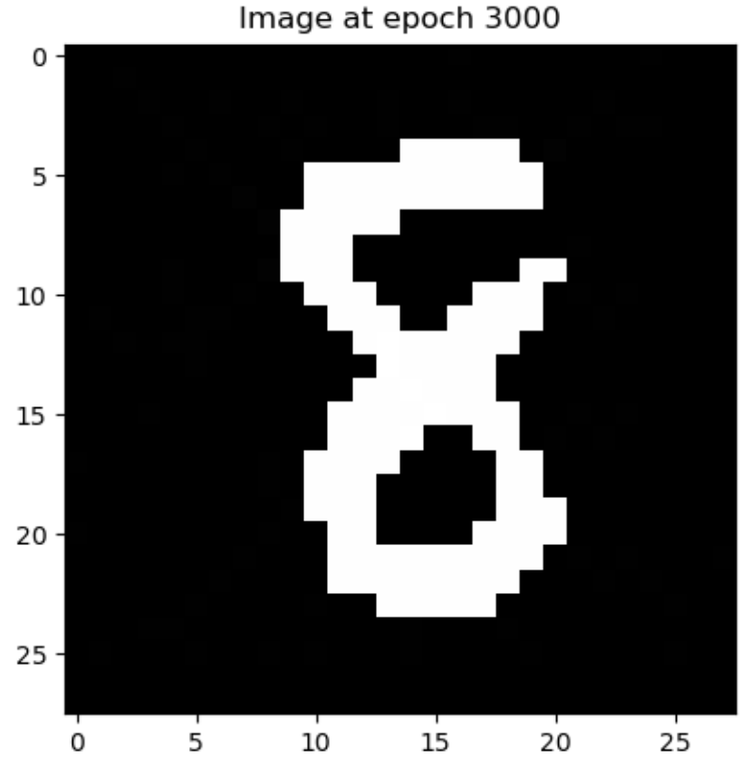
$$\nabla_\theta p = \left[\frac{\partial}{\partial\theta} f_\theta(v_{i^*}), \frac{\partial}{\partial\theta} f_\theta(w_{a^*})\right]$$

Let's say we want to remove the stains in that image.

We can use gradient descent to minimize loss

$$\mathcal{L}(X) = \sum_p \|p\|_2^2 + \sum_{P\in I} \max\{|P|, |1-P|\},$$

with $p \in D_{\text{Pixel}}(I)$.



Topological gradient descent

[*Optimizing persistent homology based functions*, C., Chazal, Glisse, Ike, Kanna, Umeda, ICML, 2021]

Topological gradient descent

[*Optimizing persistent homology based functions*, C., Chazal, Glisse, Ike, Kanna, Umeda, ICML, 2021]

For a fixed ordering of the simplices in a simplicial complex K, the corresponding persistence diagram always has the same number of points: its gradient is well-defined!

Topological gradient descent

[*Optimizing persistent homology based functions*, C., Chazal, Glisse, Ike, Kanna, Umeda, ICML, 2021]

For a fixed ordering of the simplices in a simplicial complex K, the corresponding persistence diagram always has the same number of points: its gradient is well-defined!

If the ordering changes, the boundary matrix can have a new reduced form and the persistence diagram can have a new, different number of points.

Topological gradient descent

[*Optimizing persistent homology based functions*, C., Chazal, Glisse, Ike, Kanna, Umeda, ICML, 2021]

For a fixed ordering of the simplices in a simplicial complex K, the corresponding persistence diagram always has the same number of points: its gradient is well-defined!

If the ordering changes, the boundary matrix can have a new reduced form and the persistence diagram can have a new, different number of points.

Prop: Let K be a simplicial complex and let $\Phi : A \to \mathbb{R}^{|K|}$ a (parameterized) filtration of K. There exists a partition $A = S \sqcup O_1 \sqcup \cdots \sqcup O_k$ s.t. all the restrictions $\Phi : O_i \to \mathbb{R}^{|K|}$ are differentiable.

The O_i's are the parts of A where the ordering of the simplices of K is preserved, and S is the boundaries of all O_i's.

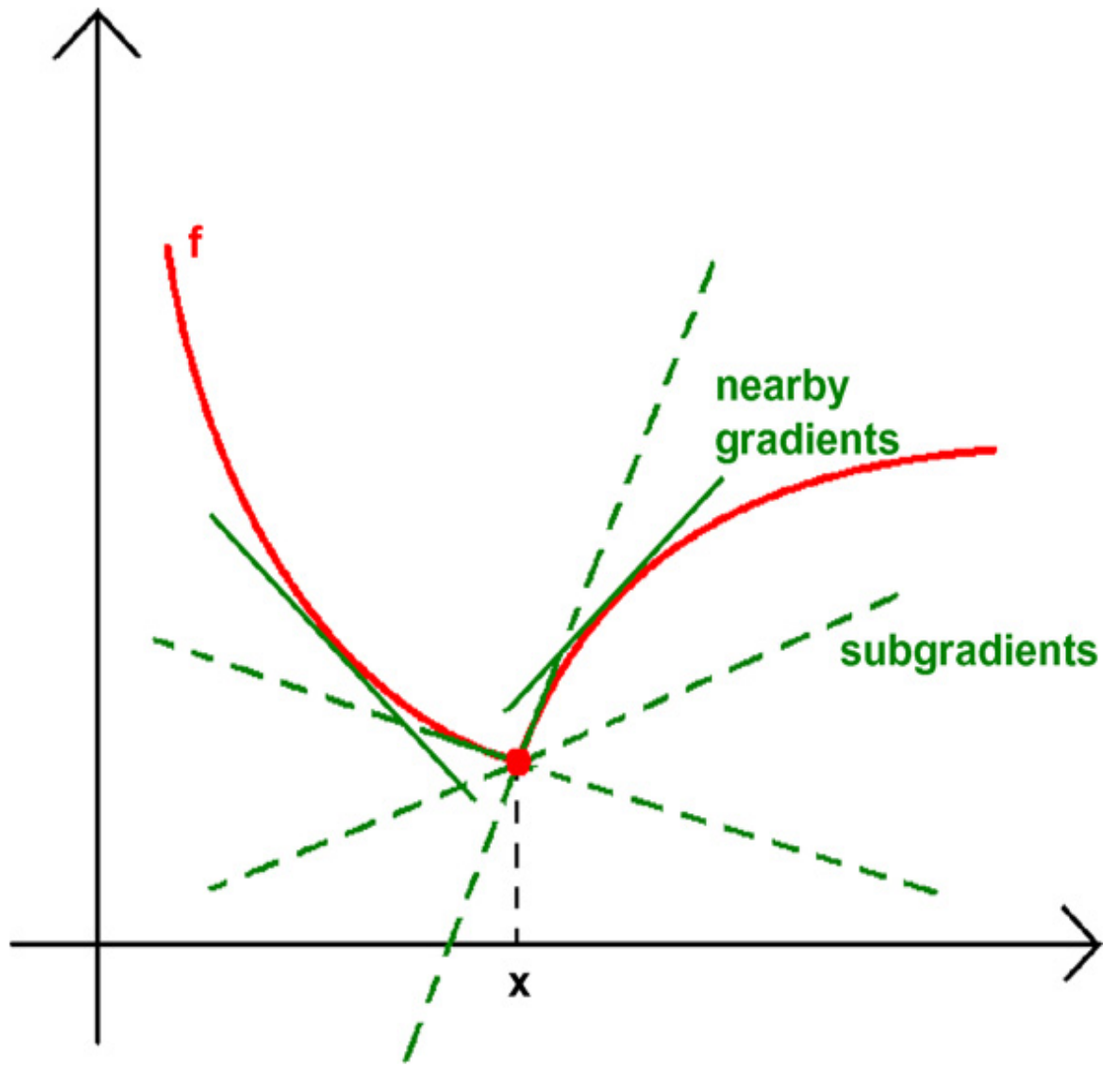
Topological gradient descent

[*Optimizing persistent homology based functions*, C., Chazal, Glisse, Ike, Kanna, Umeda, ICML, 2021]

Def: The *Clarke subdifferential* $\partial\mathcal{L}$ of $\mathcal{L}$ is the set:

$$\partial_x\mathcal{L} = \mathrm{conv}\{\lim_{x_i\to x}\nabla\mathcal{L}(x_i) : \mathcal{L} \text{ is diff. at } x_i\},$$

where conv denotes the convex hull.



Topological gradient descent

[*Optimizing persistent homology based functions*, C., Chazal, Glisse, Ike, Kanna, Umeda, ICML, 2021]

Let $\{\alpha_k\}_k$, $\{\zeta_k\}_k$ s.t.

$\alpha_k \geq 0$, $\sum_k \alpha_k = +\infty$ and $\sum_k \alpha_k^2 < +\infty$

ζ_k random variables s.t. $E[\zeta_k] = 0$ and $E[\|\zeta_k\|^2] < C$ for some $C > 0$

Thm: As long as $\mathcal{L} \circ \mathrm{Pers} \circ \Phi$ is locally Lipschitz, the sequence

$$a_{k+1} = a_k - \alpha_k(g_k + \zeta_k),$$

where $g_k \in \partial_{a_k}(\mathcal{L} \circ \mathrm{Pers} \circ \Phi)$, converges to a critical point of $\mathcal{L} \circ \mathrm{Pers} \circ \Phi$.

Topological stratified gradient descent

[*A gradient sampling algorithm for stratified maps with applications to topological data analysis*, Leygonie, C., Lacombe, Oudot, 2021]

Better guarantees can be obtained by smoothing the gradient definition.

Def: The *smoothed topological gradient* of $\mathrm{Pers} \circ \Phi$ is defined as:

$$\tilde{\nabla}_a = \mathrm{argmin}\{\|g\| \ : \ g \in \mathrm{conv}(S_a)\}$$

where $S_a = \{\nabla_{a'} \ : \ a' \in O_i, O_i \in \mathcal{N}(O_a)\}$, where O_a is the stratum associated to a, and $\mathcal{N}(O_a)$ is the set of strata that are close to O_a.

Intuitively, close strata means that their corresponding orderings are very similar, e.g., they differ by single swaps, or their distance is bounded by $\epsilon > 0$.

Topological stratified gradient descent

[*A gradient sampling algorithm for stratified maps with applications to topological data analysis*, Leygonie, C., Lacombe, Oudot, 2021]

Better guarantees can be obtained by smoothing the gradient definition.

Def: The *smoothed topological gradient* of $\mathrm{Pers} \circ \Phi$ is defined as:

$$\tilde{\nabla}_a = \mathrm{argmin}\{\|g\| \ : \ g \in \mathrm{conv}(S_a)\}$$

where $S_a = \{\nabla_{a'} \ : \ a' \in O_i, O_i \in \mathcal{N}(O_a)\}$, where O_a is the stratum associated to a, and $\mathcal{N}(O_a)$ is the set of strata that are close to O_a.

Intuitively, close strata means that their corresponding orderings are very similar, e.g., they differ by single swaps, or their distance is bounded by $\epsilon > 0$.

Thm: Let $\epsilon > 0$. As long as $\mathcal{L} \circ \mathrm{Pers} \circ \Phi$ is Lipschitz, the sequence

$$a_{k+1} = a_k - \epsilon \cdot \tilde{\nabla}_{a_k} / \|\tilde{\nabla}_{a_k}\|,$$

converges in **finitely many** iterations to $\tilde{a}$ s.t. $\exists \bar{a} \ : \ \tilde{\nabla}_{\bar{a}} = 0$ and $\|\tilde{a} - \bar{a}\| \leq \epsilon$.

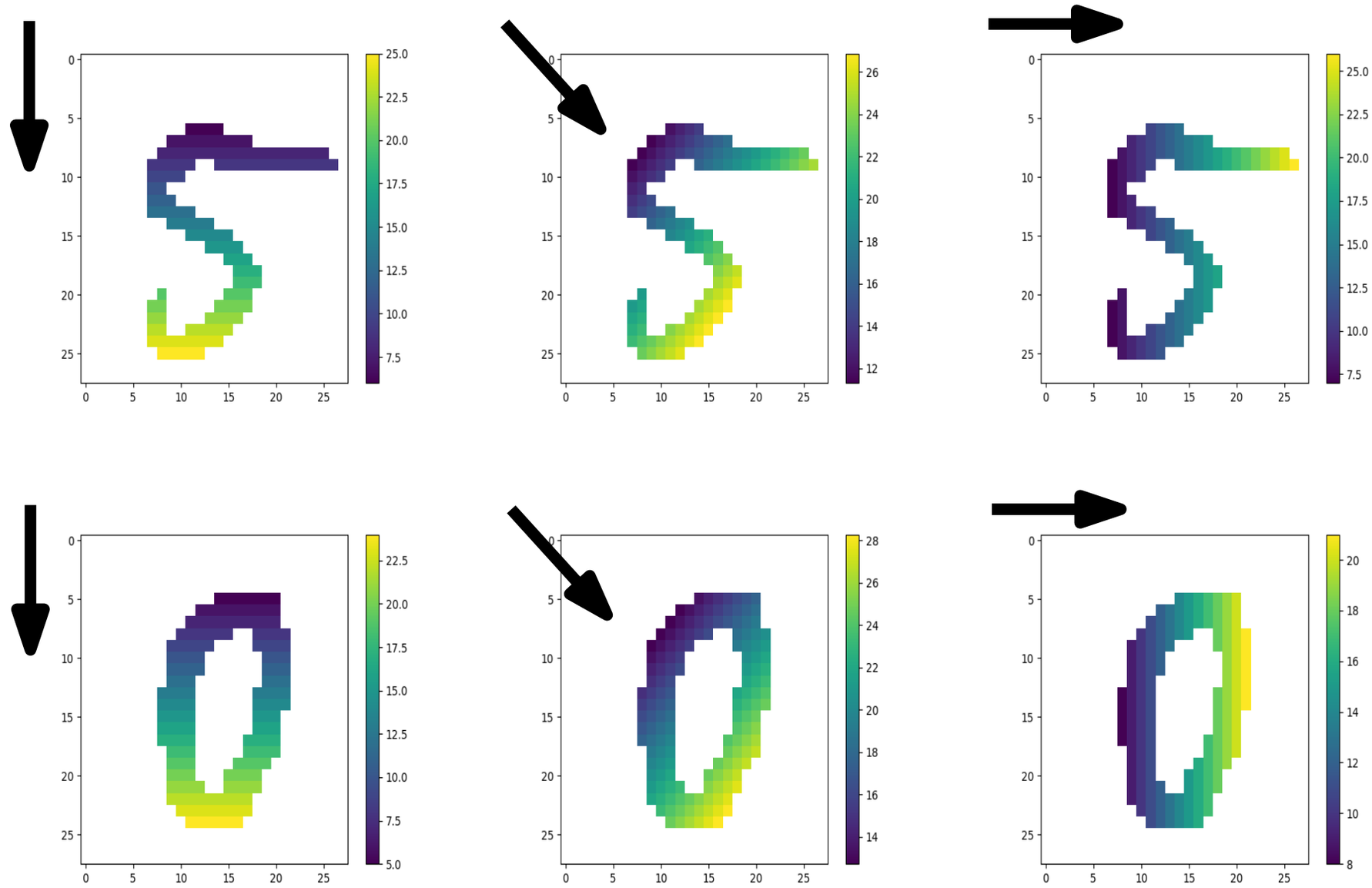
Example: filter selection

Assume we have a supervised classification task. The goal is to find a filtration from a family $\mathcal{F}$ such that the corresponding persistence diagrams give the best classification score.

Example: filter selection

Assume we have a supervised classification task. The goal is to find a filtration from a family $\mathcal{F}$ such that the corresponding persistence diagrams give the best classification score.

Ex: images filtered by a direction parameterized by angle.



Example: filter selection

Assume we have a supervised classification task. The goal is to find a filtration from a family $\mathcal{F}$ such that the corresponding persistence diagrams give the best classification score.

Idea: minimize:

$$\mathcal{L}(f) = \sum_l \frac{\sum_{y_i=y_j=l} d_p(D_f(x_i), D_f(x_j))}{\sum_{y_i=l} d_p(D_f(x_i), D_f(x_j))},$$

one can also use Sliced Wasserstein for speedup.

Example: filter selection

Assume we have a supervised classification task. The goal is to find a filtration from a family $\mathcal{F}$ such that the corresponding persistence diagrams give the best classification score.

Idea: minimize:

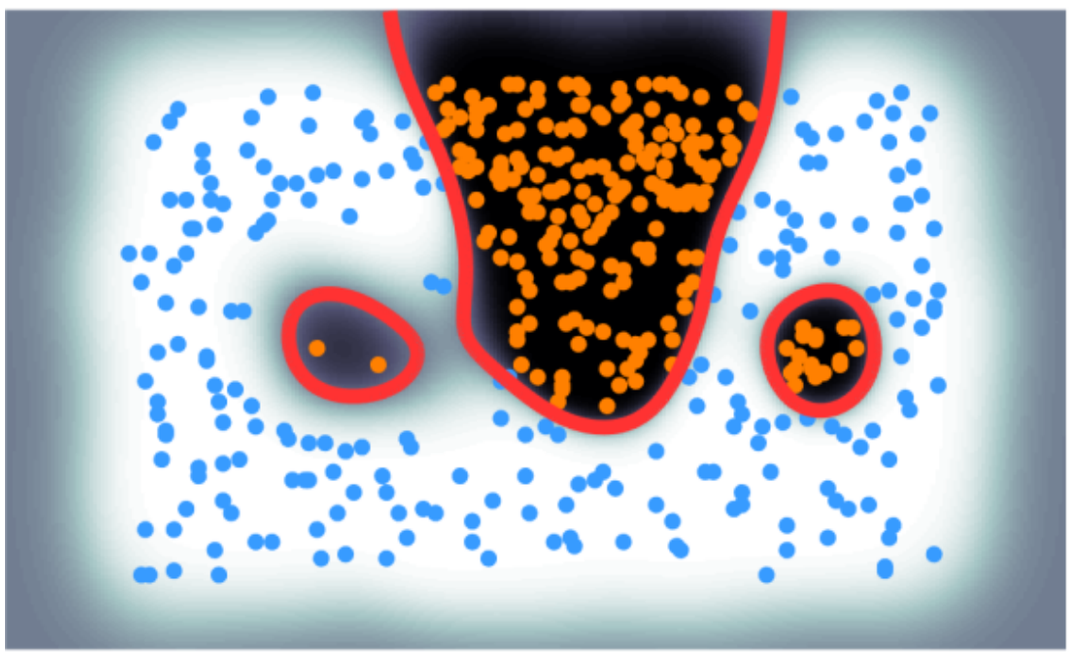
$$\mathcal{L}(f) = \sum_l \frac{\sum_{y_i=y_j=l} d_p(D_f(x_i), D_f(x_j))}{\sum_{y_i=l} d_p(D_f(x_i), D_f(x_j))},$$

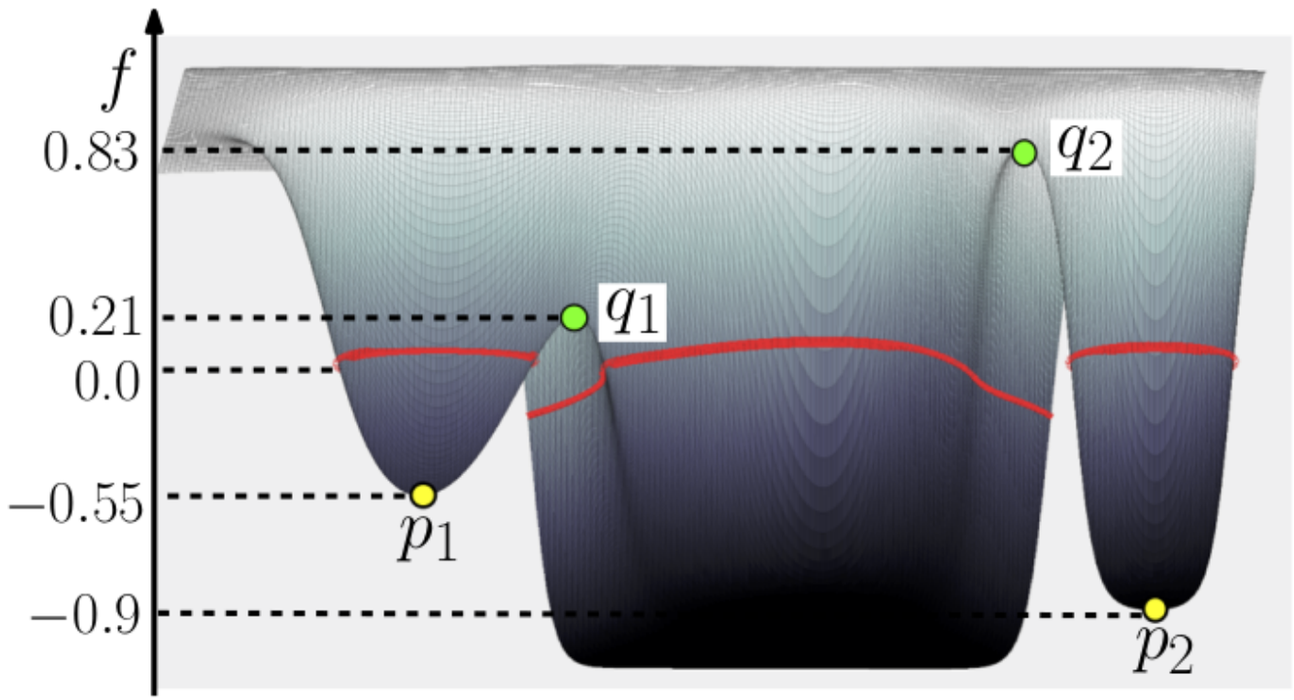
one can also use Sliced Wasserstein for speedup.

Dataset	Baseline	Before	After	Difference	Dataset	Baseline	Before	After	Difference
vs01	100.0	61.3	99.0	**+37.6**	vs26	99.7	98.8	98.2	-0.6
vs02	99.4	98.8	97.2	-1.6	vs28	99.1	96.8	96.8	0.0
vs06	99.4	87.3	98.2	**+10.9**	vs29	99.1	91.6	98.6	**+7.0**
vs09	99.4	86.8	98.3	**+11.5**	vs34	99.8	99.4	99.1	-0.3
vs16	99.7	89.0	97.3	**+8.3**	vs36	99.7	99.3	99.3	-0.1
vs19	99.6	84.8	98.0	**+13.2**	vs37	98.9	94.9	97.5	**+2.6**
vs24	99.4	98.7	98.7	0.0	vs57	99.7	90.5	97.2	**+6.7**
vs25	99.4	80.6	97.2	**+16.6**	vs79	99.1	85.3	96.9	**+11.5**

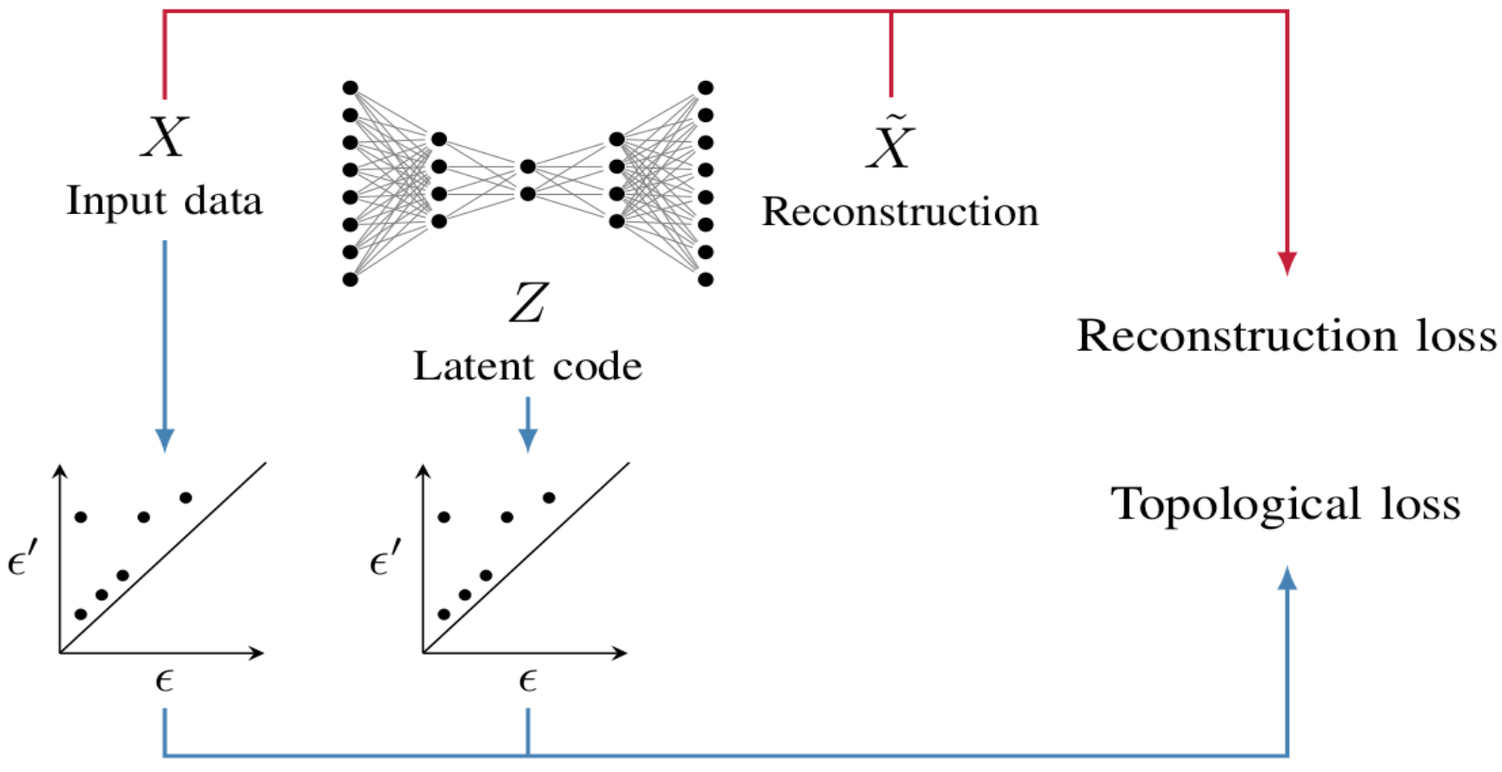
More examples

[*A Topological Regularizer for Classifiers via Persistent Homology*, Chen, Ni, Bai, Wang, AISTATS, 2019]





[*Topological autoencoders*, Moor, Horn, Rieck, Borgwardt, ICML, 2020]



I. Topological Inference on Persistence Diagrams for Point Clouds

II. Topological Vectorizations for Inference and Machine Learning

(III. A Gaussian Kernel for Persistence Diagrams)

IV. Deep Learning for Persistence Diagrams

Thanks! Questions?