Contributions to geometric inference for manifolds and to the statistical study of persistence diagrams

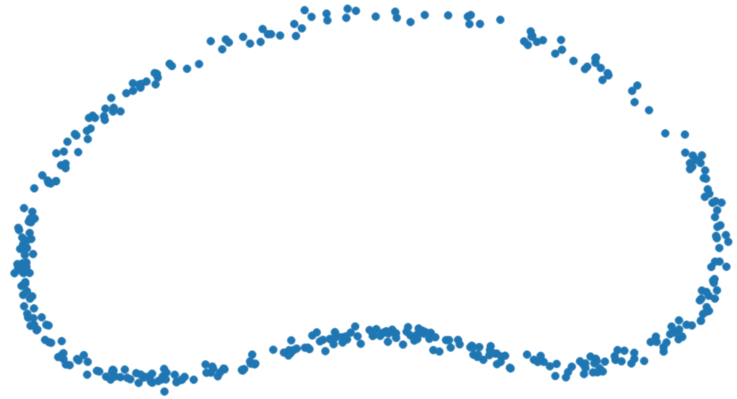
PhD Defense - August 30 2021

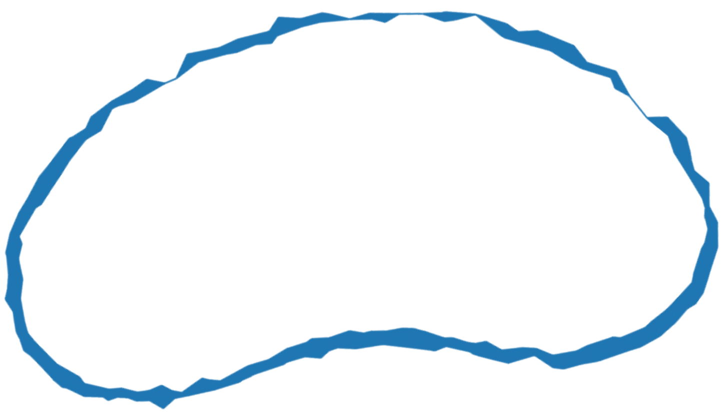
Vincent Divol

vincent.divol@nyu.edu

vincentdivol.github.io

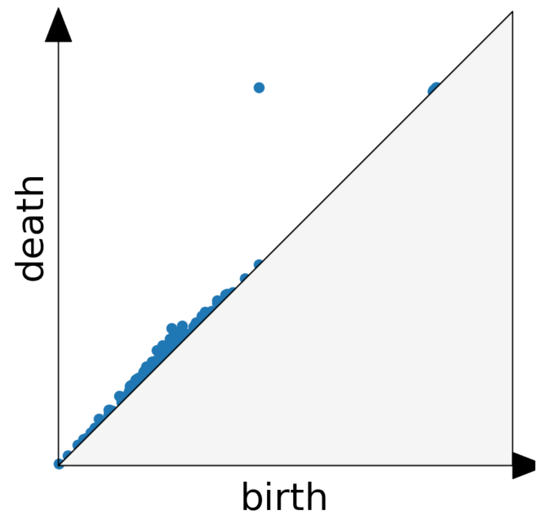
DataShape
Inria Saclay / Laboratoire Mathématique d'Orsay





Part I

Manifold inference

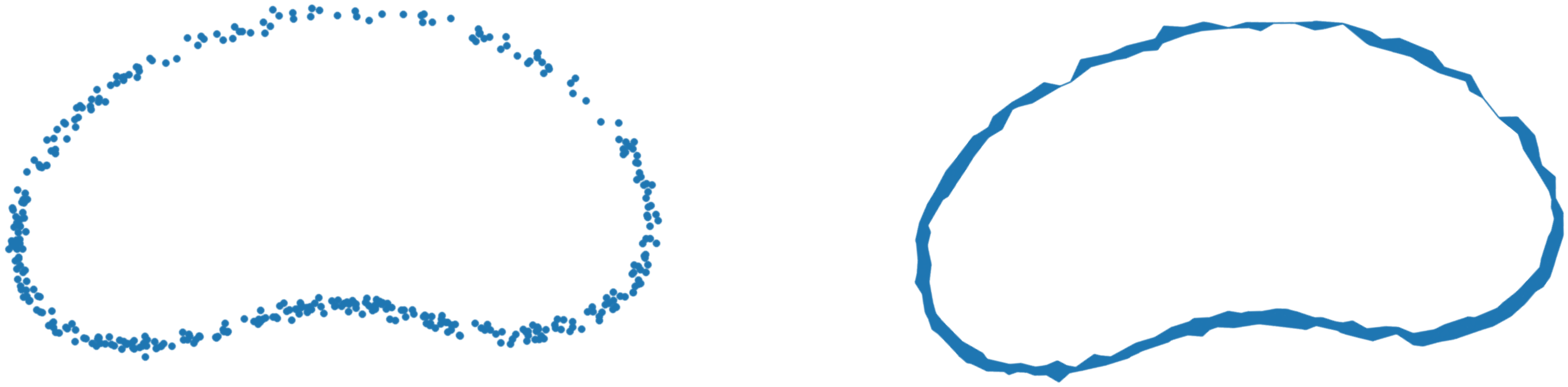


Part II

Persistence diagrams

Part I

Manifold inference



The manifold assumption

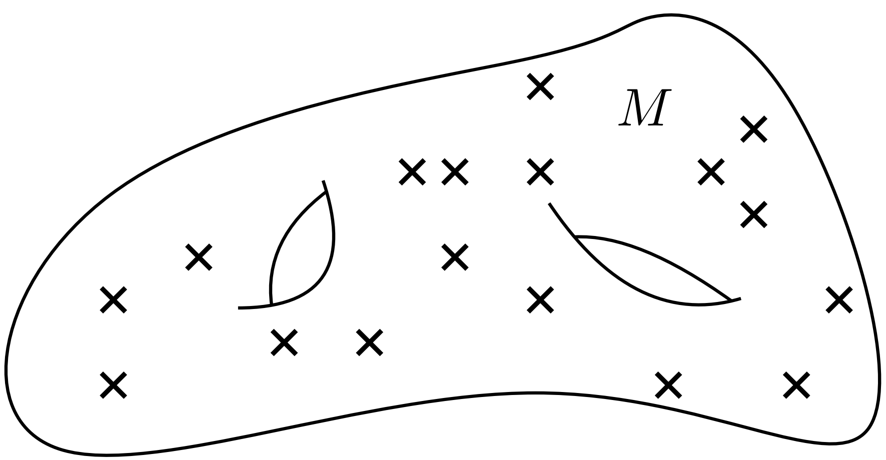
$\mathcal{X}_n = \{X_1, \dots, X_n\}$ = set of n random observations in $\mathbb{R}^D$

$n \ll D$

Key assumption:

There is a low dimensional structure underlying the observations $\mathcal{X}_n$.

$\mathcal{X}_n$ lies close to a **manifold** M.

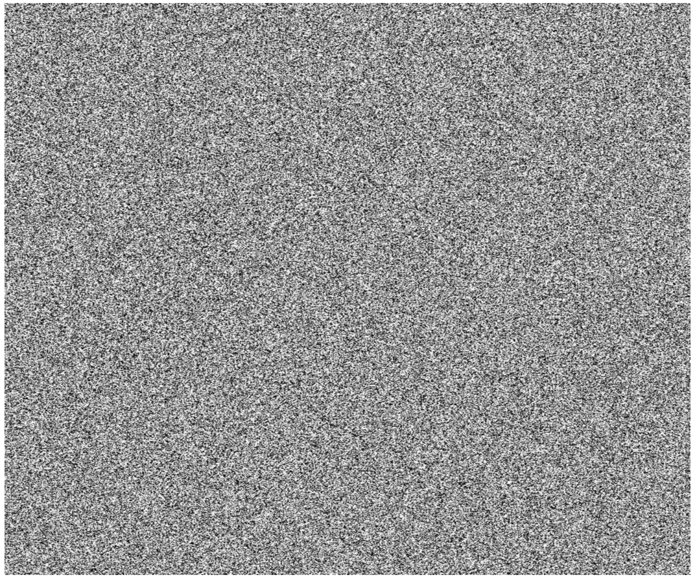


Goal: propose a reconstruction $\hat{M}$ that is close to M for the Hausdorff distance d_H

The manifold assumption

$\mathcal{X}_n = \{X_1, \ldots, X_n\}$ = set of n random observations in $\mathbb{R}^D$

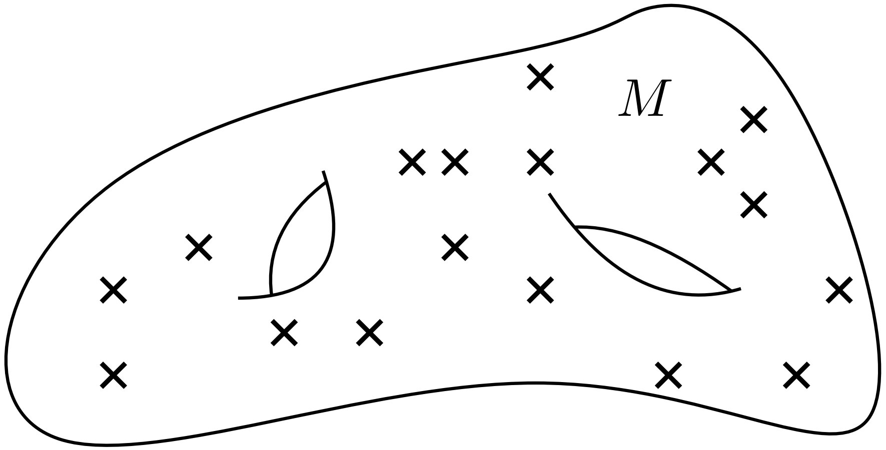
$n \ll D$



Key assumption:

There is a low dimensional structure underlying the observations $\mathcal{X}_n$.

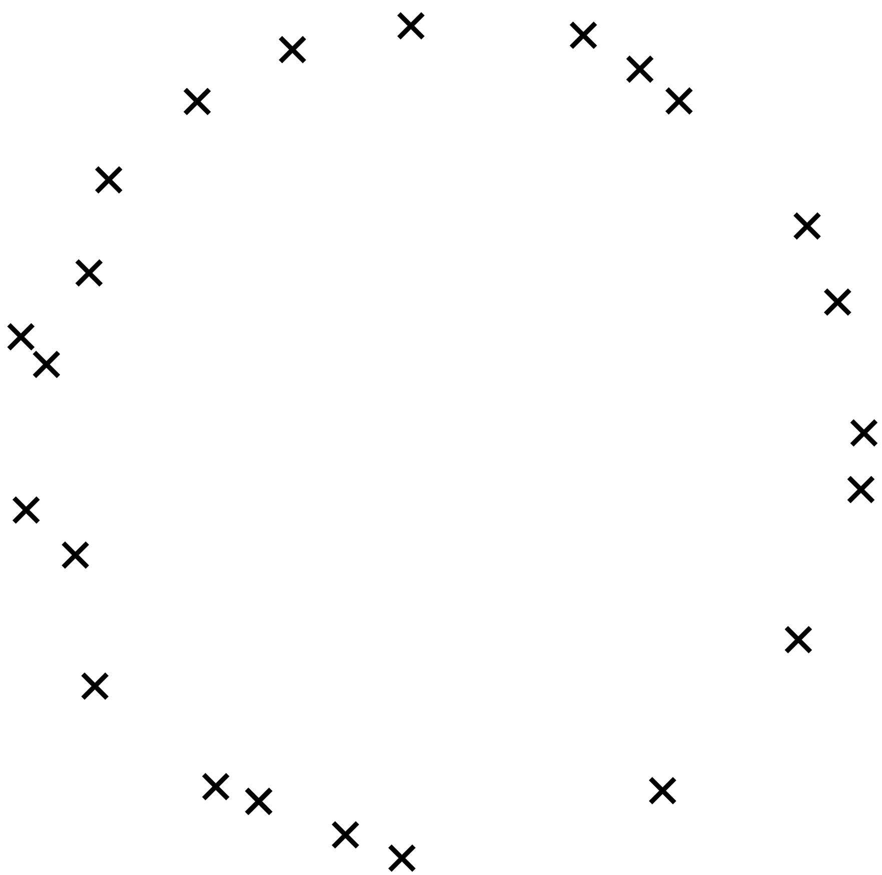
$\mathcal{X}_n$ lies close to a **manifold** M.



Goal: propose a reconstruction $\hat{M}$ that is close to M for the Hausdorff distance d_H

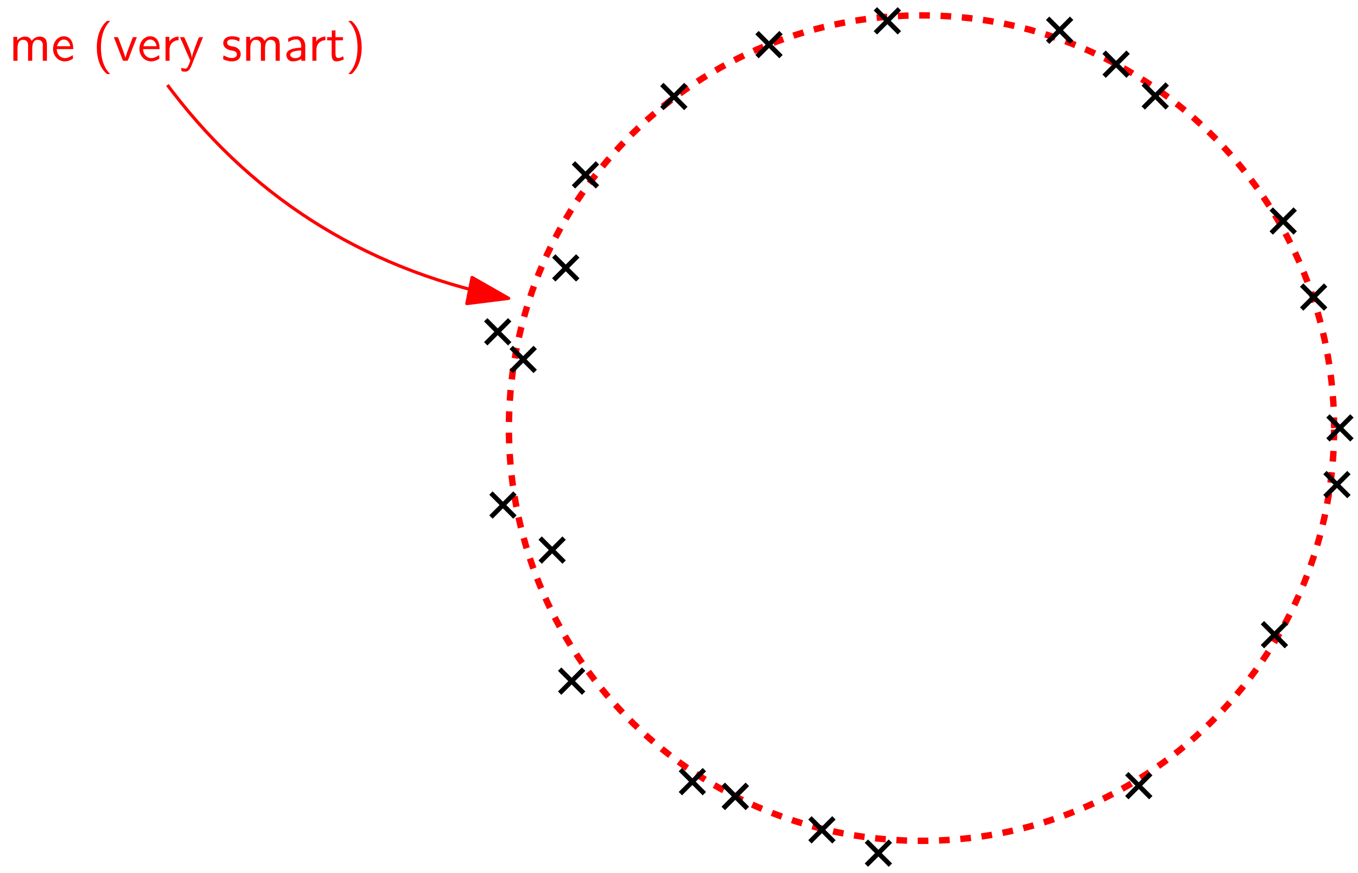
Sampling hypotheses

Find the curve fitting the points!



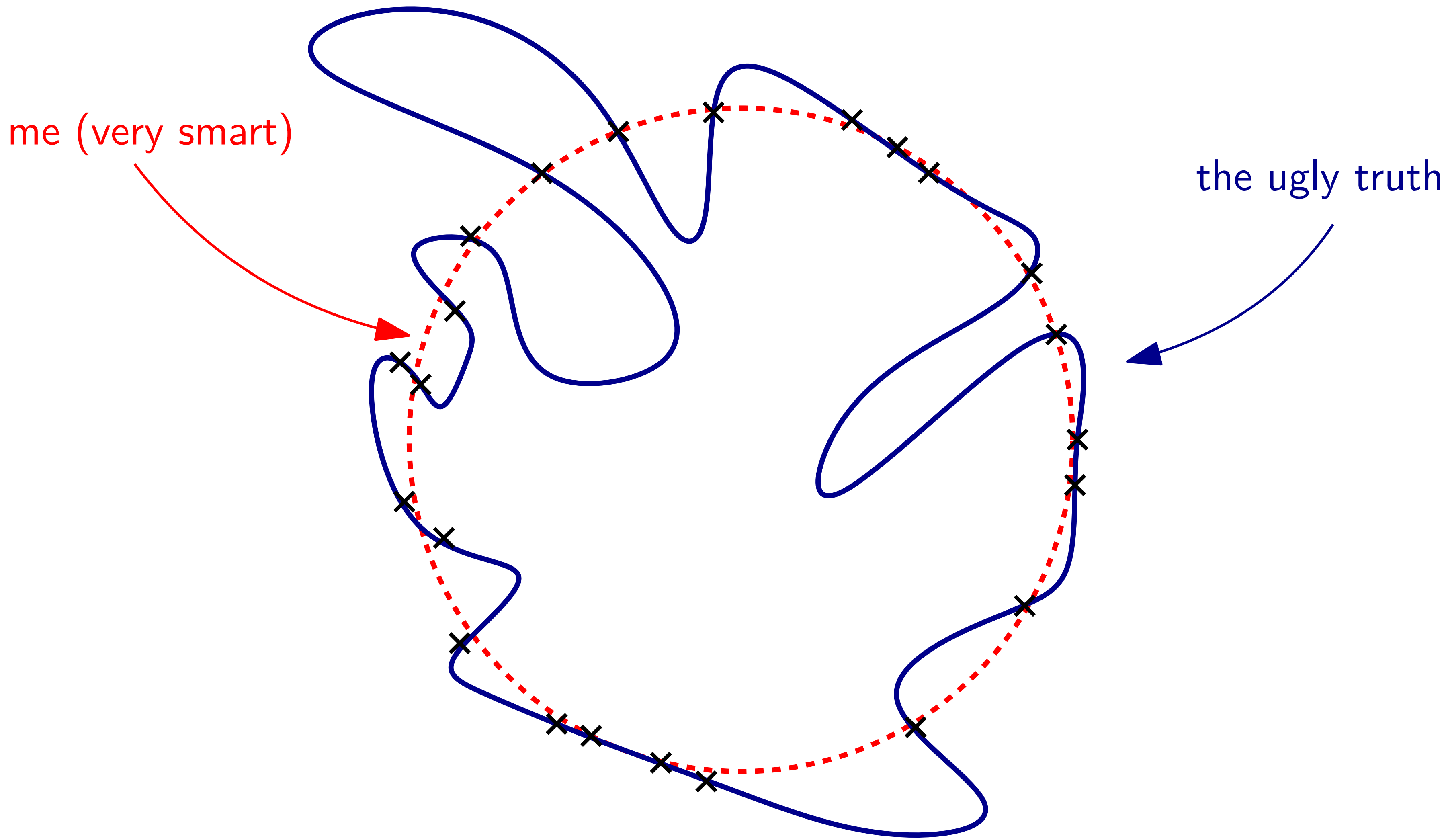
Sampling hypotheses

Find the curve fitting the points!



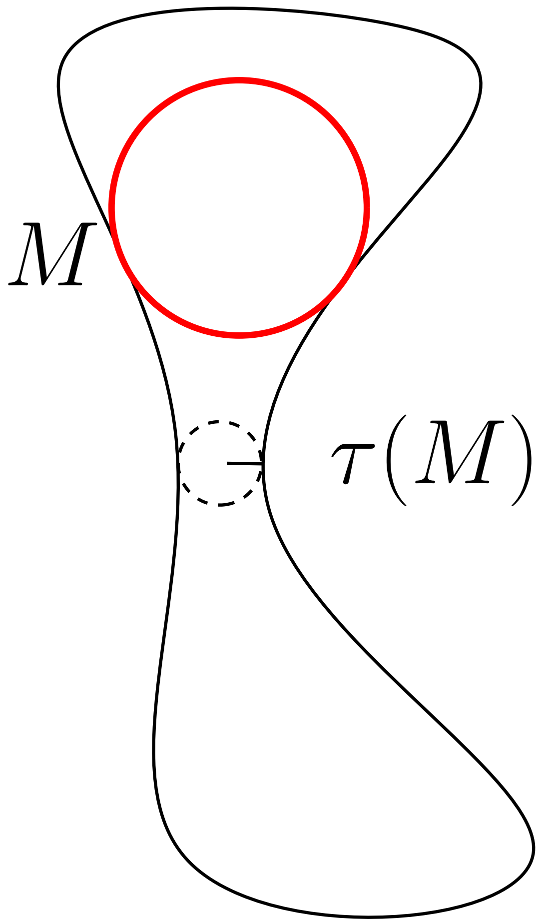
Sampling hypotheses

Find the curve fitting the points!

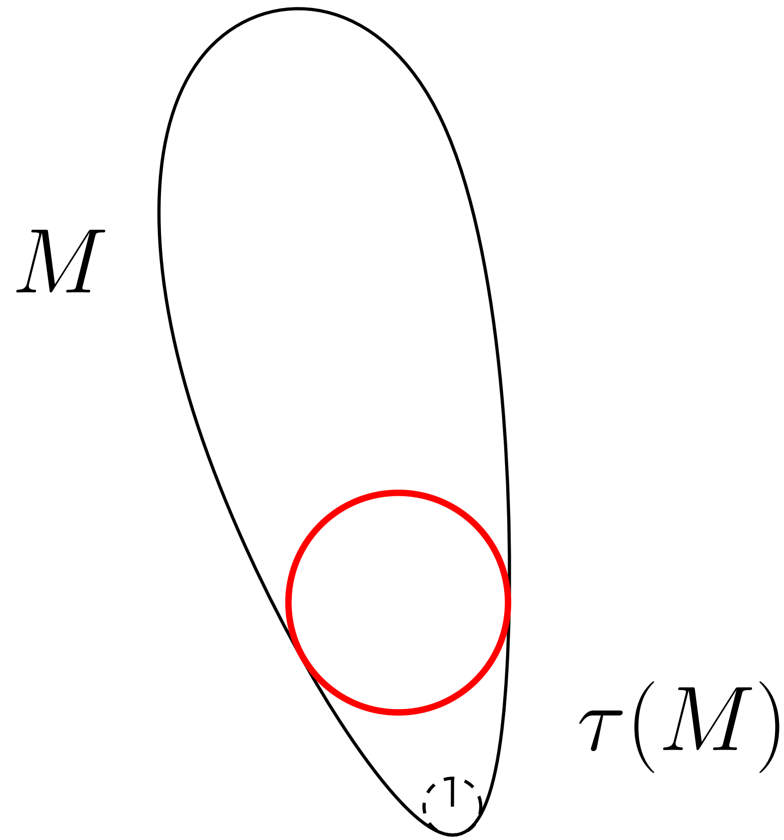


Sampling hypotheses

The **reach** of M is the radius of the largest ball one can make roll freely around M without bumping into it. [Federer '59]



controls the size of the minimal bottleneck



controls the curvature radius

Minimax rates

Definition: $\mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max}}$ is the set of distributions μ supported on a d-dimensional manifold M with reach $\tau(M) \geq \tau_{\min}$ and density f satisfying

$$\forall x \in M, \quad 0 < f_{\min} \leq f(x) \leq f_{\max} < \infty.$$

[Genovese & al. '12]

$$\mathrm{Risk}_n(\mu, \hat{M}) := \mathbb{E}_{\mu^{\otimes n}}[d_H(\hat{M}(\mathcal{X}_n), M)]$$

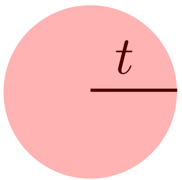
$\mathcal{R}_n(\mathcal{P}) :=$ the best average precision in the worst case

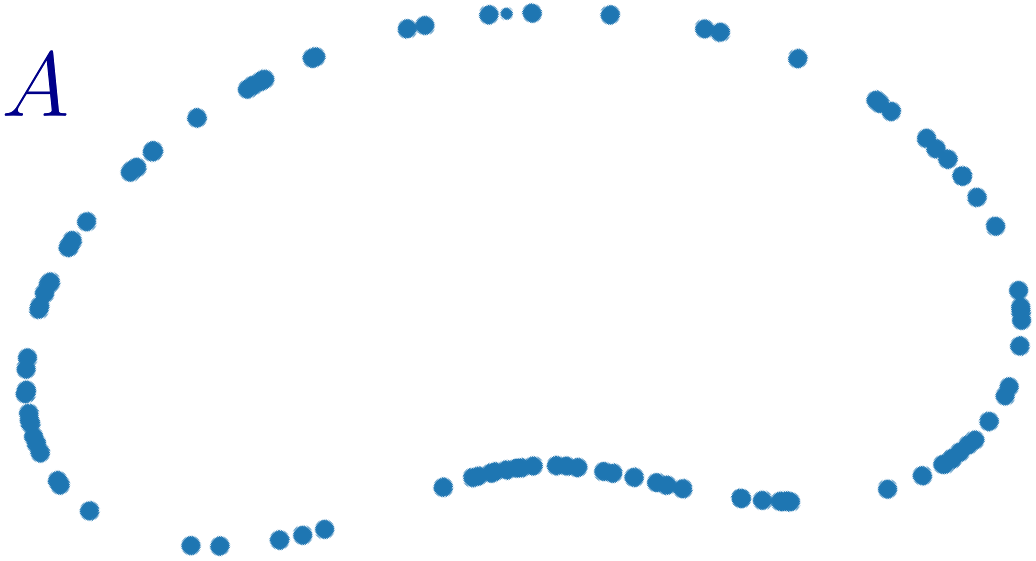
$$:= \inf_{\hat{M}} \sup_{\mu \in \mathcal{P}} \mathrm{Risk}_n(\mu, \hat{M})$$

$$\mathcal{R}_n(\mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max}}) \asymp \left(\frac{\ln n}{n}\right)^{2/d}$$

[Genovese & al. '12]
[Kim Zhou '15]

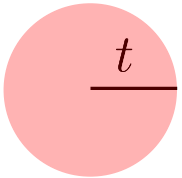
Local convex hull

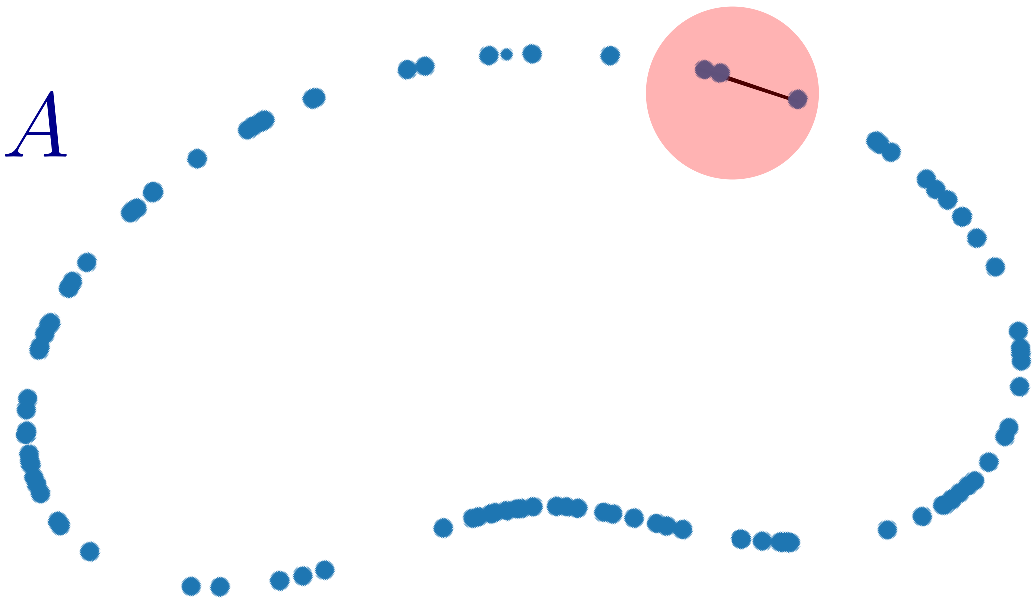
Choose a scale t 



Local convex hull

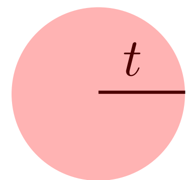
Choose a scale t

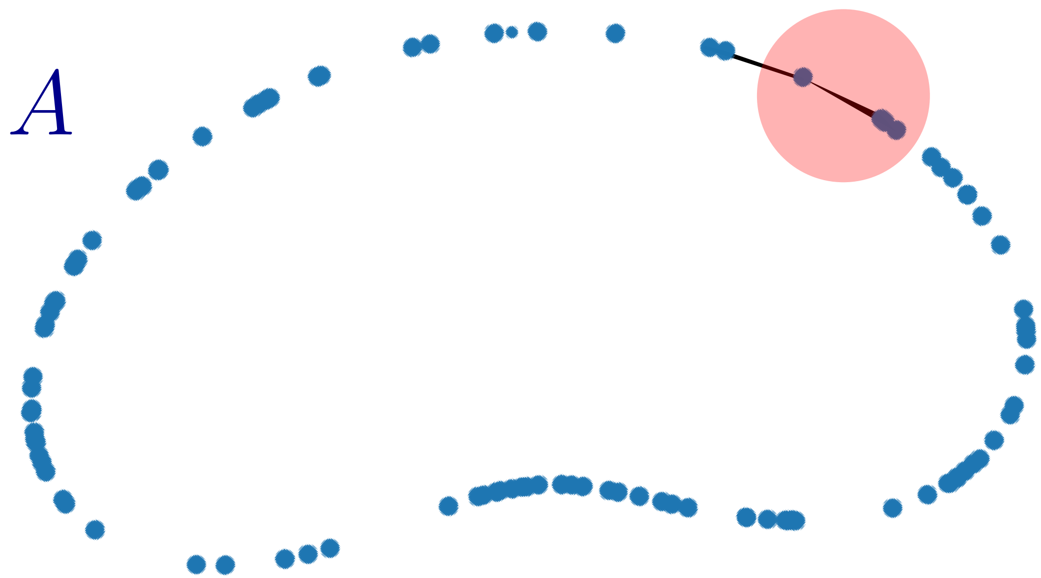




Local convex hull

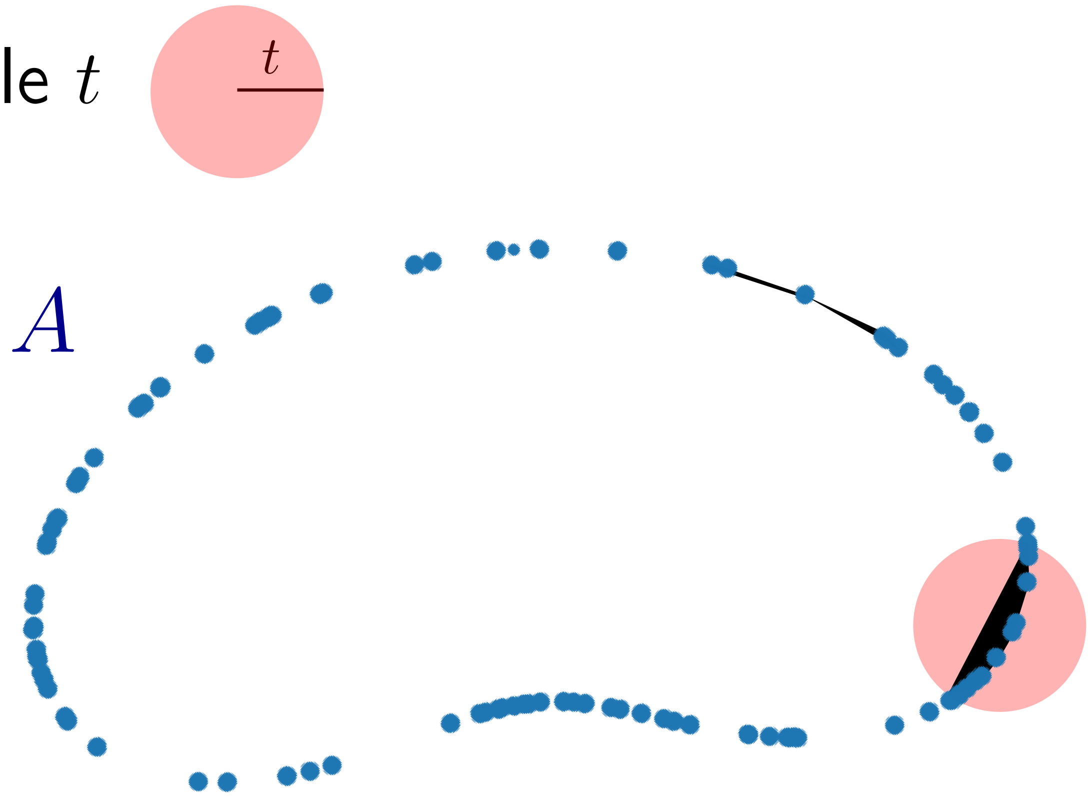
Choose a scale t





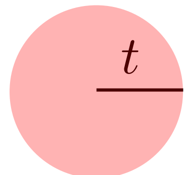
Local convex hull

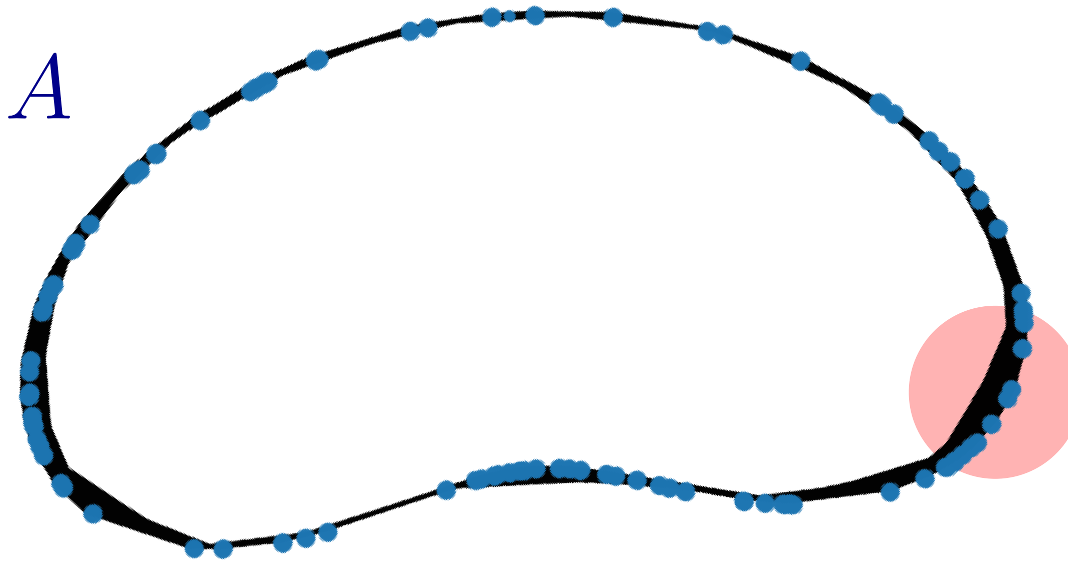
Choose a scale t



Local convex hull

Choose a scale t



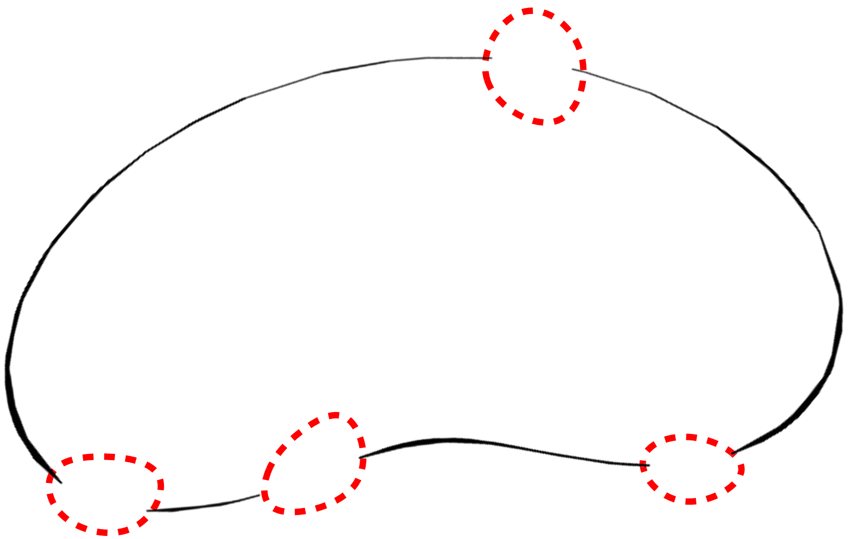


$$\mathrm{Conv}(t, A) = \bigcup_{\substack{\sigma \subset A \\ r(\sigma) \leq t}} \mathrm{Conv}(\sigma)$$

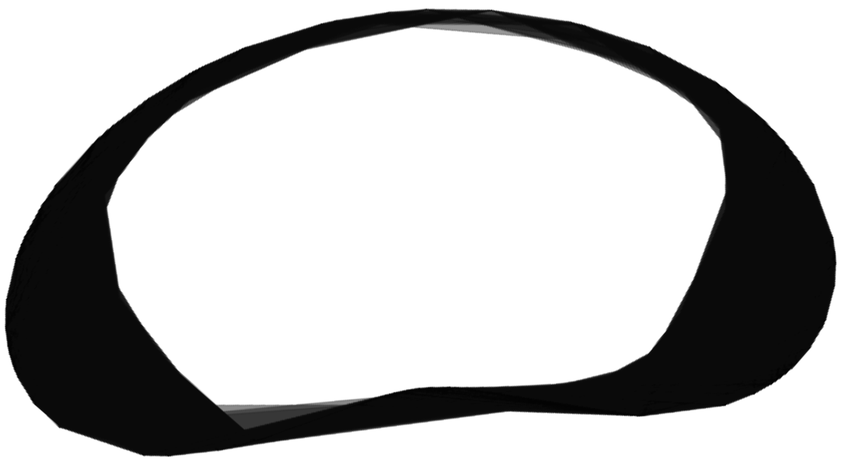
radius of the smallest enclosing ball of σ

Local convex hull

Let $t^*(A) := \inf\{t < \tau(M) : \ \pi_M(\mathrm{Conv}(t, A)) = M\}$.



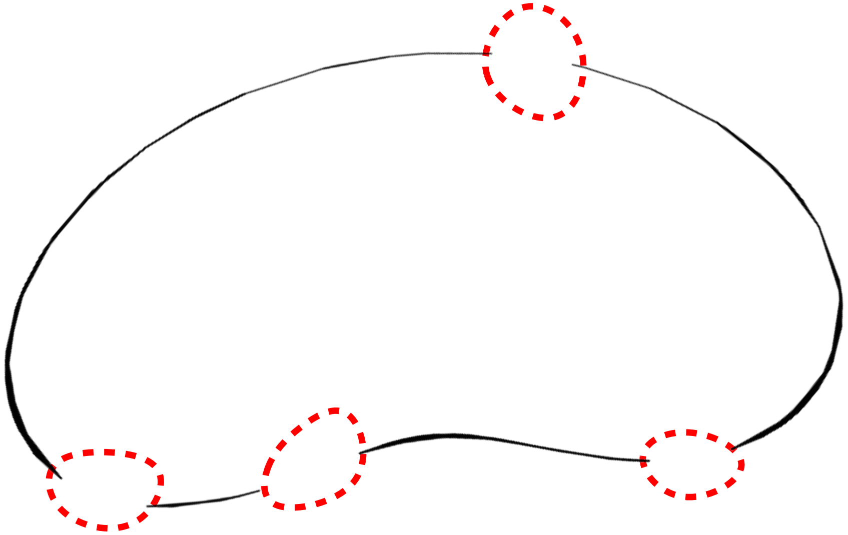
$t < t^*(A)$



$t \geq t^*(A)$

Local convex hull

Let $t^*(A) := \inf\{t < \tau(M) : \ \pi_M(\mathrm{Conv}(t, A)) = M\}$.



$t < t^*(A)$

$t \geq t^*(A)$

$\rightarrow$ Choose $t > t^*(A)$, but as small as possible.

Proposition: [D.] If $t \geq t^*(A)$, then $d_H(\mathrm{Conv}(t, A), M) \leq \frac{t^2}{\tau(M)}$.

Local convex hull

Theorem: [D.] Choose $t = C_{f_{\min},d}\left(\frac{\ln n}{n}\right)^{1/d}$. Let $\mu \in \mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max}}$. If $\mathcal{X}_n$ is a n-sample from law μ, then, for n large enough,

$$\mathbb{E}[d_H(\mathrm{Conv}(t,\mathcal{X}_n),M)] \lesssim \left(\frac{\ln n}{n}\right)^{2/d}.$$

Local convex hull

Theorem: [D.] Choose $t = C_{f_{\min},d}\left(\frac{\ln n}{n}\right)^{1/d}$. Let $\mu \in \mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max}}$. If $\mathcal{X}_n$ is a n-sample from law μ, then, for n large enough,

$$\mathbb{E}[d_H(\mathrm{Conv}(t, \mathcal{X}_n), M)] \lesssim \left(\frac{\ln n}{n}\right)^{2/d}.$$

the devil is in the details!

→ Same problem for all minimax manifold estimators [Genovese & al '12] [Aamari Levrard '18 '19] [Puchkin Spokoiny '19] [Sober Levin '19]

→ Problem of parameter tuning is a classical problem: model/parameter selection

- cross-validation [Arlot Celisse '09]
- penalization (e.g. ridge, Lasso, BIC/AIC)
- Goldenshluger-Lepski method/ PCO method [Lacour Massart Rivoirard '17]

Local convex hull

Theorem: [D.] Choose $t = C_{f_{\min},d}\left(\frac{\ln n}{n}\right)^{1/d}$. Let $\mu \in \mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max}}$. If $\mathcal{X}_n$ is a n-sample from law μ, then, for n large enough,

$$\mathbb{E}[d_H(\mathrm{Conv}(t,\mathcal{X}_n),M)] \lesssim \left(\frac{\ln n}{n}\right)^{2/d}.$$

the devil is in the details!

→ Same problem for all minimax manifold estimators [Genovese & al '12] [Aamari Levrard '18 '19] [Puchkin Spokoiny '19] [Sober Levin '19]

→ Problem of parameter tuning is a classical problem: model/parameter selection

- cross-validation [Arlot Celisse '09]
- penalization (e.g. ridge, Lasso, BIC/AIC)
- Goldenshluger-Lepski method/ PCO method [Lacour Massart Rivoirard '17]

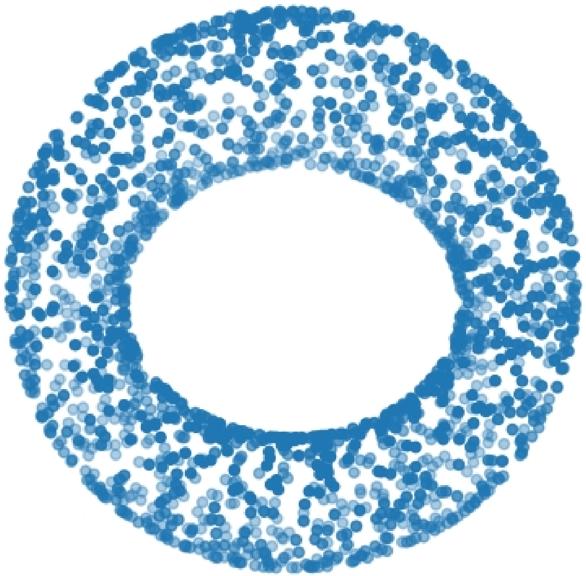
"Compare each estimator of the family with the most overfitted estimator of the family"

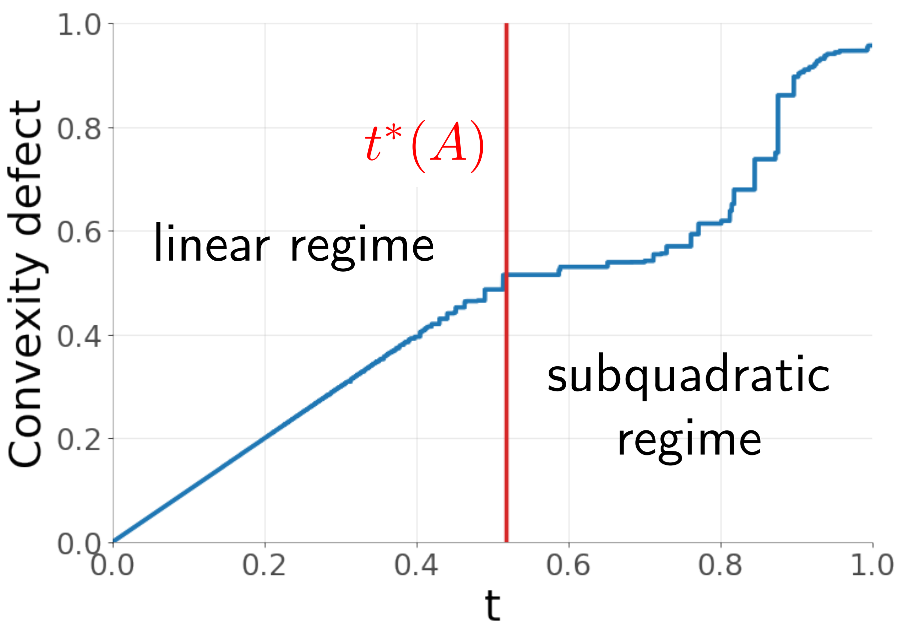
Convexity defect function

Definition: [Attali, Lieutier, Salinas '12] Let $A \subset \mathbb{R}^D$. The convexity defect function of A at scale t is defined by $h(t, A) := d_H(\mathrm{Conv}(t, A), A)$.

→ For M a manifold, $h(t, M) \leq t^2/\tau(M)$.

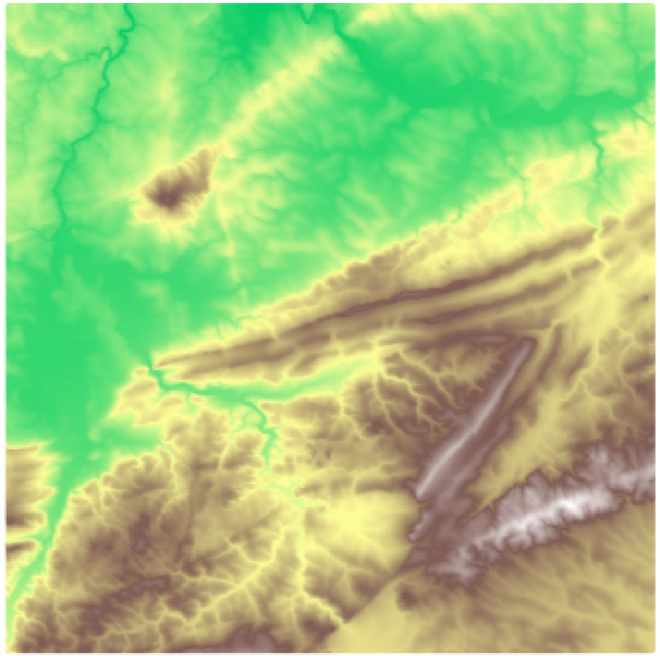
→ And for $\mathcal{X}_n$?

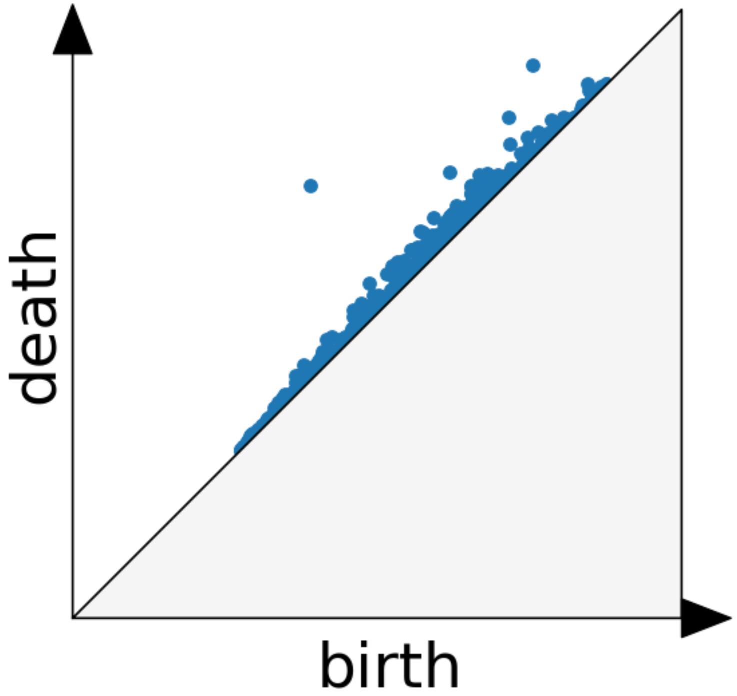




Part II

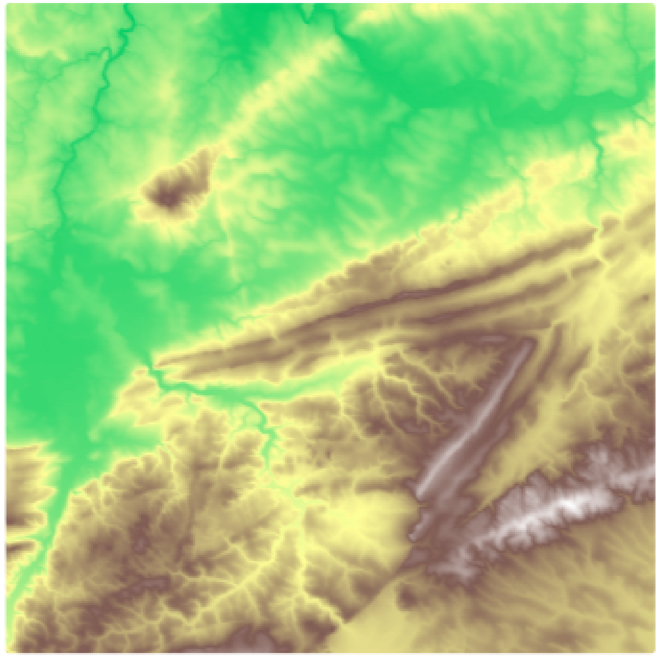
Statistics and persistence diagrams





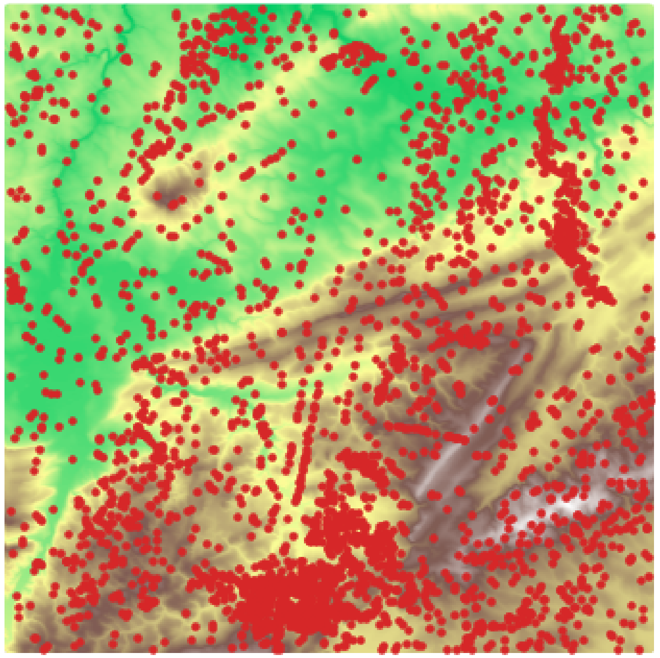
Alpinism

What is a peak?



Alpinism

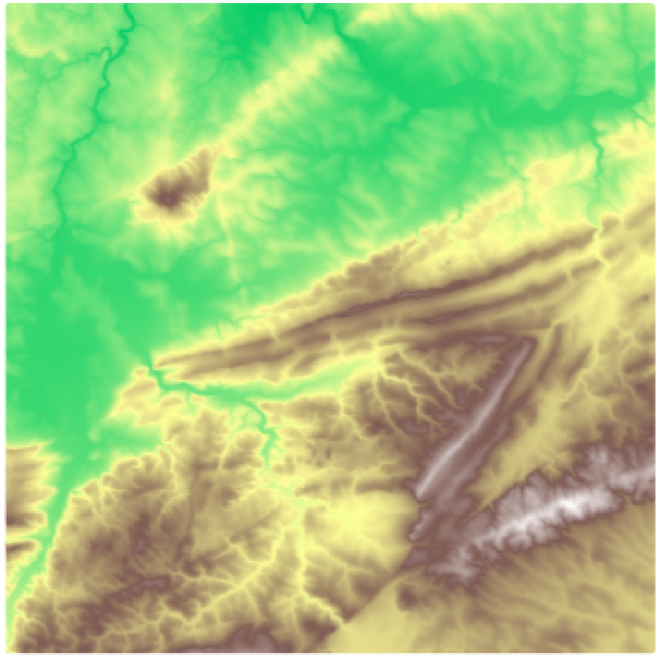
What is a peak?



A local maximum of the elevation function?

Alpinism

What is a peak?

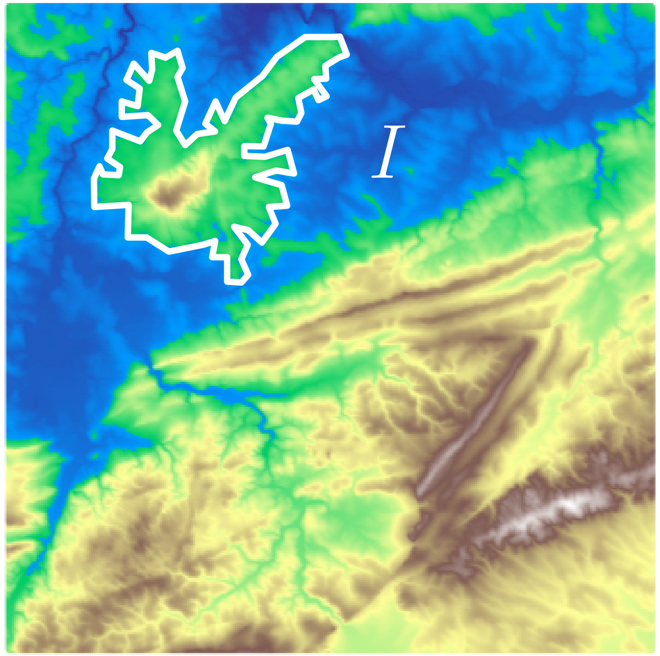


A local maximum of the elevation function?

Alpinism

What is a peak?

The island I appears at sea level b (its **birth time**) ...

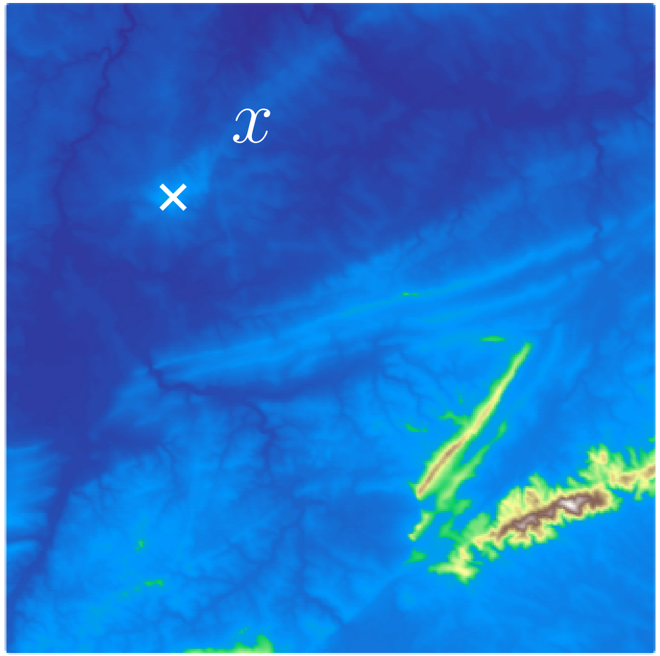


Alpinism

What is a peak?

The island I appears at sea level b (its **birth time**) ...

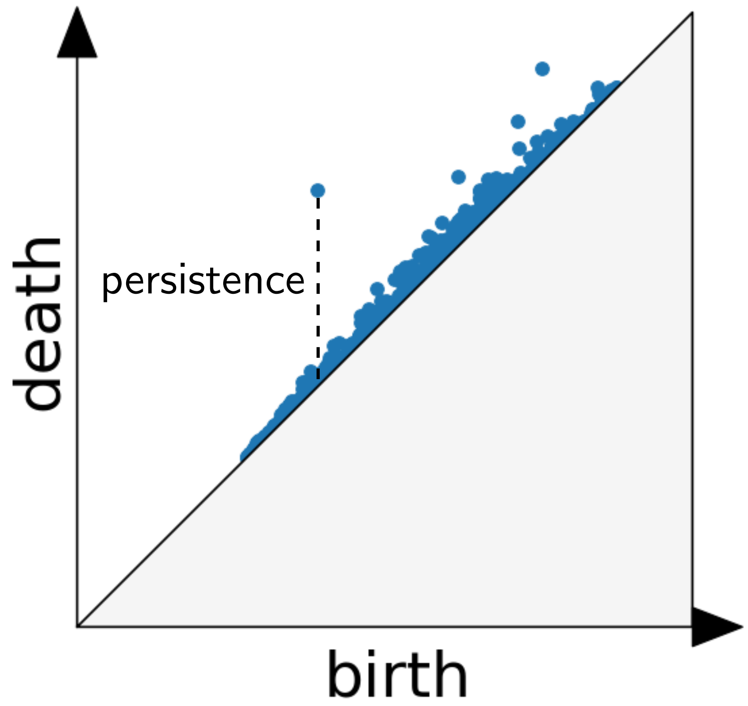
... and disapears at sea level d (its **death time**) at local maximum x.



The point x is a peak if the **persistence** $:= d - b$ of the island I is larger than 91m $(= 300\text{ft})$.

Alpinism

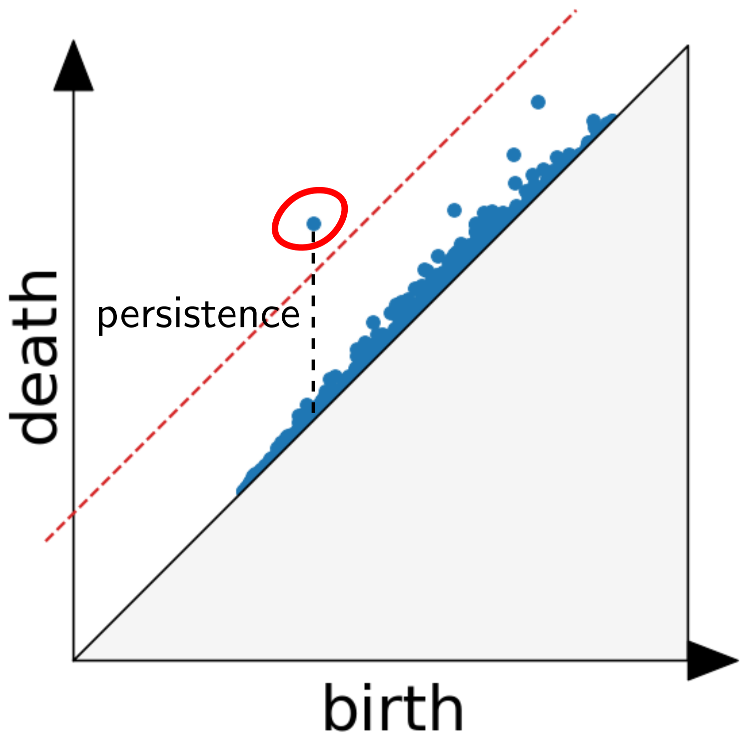
What is a peak?



The persistence diagram (PD) of the elevation function is the collection of the points (b, d), where (b, d) corresponds to the birth/death of an island.

Alpinism

What is a peak?

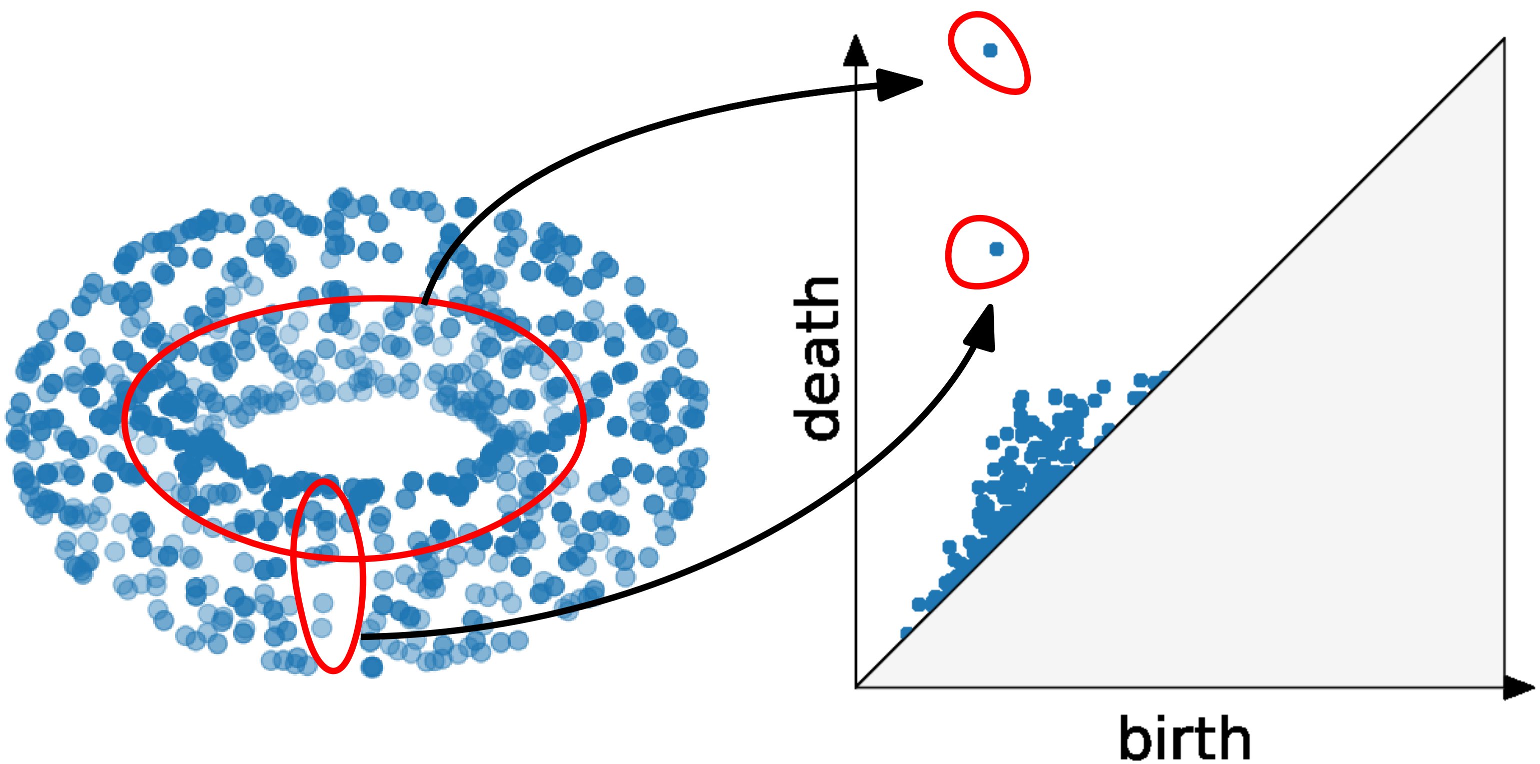


The persistence diagram (PD) of the elevation function is the collection of the points (b, d), where (b, d) corresponds to the birth/death of an island.

Persistence diagrams

- Let $\mathcal{X}$ be a loc. finite simplicial complex. Then, the persistence diagram $\mathrm{dgm}(\phi)$ is defined for any proper continuous function $\phi : \mathcal{X} \to [0, \infty)$.

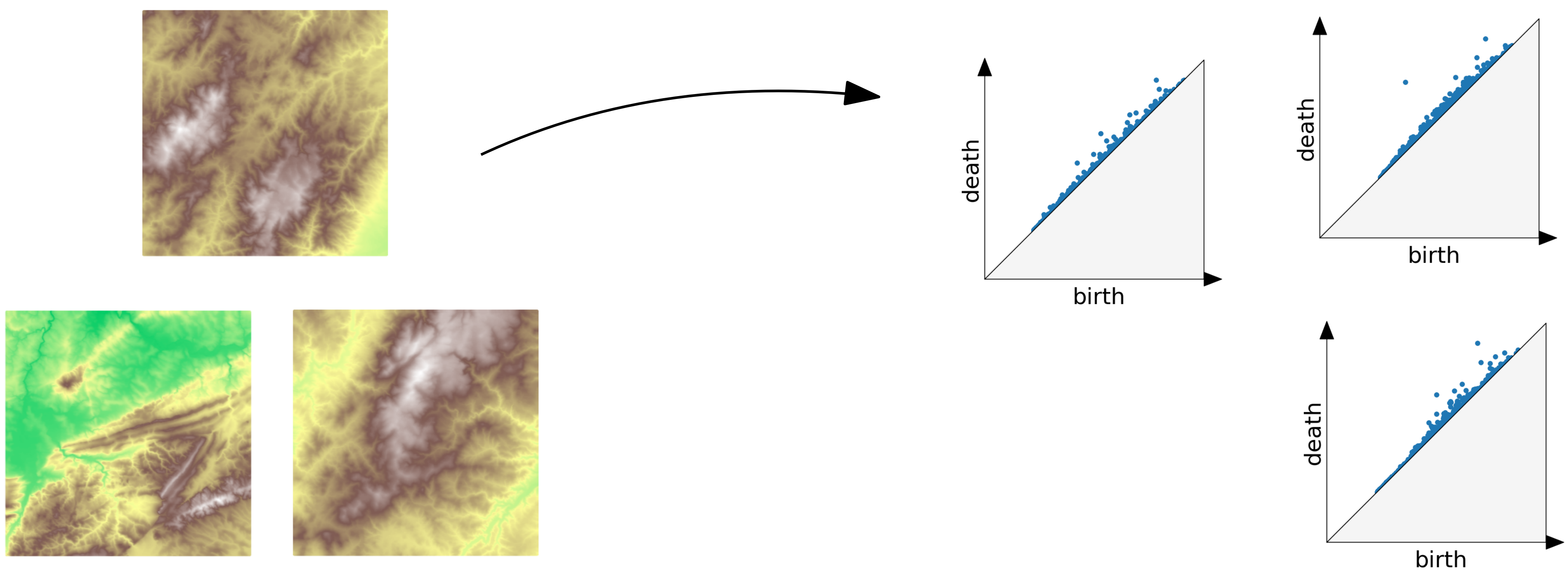
  Ex: ϕ is the distance function to a set A.



Persistence diagrams

- Let $\mathcal{X}$ be a loc. finite simplicial complex. Then, the persistence diagram $\mathrm{dgm}(\phi)$ is defined for any proper continuous function $\phi : \mathcal{X} \to [0, \infty)$.

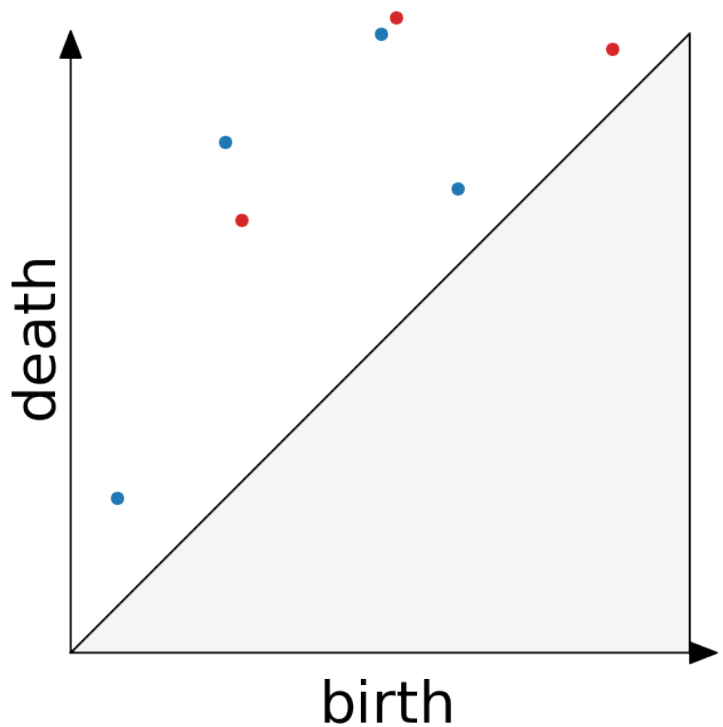
  Ex: ϕ is the distance function to a set A.



Distance between persistence diagrams

Let a and b be two persistence diagrams. Let $1 \leq p \leq \infty$. Let $\Gamma(a, b)$ be the set of bijections between $a \cup \partial\Omega$ and $b \cup \partial\Omega$.

$$\boxed{d_p(a, b) := \inf_{\gamma \in \Gamma(a,b)} \left(\sum_{x \in a \cup \partial\Omega} |x - \gamma(x)|^p \right)^{1/p}}$$



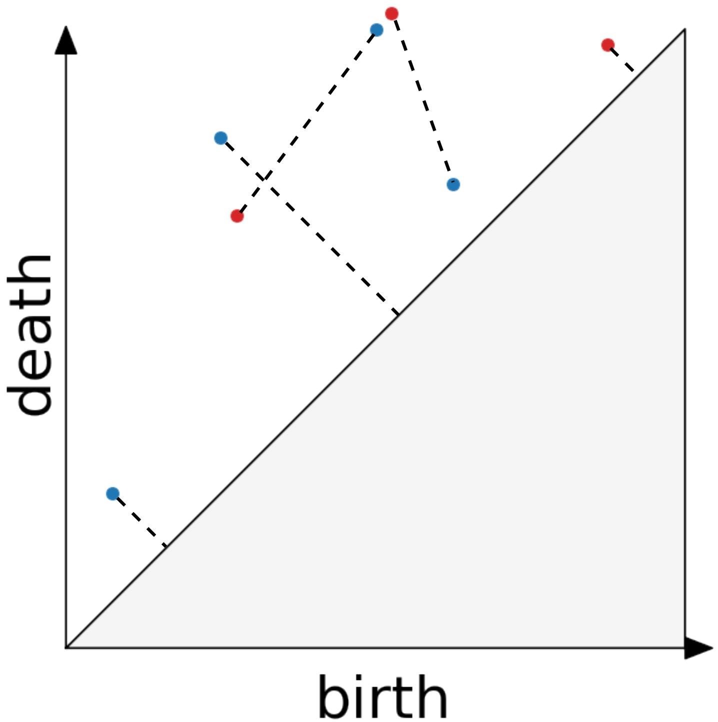
$$\boxed{\mathrm{Pers}_p(a) := \sum_{x \in a} \mathrm{pers}(x)^p}$$

$$\boxed{\mathcal{D}^p := \{a : \ \mathrm{Pers}_p(a) < \infty\}}$$

Distance between persistence diagrams

Let a and b be two persistence diagrams. Let $1 \leq p \leq \infty$. Let $\Gamma(a, b)$ be the set of bijections between $a \cup \partial\Omega$ and $b \cup \partial\Omega$.

$$\boxed{d_p(a, b) := \inf_{\gamma \in \Gamma(a,b)} \left(\sum_{x \in a \cup \partial\Omega} |x - \gamma(x)|^p \right)^{1/p}}$$



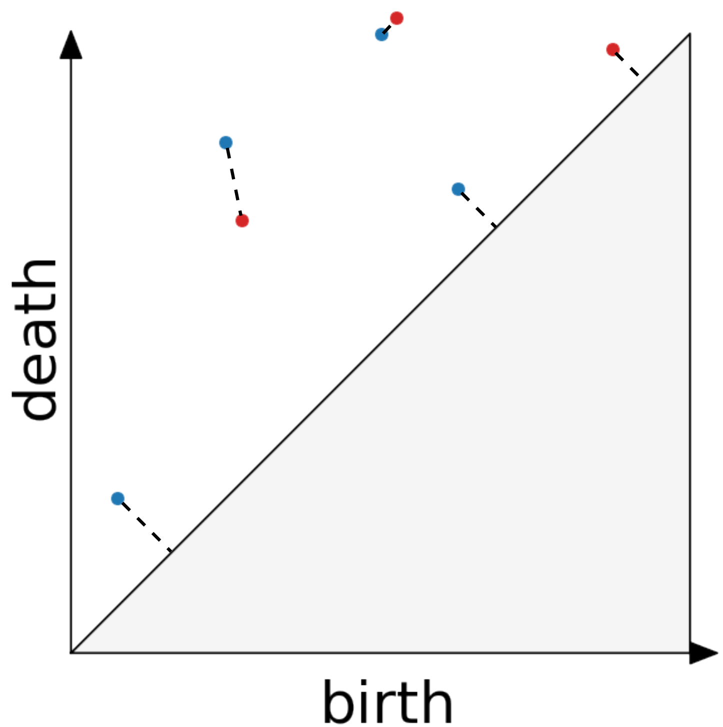
$$\boxed{\mathrm{Pers}_p(a) := \sum_{x \in a} \mathrm{pers}(x)^p}$$

$$\boxed{\mathcal{D}^p := \{a : \ \mathrm{Pers}_p(a) < \infty\}}$$

Distance between persistence diagrams

Let a and b be two persistence diagrams. Let $1 \leq p \leq \infty$. Let $\Gamma(a, b)$ be the set of bijections between $a \cup \partial\Omega$ and $b \cup \partial\Omega$.

$$\boxed{d_p(a, b) := \inf_{\gamma \in \Gamma(a,b)} \left(\sum_{x \in a \cup \partial\Omega} |x - \gamma(x)|^p \right)^{1/p}}$$



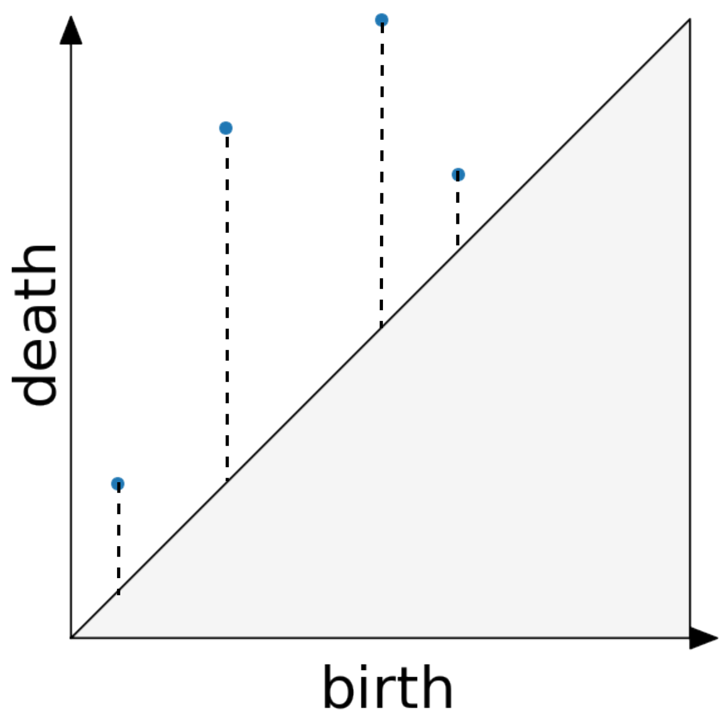
$$\boxed{\mathrm{Pers}_p(a) := \sum_{x \in a} \mathrm{pers}(x)^p}$$

$$\boxed{\mathcal{D}^p := \{a : \ \mathrm{Pers}_p(a) < \infty\}}$$

Distance between persistence diagrams

Let a and b be two persistence diagrams. Let $1 \leq p \leq \infty$. Let $\Gamma(a, b)$ be the set of bijections between $a \cup \partial\Omega$ and $b \cup \partial\Omega$.

$$\boxed{d_p(a, b) := \inf_{\gamma \in \Gamma(a,b)} \left(\sum_{x \in a \cup \partial\Omega} |x - \gamma(x)|^p \right)^{1/p}}$$



$$\boxed{\mathrm{Pers}_p(a) := \sum_{x \in a} \mathrm{pers}(x)^p}$$

$$\boxed{\mathcal{D}^p := \{a : \ \mathrm{Pers}_p(a) < \infty\}}$$

The measure point of view

$a =$ (multi) set $\longleftrightarrow$ $\sum_{x \in a} \delta_x =$ measure

$\mathcal{D}^p \subset \mathcal{M}^p := \{\mu$ Radon measure : $\mathrm{Pers}_p(\mu) < \infty\}$

$=$ space of **persistent measures**

[Figalli Gigli '10]

There exists an optimal transport metric FG_p on $\mathcal{M}^p$ that extends d_p.

[D. Lacombe '20] $\boxed{\forall a, b \in \mathcal{D}^p, \quad d_p(a,b) = \mathrm{FG}_p(a,b)}$

The measure point of view

$a =$ (multi) set $\longleftrightarrow$ $\sum_{x\in a}\delta_x =$ measure

$$\mathcal{D}^p \subset \mathcal{M}^p := \{\mu \text{ Radon measure}: \ \mathrm{Pers}_p(\mu) < \infty\}$$
$=$ space of **persistent measures**

[Figalli Gigli '10]

There exists an optimal transport metric FG_p on $\mathcal{M}^p$ that extends d_p.

[D. Lacombe '20] $\boxed{\forall a, b \in \mathcal{D}^p, \quad d_p(a,b) = \mathrm{FG}_p(a,b)}$

Limit theorems

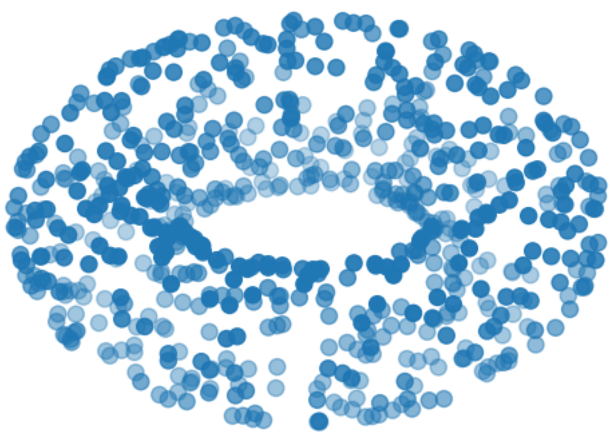
Expectations

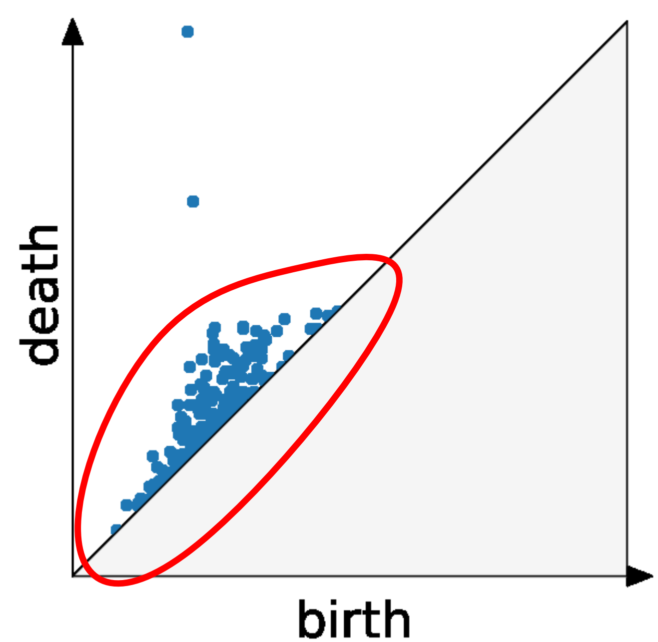
The structure of the topological noise

Theorem: [D. Polonik '19] Let f be a density on $[0,1]^d$ satisfying $0 < f_{\min} \leq f \leq f_{\max} < \infty$. Let $\mathcal{X}_n$ be a n-sample of density f, and a_n the persistence diagram of $n^{1/d}\mathcal{X}_n$. Then, there exists $\mu \neq 0$ in $\mathcal{M}^p$ such that

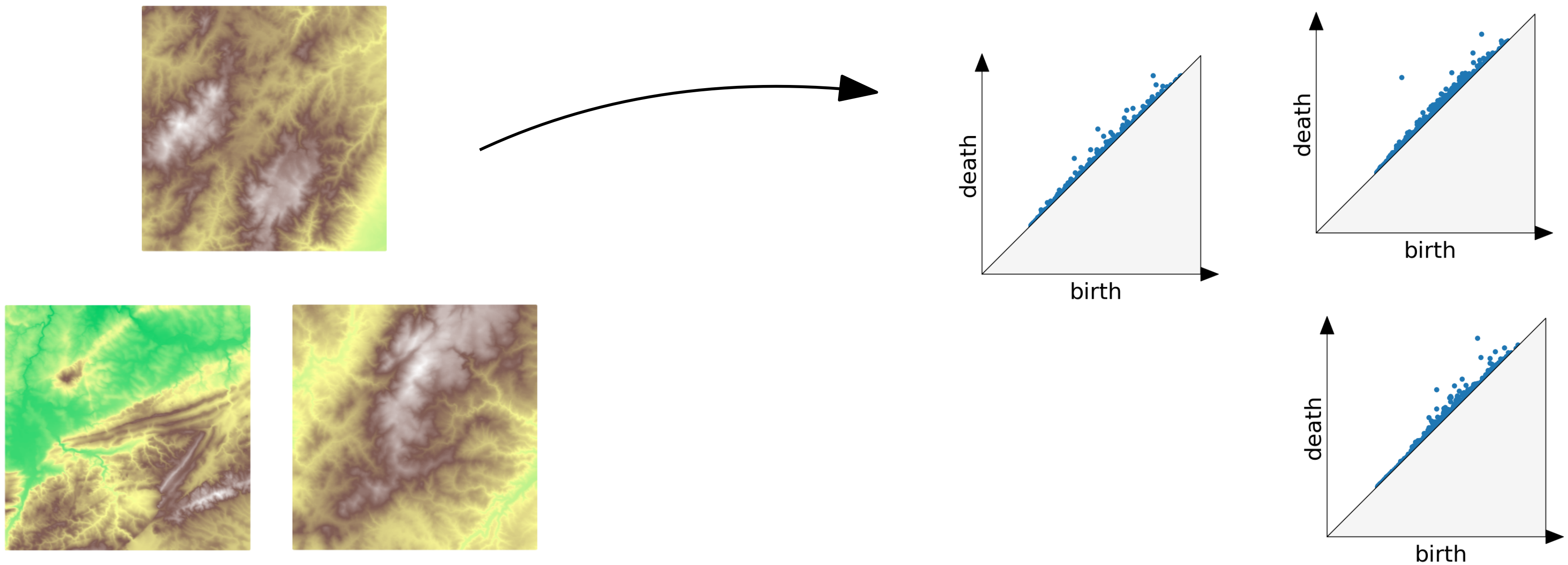
$$\mathrm{FG}_p\left(\frac{a_n}{n}, \mu\right) \to 0$$

$$\implies \mathrm{Pers}_p(a_n) \simeq n^{1-p/d}$$





The expected persistence diagram



$$C_1, \ldots, C_K \qquad\qquad a_1, \ldots, a_K$$

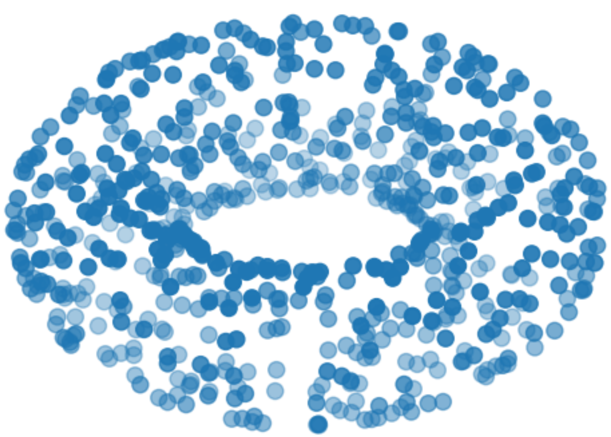
"average topology" $= \overline{a}_K = \frac{a_1 + \cdots + a_K}{K}$

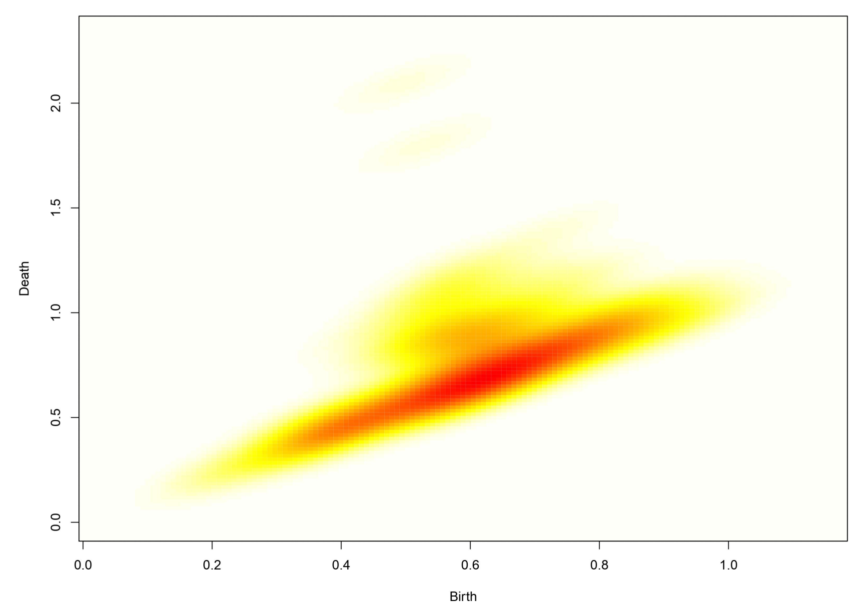
The expected persistence diagram

Let P be a probability measure on $\mathcal{D}^p$. The **expected persistence diagram** $E(P)$ is the element of $\mathcal{M}^p$ defined by the relation

$$\forall B \text{ measurable set,} \quad E(P)(B) = \mathbb{E}_{a \sim P}[a(B)]$$

Theorem 1: [Chazal D. '19] Let P be the distribution of the random persistence diagram obtained by sampling n points on a manifold M with (smooth) density f. Then, $E(P)$ is a measure with a (smooth) density.





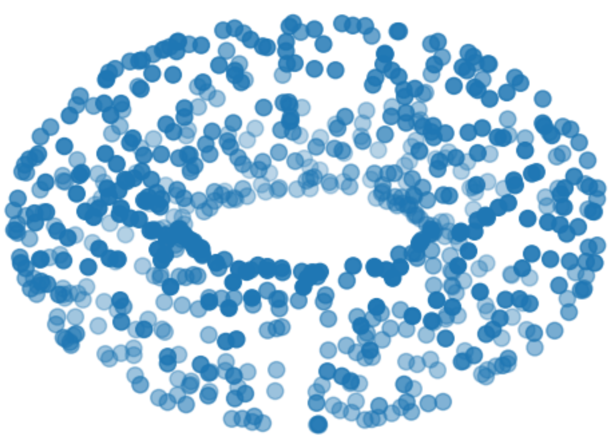
The expected persistence diagram

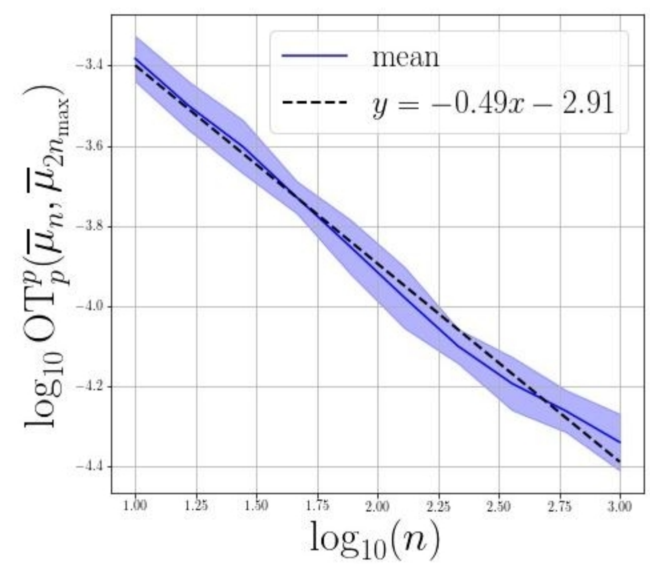
Let P be a probability measure on $\mathcal{D}^p$. The **expected persistence diagram** $E(P)$ is the element of $\mathcal{M}^p$ defined by the relation

$$\forall B \text{ measurable set,} \quad E(P)(B) = \mathbb{E}_{a\sim P}[a(B)]$$

Theorem 2: [D. Lacombe '21] Let $a_1, \ldots, a_K$ be a K-sample of distribution P with $\mathrm{Card}(a_i) \leq M$ a.s. and a_i supported on $\mathcal{B}(0, L)$ a.s. Then,

$$\mathbb{E}[\mathrm{FG}_p^p(\overline{a}_K, E(P))] \lesssim M L^p K^{-1/2}.$$





Conclusion

• In Part I, we proposed an **adaptive** manifold estimator

$\rightarrow$ and in the presence of outliers?

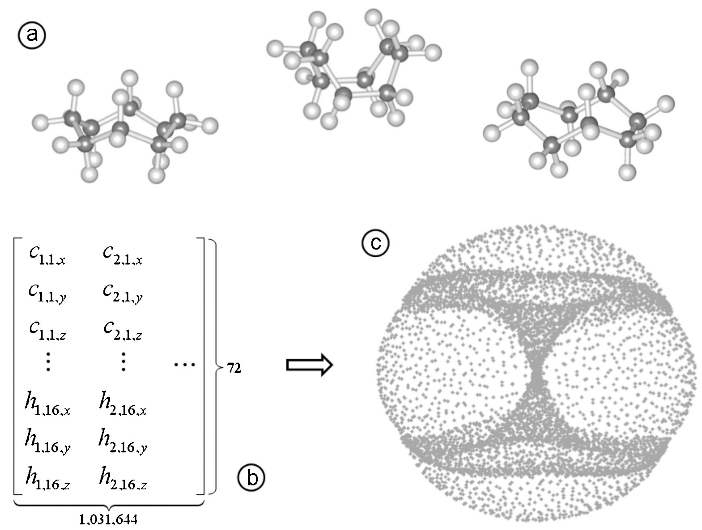
• In Part II, we took a measure point of view to study the space of persistence diagrams.

$\boxed{\text{"Any optimal transport related ML technique can be translated to the persistence diagram setting."}}$

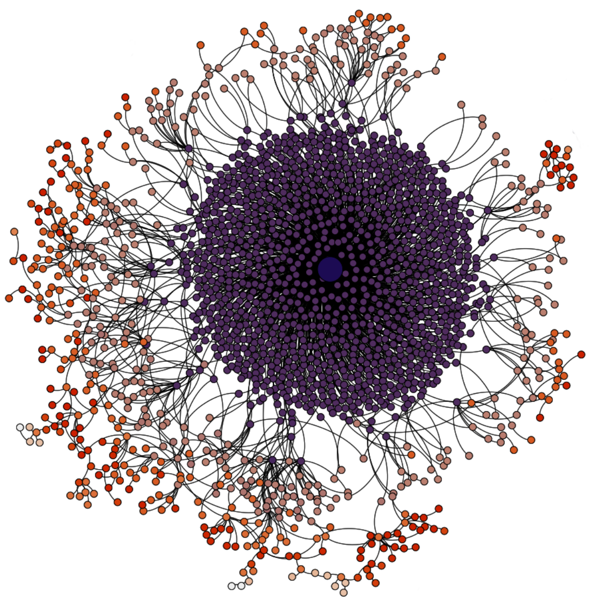
$\rightarrow$ quantization, entropic regularization, differentiation, ...

Geometry and topology in data

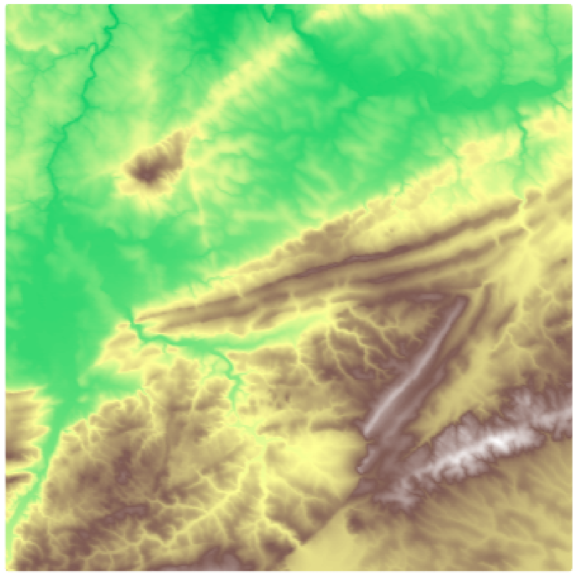
[Pickup & al, '14]



[Martin & al, '10]



[Yanardag & al., '15]



[IGN elevation dataset]

Manifold inference

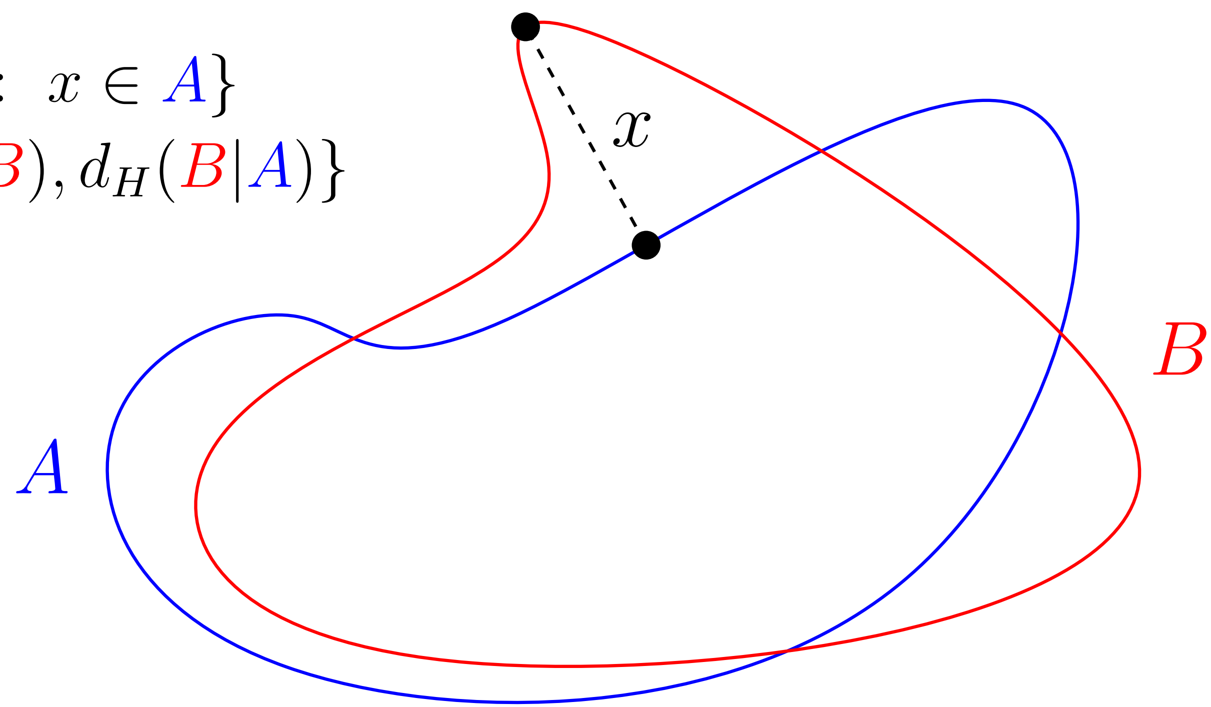
- $\mathcal{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ is a set of observations close to a manifold M (dimension d, compact, without boundary)

- Goal: reconstruct a geometric invariant of M. (ex: dimension, tangent spaces, curvature, M itself)

Question 1: How to quantify the quality of a given reconstruction?

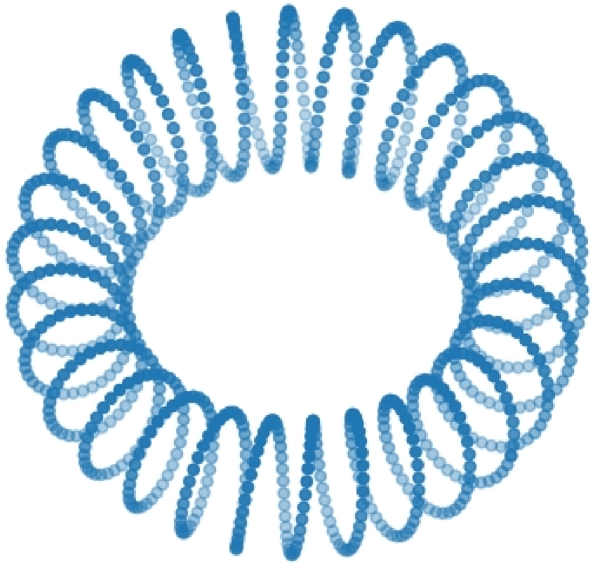
- The Hausdorff distance between A and $B \subset \mathbb{R}^D$ is defined by:

$$d_H(A|B) := \sup\{d(x, B) : \ x \in A\}$$
$$d_H(A, B) := \max\{d_H(A|B), d_H(B|A)\}$$



Sampling hypotheses

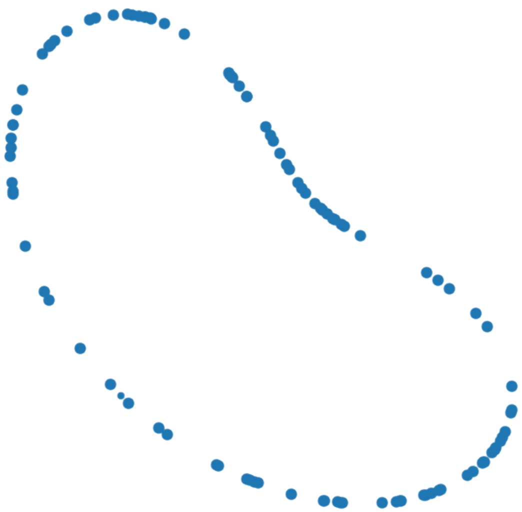
A trickier example: $M =$ spire or torus?

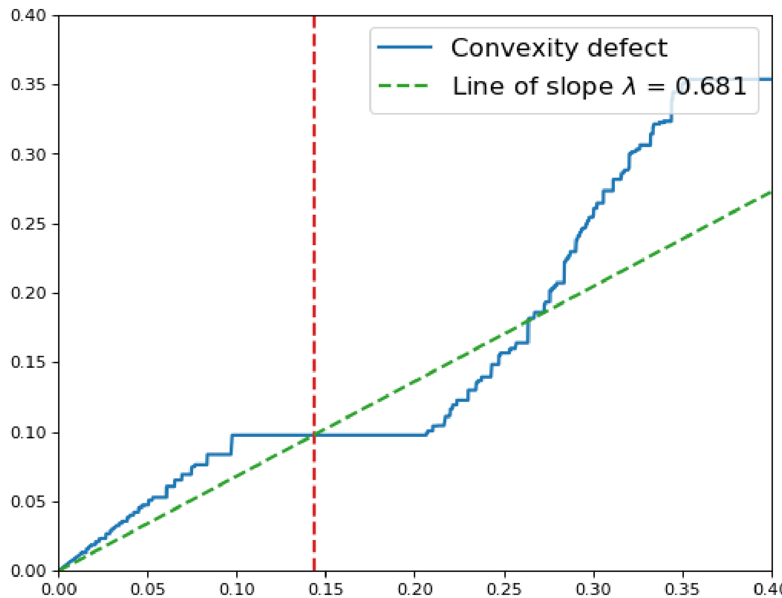


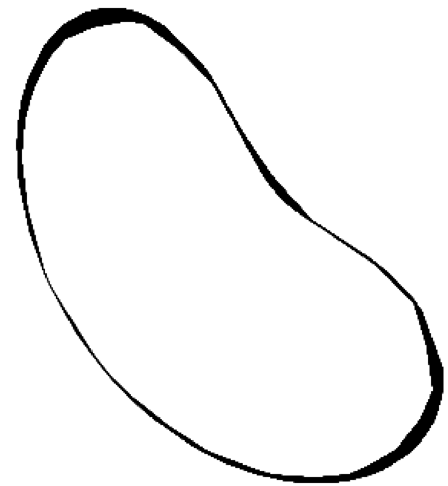
Convexity defect function

Choose $0 < \lambda < 1$ and $t_{\max} = 1/\ln n$.

$$t_\lambda(A) := \sup\{t < t_{\max} : \ h(t, A) \geq \lambda t\}$$







Theorem: [D.] Let $\mu \in \mathcal{P}^d_{\tau_{\min}, f_{\min}, f_{\max}}$. If $\mathcal{X}_n$ is a n-sample from law μ, then, for n large enough,

$$\mathbb{E} d_H(\mathrm{Conv}(t_\lambda(\mathcal{X}_n), \mathcal{X}_n), M) \lesssim \left(\frac{\ln n}{n}\right)^{2/d}.$$