Probabilistic Fréchet Means on Persistence Diagrams

Paul Bendich

Duke University :: Dept of Mathematics

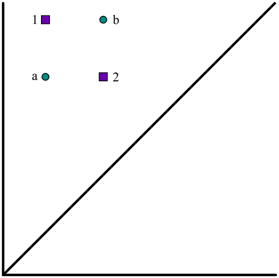
July 15, 2013

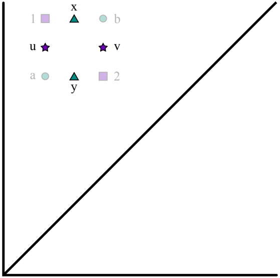
Collaborators

- This is joint work with:
  - Liz Munch (Duke)
  - Kate Turner (Chicago)
  - John Harer (Duke)
  - Sayan Mukherjee (Duke)
  - Jonathan Mattingly (Duke)

Main Idea and Results

- New definition of mean for a set X of diagrams in (D_p, W_p)
- Mileyko et. al.:
  - μ_X is itself a (set of) diagram(s) in D_p.
  - Problem: non-uniqueness leads to discontinuity issues.
- Our approach:
  - Definition: $\mu_X \in \mathcal{P}(D_p)$: (atomic) prob. dist. on diagrams.
  - Theorem: $X \to \mu_X$ is Hölder continuous (with exponent $\frac{1}{2}$)





Persistence modules

- A persistence module $\mathcal{F}$ is:
  - family of vector spaces $\{F_\alpha\}, \alpha \in \mathbb{R}$, over a fixed field
  - family of linear transformations $f_\alpha^\beta : F_\alpha \to F_\beta$, for all $\alpha \leq \beta$, s.t $\alpha \leq \gamma \leq \beta$ implies $f_\alpha^\beta = f_\gamma^\beta \circ f_\alpha^\gamma$.
- The number α is a regular value of the module if:
  - There exists $\delta > 0$ such that $f_{\alpha-\epsilon}^{\alpha+\epsilon}$ is iso. for all $\epsilon < \delta$.
- If α is not a r.v., then it is a critical value of the module.
- Module is tame if only finitely many c.v's, and each v.s is of finite rank.

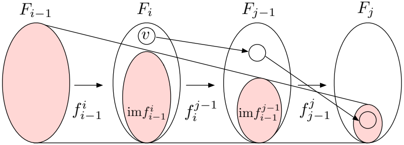
Persistence Modules

- Given finitely many c.v's $c_1 < c_2 < \ldots < c_n$.
- Interleave r.v's $a_0 < c_1 < a_1 < \ldots < c_n < a_n$.
- Set $F_i = F_{a_i}$:

$$F_0 \to F_1 \to F_2 \ldots \to F_{n-1} \to F_n$$

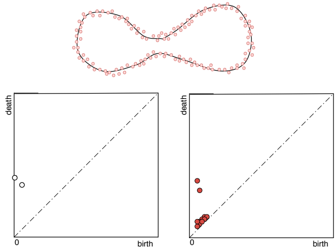
Birth and Death

- A vector $v \in F_i$ is born at c_i if $v \notin \operatorname{im} f_{i-1}^{i}$
- Such a v dies at c_j if:
  - $f_i^j(v) \in \operatorname{im} f_{i-1}^{j}$
  - $f_i^{j-1}(v) \notin \operatorname{im} f_{i-1}^{j-1}$.
- The persistence of v is $c_j - c_i$.



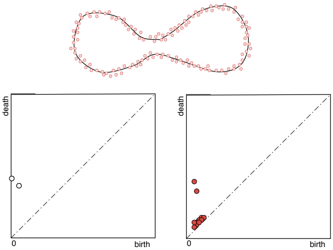
Persistence Diagrams

- Let $P^{i,j}$ be v.s of classes born at c_i and dead at c_j, and $\beta^{i,j}$ its rank.
- Plot a dot of multiplicity $\beta^{i,j}$ at (c_i, c_j) in plane.
- Plot a dot of infinite multiplicity at all $y = x$ diagonal points.
- Result is $\mathrm{Dgm}(\mathcal{F})$.

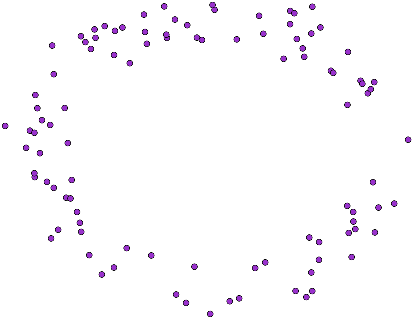


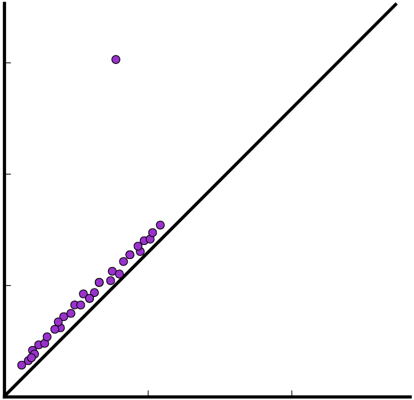
Example: persistent homology

- Let $\mathbb{Y} \subseteq \mathbb{R}^D$ be compact space.
- For $\alpha \geq 0$, define
$$Y_\alpha = d_Y^{-1}[0, \alpha]$$
- For each k, get module $\{H_k(Y_\alpha)\}$, with maps induced by inclusion.

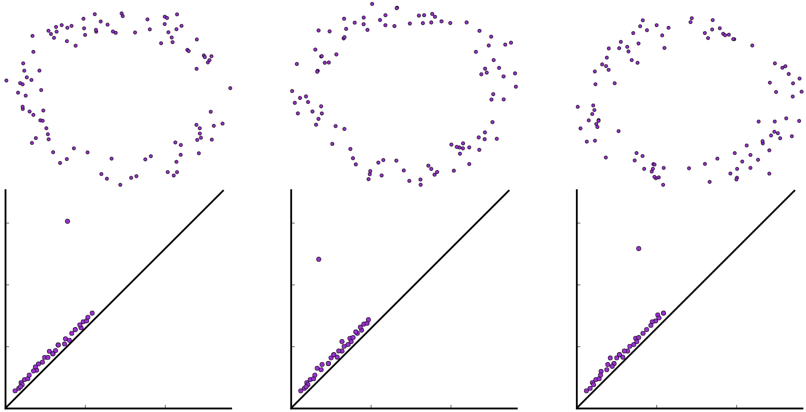


Relate Multiple Samples

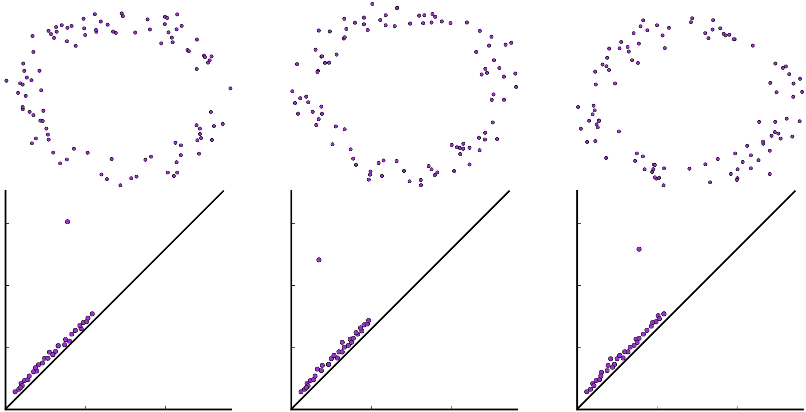




Relate Multiple Samples



Relate Multiple Samples

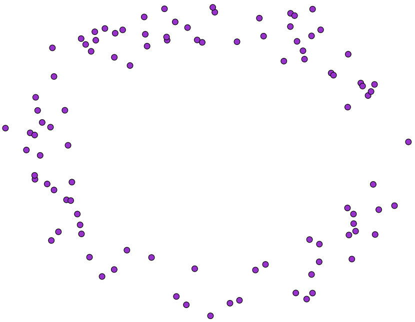


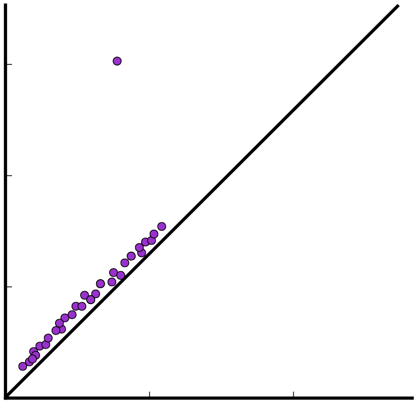
How do we give a summary of the data?

Will it play nicely with time varying persistence diagrams?

Significance Testing

- Suppose we obtain N points X in unit d-ball.
- We compute the diagram and are impressed with a feature.
- Should we be impressed?

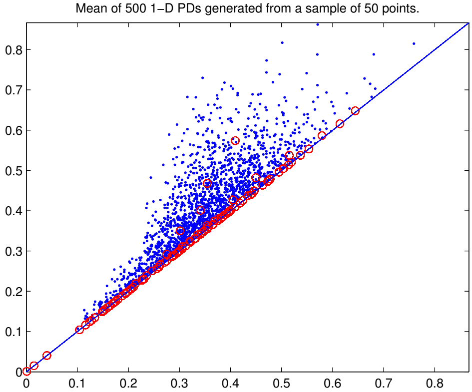




Towards Topological Null Hypothesis

- Experiment: draw N points uniformly from d-ball and compute diagram.
- Question: what is expected diagram?
- Hope: repeat experiment many times, take mean diagram as answer.

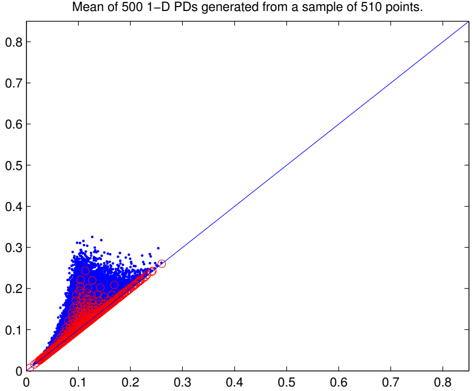
[redacted]



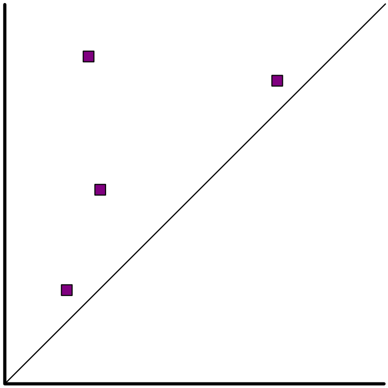
Towards Topological Null Hypothesis

- Experiment: draw N points uniformly from d-cube and compute diagram.
- Question: what is expected diagram?
- Hope: repeat experiment many times, take mean diagram as answer.

[redacted]



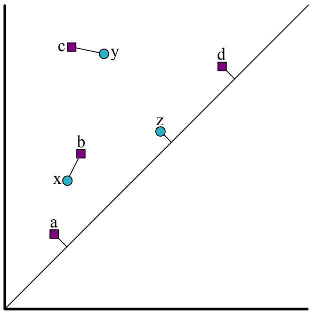
Diagrams in the Abstract



Abstract Persistence Diagram

An abstract persistence diagram is a countable multiset of points along with the diagonal, $\Delta = \{(x, x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$, with points in Δ having infinite multiplicity.

Wasserstein Distance on D_p



p-Wasserstein distance for diagrams

Given diagrams X and Y, the distance between them is

$$W_p[L_q](X, Y) = \inf_{\varphi : X \to Y} \left(\sum_{x \in X} \left(\|x - \varphi(x)\|_q \right)^p \right)^{1/p}.$$

Discrete vs continuous Wasserstein

Discrete

Given diagrams X and Y, the distance between them is

$$W_p[L_q](X,Y) = \inf_{\varphi:X\to Y}\left(\sum_{x\in X}(\|x-\varphi(x)\|_q)^p\right)^{1/p}.$$

Continuous

Given probability distributions, ν and η, on metric space $(\mathbb{X}, d_{\mathbb{X}})$ is

$$W_p[d_{\mathbb{X}}](\nu,\eta) = \left[\inf_{\gamma\in\Gamma(\nu,\eta)}\int_{\mathbb{X}\times\mathbb{X}} d_{\mathbb{X}}(x,y)^p\, d\gamma(x,y)\right]^{1/p}$$

where $\Gamma(\nu,\eta)$ is the space of distributions on $\mathbb{X}\times\mathbb{X}$ with marginals ν and η respectively.

The metric space (D_p, W_p)

- The space of persistence diagrams is

$$D_p = \{X \mid W_p[L_2](X, d_\emptyset) < \infty\}$$

  along with the p-Wasserstein metric, $W_p[L_2]$.
- Theorem (Mileyko et. al.): (D_p, W_p) is complete and separable.

Fréchet means

- Let ν be a measure on a metric space (Y, d).
- The Fréchet variance of ν is:

$$\mathrm{Var}_\nu = \inf_{x \in Y} \left[F_\nu(x) = \int_Y d(x,y)^2 \, d\nu(y) < \infty \right]$$

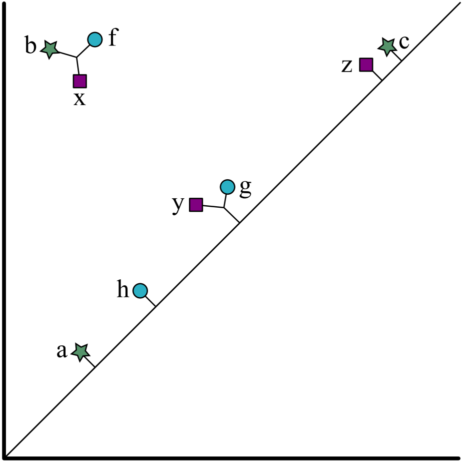
- The set at which the value is obtained

$$\mathbb{E}(\nu) = \{x | F_\nu(X) = \mathrm{Var}_\nu\}$$

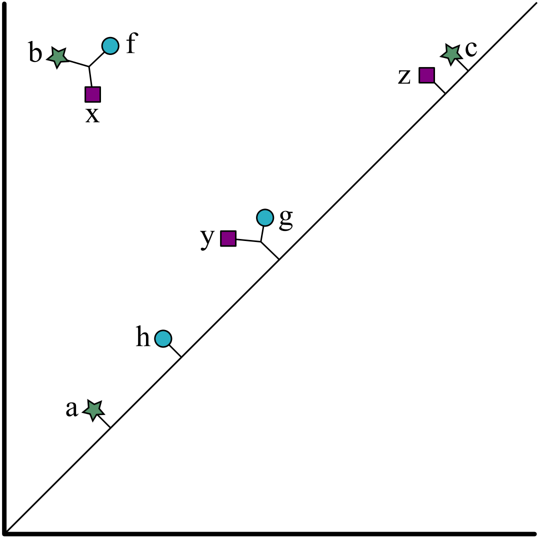
is the Fréchet expectation of ν, also called Fréchet mean.

Fréchet means in D_p: Existence

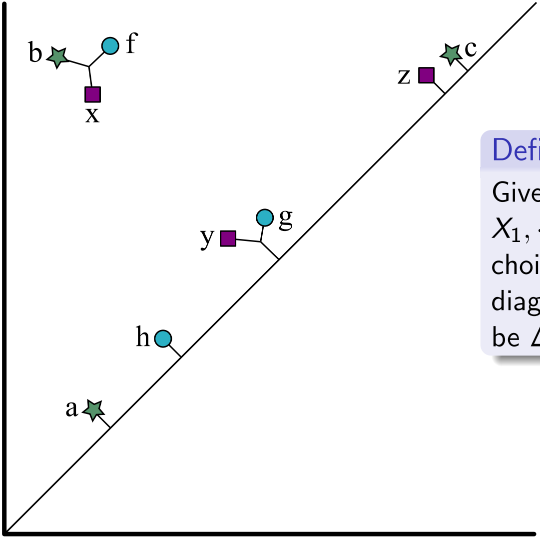
- Theorem (Mileyko et. al.): Let ν be a probability measure on $(D_p, \mathcal{B}(D_p))$ with a finite second moment. If ν has compact support, then $\mathbb{E}(\nu) \neq \emptyset$.
- In particular, Fréchet means of finite sets of diagrams exist.



Algorithm for Computation - Selections and Matchings



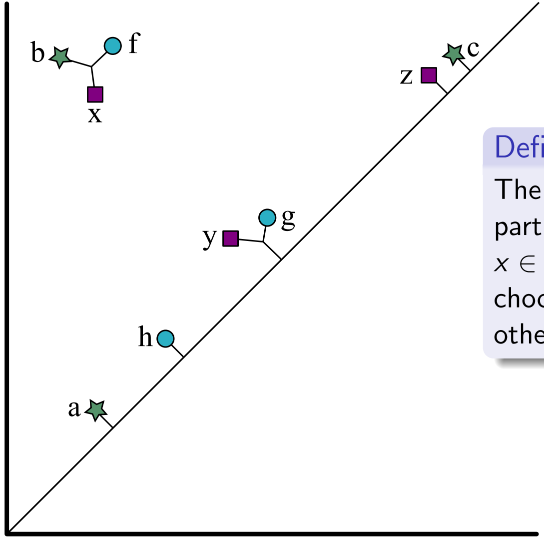
Algorithm for Computation - Selections and Matchings



Definition

Given a set of diagrams $X_1, \cdots, X_N$, a **selection** is a choice of one point from each diagram, where that point could be Δ.

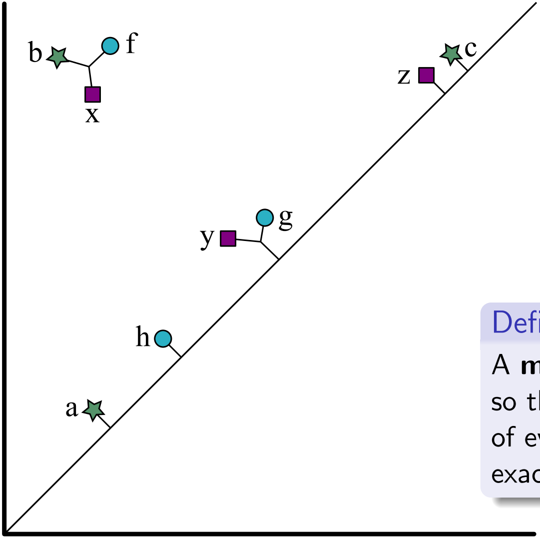
Algorithm for Computation - Selections and Matchings



Definition

The **trivial selection** for a particular off-diagonal point $x \in X_i$ is the selection s_x which chooses x for X_i and Δ for every other diagram.

Algorithm for Computation - Selections and Matchings

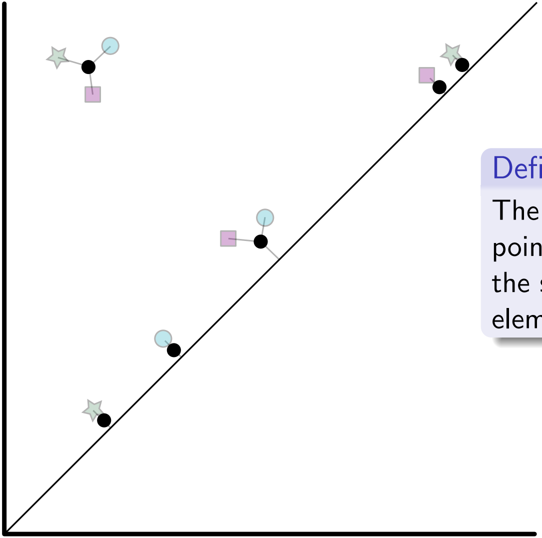


$$
\begin{array}{c}
\\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6
\end{array}
\begin{array}{c}
\begin{array}{ccc} d_\bigstar & d_\blacksquare & d_\bullet \end{array} \\
\begin{pmatrix}
b & x & f \\
a & \Delta & \Delta \\
\Delta & y & g \\
\Delta & z & \Delta \\
\Delta & \Delta & h \\
c & \Delta & \Delta
\end{pmatrix}
\end{array}
$$

Definition

A **matching** is a set of selections so that every off-diagonal point of every diagram is part of exactly one selection.

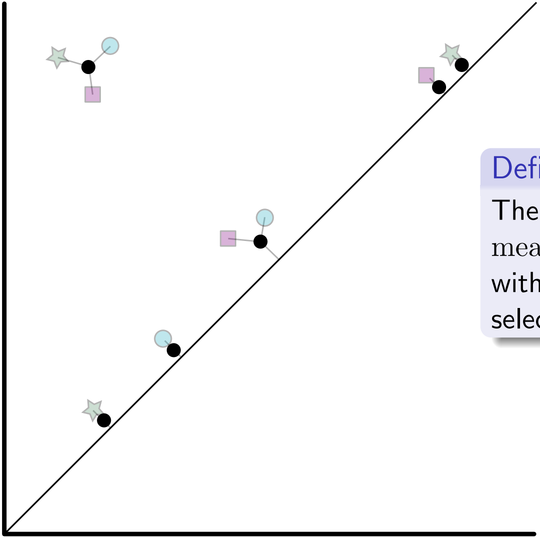
Algorithm for Computation - Selections and Matchings



Definition

The **mean of a selection** is the point which minimizes the sum of the square distances to the elements of the selection.

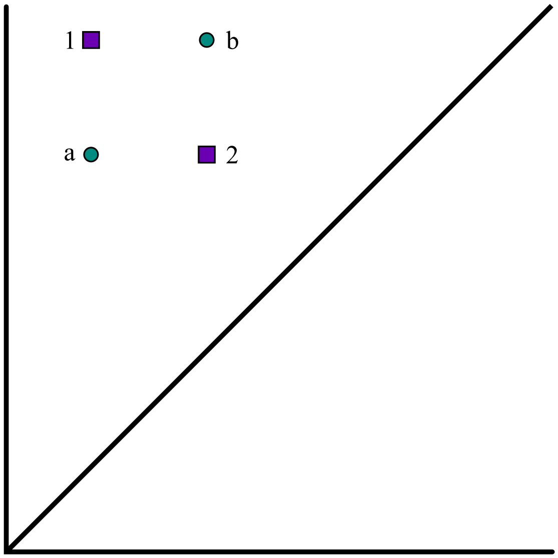
Algorithm for Computation - Selections and Matchings



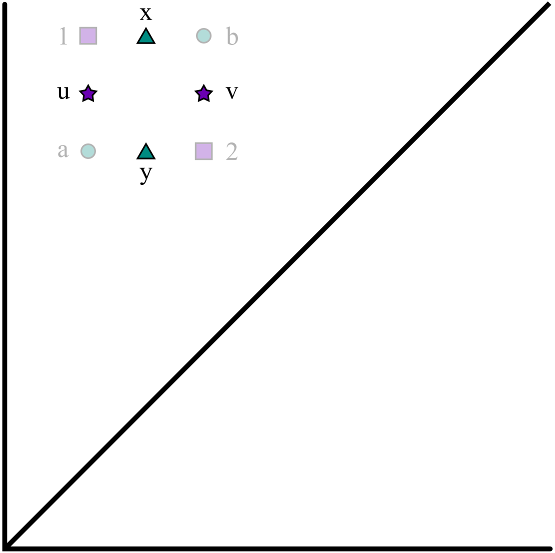
Definition

The **mean of a matching**, $\mathrm{mean}_X(G)$, is a diagram in D_p with a point at the mean of each selection

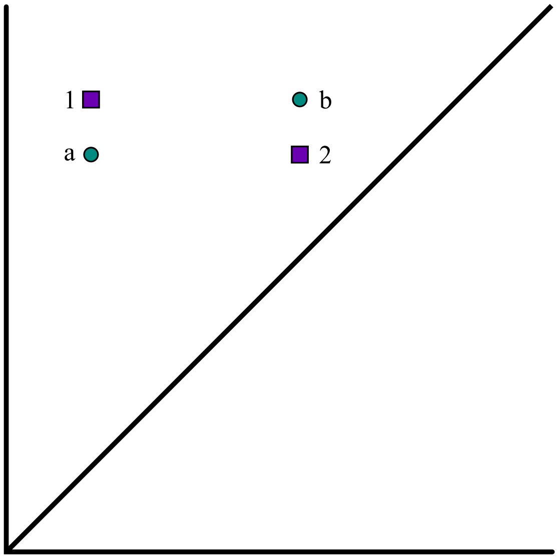
Problem: Fréchet means need not be unique!



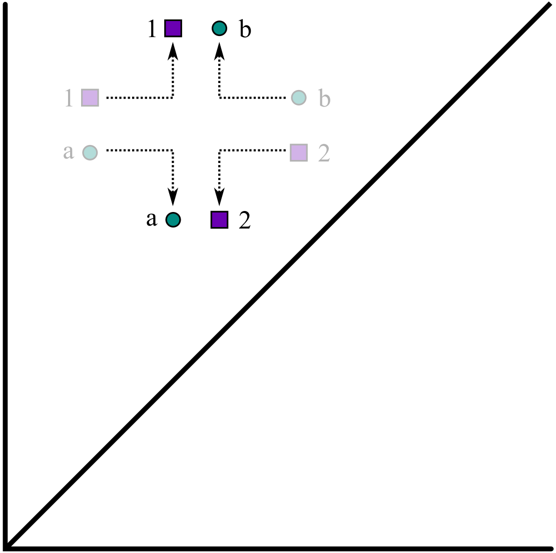
Problem: Fréchet means need not be unique!



Problem: Fréchet means need not be unique!



Problem: Fréchet means need not be unique!

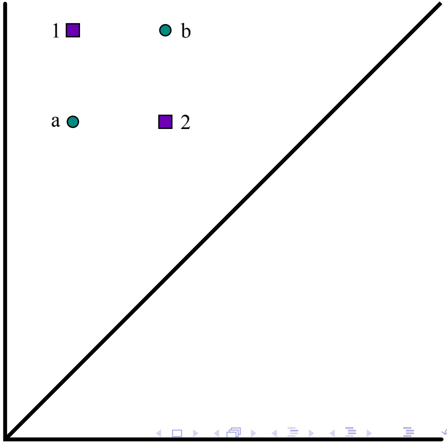


Solution: Randomize Matchings!

- Note: non-uniquness of mean caused by non-uniqueness of optimal matching.
- Idea: consider all matchings, with probability weights.
- Formally: if $X = \{X_1, \ldots, X_N\} \subseteq D_p$, then $\mu_X \in \mathcal{P}(D_p)$, with:

Definition

$$\mu_X = \sum_G \mathbb{P}(\mathcal{H} = G)\ \delta_{\mathrm{mean}_X(G)}$$

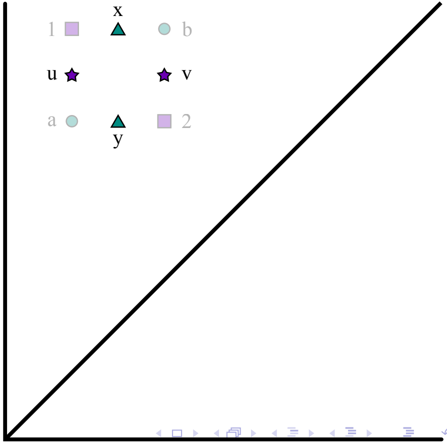


Solution: Randomize Matchings!

- Note: non-uniquness of mean caused by non-uniqueness of optimal matching.
- Idea: consider all matchings, with probability weights.
- Formally: if $X = \{X_1, \ldots, X_N\} \subseteq D_p$, then $\mu_X \in \mathcal{P}(D_p)$, with:

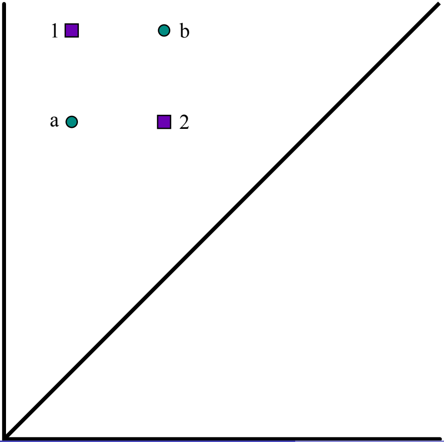
Definition

$$\mu_X = \sum_G \mathbb{P}(\mathcal{H} = G)\ \delta_{\mathrm{mean}_X(G)}$$



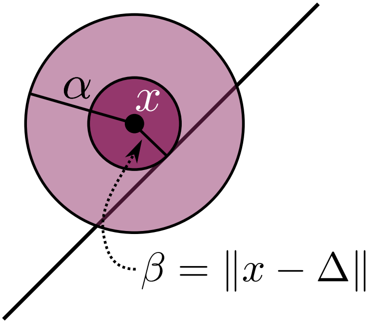
What is $\mathcal{H}$?

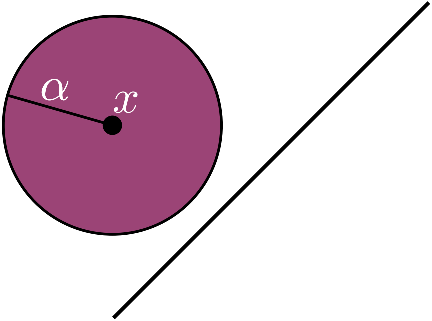
- $\mathcal{H}$ is a matching-valued random variable (randomized coupling).
- Perturb each diagram X_i to create random diagram X_i'.
- Associate the optimal matching among the drawn diagrams to one of the original matchings.
- This defines a probability weight on each possible matching.



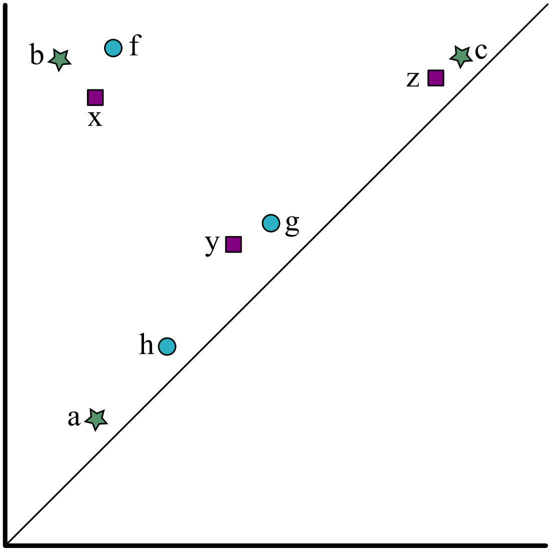
The random diagram

- Pick $\alpha > 0$
- Let $\eta \in \mathcal{P}(\mathbb{R}^2)$ be uniform on $B_\alpha(0)$ (other choices also work).
- Define η_x to be the translation of η to x.
- For each $x \in X_i$, make X_i' by:
  1. Draw point from η_x
  2. If contained in $B_{\|x-\Delta\|}(x)$, add it to X_i'.

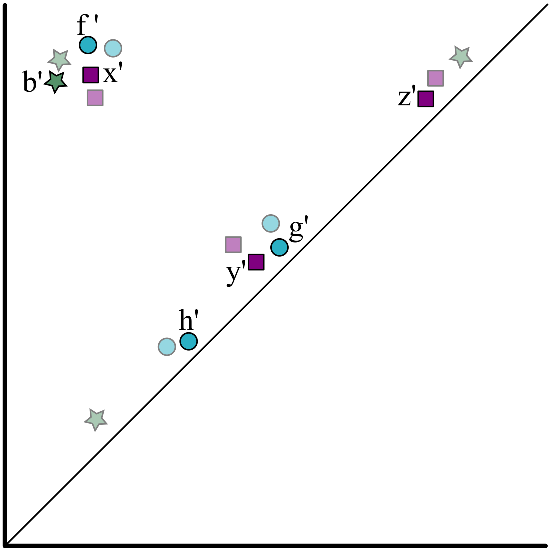




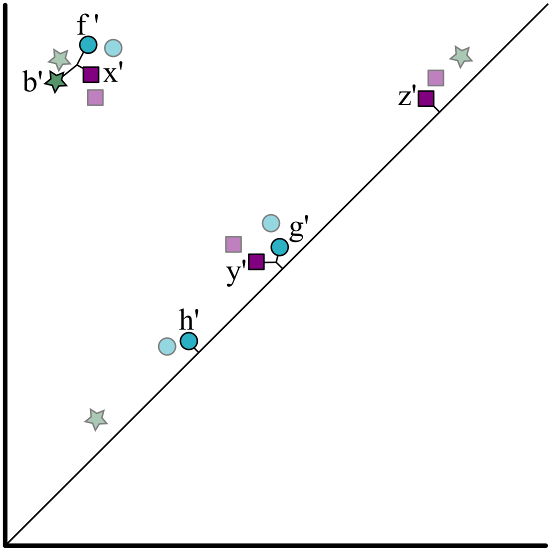
Example



Example

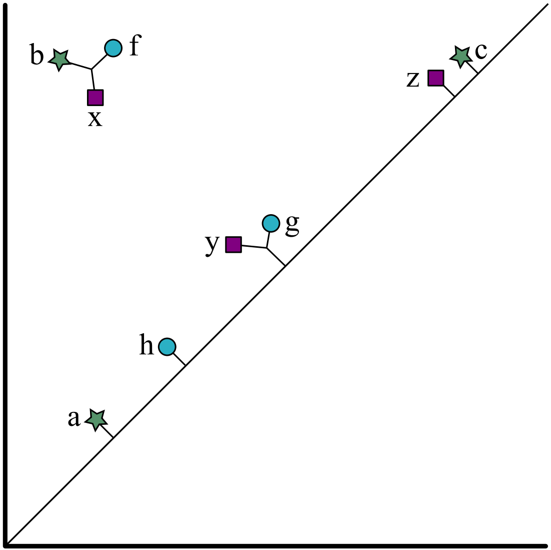


Example



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} d'_{\bigstar} & d'_{\square} & d'_{\bigcirc} \end{array} \\ \begin{pmatrix} b' & x' & f' \\ \Delta & y' & g' \\ \Delta & \Delta & h' \\ \Delta & z' & \Delta \end{pmatrix} \end{array}.$$

Example



$$\begin{array}{c|ccc} & d'_{\bigstar} & d'_{\square} & d'_{\bigcirc} \\ \hline 1 & b' & x' & f' \\ 2 & \Delta & y' & g' \\ 3 & \Delta & \Delta & h' \\ 4 & \Delta & z' & \Delta \end{array}.$$

$$\begin{array}{c|ccc} & d_{\bigstar} & d_{\square} & d_{\bigcirc} \\ \hline 1 & b & x & f \\ 2 & \Delta & y & g \\ 3 & \Delta & \Delta & h \\ 4 & \Delta & z & \Delta \\ 5 & a & \Delta & \Delta \\ 6 & c & \Delta & \Delta \end{array}.$$

Main Theorem

- Let $S_{M,K} \subseteq D_p$ be diagrams with at most K dots, each with persistence at most M.

Theorem

The map

$$\begin{array}{ccc} (S_{M,K})^N & \longrightarrow & \mathcal{P}(S_{M,NK}) \\ X = \{X_1, \ldots, X_N\} & \longmapsto & \mu_X \end{array}$$

is Hölder continuous with exponent $\frac{1}{2}$. That is, there exists a constant C such that the inequality

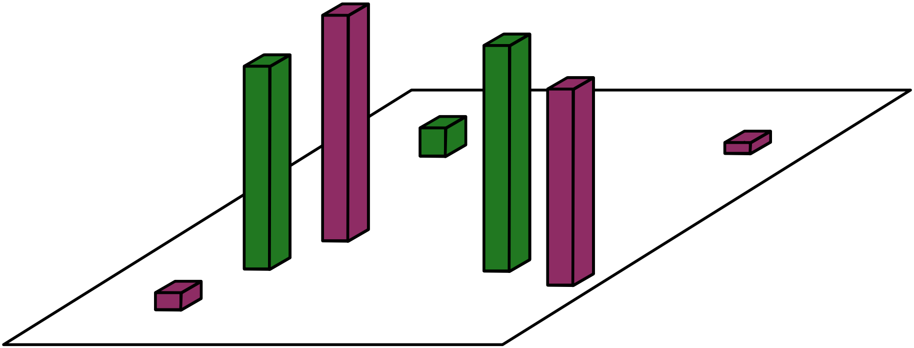
$$W_2(\mu_X, \mu_Y) \leq C\sqrt{W_2(X, Y)}$$

holds for all pairs of sets of N diagrams.

Outline of the Proof

Wasserstein distance on $\mathcal{P}(D_p)$

$$W_p(\nu,\eta) = \left[\inf_{\gamma \in \Gamma(\nu,\eta)} \int_{D_p \times D_p} W_2(X,Y)^p \, d\gamma(X,Y) \right]^{1/p}$$



Outline of the Proof - Pairing

The problem

It's easy to associate parts of the matching if a point $x \in X_i$ is matched with and off-diagonal point $y \in Y_i$ under $\varphi_i : X_i \to Y_i$.
What do you do with the rest of the points?

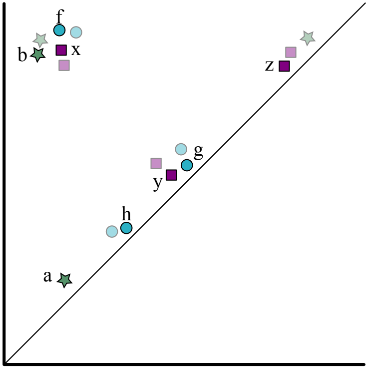
Definition

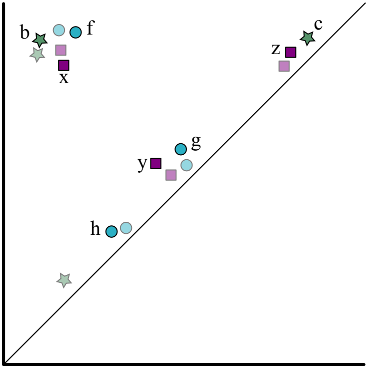
$$\widetilde{X}_i = \{x \in X_i \mid \varphi_i(x) \neq \Delta\}$$
$$\widetilde{Y}_i = \{y \in Y_i \mid \varphi_i^{-1}(y) \neq \Delta\}$$
$$\mathcal{G}_X = \text{matchings on } X_1, \cdots, X_N$$

$$\begin{tikzcd}
\mathcal{G}_{\widetilde{X}} \arrow[r] \arrow[d, "i_{\widetilde{X}}"] & \mathcal{G}_{\widetilde{Y}} \arrow[d, "i_{\widetilde{Y}}"] \\
\mathcal{G}_X \arrow[r, dotted] & \mathcal{G}_Y
\end{tikzcd}$$

$$\mathrm{Im}\,(i_{\widetilde{X}}) \leftrightarrow \mathrm{Im}\,(i_{\widetilde{X}})$$

Outline of the Proof - Pairing





$$\begin{array}{cc} & \begin{array}{ccc} d_\bigstar & d_\square & d_\bigcirc \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{pmatrix} b & x & f \\ \Delta & y & g \\ \Delta & \Delta & h \\ \Delta & z & \Delta \\ a & \Delta & \Delta \end{pmatrix} \end{array}.$$

$$\begin{array}{cc} & \begin{array}{ccc} d_\bigstar & d_\square & d_\bigcirc \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{pmatrix} b & x & f \\ \Delta & y & g \\ \Delta & \Delta & h \\ \Delta & z & \Delta \\ c & \Delta & \Delta \end{pmatrix} \end{array}.$$

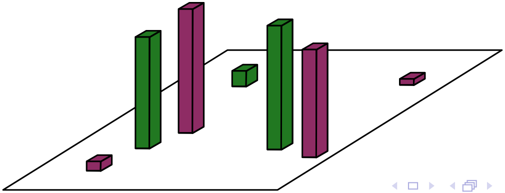
Outline of the Proof - Big Inequality

$$W_p(\mu_X, \mu_Y) \leq$$

$$\sum_{\substack{(G,H)\\ \in \mathcal{G}_X \times \mathcal{G}_Y \\ \text{Paired}}} \min\{\mathbb{P}(\mathcal{H}_X = G), \mathbb{P}(\mathcal{H}_Y = H)\} \cdot W_p(\text{mean}_X(G), \text{mean}_Y(H))$$

$$+ \sum_{\substack{(G,H) \in \mathcal{G}_X \times \mathcal{G}_Y \\ \text{Paired}}} |\mathbb{P}(\mathcal{H}_X = G) - \mathbb{P}(\mathcal{H}_Y = H)| \cdot \overline{M}$$

$$+ \sum_{G \in \mathcal{G}_X \text{ unpaired}} |\mathbb{P}(\mathcal{H}_X = G)| \cdot \overline{M} + \sum_{H \in \mathcal{G}_Y \text{ unpaired}} |\mathbb{P}(\mathcal{H}_Y = H)| \cdot \overline{M}$$



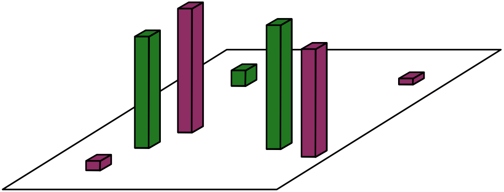
Outline of the Proof - Big Inequality

$$W_p(\mu_X, \mu_Y) \leq$$

$$\sum_{\substack{(G,H)\\ \in \mathcal{G}_X \times \mathcal{G}_Y \\ \text{Paired}}} \min\{\mathbb{P}(\mathcal{H}_X = G), \mathbb{P}(\mathcal{H}_Y = H)\} \cdot W_p(\text{mean}_X(G), \text{mean}_Y(H))$$

$$+ \sum_{\substack{(G,H) \in \mathcal{G}_X \times \mathcal{G}_Y \\ \text{Paired}}} |\mathbb{P}(\mathcal{H}_X = G) - \mathbb{P}(\mathcal{H}_Y = H)| \cdot \overline{M}$$

$$+ \sum_{G \in \mathcal{G}_X \text{ unpaired}} |\mathbb{P}(\mathcal{H}_X = G)| \cdot \overline{M} + \sum_{H \in \mathcal{G}_Y \text{ unpaired}} |\mathbb{P}(\mathcal{H}_Y = H)| \cdot \overline{M}$$



Further Goals

- Find explicit relation between older definition and ours.
- Do some honest statistics (laws of large numbers, ...)
- Get rid of $S_{M,K}$ crutch.