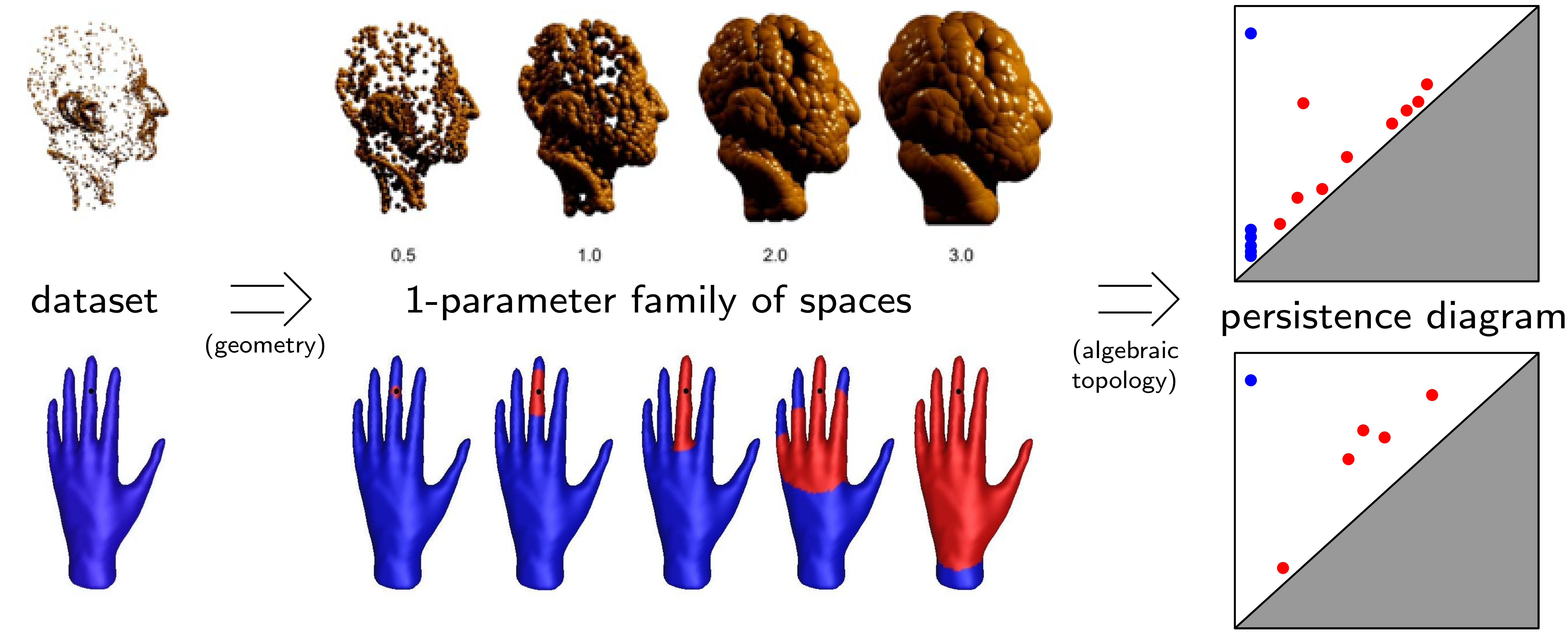
Statistics on Persistence Diagrams

Persistence diagrams as descriptors for data



Pros:

- strong invariance and stability:

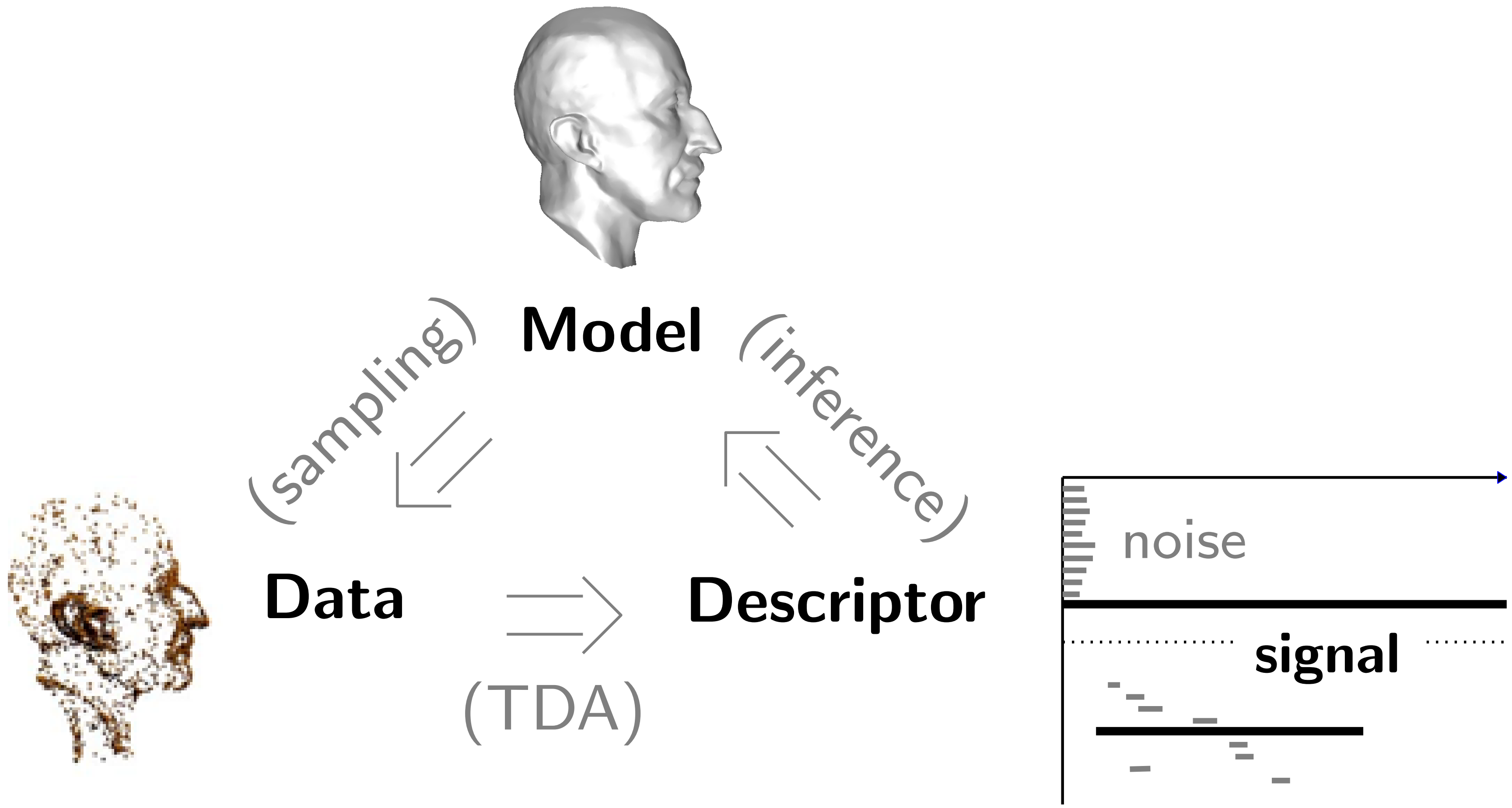
  $\mathrm{d}_p(\mathrm{Dg}\, X, \mathrm{Dg}\, Y) \leq \mathrm{cst}\, \mathrm{d}_{\mathrm{GH}}(X, Y)$

- information of a different nature
- flexible and versatile

Cons:

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

Statistics for persistence diagrams



Statistics:

- signal vs noise discrimination
- convergence rates
- confidence indices/intervals, principal components, etc.

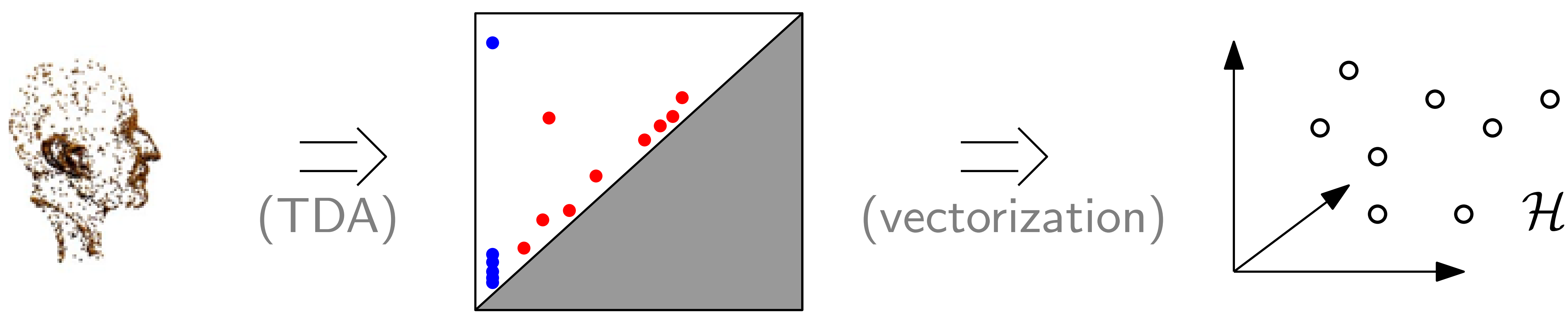
Statistics for persistence diagrams

3 approaches for statistics:

- Fréchet means in diagrams space

- embedding into Hilbert spaces

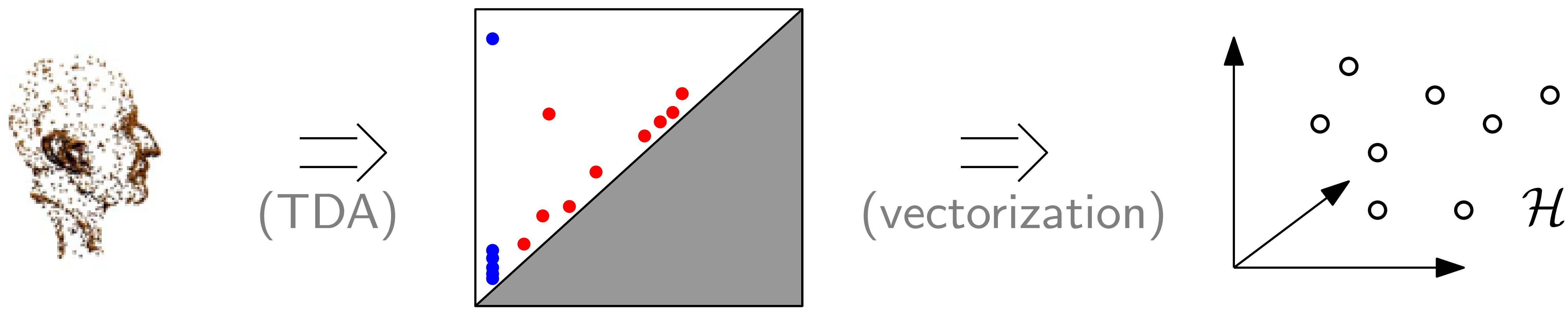
- push-forwards from data space



Statistics for persistence diagrams

3 approaches for statistics:

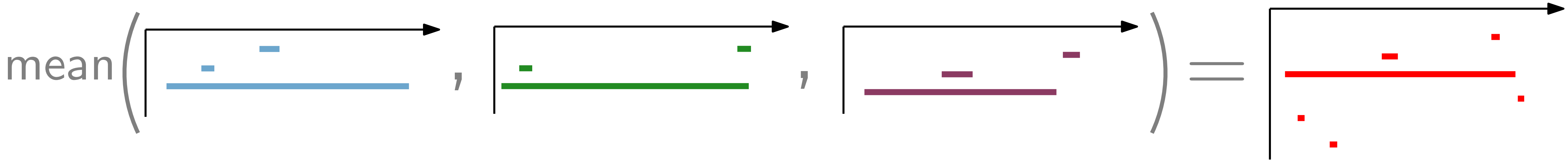
- Fréchet means in diagrams space — 1
- embedding into Hilbert spaces — 2 — today's menu
- push-forwards from data space — 3

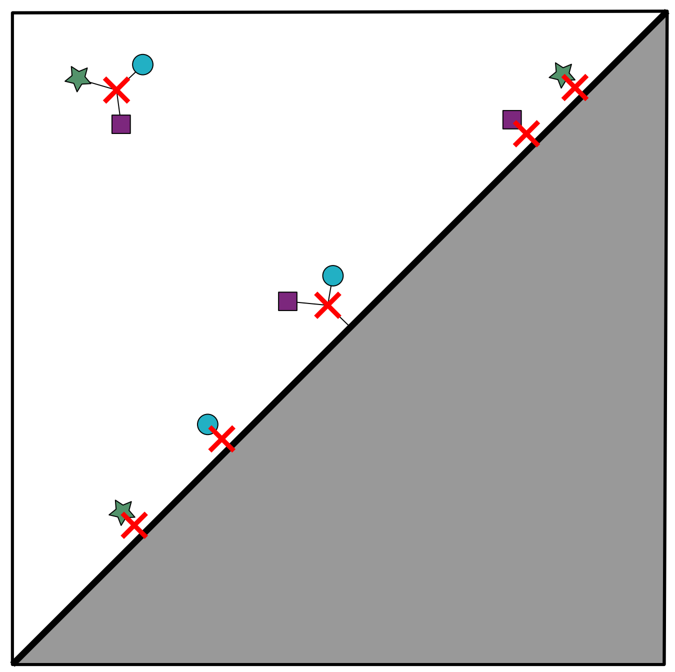


1. Fréchet means in diagrams space

Means as a gateway to statistical analysis

central limit theorems, confidence intervals, geodesic PCA,
clustering (k-means, EM, Mean-Shift, etc.)

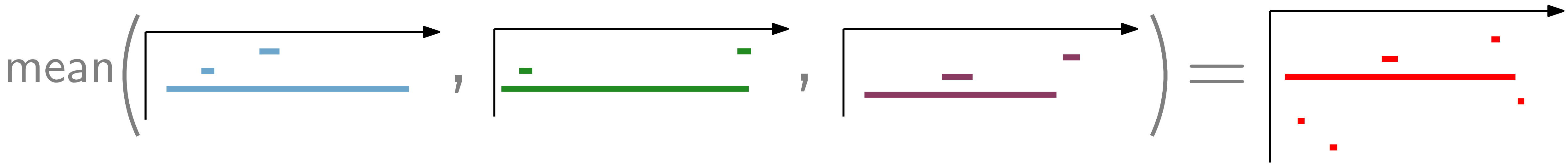




1. Fréchet means in diagrams space

Means as a gateway to statistical analysis

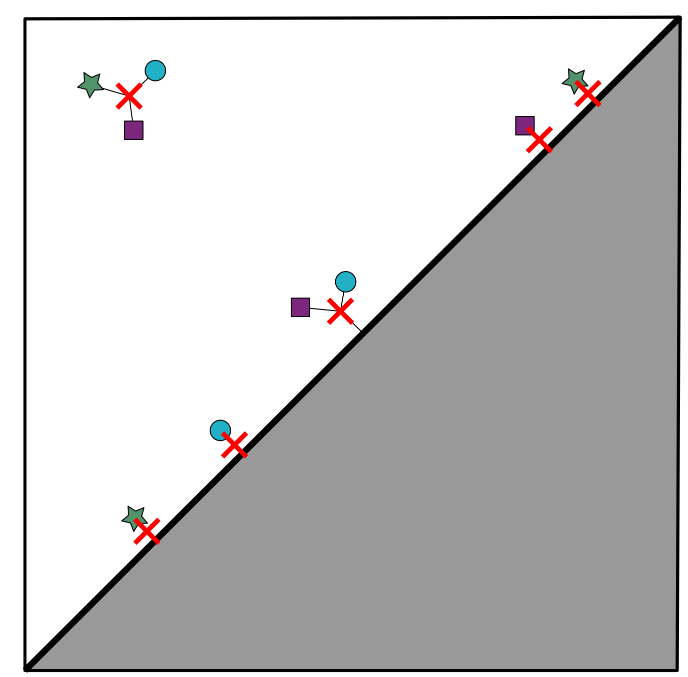
central limit theorems, confidence intervals, geodesic PCA,
clustering (k-means, EM, Mean-Shift, etc.)



No coordinates $\rightsquigarrow$ means as minimizers of variance (Fréchet means)

Given diagrams $D_1, \cdots, D_n$:

$$\bar{D} \in \arg\min_{D} \frac{1}{n} \sum_i \mathrm{d}_p(D,\, D_i)^2$$



Prop.: minimizers do exist

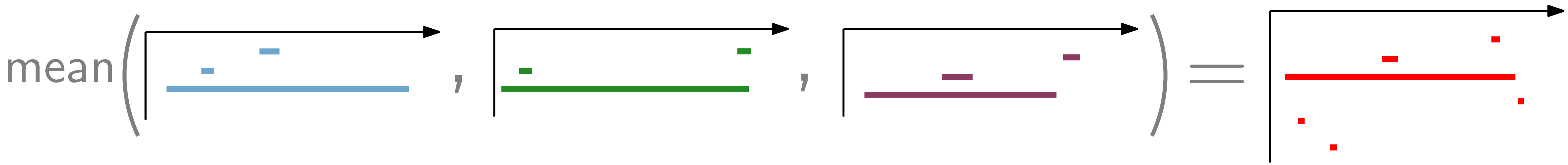
(diagram space is complete and separable)

[K. Turner et al.: "Fréchet means for distributions of persistence diagrams", 2012]

1. Fréchet means in diagrams space

Means as a gateway to statistical analysis

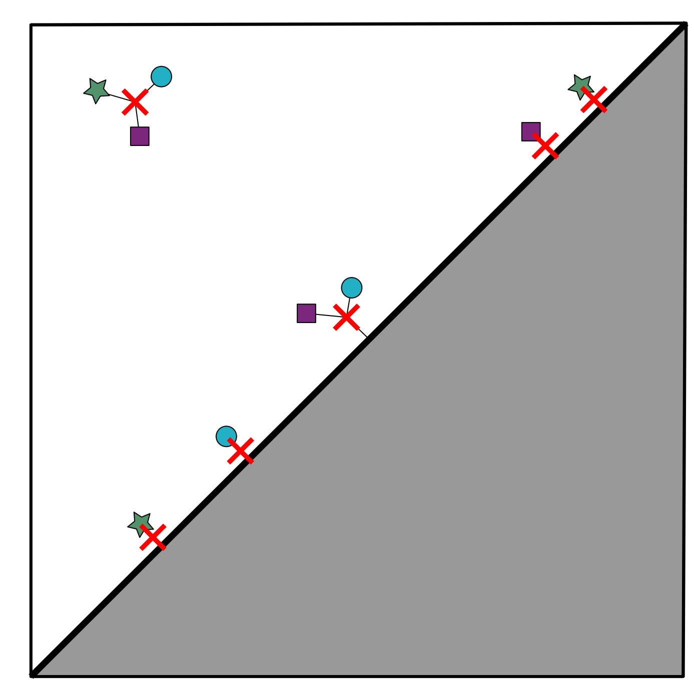
central limit theorems, confidence intervals, geodesic PCA,
clustering (k-means, EM, Mean-Shift, etc.)



No coordinates $\rightsquigarrow$ means as minimizers of variance (Fréchet means)

Given diagrams $D_1, \cdots, D_n$:

$$\bar{D} \in \arg\min_{D} \frac{1}{n} \sum_i \mathrm{d}_p(D,\, D_i)^2$$



Problem: non-unique argmin, local minima, num. issues

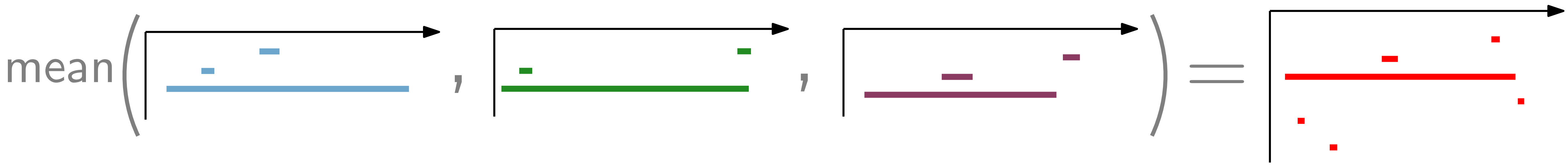
(non-convex energy, highly curved space)

[K. Turner et al.: "Fréchet means for distributions of persistence diagrams", 2012]

1. Fréchet means in diagrams space

Means as a gateway to statistical analysis

central limit theorems, confidence intervals, geodesic PCA,
clustering (k-means, EM, Mean-Shift, etc.)



No coordinates $\rightsquigarrow$ means as minimizers of variance (Fréchet means)

Given diagrams $D_1, \cdots, D_n$:

$$\bar{D} \in \arg\min_{D} \; \frac{1}{n} \sum_i \mathrm{d}_p(D,\, D_i)^2$$

barcode distance is a transportation type distance $\rightsquigarrow$ connection to Optimal Transport

Problem: non-unique argmin, local minima, num. issues

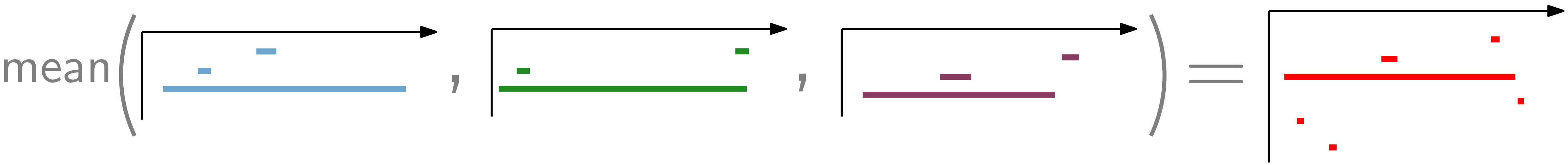
(non-convex energy, highly curved space)

[K. Turner et al.: "Fréchet means for distributions of persistence diagrams", 2012]

1. Fréchet means in diagrams space

Means as a gateway to statistical analysis

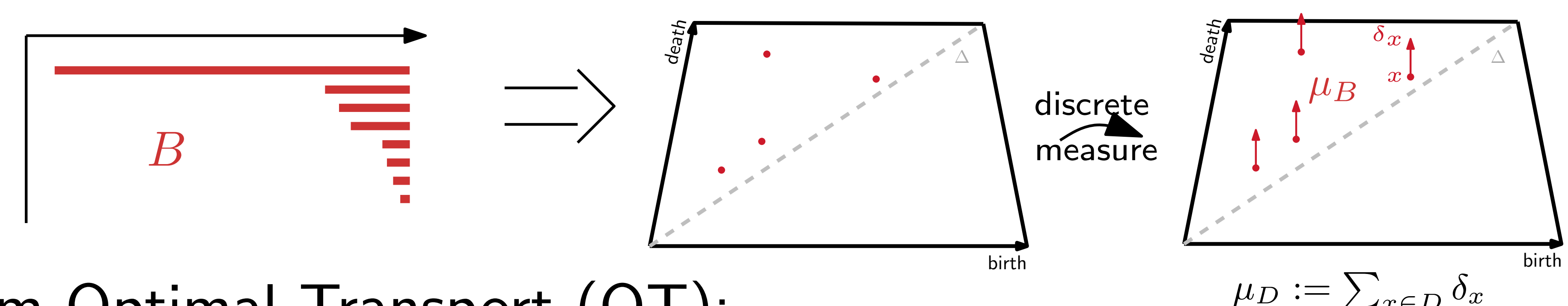
central limit theorems, confidence intervals, geodesic PCA,
clustering (k-means, EM, Mean-Shift, etc.)



OT Approach: recast problem in measure space

[Lacombe, Cuturi, O. 2018]

$B \mapsto \mu_B$



$\rightsquigarrow$ use relaxations from Optimal Transport (OT):

measures: $\mu_B \mapsto \mu_B * \mathcal{U}_{[0,\varepsilon]^2}$

[M. Agueh, G. Carlier: "Barycenters in the Wasserstein Space", 2011]

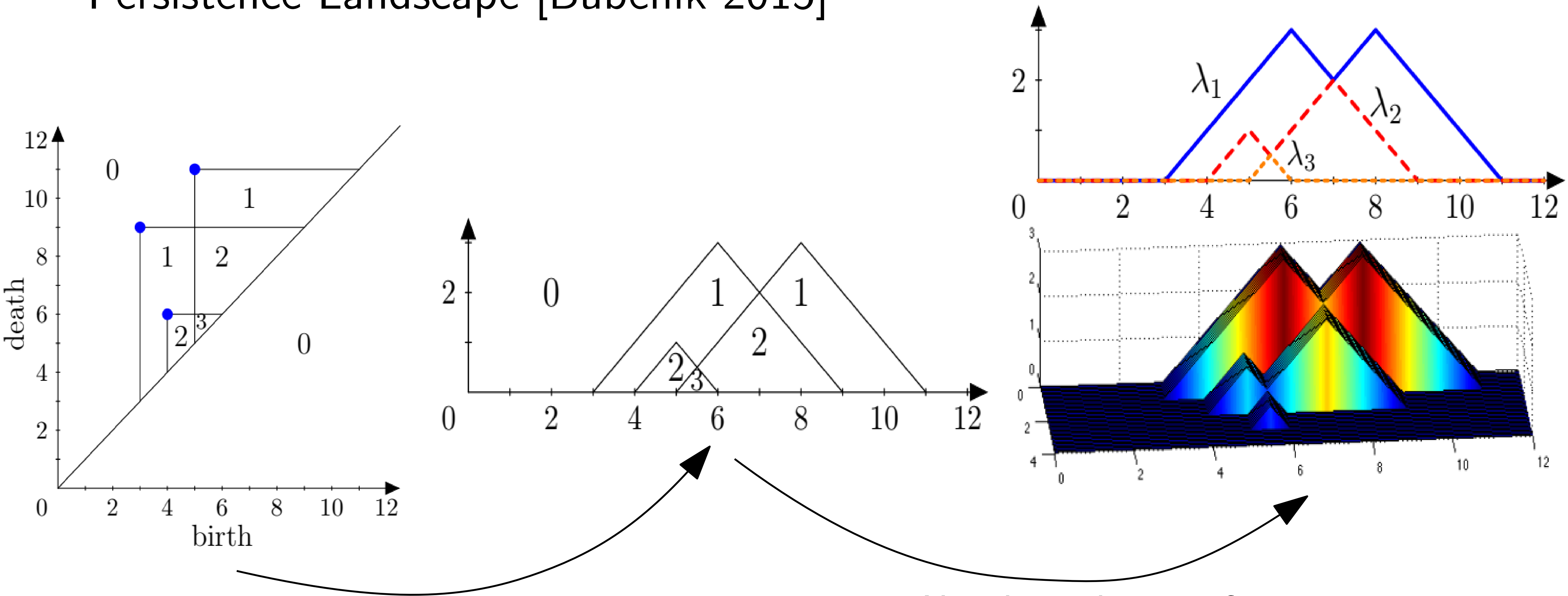
metric: $W_{2,\gamma}(\mu_{B_i}, \mu_{B_j})^2 := \inf_{\nu} \int \|x-y\|^2 d\nu(x,y) + \gamma\, H(\nu)$

[M. Cuturi, A. Doucet: "Fast computation of Wasserstein barycenters", 2014]

strictly convex problem
$\Rightarrow$ unique mean
easy to compute

2. Hilbert space embedding

Persistence Landscape [Bubenik 2015]

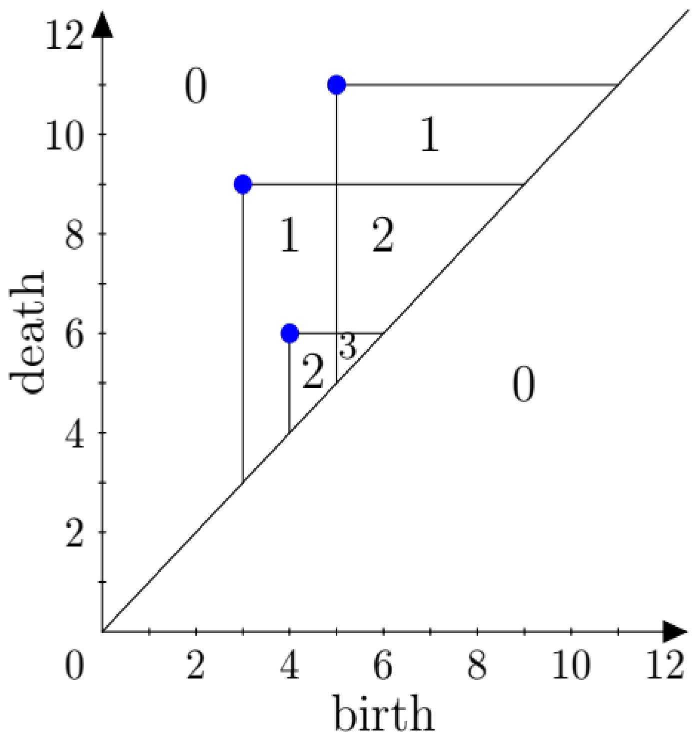


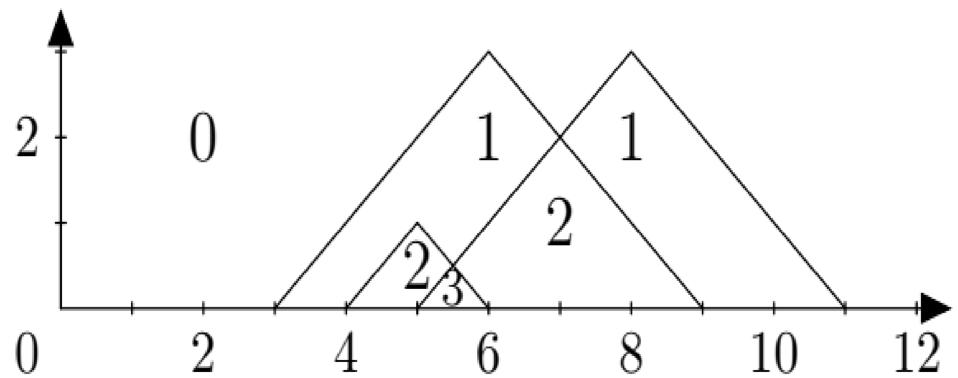
Rotate PD
Compute rank function

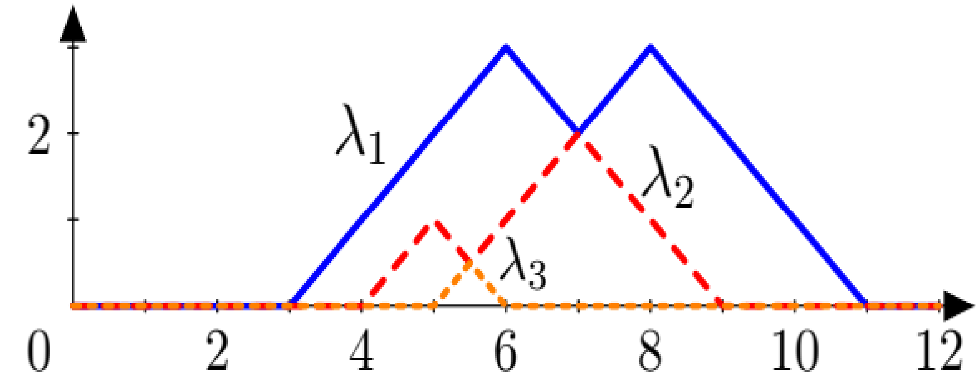
Use boundaries of
rank function

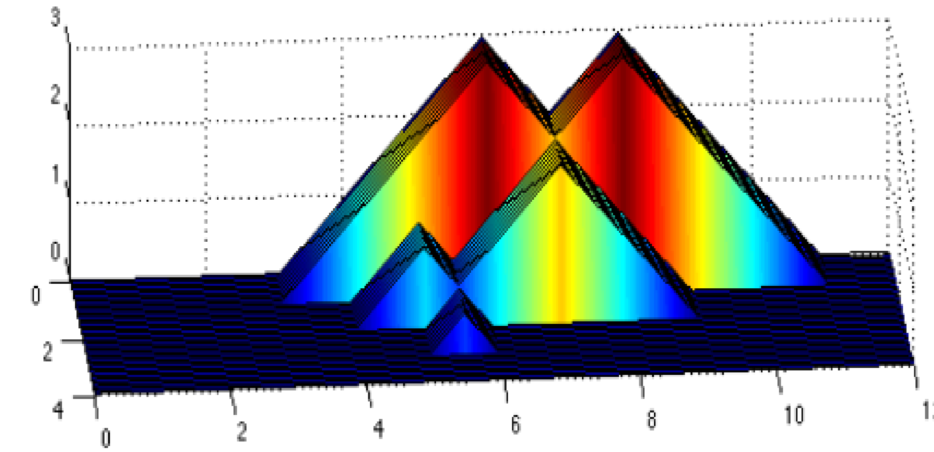
2. Hilbert space embedding

Persistence Landscape [Bubenik 2015]









Rotate PD
Compute rank function

Use boundaries of rank function

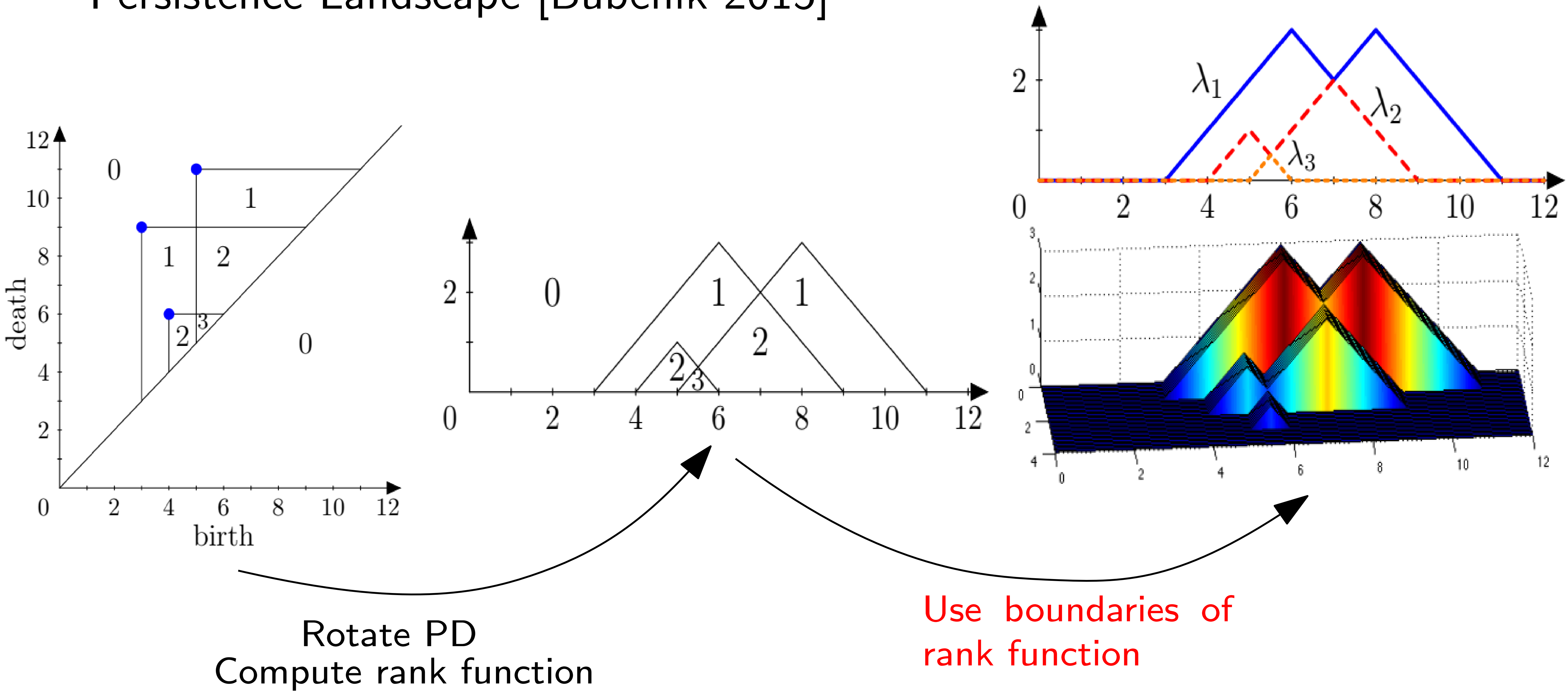
$$x \leq y \Longrightarrow f^{-1}(-\infty, x) \subseteq f^{-1}(-\infty, y)$$

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

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

2. Hilbert space embedding

Persistence Landscape [Bubenik 2015]



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: $\Lambda(i,t) = \lambda_{\lfloor i \rfloor}(t)$

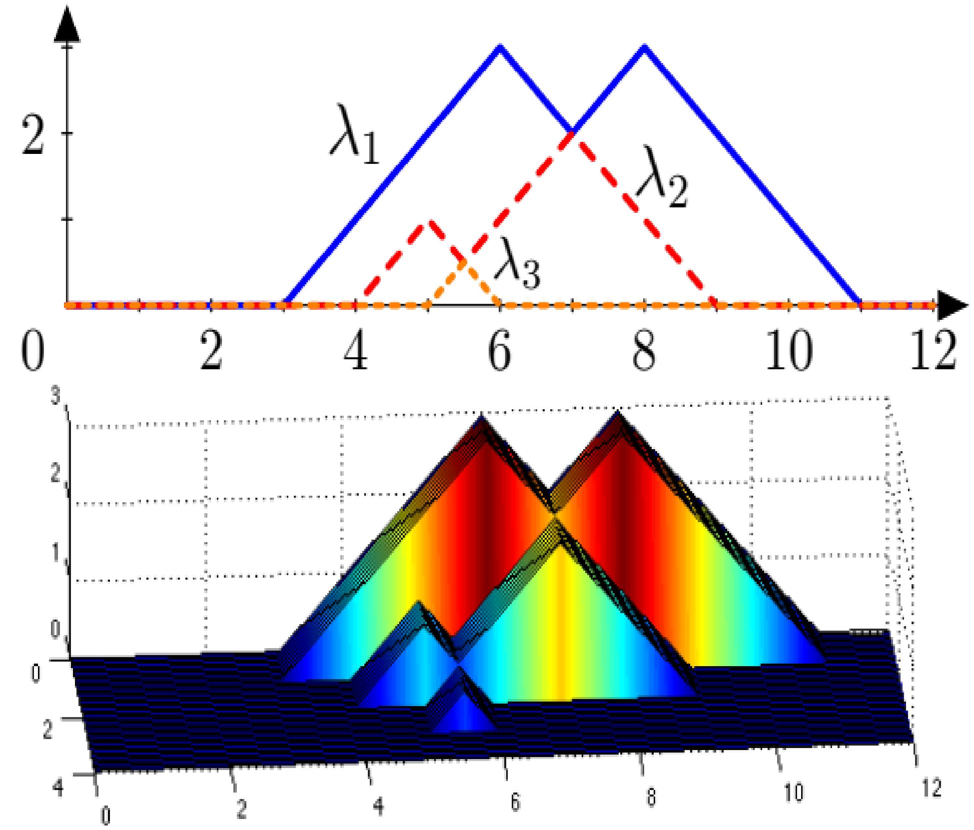
2. Hilbert space embedding

Persistence Landscape [Bubenik 2015]

Prop: [Bubenik 2015]

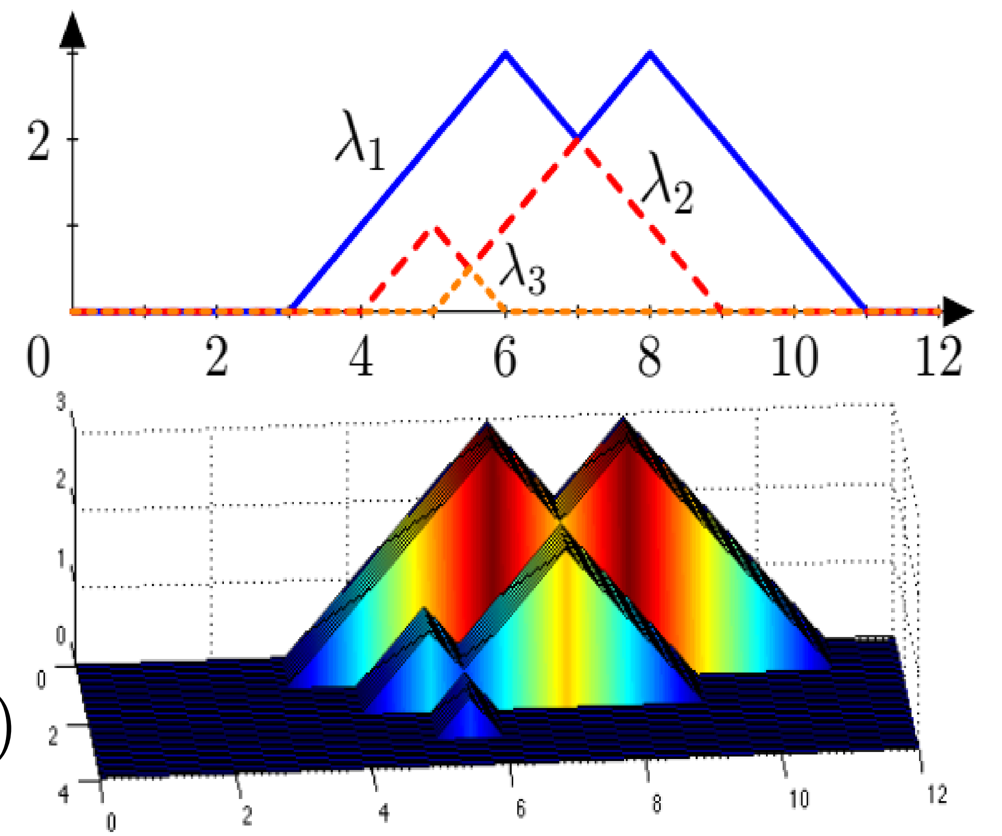
$$\|\Lambda(\mathrm{Dg}) - \Lambda(\mathrm{Dg}')\|_\infty \leq \mathrm{d}_\infty(\mathrm{Dg}, \mathrm{Dg}')$$

$\Longrightarrow$ Λ is Lipschitz hence Borel measurable



2. Hilbert space embedding

Persistence Landscape [Bubenik 2015]



Prop: [Bubenik 2015]

$\|\Lambda(\mathrm{Dg}) - \Lambda(\mathrm{Dg}')\|_\infty \leq \mathrm{d}_\infty(\mathrm{Dg}, \mathrm{Dg}')$

$\Longrightarrow$ Λ is Lipschitz hence Borel measurable

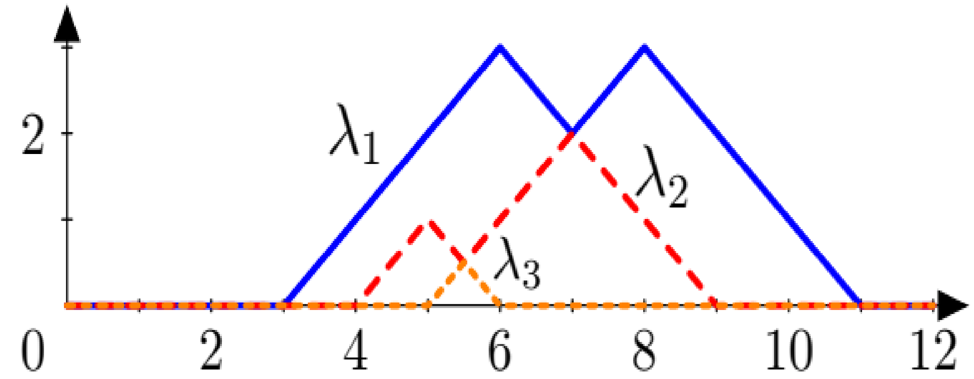
Given $D_1, \cdots, D_n \sim \mu$ iid, let $\bar{\Lambda}^n = \frac{1}{n}\sum_{i=1}^n \Lambda(D_i)$

Thm: (strong law of large numbers) [Bubenik 2015]

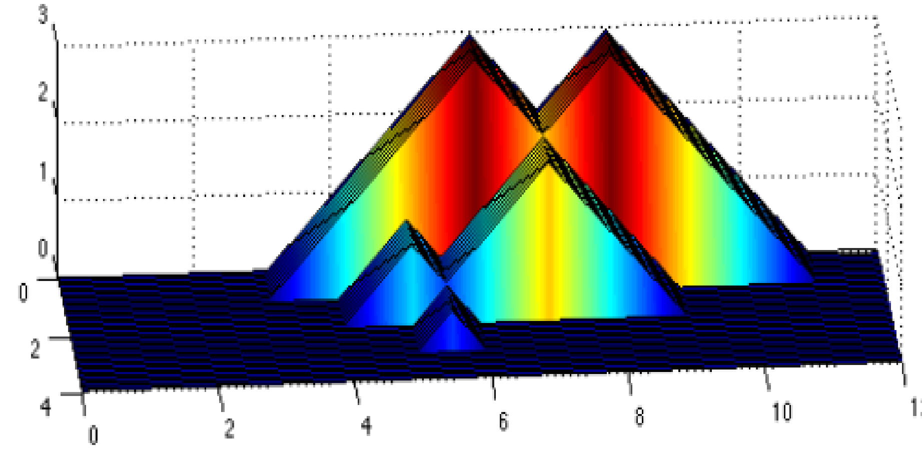
If $\mathbb{E}(\|\Lambda(\mu)\|) < +\infty$, then $\bar{\Lambda}^n \xrightarrow{\mathrm{a.s.}} \mathbb{E}(\Lambda(\mu))$.

2. Hilbert space embedding

Persistence Landscape [Bubenik 2015]



Prop: [Bubenik 2015]
$\|\Lambda(\mathrm{Dg}) - \Lambda(\mathrm{Dg}')\|_\infty \leq \mathrm{d}_\infty(\mathrm{Dg}, \mathrm{Dg}')$



$\Longrightarrow$ Λ is Lipschitz hence Borel measurable

Given $D_1, \cdots, D_n \sim \mu$ iid, let $\bar{\Lambda}^n = \frac{1}{n}\sum_{i=1}^n \Lambda(D_i)$

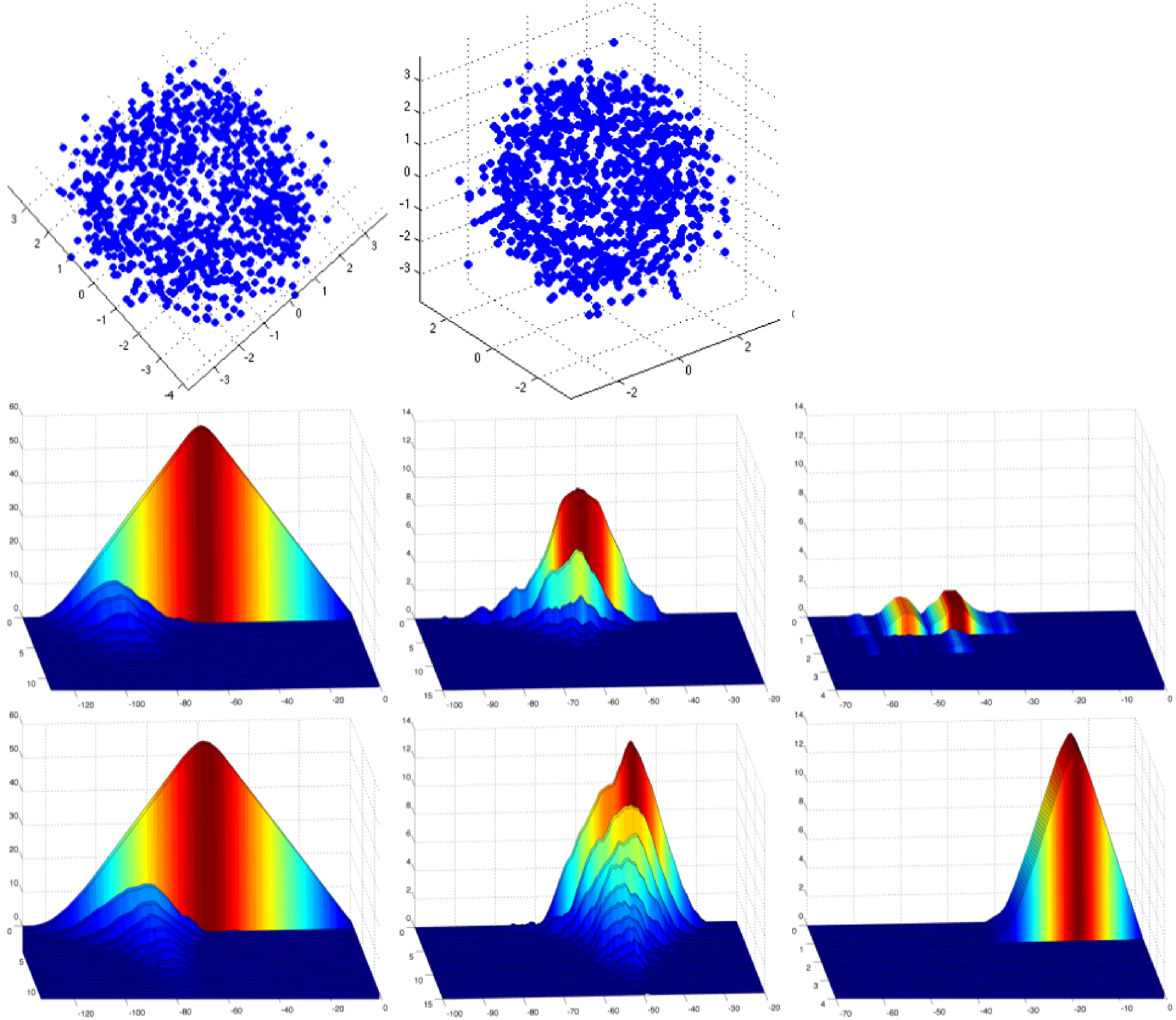
Thm: (strong law of large numbers) [Bubenik 2015]
If $\mathbb{E}(\|\Lambda(\mu)\|) < +\infty$, then $\bar{\Lambda}^n \xrightarrow{\mathrm{a.s.}} \mathbb{E}(\Lambda(\mu))$.

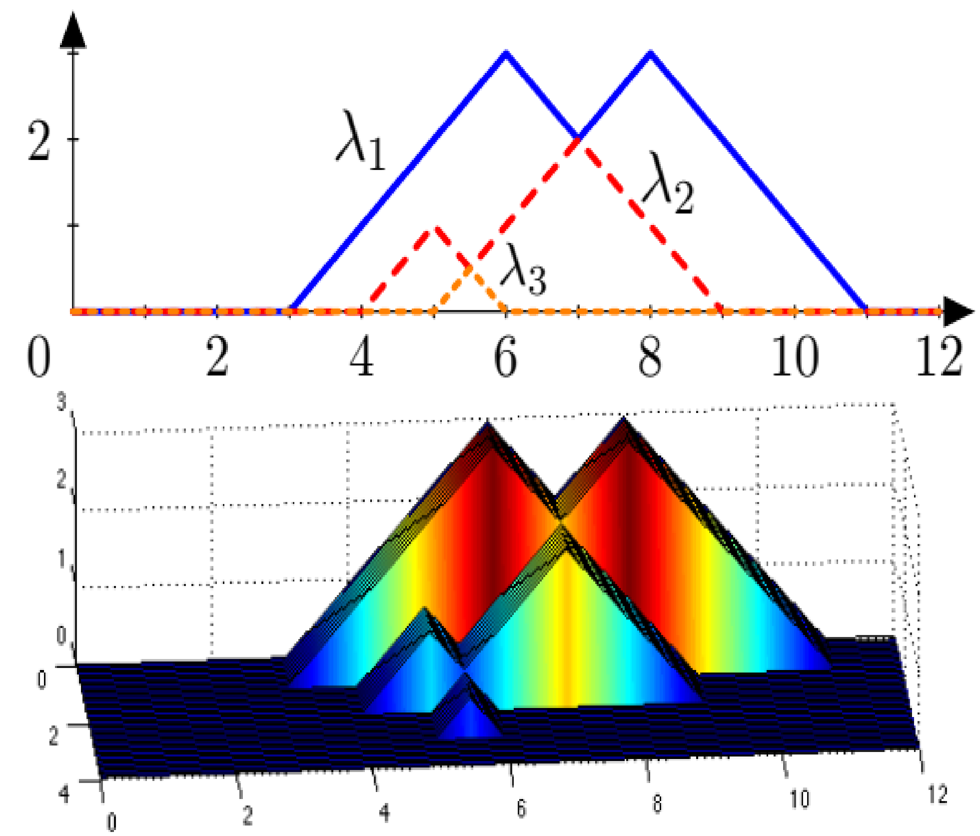
Thm: (central limit theorem) [Bubenik 2015]
If $\mathbb{E}(\|\Lambda(\mu)\|) < +\infty$ and $\mathbb{E}(\|\Lambda(\mu)\|^2) < +\infty$, then
$\sqrt{n}\left(\bar{\Lambda}^n - \mathbb{E}(\Lambda(\mu))\right) \xrightarrow{d} \mathcal{N}(0, \Sigma(\Lambda(\mu)))$.

2. Hilbert space embedding

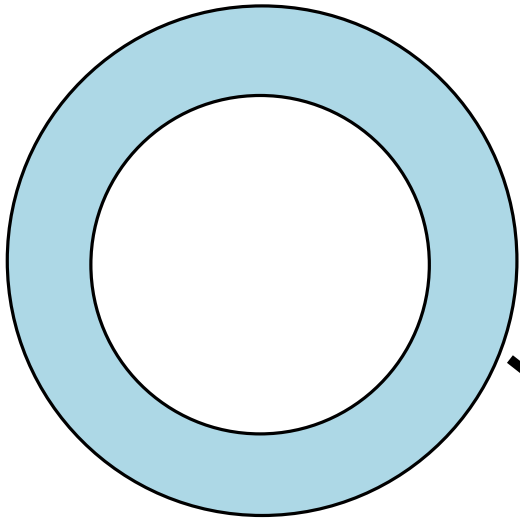
Persistence Landscape [Bubenik 2015]

Problem: mean landscape is not a landscape





3. Push-forwards from data space



(X, d_X) compact metric space

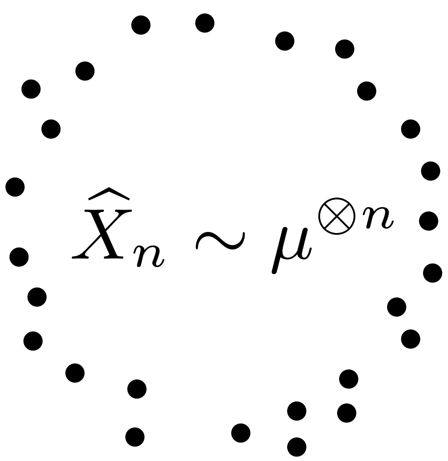
μ probability measure with $\operatorname{supp} \mu = X$

[Chazal et al.] [Wasserman et al.]

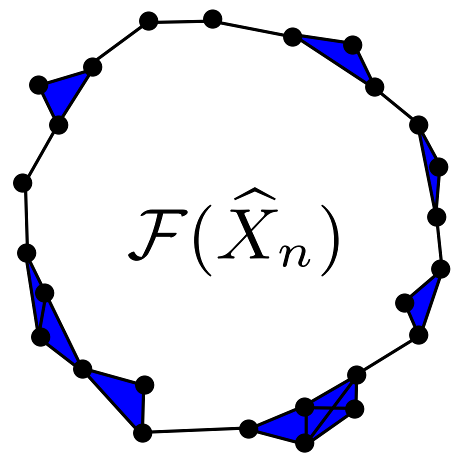
$\mathcal{F}(X)$

Sample n points iid according to μ.

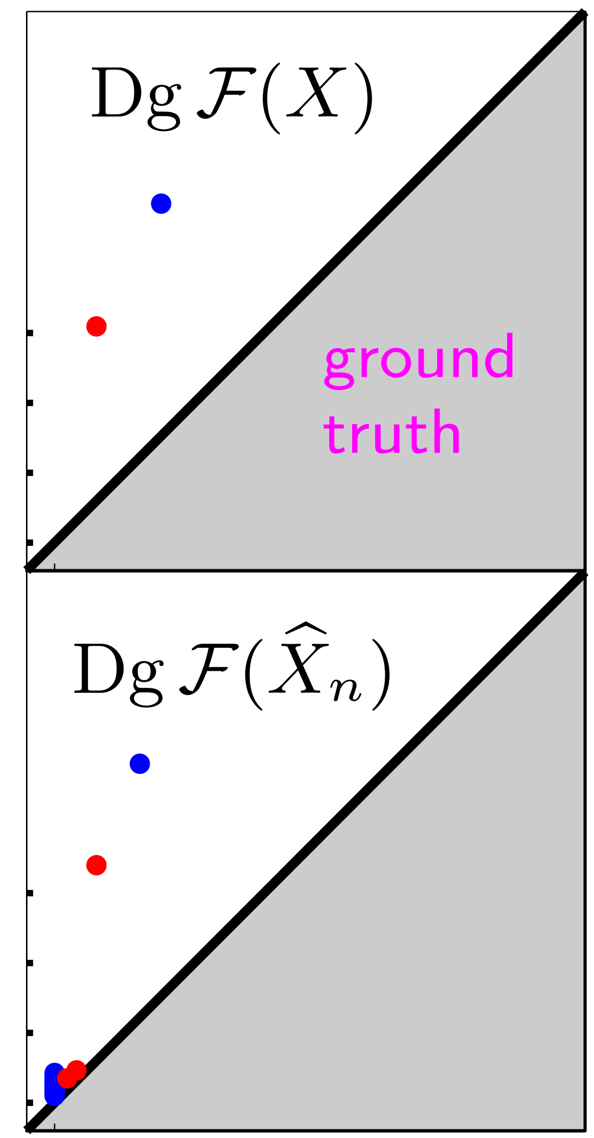
(Rips) filtration



$\widehat{X}_n \sim \mu^{\otimes n}$



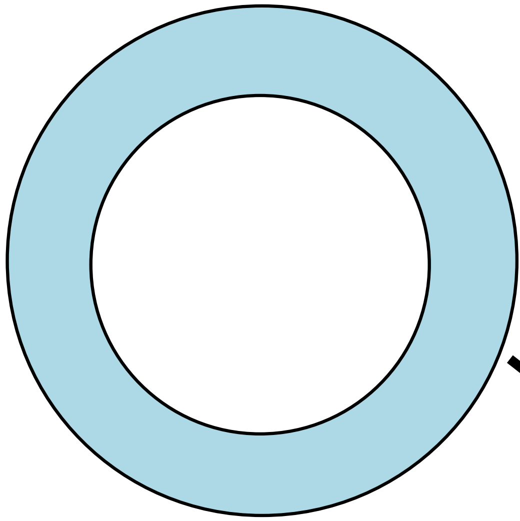
$\mathcal{F}(\widehat{X}_n)$



Questions:

- Statistical properties of the estimator $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$?
- Convergence to the ground truth $\mathrm{Dg}\,\mathcal{F}(X)$? Deviation bounds?

3. Push-forwards from data space



(X, d_X) compact metric space

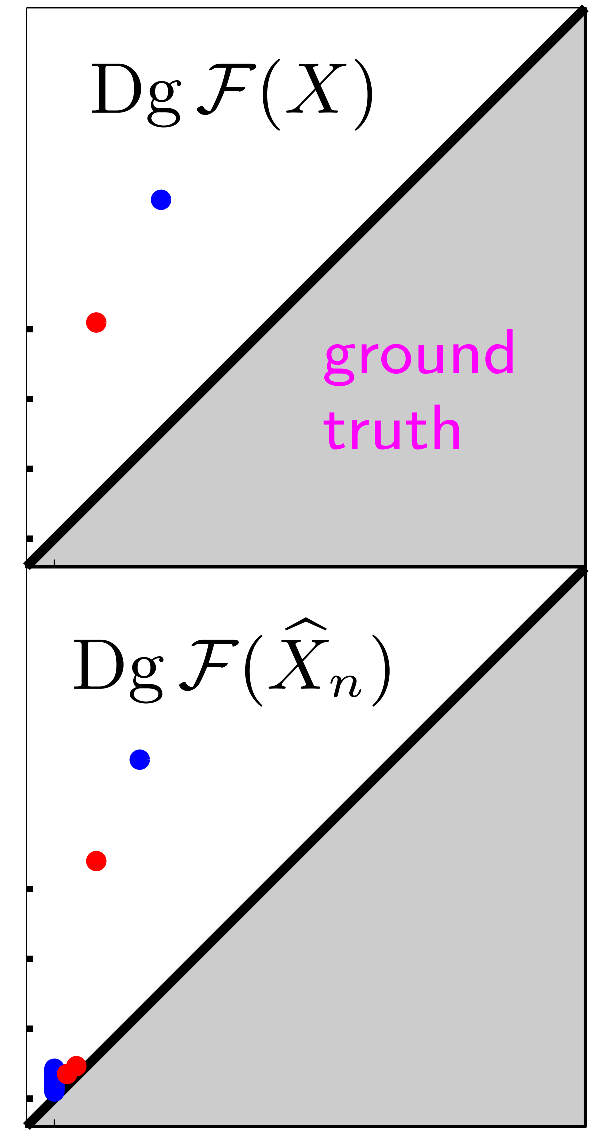
[Chazal et al.] [Wasserman et al.]

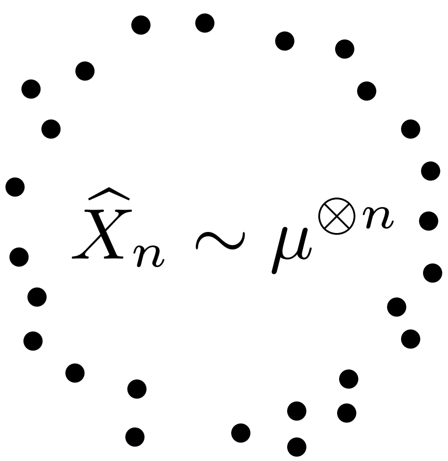
μ probability measure with $\operatorname{supp} \mu = X$

$\mathcal{F}(X)$

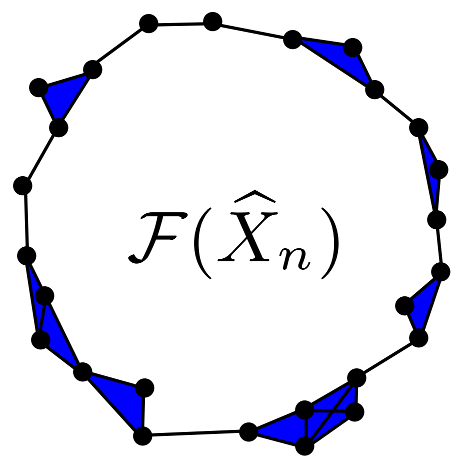
Sample n points iid according to μ.

(Rips) filtration





$\widehat{X}_n \sim \mu^{\otimes n}$



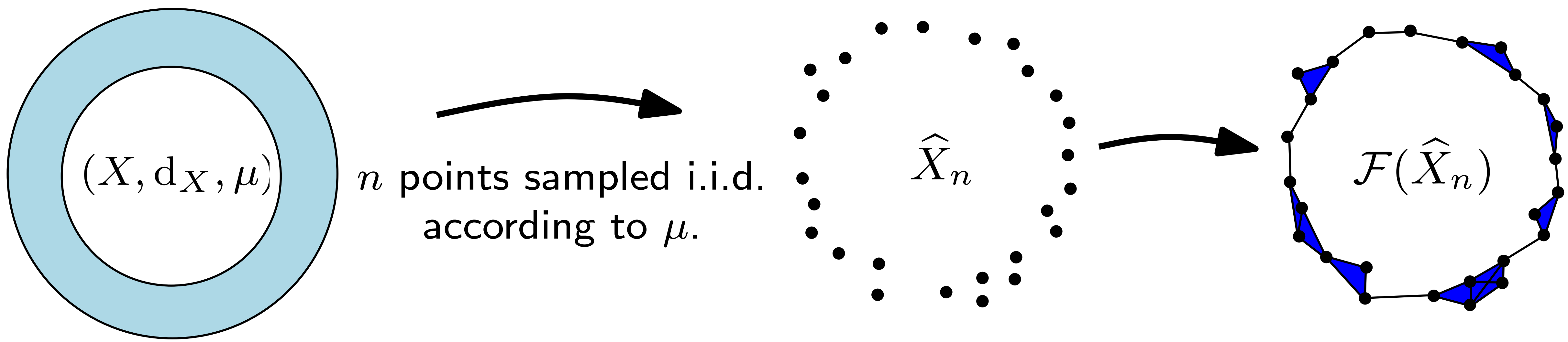
$\mathcal{F}(\widehat{X}_n)$

Stability thm: $\mathrm{d}_\infty(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)) \leq 2\mathrm{d}_\mathrm{H}(\widehat{X}_n, X)$

$\Rightarrow$ for any $\varepsilon > 0$,

$$\mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X),\right) > \varepsilon\right) \leq \mathbb{P}\left(\mathrm{d}_\mathrm{H}(\widehat{X}_n, X) > \frac{\varepsilon}{2}\right)$$

Deviation inequality



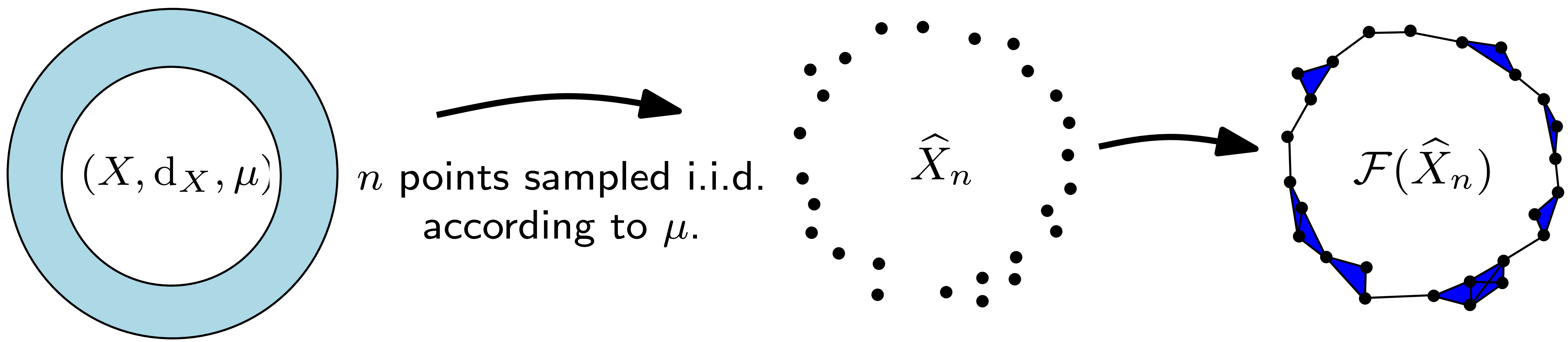
For $a, b > 0$, μ satisfies the (a,b)-**standard** assumption if for any $x \in X$ and any $r > 0$, we have $\mu(B(x,r)) \geq \min(ar^b, 1)$.

Theorem [Chazal, Glisse, Labruère, Michel 2014-15]:

If μ is (a,b)-standard then for any $\varepsilon > 0$:

$$\mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > \varepsilon\right) \leq \frac{8^b}{a\varepsilon^b}\exp(-na\varepsilon^b)$$

Deviation inequality / rate of convergence



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

Theorem [Chazal, Glisse, Labruère, Michel 2014-15]:

If μ is (a, b)-standard then for any $\varepsilon > 0$:

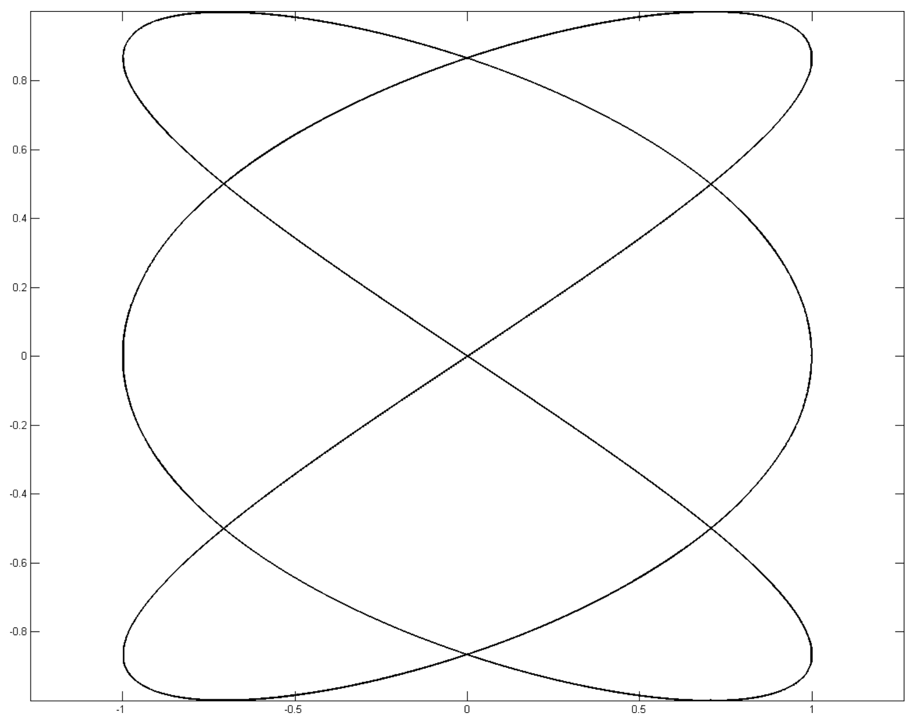
$$\mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > \varepsilon\right) \leq \frac{8^b}{a\varepsilon^b}\exp(-na\varepsilon^b)$$

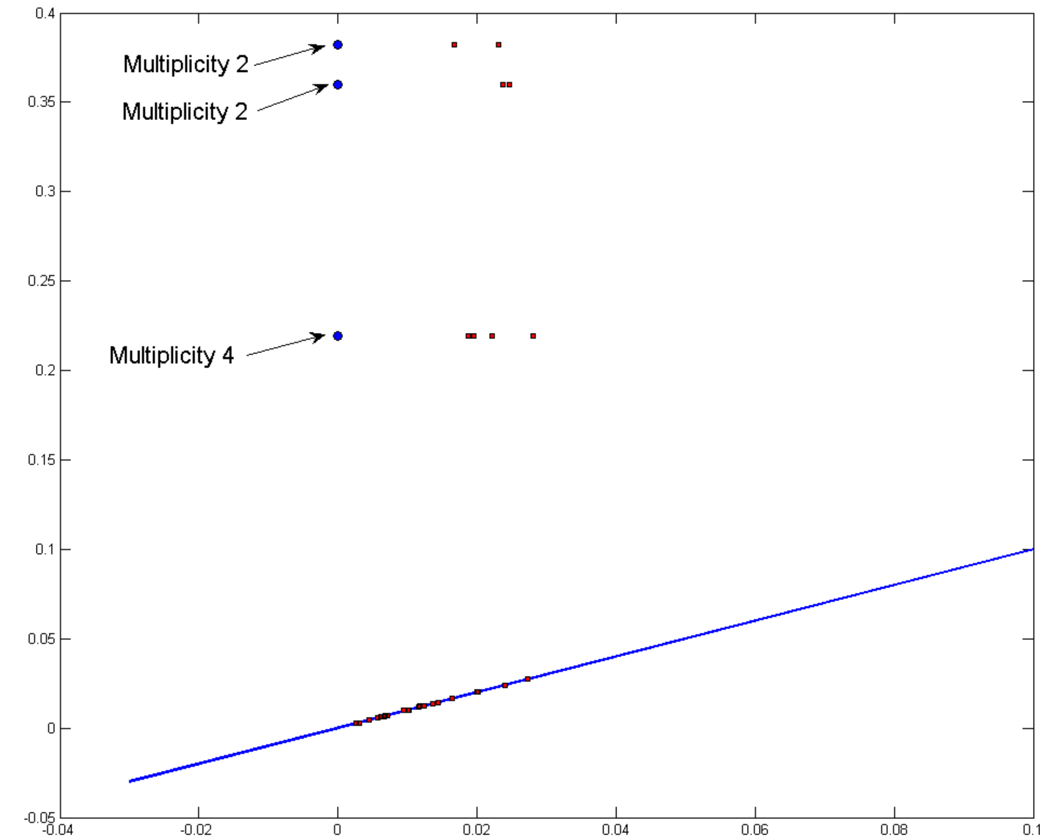
Corollary [Chazal, Glisse, Labruère, Michel 2014-15]:

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}\left[\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n),\ \mathrm{Dg}\,\mathcal{F}(X)\right)\right] \leq C\ \left(\frac{\log n}{n}\right)^{1/b},$$

where C depends only on a, b. Moreover, the estimator $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ is **minimax optimal** (up to a $\log n$ factor) on the space $\mathcal{P}$ of (a, b)-standard probability measures on X.

Numerical illustrations

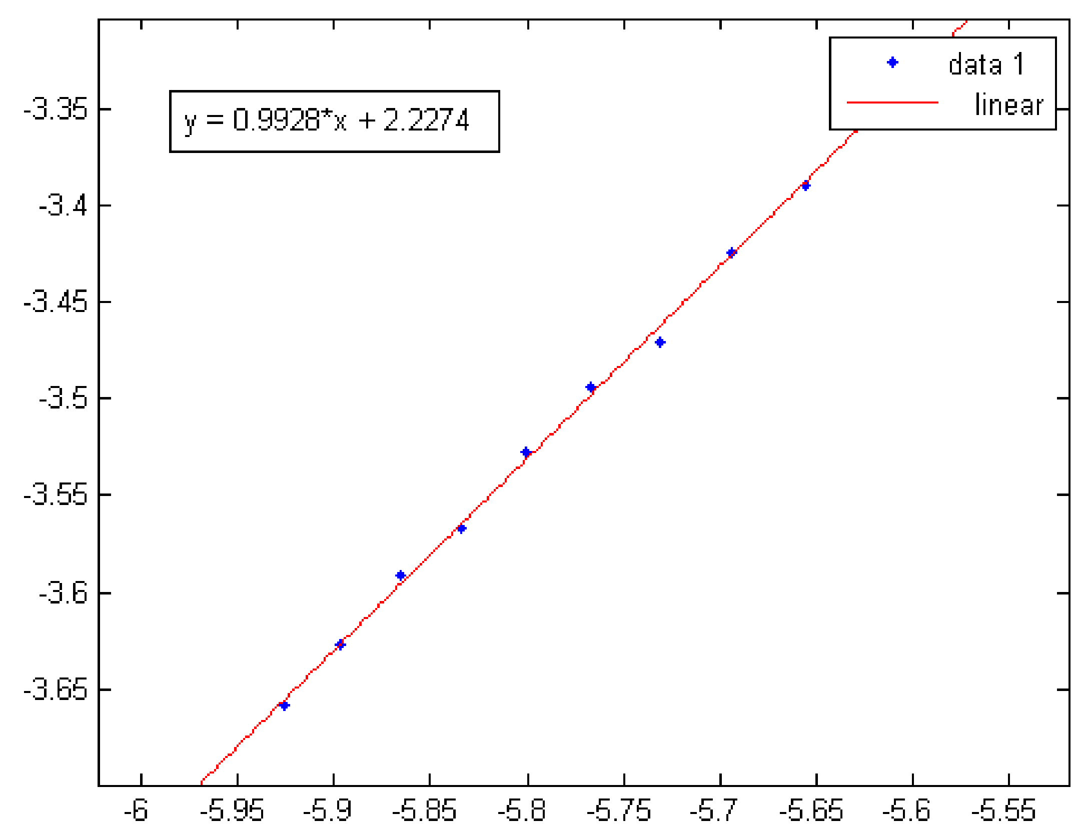




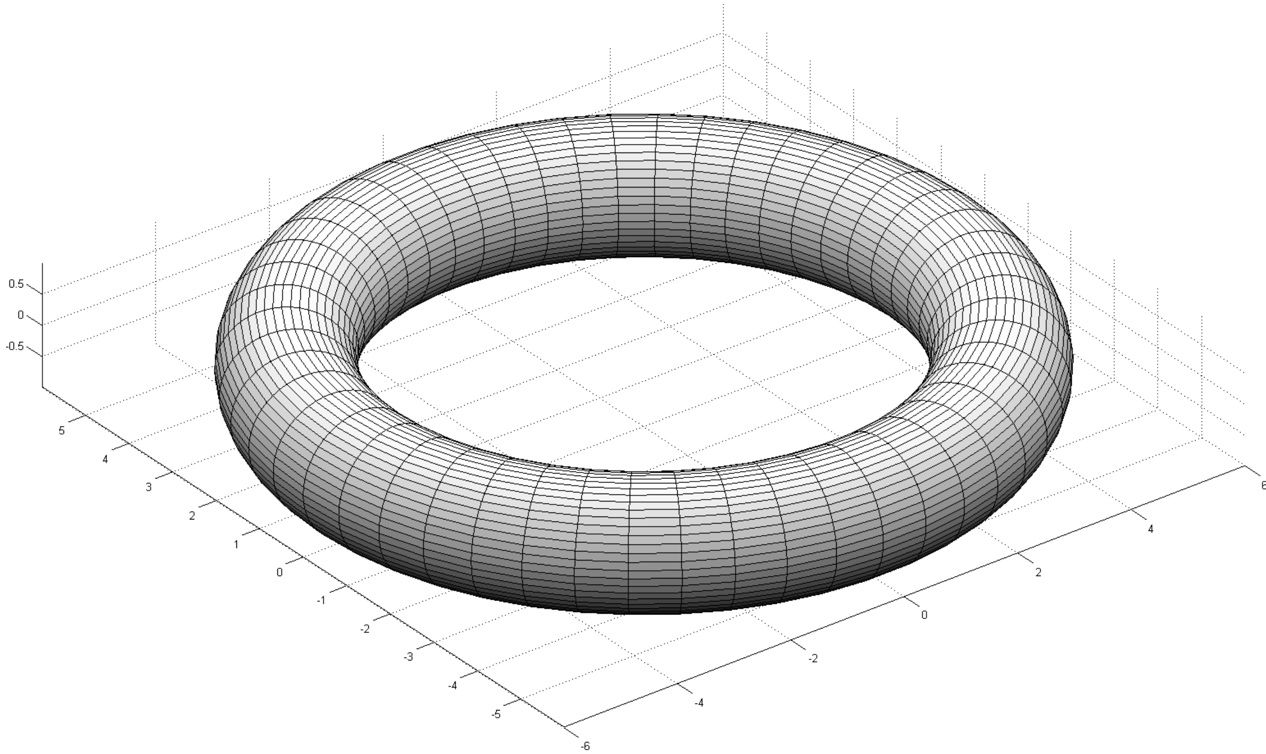
- μ: unif. measure on Lissajous curve X.
- $\mathcal{F}$: distance to X in $\mathbb{R}^2$.
- sample $k = 300$ sets of n points for $n = [2100 : 100 : 3000]$.
- compute

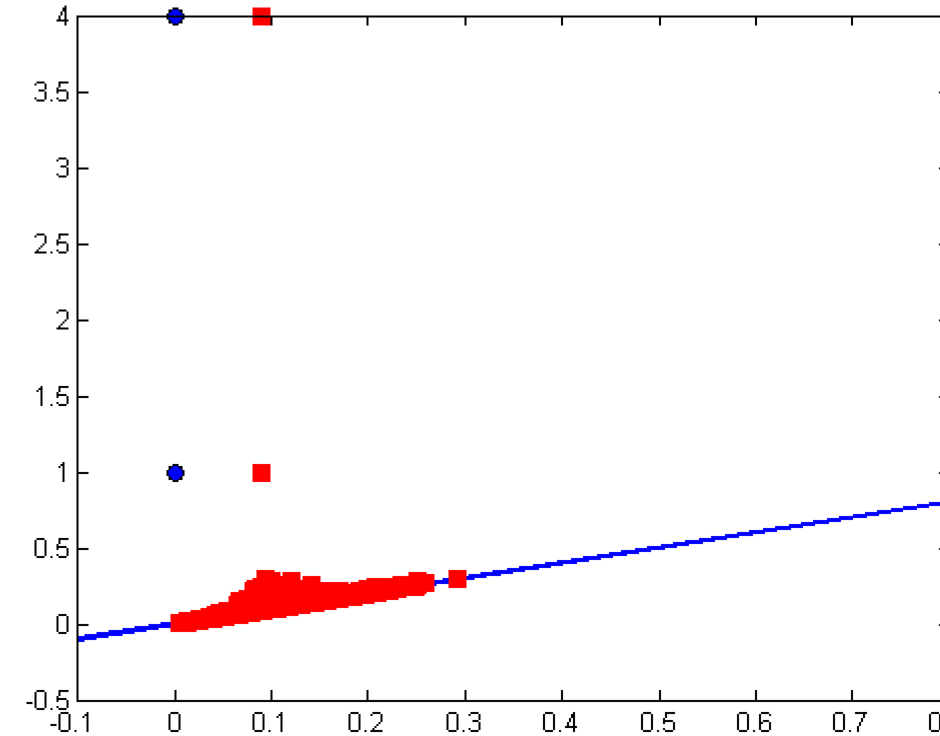
$$\widehat{\mathbb{E}}_n = \widehat{\mathbb{E}}[\mathrm{d}_\infty(\mathrm{Dg}\,\mathcal{F}(\widehat{X_n}), \mathrm{Dg}\,\mathcal{F}(X))].$$

- plot $\log(\widehat{\mathbb{E}}_n)$ as a function of $\log(\log(n)/n)$.



Numerical illustrations

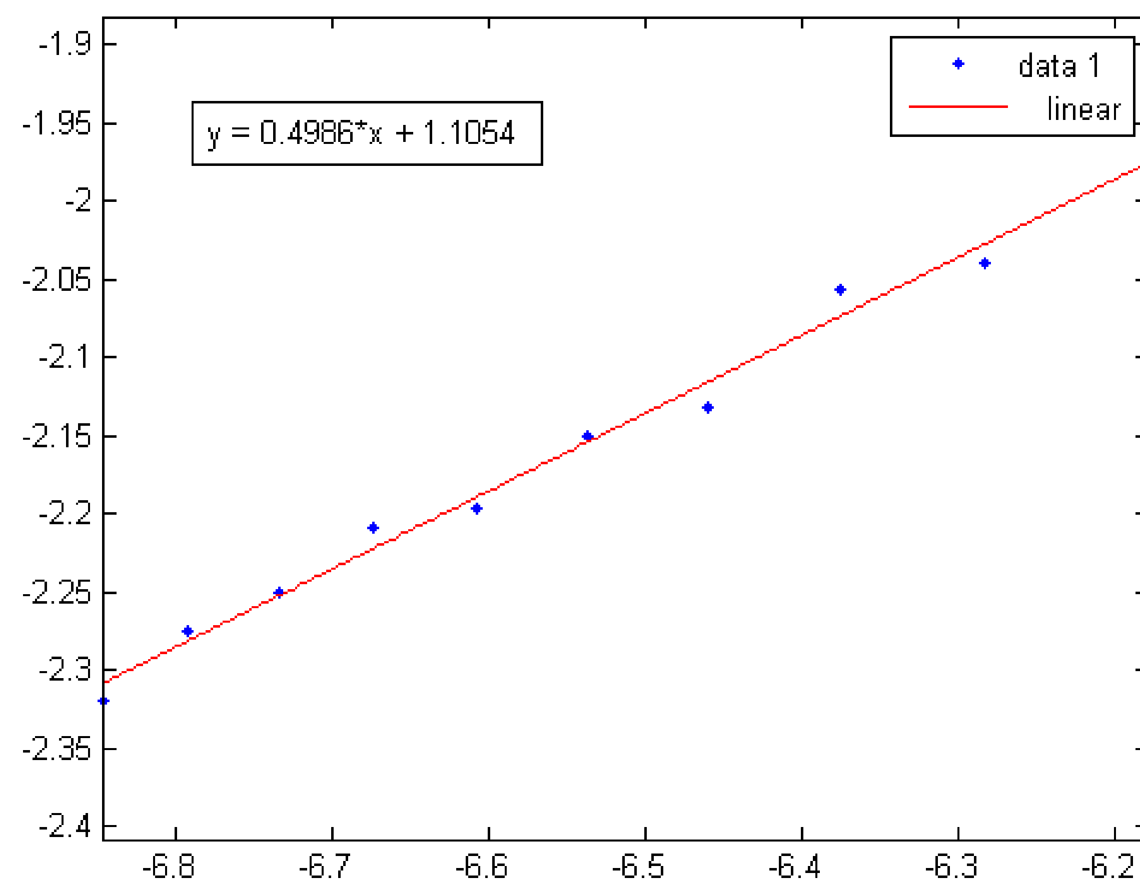




- μ: unif. measure on a torus X.
- $\mathcal{F}$: distance to X in $\mathbb{R}^3$.
- sample $k = 300$ sets of n points for $n = [12000 : 1000 : 21000]$.
- compute

$$\widehat{\mathbb{E}}_n = \widehat{\mathbb{E}}[\mathrm{d}_\infty(\mathrm{Dg}\,\mathcal{F}(\widehat{X_n}), \mathrm{Dg}\,\mathcal{F}(X))].$$

- plot $\log(\widehat{\mathbb{E}}_n)$ as a function of $\log(\log(n)/n)$.



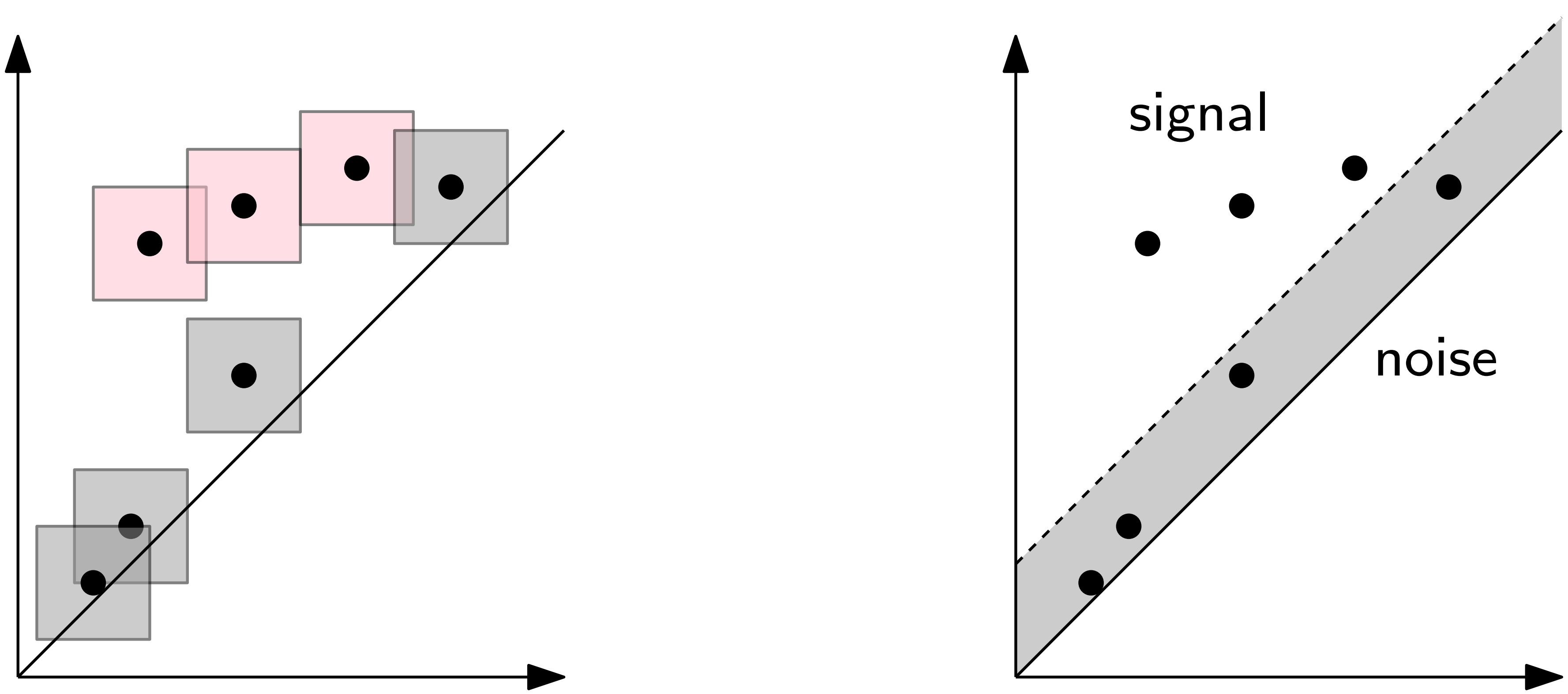
Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n \to \mathcal{F}(\widehat{X}_n) \to \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > c_n(\alpha)\right) \leq \alpha$$

$\to$ confidence region: d_∞-ball of radius $c_n(\alpha)$ around $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$



Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n \to \mathcal{F}(\widehat{X}_n) \to \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > c_n(\alpha)\right) \leq \alpha$$

Note: we already have an inequality of this kind but...

$$\mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > \varepsilon\right) \leq \frac{8^b}{a\varepsilon^b}\exp(-na\varepsilon^b)$$

unknown

Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n \to \mathcal{F}(\widehat{X}_n) \to \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > c_n(\alpha)\right) \leq \alpha$$

Bootstrap: (ideally)

- draw $X^* = X_1^*, \cdots, X_n^*$ iid from $\mu_{\widehat{X}_n}$ (empirical measure on $\widehat{X}_n$)
- compute $d^* = \mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(X^*), \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)\right)$
- repeat N times to get $d_1^*, \cdots, d_N^*$
- let q_α be the $(1-\alpha)$ quantile of $\frac{1}{N}\sum_{i=1}^N \delta_{d_i^*}$

Principle [Efron 1979]: variations of $\mathrm{Dg}\,\mathcal{F}(X^*)$ around $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ are same as variations of $\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$ around $\mathrm{Dg}\,\mathcal{F}(X)$.

Note: requires some conditions on (X, d_X, μ)...

Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n \to \mathcal{F}(\widehat{X}_n) \to \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > c_n(\alpha)\right) \leq \alpha$$

Bootstrap: (in fact)

- draw $X^* = X_1^*, \cdots, X_n^*$ iid from $\mu_{\widehat{X}_n}$ (empirical measure on $\widehat{X}_n$)
- compute $d^* = \cancel{\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(X^*), \mathrm{Dg}\,\mathcal{F}(\widehat{X}_n)\right)}$ $\mathrm{d_H}(X^*, \hat{X}_n)$
- repeat N times to get $d_1^*, \cdots, d_N^*$
- let q_α be the $(1-\alpha)$ quantile of $\frac{1}{N}\sum_{i=1}^N \delta_{d_i^*}$

Theorem [Balakrishnan et al. 2013] + [Chazal et al. 2014]:

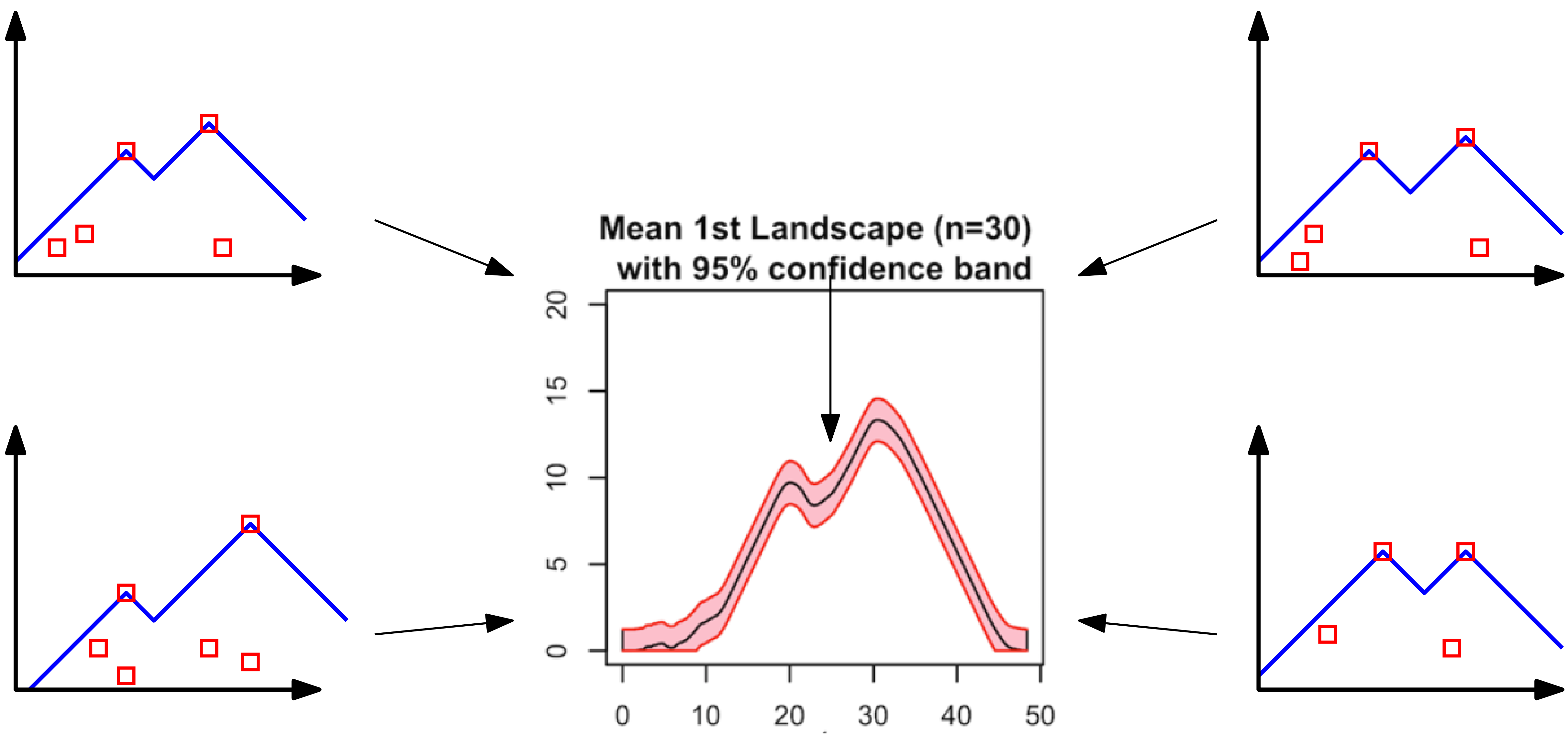
$$\limsup_{n\to\infty} \mathbb{P}\left(\mathrm{d}_\infty\left(\mathrm{Dg}\,\mathcal{F}(\widehat{X}_n), \mathrm{Dg}\,\mathcal{F}(X)\right) > q_\alpha\right) \leq \alpha.$$

Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n^1, \cdots, \widehat{X}_n^r \to \phi(D_n^1), \cdots, \phi(D_n^r)$

$$\downarrow$$

empirical mean feature vector $\longrightarrow$ $\bar{v} = \frac{1}{r}\sum_{i=1}^{r} \phi(D_n^i)$



Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n^1, \cdots, \widehat{X}_n^r \to \phi(D_n^1), \cdots, \phi(D_n^r)$

$$\downarrow$$

empirical mean feature vector $\longrightarrow$ $\bar{v} = \frac{1}{r} \sum_{i=1}^r \phi(D_n^i)$

Goal: given $\alpha \in (0,1)$, estimate $c_n(\alpha) \geq 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\left\| \bar{v} - \underbrace{\mathbb{E}_{(\phi \circ \mathrm{Dg} \circ \mathcal{F})^*(\mu^{\otimes n})}[v]}_{} \right\|_{\mathcal{H}} > c_n(\alpha) \right) \leq \alpha$$

mean feature vector according to the measure induced by $\mu^{\otimes n}$

(call it $\Lambda_{\mu,n}$ for landscapes)

Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n^1, \cdots, \widehat{X}_n^r \to \Lambda(D_n^1), \cdots, \Lambda(D_n^r)$

$$\downarrow$$

$$\bar{\Lambda} = \frac{1}{r} \sum_{i=1}^{r} \Lambda(D_n^i)$$

Bootstrap with landscapes:

- draw $\Lambda_1^*, \cdots, \Lambda_r^*$ iid from $\frac{1}{r} \sum_{i=1}^{r} \delta_{\Lambda(D_n^i)}$
- compute $\bar{\Lambda}^* = \frac{1}{r} \sum_{i=1}^{r} \Lambda_i^*$ and $d^* = \|\bar{\Lambda}^* - \bar{\Lambda}\|_\infty$
- repeat N times to get $d_1^*, \cdots, d_N^*$
- let q_α be the $(1 - \alpha)$ quantile of $\frac{1}{N} \sum_{i=1}^{N} \delta_{d_i^*}$

Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n^1, \cdots, \widehat{X}_n^r \to \Lambda(D_n^1), \cdots, \Lambda(D_n^r)$

$$\downarrow$$

$$\bar{\Lambda} = \frac{1}{r}\sum_{i=1}^{r} \Lambda(D_n^i)$$

Bootstrap with landscapes:

- draw $\Lambda_1^*, \cdots, \Lambda_r^*$ iid from $\frac{1}{r}\sum_{i=1}^{r} \delta_{\Lambda(D_n^i)}$
- compute $\bar{\Lambda}^* = \frac{1}{r}\sum_{i=1}^{r} \Lambda_i^*$ and $d^* = \|\bar{\Lambda}^* - \bar{\Lambda}\|_\infty$
- repeat N times to get $d_1^*, \cdots, d_N^*$
- let q_α be the $(1-\alpha)$ quantile of $\frac{1}{N}\sum_{i=1}^{N} \delta_{d_i^*}$

Theorem [Chazal et al. 2014]:

$$\limsup_{r\to\infty} \mathbb{P}\left(\left\|\bar{\Lambda} - \Lambda_{\mu,n}\right\|_\infty > q_\alpha\right) \leq \alpha.$$

Confidence regions

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_n^1, \cdots, \widehat{X}_n^r \to \Lambda(D_n^1), \cdots, \Lambda(D_n^r)$

$$\downarrow$$

$$\bar{\Lambda} = \frac{1}{r}\sum_{i=1}^r \Lambda(D_n^i)$$

Bootstrap with landscapes:

- draw $\Lambda_1^*, \cdots, \Lambda_r^*$ iid from $\frac{1}{r}\sum_{i=1}^r \delta_{\Lambda(D_n^i)}$
- compute $\bar{\Lambda}^* = \frac{1}{r}\sum_{i=1}^r \Lambda_i^*$ and $d^* = \|\bar{\Lambda}^* - \bar{\Lambda}\|_\infty$
- repeat N times to get $d_1^*, \cdots, d_N^*$
- let q_α be the $(1-\alpha)$ quantile of $\frac{1}{N}\sum_{i=1}^N \delta_{d_i^*}$

Theorem [Chazal et al. 2015]:

$$\left\|\bar{\Lambda} - \Lambda(\mathrm{Dg}\,\mathcal{F}(X))\right\|_\infty \leq \underbrace{\left\|\bar{\Lambda} - \Lambda_{\mu,n}\right\|_\infty}_{\text{variance term}} + \underbrace{\|\Lambda_{\mu,n} - \Lambda(\mathrm{Dg}\,\mathcal{F}(X))\|_\infty}_{\text{bias term } \leq C\left(\frac{\log n}{an}\right)^{1/b}}$$

variance term

bias term $\leq C\left(\frac{\log n}{an}\right)^{1/b}$ when μ is (a,b)-standard

Subsampling

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_m$ with m large (e.g. $m \geq 10^6$ or 10^9 or 10^{12})

Subsampling

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_m$ with m large (e.g. $m \geq 10^6$ or 10^9 or 10^{12})

Subsampling with landscapes: Let $n \ll m$

- draw X^* from $\mu_{\widehat{X}_m}^{\otimes n}$ (n points iid from empirical measure on $\widehat{X}_m$)
- compute $\Lambda^* = \Lambda(\mathrm{Dg}\,\mathcal{F}(X^*))$
- repeat N times to get $\Lambda_1^*, \cdots, \Lambda_N^*$
- compute $\bar{\Lambda}^* = \frac{1}{N}\sum_{i=1}^{N} \Lambda_i^*$

Subsampling

Setup: $(X, \mathrm{d}_X, \mu) \to \widehat{X}_m$ with m large (e.g. $m \geq 10^6$ or 10^9 or 10^{12})

Subsampling with landscapes: Let $n \ll m$

- draw X^* from $\mu_{\widehat{X}_m}^{\otimes n}$ (n points iid from empirical measure on $\widehat{X}_m$)
- compute $\Lambda^* = \Lambda(\mathrm{Dg}\,\mathcal{F}(X^*))$
- repeat N times to get $\Lambda_1^*, \cdots, \Lambda_N^*$
- compute $\bar{\Lambda}^* = \frac{1}{N}\sum_{i=1}^N \Lambda_i^*$

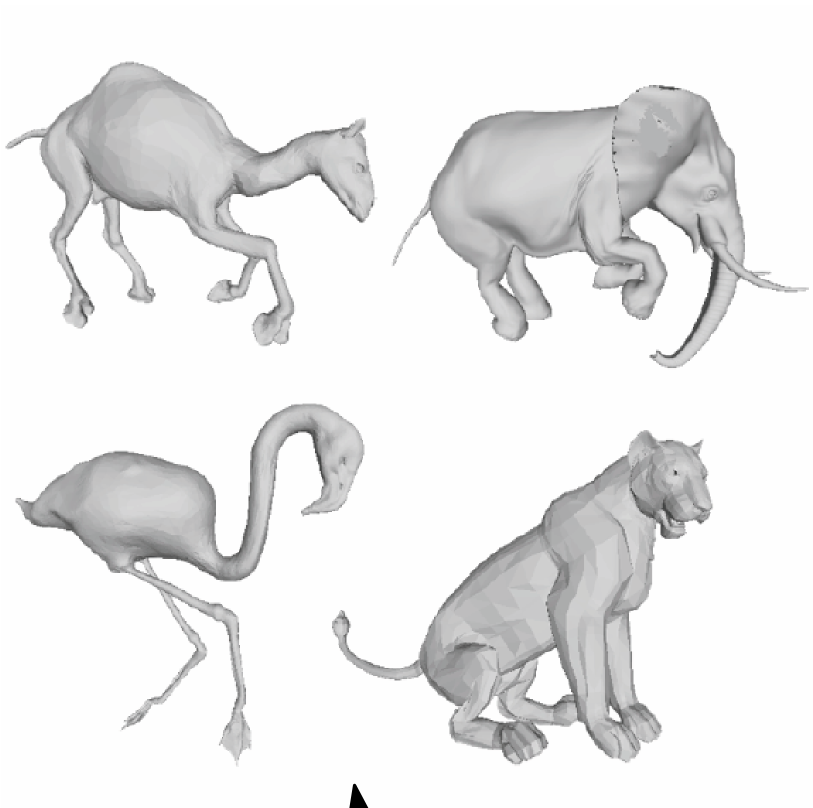
Theorem [Chazal et al. 2015]:

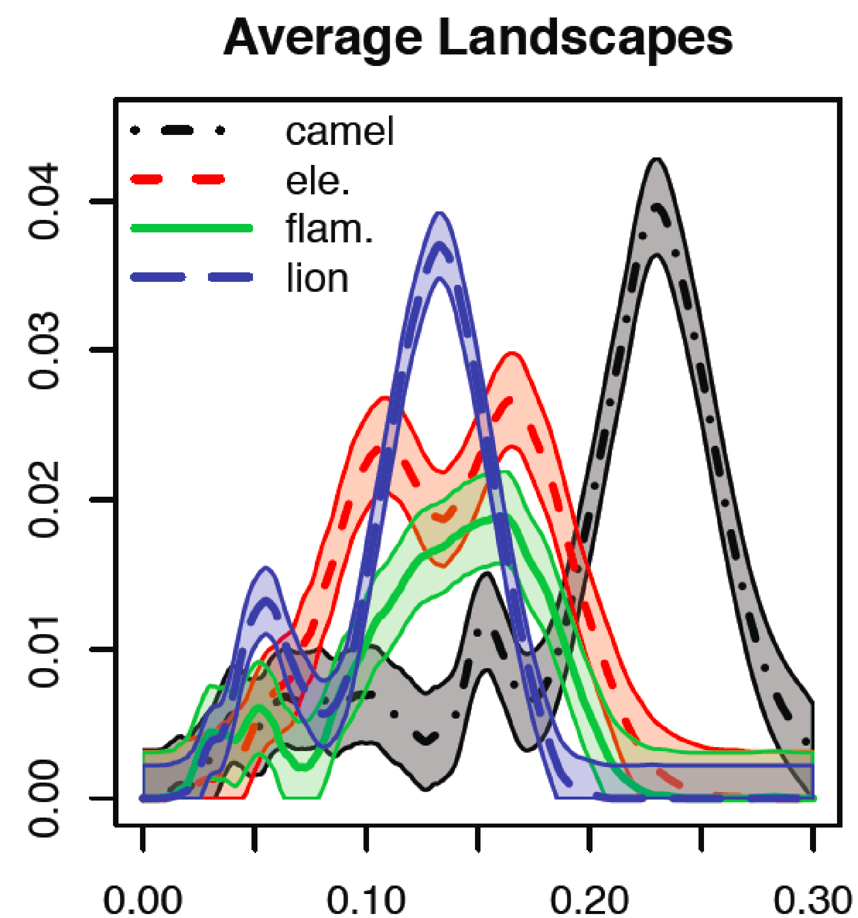
$$\left\| \Lambda_{\mu_{\widehat{X}_m}, n} - \Lambda_{\mu, n} \right\|_\infty \leq n^{1/p}\, W_p(\mu_{\widehat{X}_m}, \mu)$$

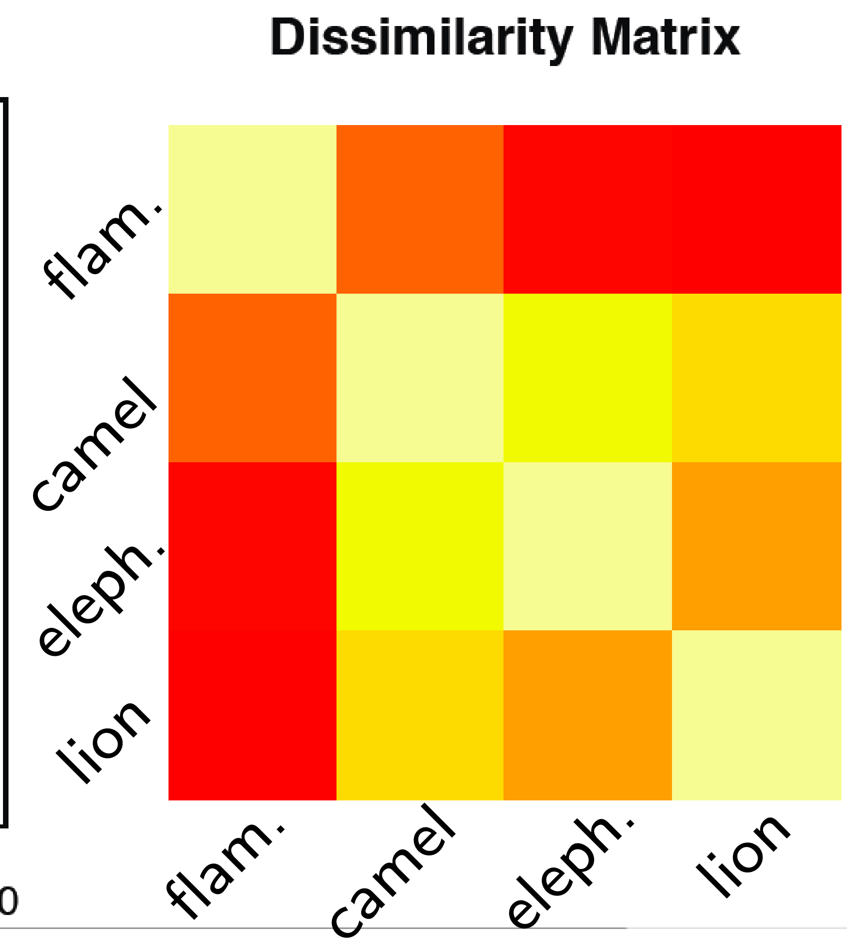
$\to$ by approximating $\Lambda_{\mu_{\widehat{X}_m}, n}$, the empirical mean $\bar{\Lambda}^*$ also approximates $\Lambda_{\mu, n}$

Some applications

Application 1: 3D shapes classification





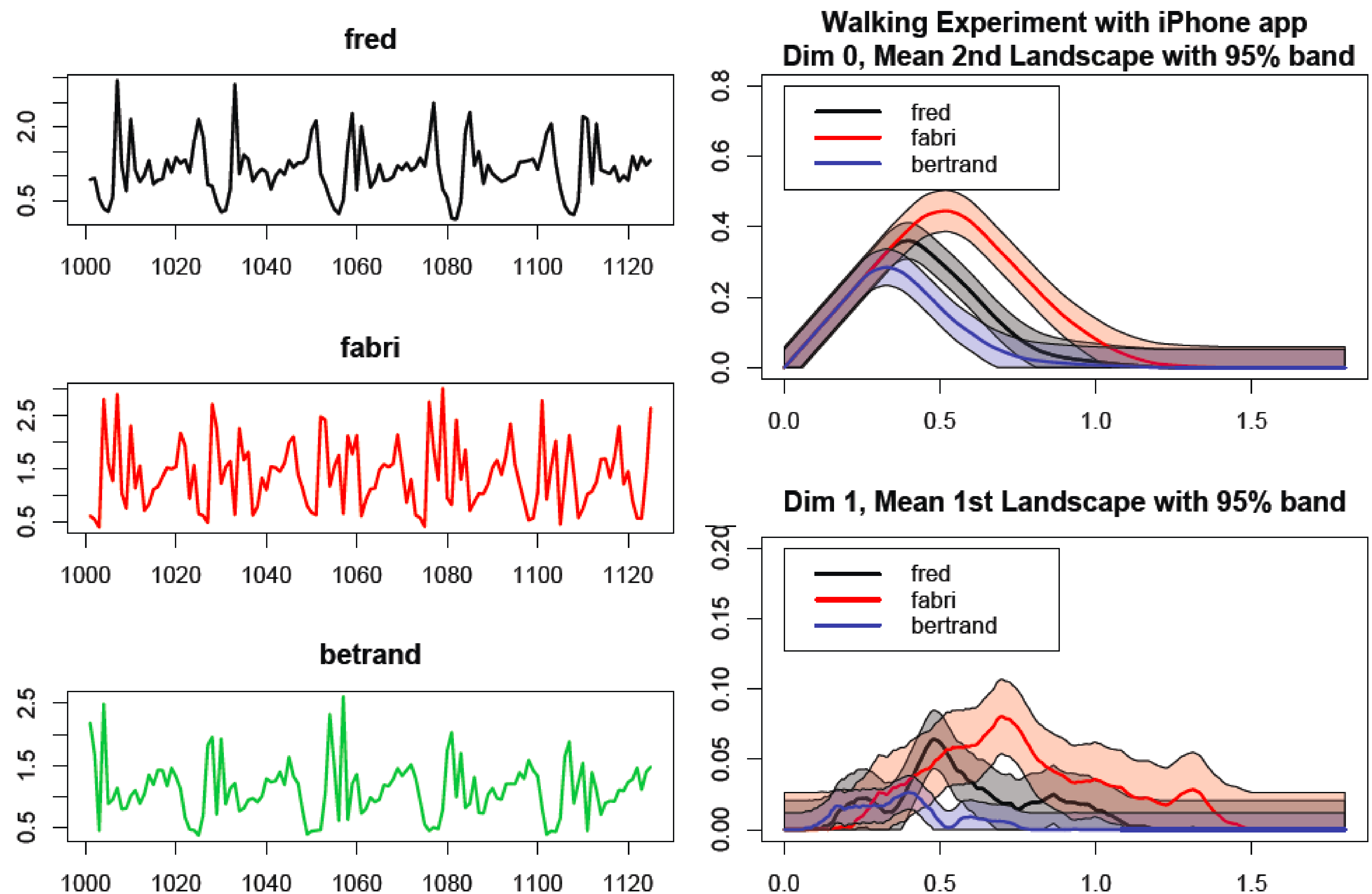


each mesh has 7K to 40K vertices

From $N = 100$ subsamples of size $m = 300$

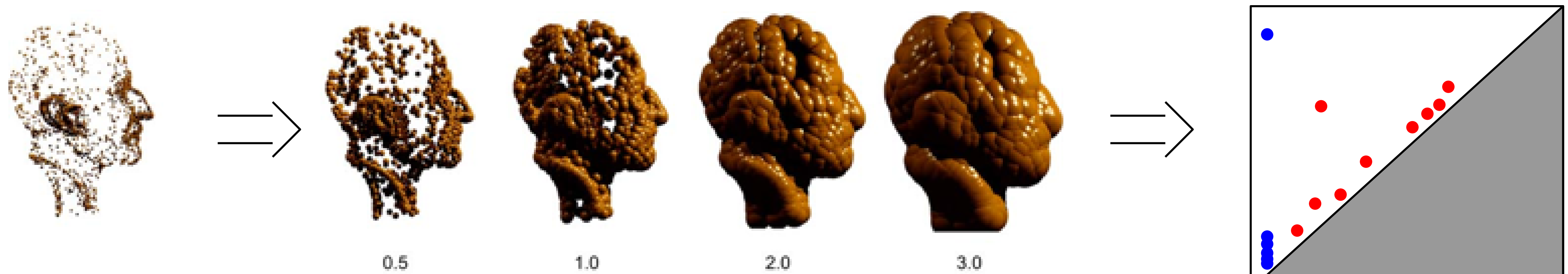
Some applications

Application 2: walking behaviors classification from smartphone accelerometer data

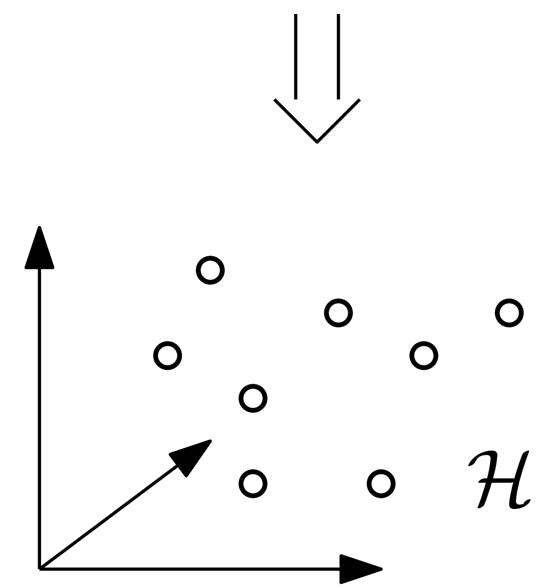


- spatial time series (accelerometer data from the smarphone of users).
- no registration/calibration preprocessing step needed to compare!

Recap'



dataset filtration persistence diagram



- statistical analysis based on stability theorem(s):
  - cvgence rates
  - confidence regions (bootstrap, subsampling)
  - stats. on diagrams (Fréchet means [Turner et al. 2012])
  - stats. on feature vectors (landscapes)