TDA - fall 2025

bottleneck distance & stability

Recap

- Welcome back!
  - ↳ How did you like AARTN talks?

- Next HW: a "paper chase"
  - Intention:
    - exposure to topics you like within TDA
    - practice reviewing
    - explore topics for final project later on

The workflow so far

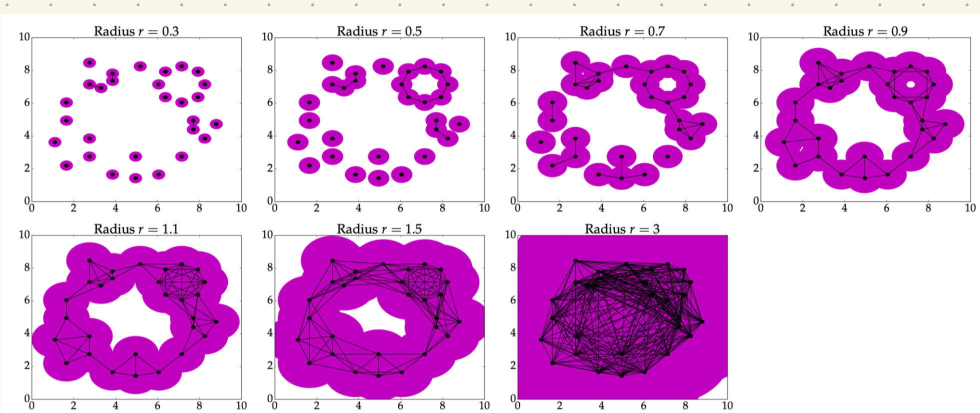
Build a filtration F from a simplicial complex

↳ usually parameterized by a function f, via sublevel sets

Example: $P \subseteq \mathbb{R}^n$

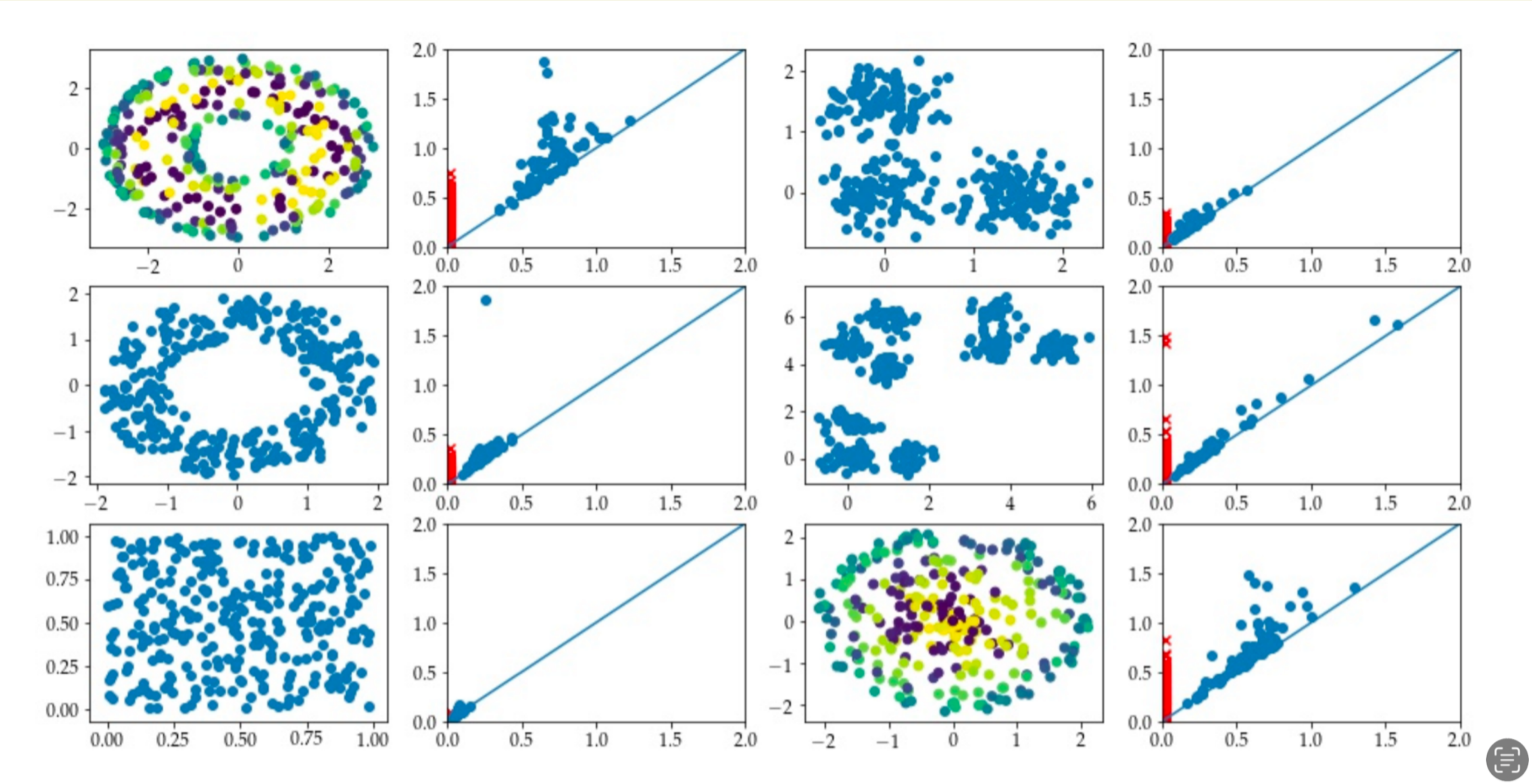
Build Rips filtration: $K_i = VR(P, r_i)$

for $0 \le r_1 \le \cdots \le r_k$,



$\Rightarrow$ Persistence diagram

Result: H_0 and H_1



... now what?

Distance measures

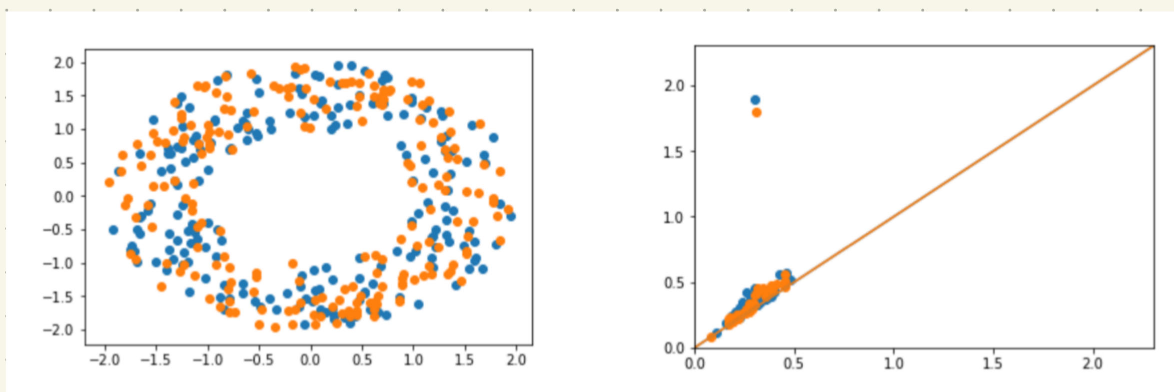
A distance on a set X is a function

$d: X \times X \to \mathbb{R}_{\geq 0}$ s.t. $\forall x, y, z \in X$

- $d(x,y) \geq 0$ & $d(x,y) = 0 \iff x = y$
- $d(x,y) = d(y,x)$
- $d(x,z) \leq d(x,y) + d(y,z)$

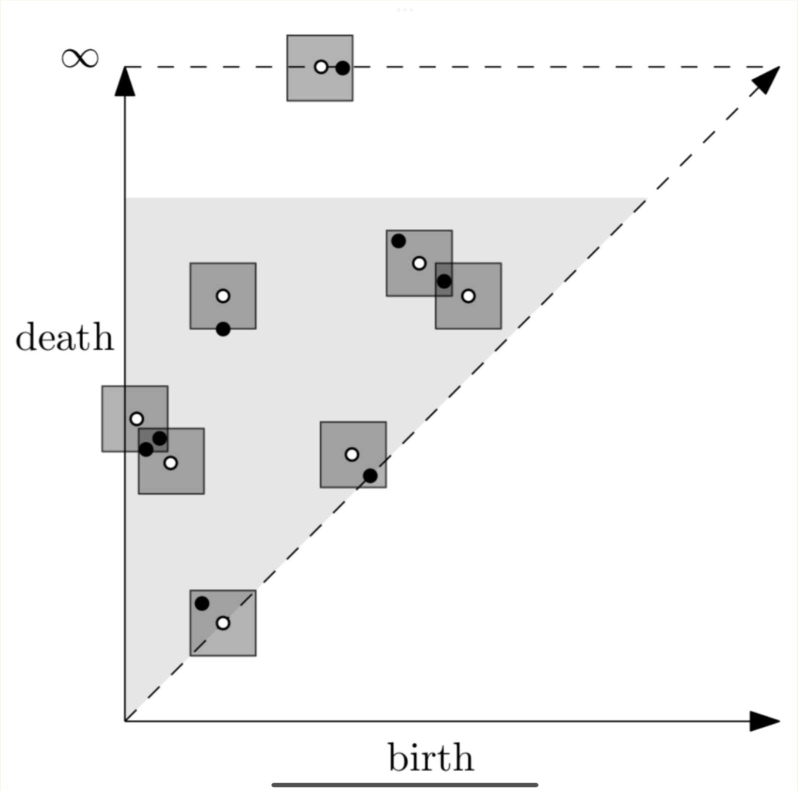
Our goal: distances for PDs

- $X = \{\mathrm{Dgm}\, F(K)\}_{F,K}$



Bottleneck distance (book's version)

Let $\Pi = \{\pi : Dgm_p(f) \to Dgm_p(g)\}$ denote the set of all bijections from PD of f to PD of g.



Let $\|\vec{x}-y\|_\infty = \max\{|x_1-x_2|, |y_1-y_2|\}$, where $\infty - \infty = 0$.

Then $d_B(Dgm(f), Dgm(g))$

$$= \inf_{\pi \in \Pi} \sup_{x \in Dgm(f)} \|x - \pi(x)\|_\infty.$$

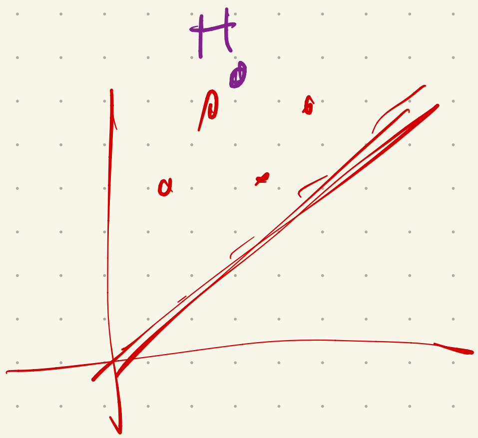
Fact: d_B is a metric. ↪ compute in $O(n^2)$

Proof:

- $d_B(X,Y) \geq 0$ ✓
- $d_B(X,Y) = 0 \iff X = Y$ ✓
- $d_B(X,Y) = d_B(Y,X)$

↪ bijection π

(b/c in $\mathbb{R}^2$ w/ L_∞ dist) ✓



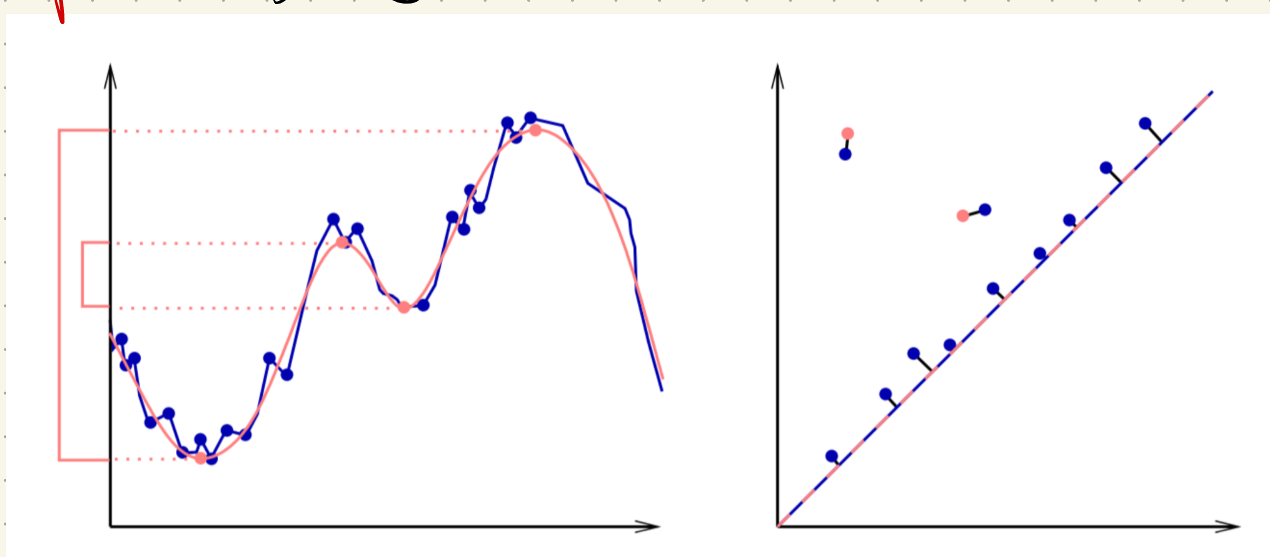
triangle inequality $x \cdot \; \cdot \; \cdot z$

Stability

Let X be a triangulatable space & $f, g : X \to \mathbb{R}$ be tame functions (eg: smooth & Lipschitz) giving rise to two space filtrations F_f & F_g. Then $\forall p \geq 0$,

$$d_B(\mathrm{Dgm}_p(F_f), \mathrm{Dgm}_p(F_g)) \leq \|f-g\|_\infty$$

pth homology diagram

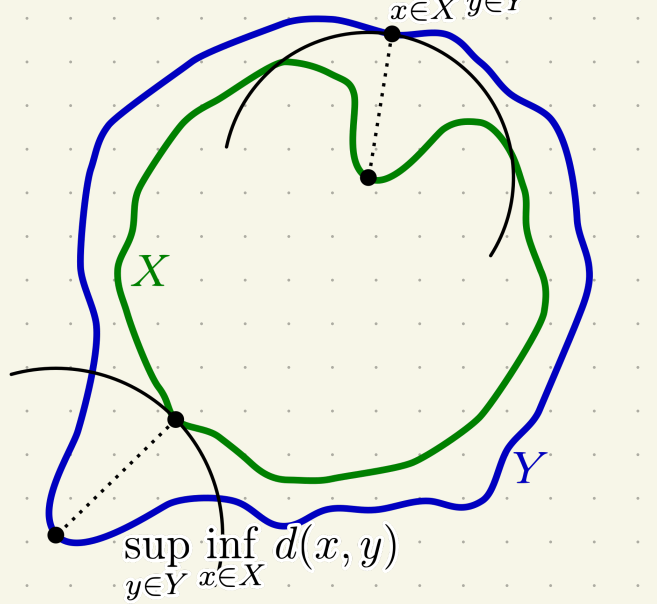


More stability:

Let d_H be Hausdorff distance:

$$d_H(X,Y) = \max\left\{ \sup_x \inf_y \|x-y\|, \sup_y \inf_x \|x-y\| \right\}$$

Translating:



For finite point clouds $X, Y \subseteq \mathbb{R}^d$, Let $\mathrm{Dgm}(C(X))$ & $\mathrm{Dgm}(C(Y))$ be the persistence diagrams of the filtration defined by the Čech complex. Then,

$$d_B(\mathrm{Dgm}(C(X)), \mathrm{Dgm}(C(Y))) \leq d_H(X, Y).$$

Proof picture:

(fixing p & doing only H_p)

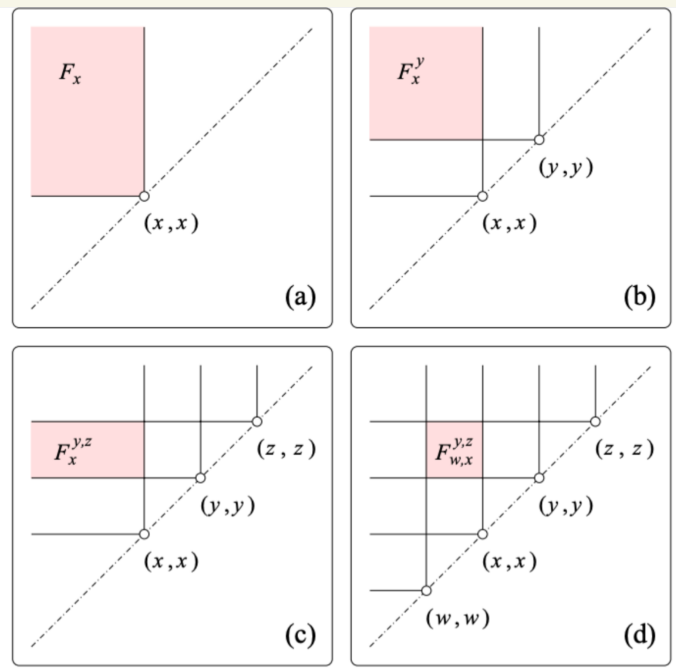


Figure 3: (a) Homology group of the sub-level set $f^{-1}(-\infty, x]$. (b) Image of F_x in F_y. (c) Kernel of surjection $F_x^y \to F_x^z$. (d) Quotient of $F_x^{y,z}$ and $F_w^{y,z}$.

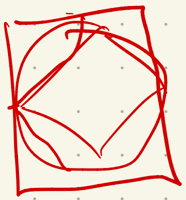
p,q-Wasserstein distance

Given diagrams X + Y,

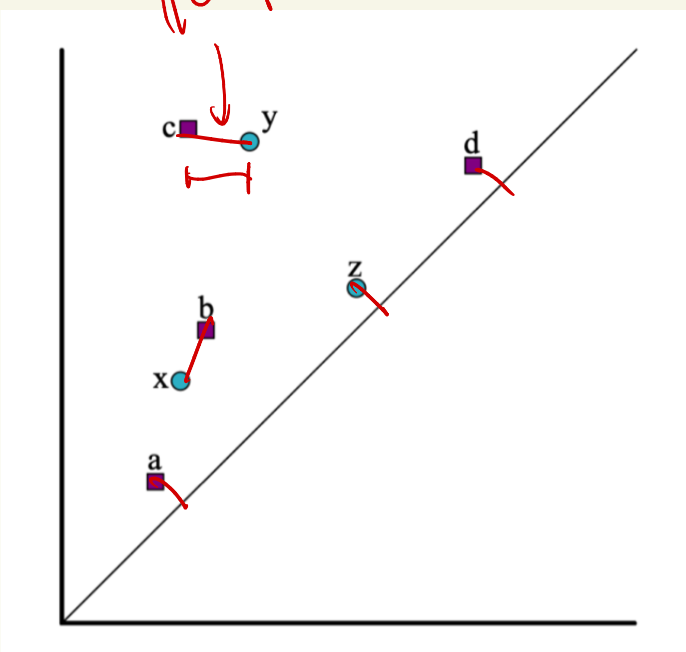
$$W_p^q(X,Y) = \inf_{e: X\to Y} \left(\sum_{x\in X} \|x - e(x)\|_p^{\,q} \right)^{1/q}$$

↑ bijection

$$= \inf_{e: X\to Y} \left(\sum_{x\in X} \left(\|x - e(x)\|_p\right)^q \right)^{1/q}$$



$\|c-y\|_2$



special cases:

$p = q = \infty$: bottleneck

$p = q = 2$: common!

Note of warning:
Notation is not consistent!

Our book: $p=q$ in Ch. 13, W_p^q

Definition 3.10 (Wasserstein distance). Let Π be the set of bijections as defined in Definition 3.9. For any $p \geq 0$, $q \geq 1$, the q-Wasserstein distance is define as

$$\mathsf{d}_{W,q}(\mathrm{Dgm}_p(\mathcal{F}_f), \mathrm{Dgm}_p(\mathcal{F}_g)) = \inf_{\pi \in \Pi} \left[\Sigma_{x \in \mathrm{Dgm}_p(\mathcal{F}_f)} \left(\|x - \pi(x)\|_q\right)^q\right]^{1/q}.$$

Other reference:

$$W_q(X,Y) = \left[\inf_{\eta: X \to Y} \sum_{x \in X} \|x - \eta(x)\|_\infty^q\right]^{1/q}$$

Nice resource: AATRN talk by Kate Turner

Nonetheless, can get some weaker notion of stability:

usually need addition of Lipschitz!

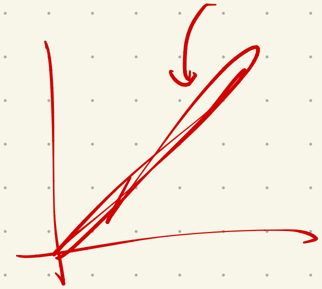
$$|f(x) - f(y)| \leq \|x - y\|_2$$ ← usually

Then: $\exists$ C & $k \geq 1$ s.t.

$$W_q^q(X, Y) \leq C \cdot \|f - g\|_\infty^{1 - \frac{k}{q}}$$

[Note: Hiding some technicalities here - I recommend Skraba + Turner 2020 if you are curious!]

Space of persistent diagrams



Let $D_\emptyset$ be the empty diagram.

The space of persistent diagrams D_p^q is the set of diagrams with finite distance to $D_\emptyset$, ie

$$D_p^q = \{X \mid W_p^q(X, D_\emptyset) < \infty\}$$

so for each $x \in X$,
can match to diagonal

Note: does not necessarily mean X is finite!

Some statistical things

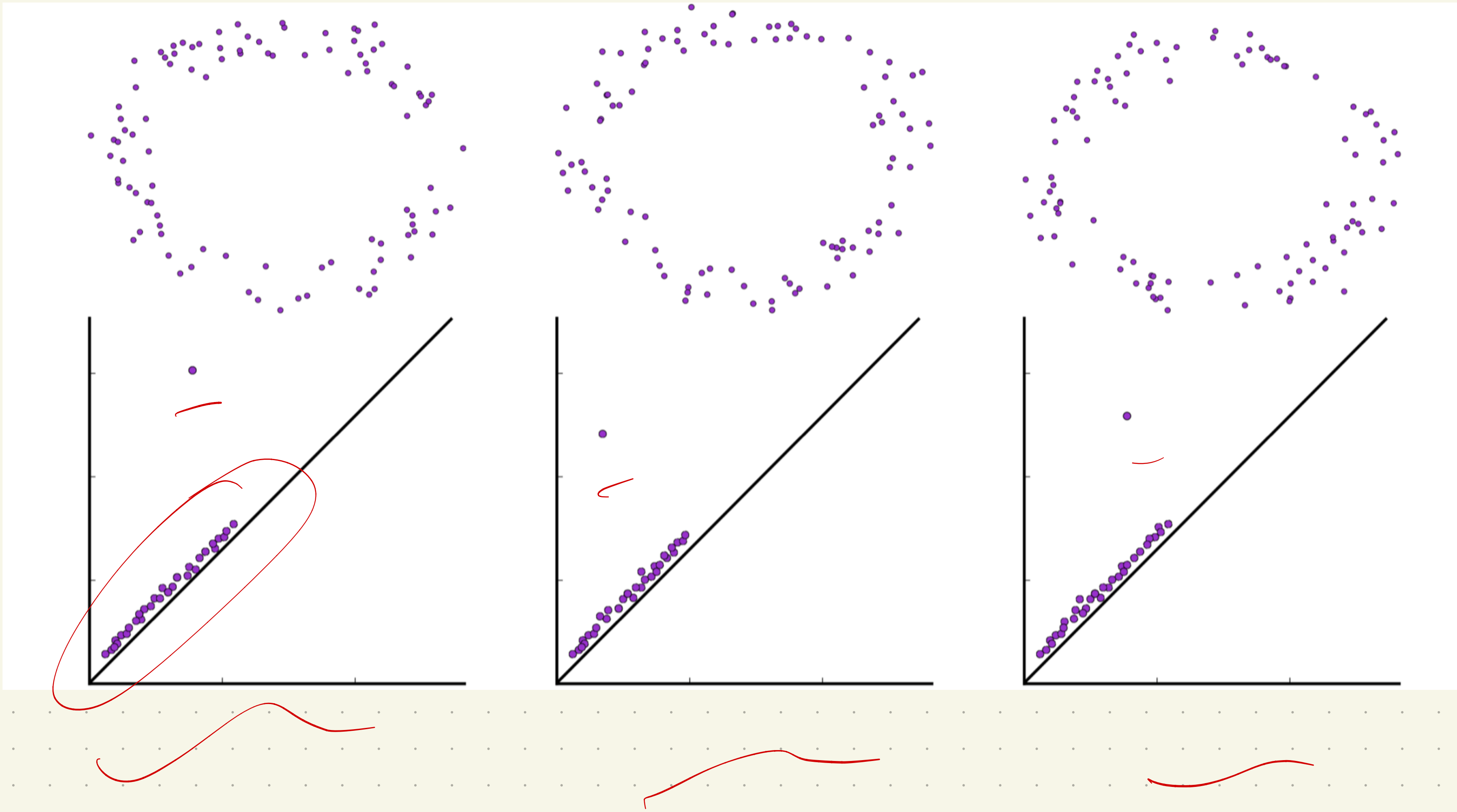
- D_p^q is complete & seperable (Polish) if $p=\infty$ & $q \in \mathbb{Z}_{\geq 1}$

  ↳ Why? probility distributions!

- Non-negatively curved Alexandrov space if $p=q=2$. D_2^2 W_2^2

  ↳ Why? gradient descent

How can we get an "average"?



Frechet means

Consider $X=\{x_1,\dots,x_n\}\subset\mathbb{R}^2$

The Frechet variance of X is

$$\mathrm{Var}_\nu = \inf_{a\in\mathbb{R}^2}\left[F_\nu(a) = \frac{1}{n}\sum_i \|x_i-a\|^2\right]$$

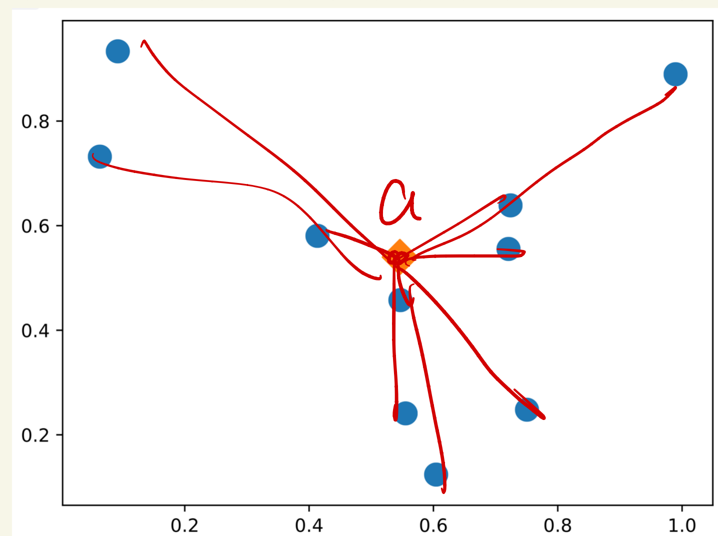
The set which realizes is

$$E(\nu)=\{a\in\mathbb{R}^2 \mid F_\nu(a)=\mathrm{Var}_\nu\}$$

↳ called Frechet mean

(or Frechet expectation)

→ Unique & computable!



↑ blue pts are X

Now let $\{X_1, \ldots, X_n\}$ be persistence diagrams in D_p^p. $\leftarrow D_p$ $+ q^{?p}$

Frechet variance

$$\mathrm{Var}_\nu = \inf_{x \in Y} \left[F_\nu(Y) = \frac{1}{n} \sum_{i=1}^{n} W_p(X_i, Y)^2 \right]$$

& Frechet mean is the set where value is obtained:

$$E(\nu) = \{ Y \mid F_\nu(Y) = \mathrm{Var}_\nu \}$$

... what??

Picture: not unique!

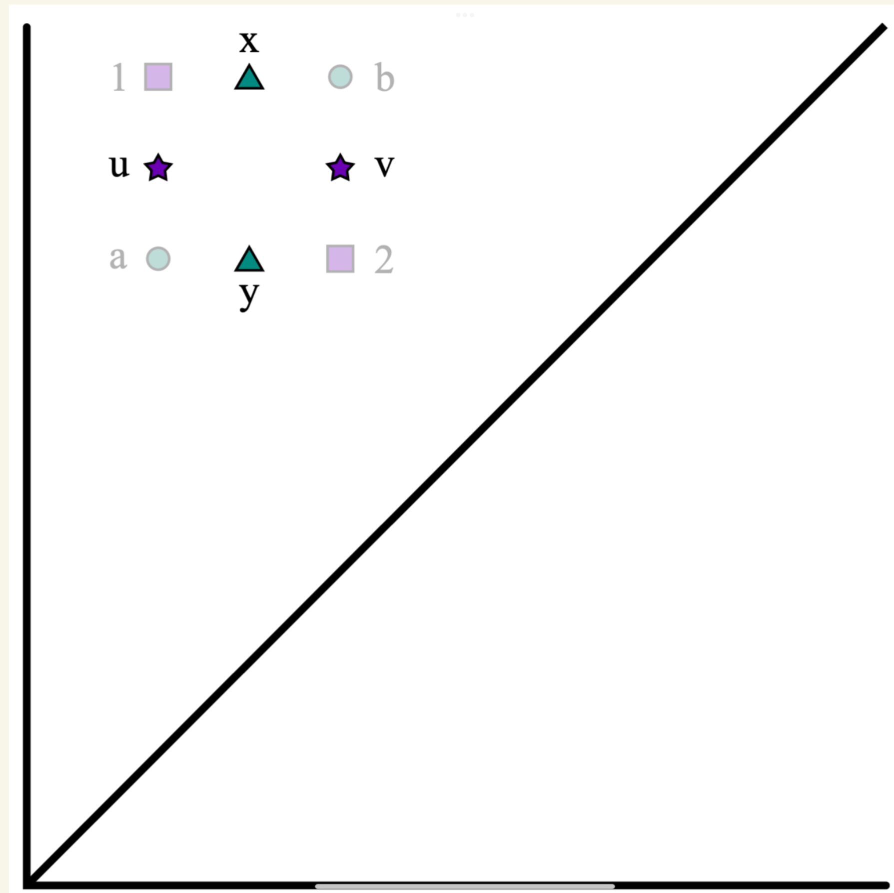
$X_1 = \{\square_1, \square_2\}$

$X_2 = \{\circ_a, \circ_b\}$

Two Frechet means:

$Y_1 = \{\star_u, \star_v\}$

$Y_2 = \{\triangle_x, \triangle_y\}$



& $E[\nu] \supseteq \{Y_1, Y_2\}$

The good news:

The Frechet mean is non-empty

$\boxed{\text{Mileyko et al 2011 \quad Turner et al 2014}}$

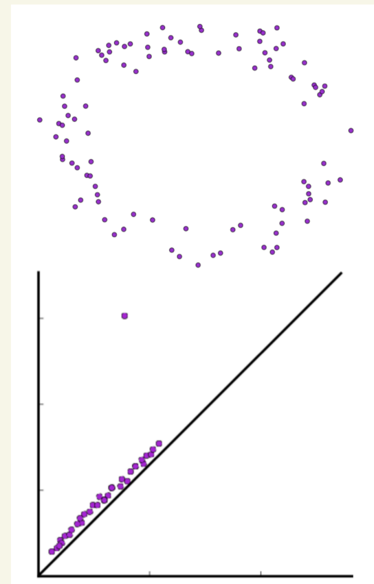
(with some mild assumptions on distribution of the set)

For D_2^2: gradient descent algorithm to compute local minimum.

In general, though: open

Changing the question: Fasy et al 2014

What is an estimate for the average (true?) diagram, & how far off am I?



- Want to estimate PD for a set $M \subseteq \mathbb{R}^d$
- Don't know M
  - ↳ but, have a sample $S_n = \{X_1, \dots, X_n\}$ drawn uniformly from M.
- Persistence diagram for S_n is used as an estimator for X → denoted $\hat{X}$

Confidence intervals

Given a collection of points $X = \{x_1, \dots, x_n\}$ from $\mathbb{R}$, the $100 \cdot (1-\alpha)\%$ confidence interval for the mean μ is the interval $[u(X), v(X)]$ such that

$$P(\mu \in [u(X), v(X)) = 1-\alpha$$

Equivalently: find c + an estimate for μ called $\hat{\mu}$ s.t.

$$P(\|\mu - \hat{\mu}\| \geq c) = \alpha$$

How to use in persistence?

Fix $\alpha \in (0,1)$

Want $c_n := c_n(X_1, \dots, X_n)$ s.t.

$$\limsup_{n\to\infty} P\left(d_B(\hat{X}, X_n) > c_n\right) \le \alpha$$

Then, $[0, c_n]$ is an asymptotic $(1-\alpha)$ confidence set for the bottleneck distance $d_B(\hat{X}, X)$.

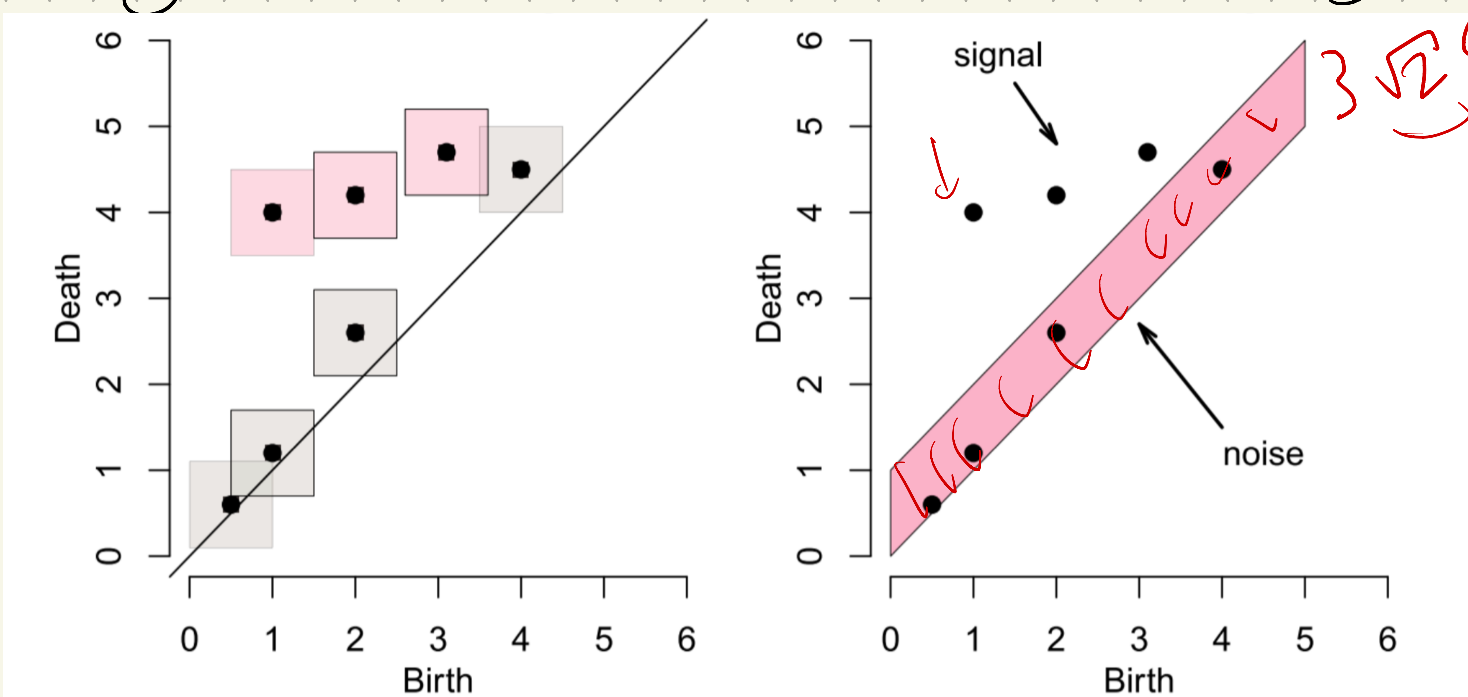
The confidence set C_n is the set of diagrams whose distance to $\hat{X}$ is $\le c_n$

$$C_n = \{ Y \mid d_B(\hat{X}, Y) \le c_n \}$$

Assume you have $\hat{X}$ + c_n!

Put a box of width $2c_n$ at every point in $\hat{X}$.

A point is noise if its box intersects the diagonal → or, put strip along diagonal!



$3\sqrt{2}\,c_n$?

How to get c_n though?

- Start with data $S = \{X_1, \dots, X_n\}$
- Choose $b = b_n$ such that $b = o\left(\frac{n}{\log n}\right)$
- Pretend we have all $N = \binom{n}{b}$ subsamples $S^1, \dots, S^N$

  ↳ "bootstrapping"

  (In reality: just do a lot)
- Calculate $d_H(S^j, S)$, $j = 1 \dots N$
- Set $L_b(t) = \frac{1}{N} \sum_{j=1}^{N} I(T_j > t)$ &

  set $c_b = 2L_b^{-1}(\alpha)$

What now??

Using a theorem here:

Theorem

For mild assumptions on the space M, and for all large n,

$$\mathbb{P}(d_B(\hat{X}, X) > c_b) \leq \mathbb{P}(d_H(\mathcal{S}_n, \mathbb{M}) > c_b) \leq \alpha + O\left(\frac{b}{n}\right)^{\frac{1}{4}}$$

[Note: there is every chance you may be better at probability than me...]

Back to pictures

(Next time—stay tuned!)