OSL 2015

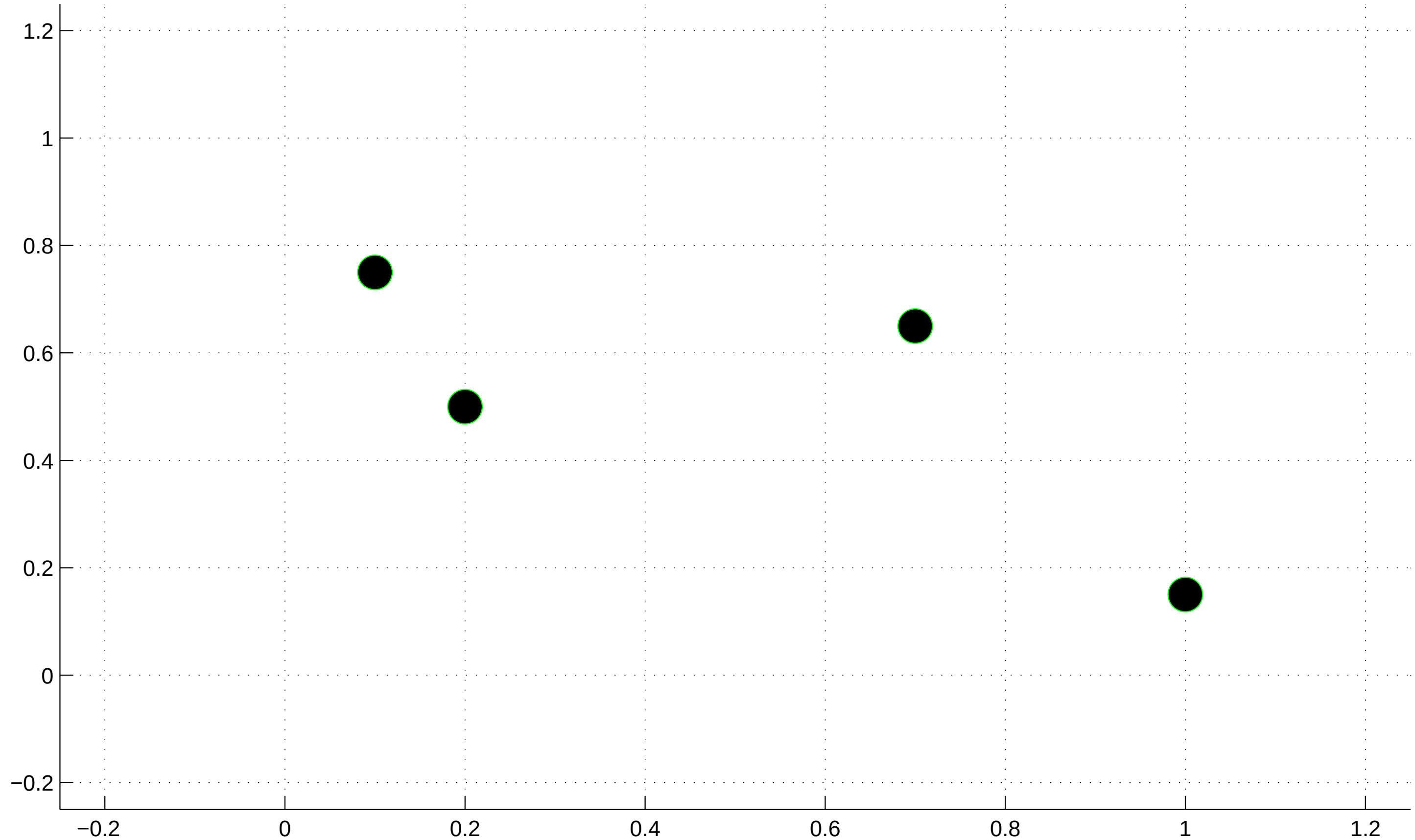
The Wasserstein Barycenter Problem

Marco Cuturi

mcuturi@i.kyoto-u.ac.jp

Joint work with G. Peyré, G. Carlier, J.D. Benamou, L. Nenna, A. Gramfort, J. Solomon, ...

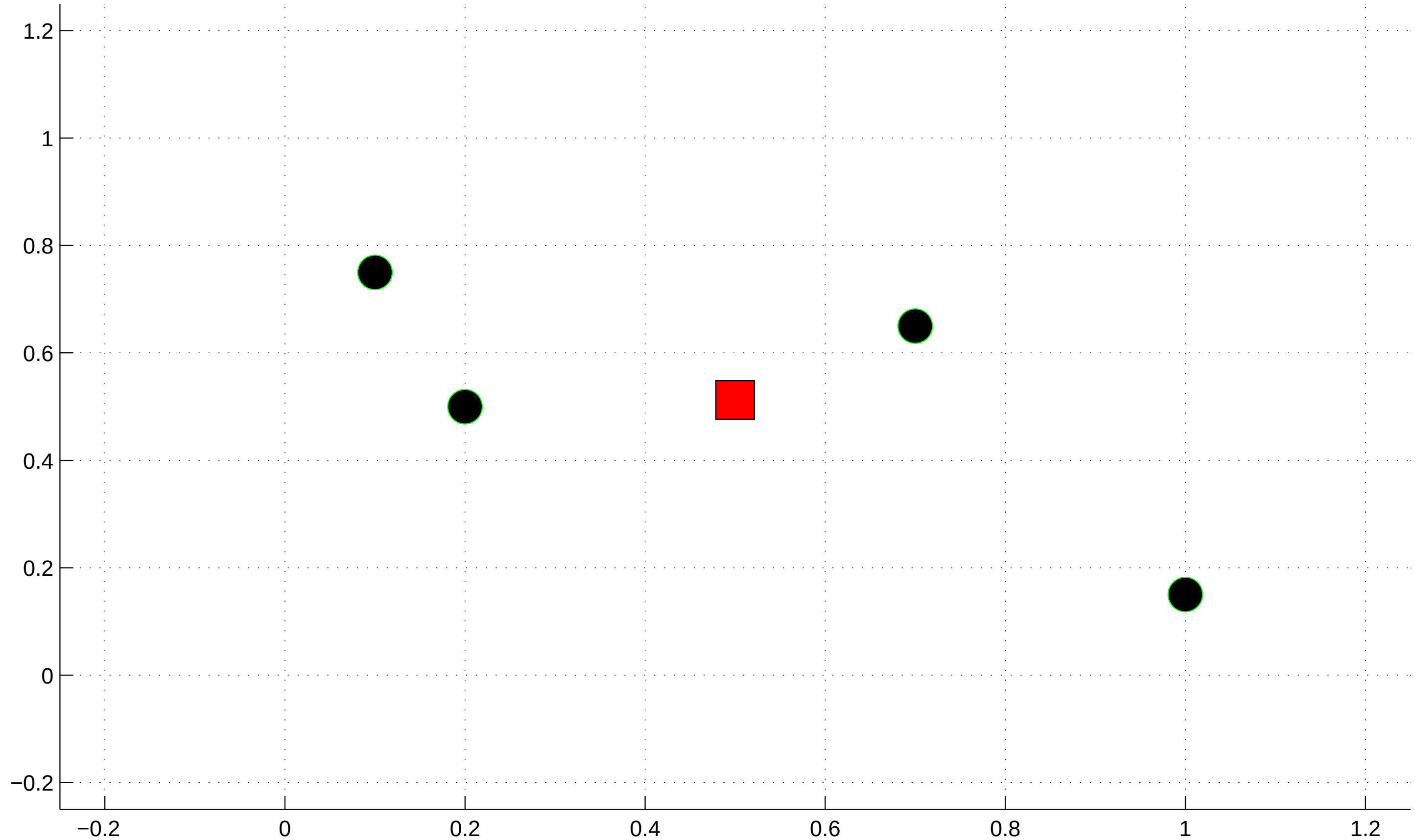
Motivation



4 points in $\mathbb{R}^2$

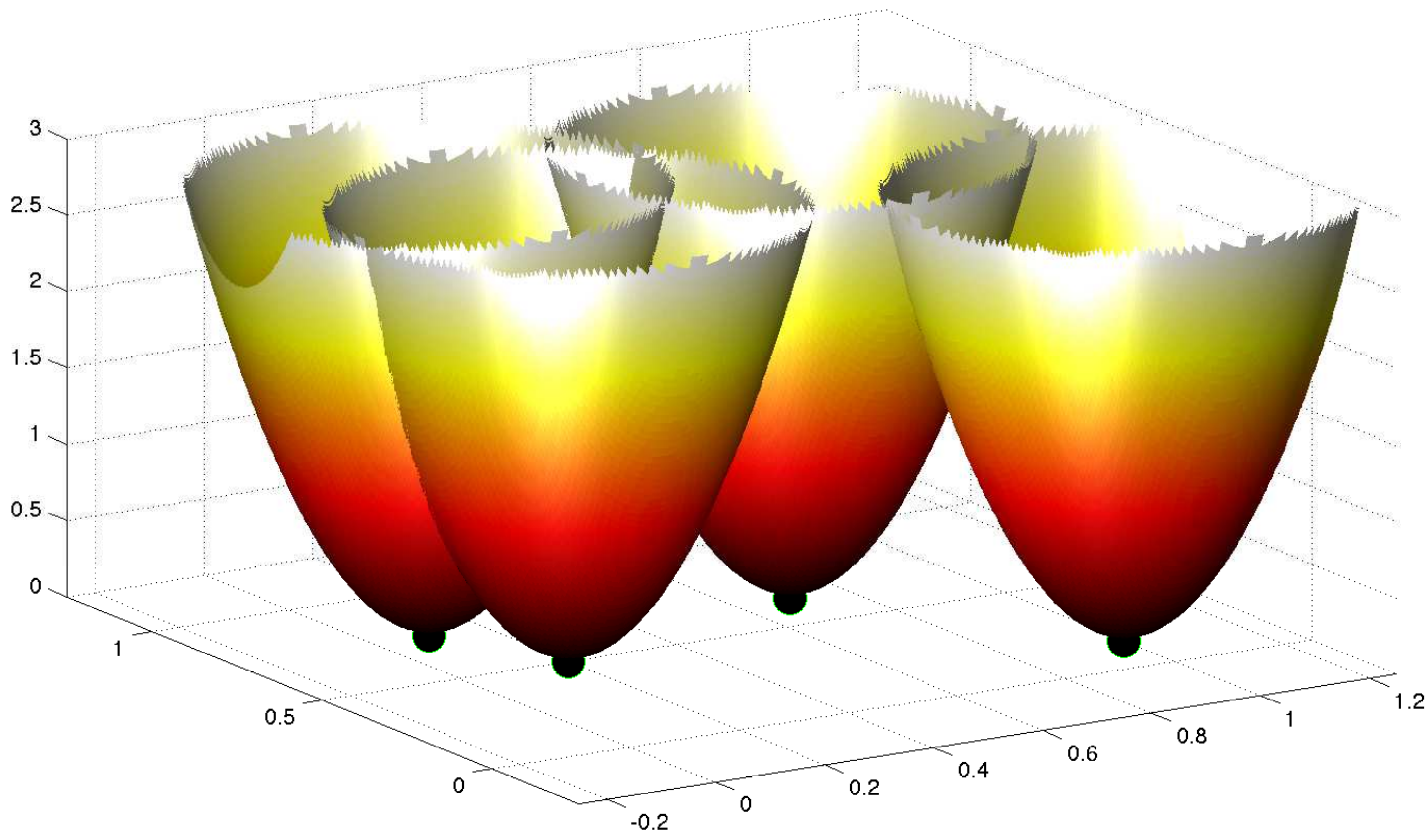
$$x_1, x_2, x_3, x_4$$

Mean



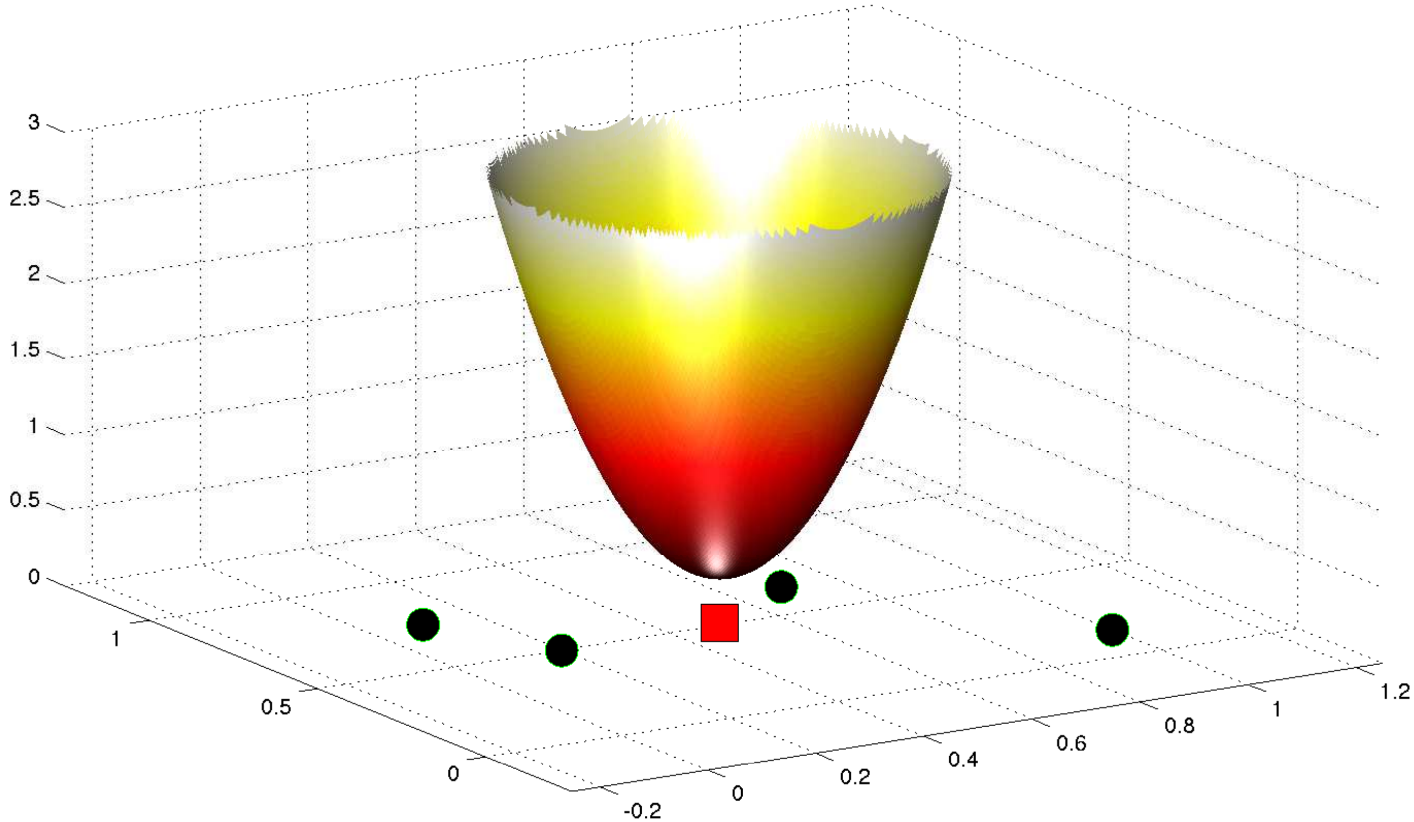
Their **mean** is $(x_1 + x_2 + x_3 + x_4)/4$.

Computing Means



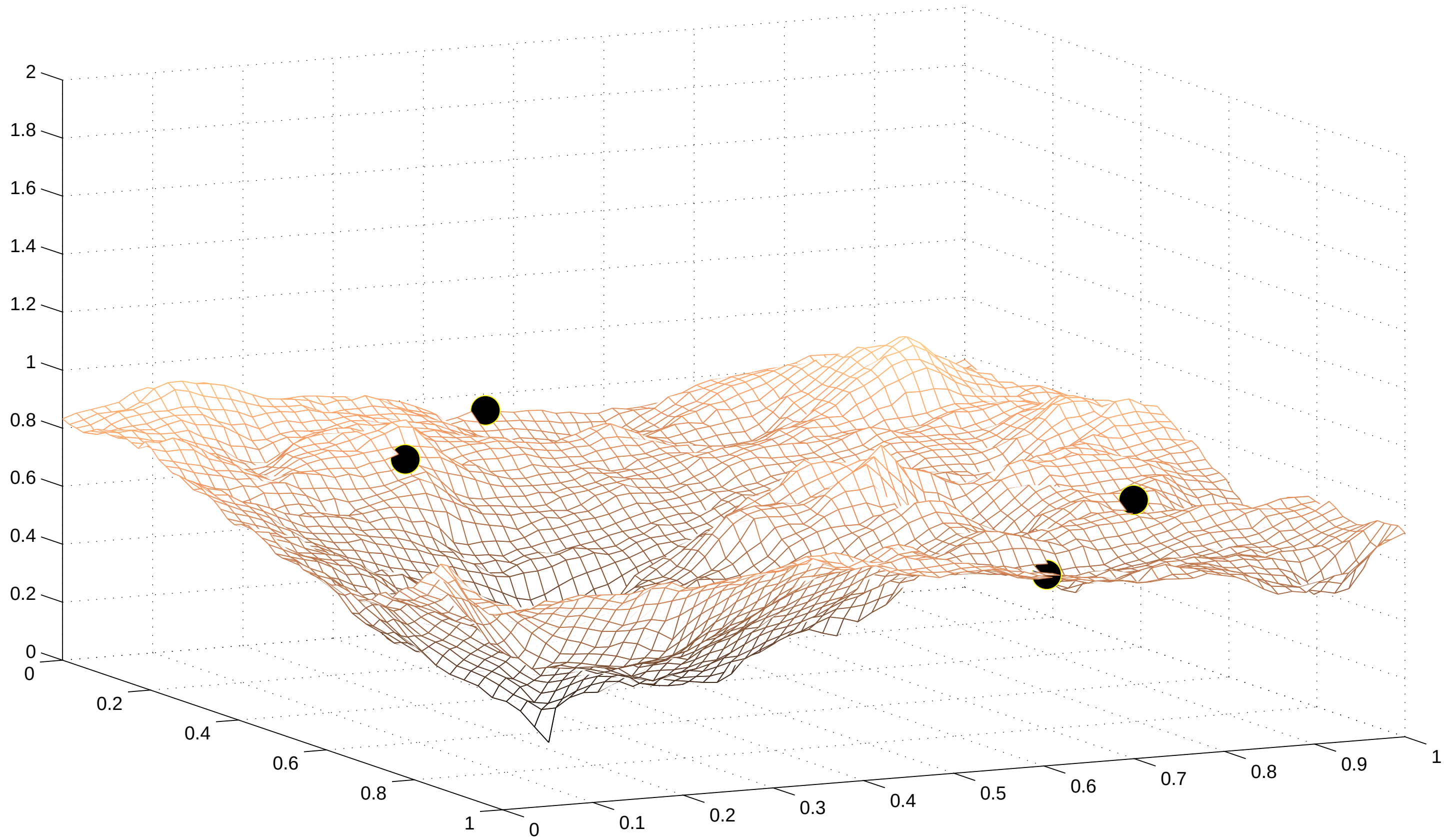
Consider for each point the function $\|\cdot - x_i\|_2^2$

Computing Means



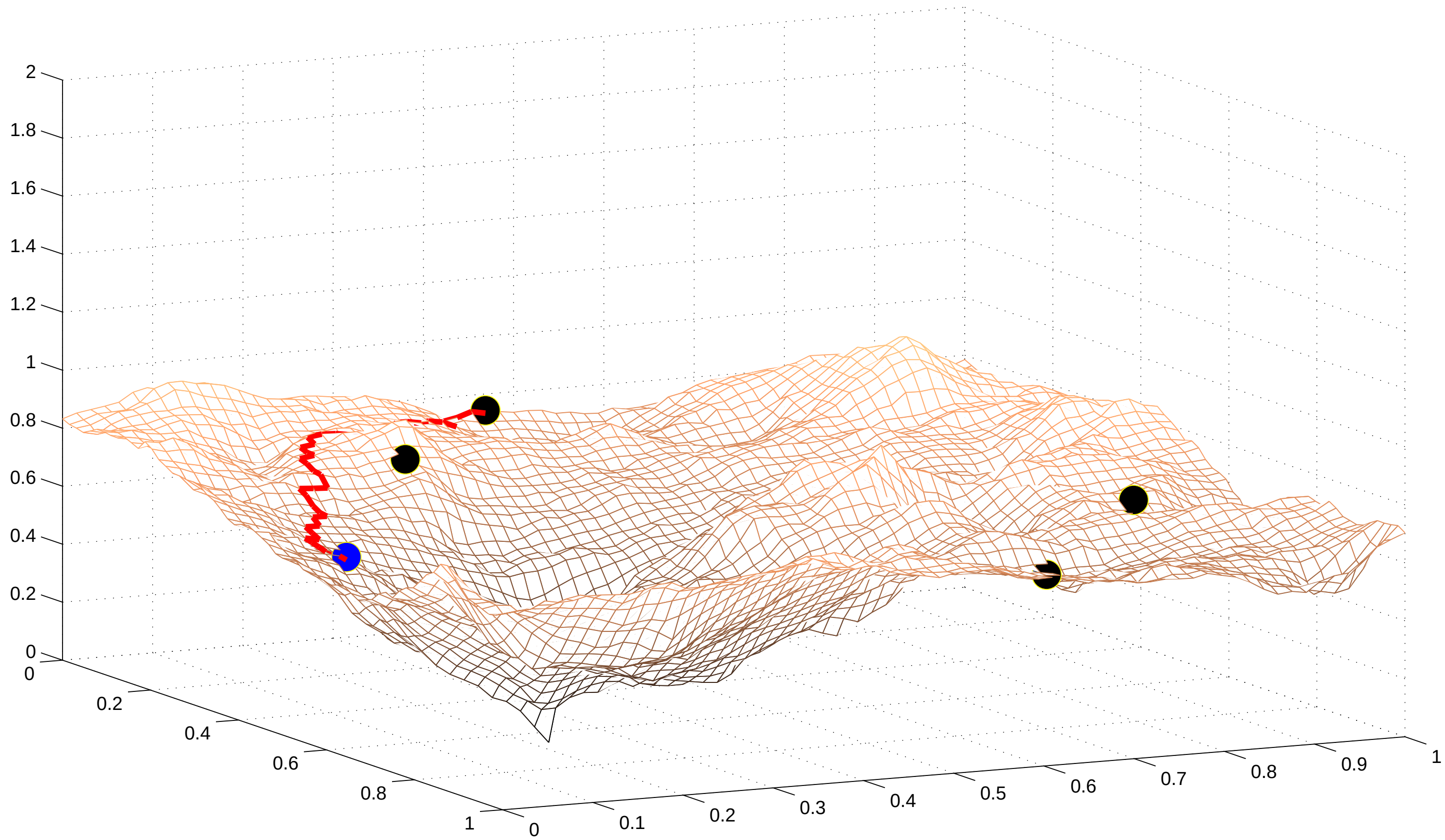
The **mean** is the argmin $\frac{1}{4}\sum_{i=1}^{4}\|\cdot - x_i\|_2^2$.

Means in Metric Spaces



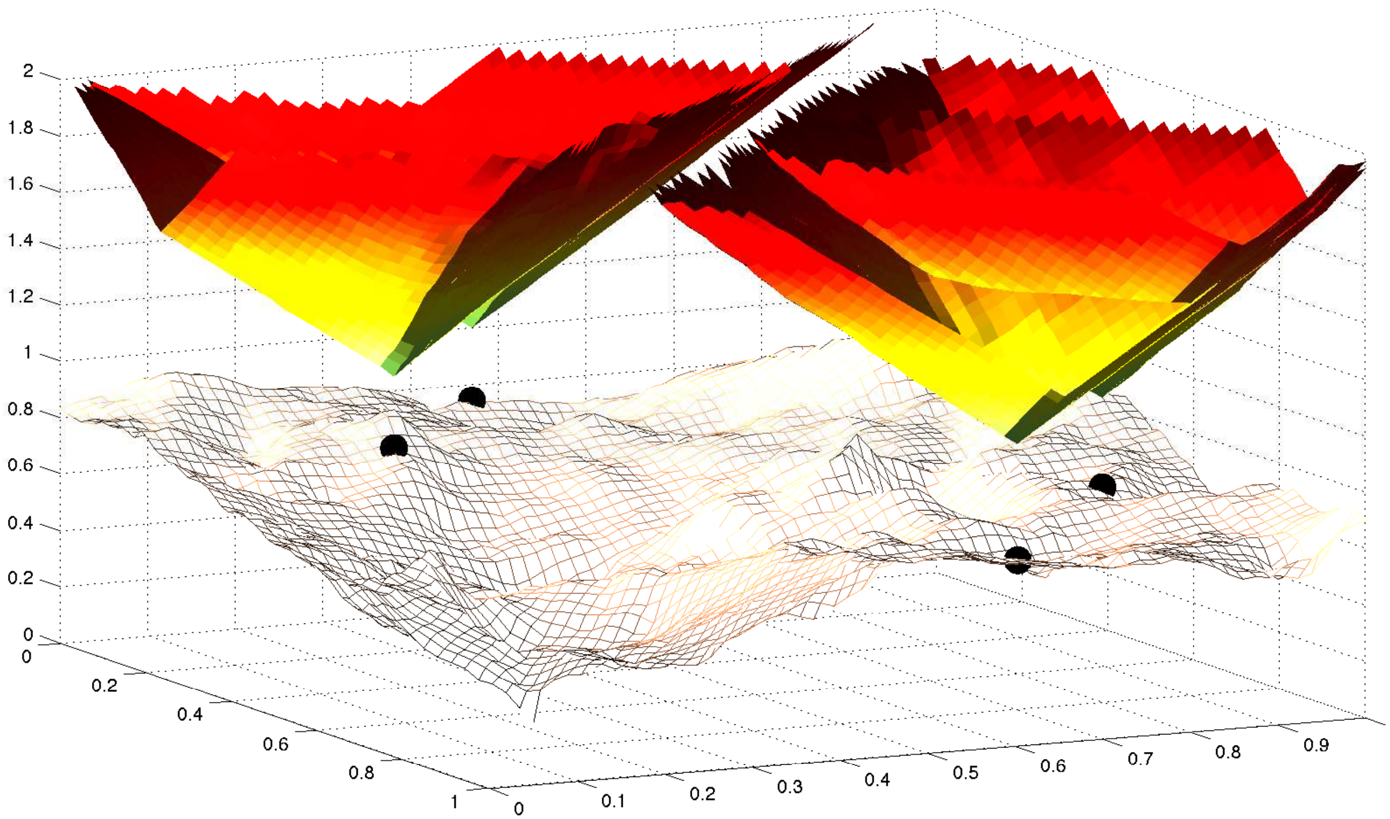
Means can be defined using any distance/divergence/discrepancy.

Means in Metric Spaces



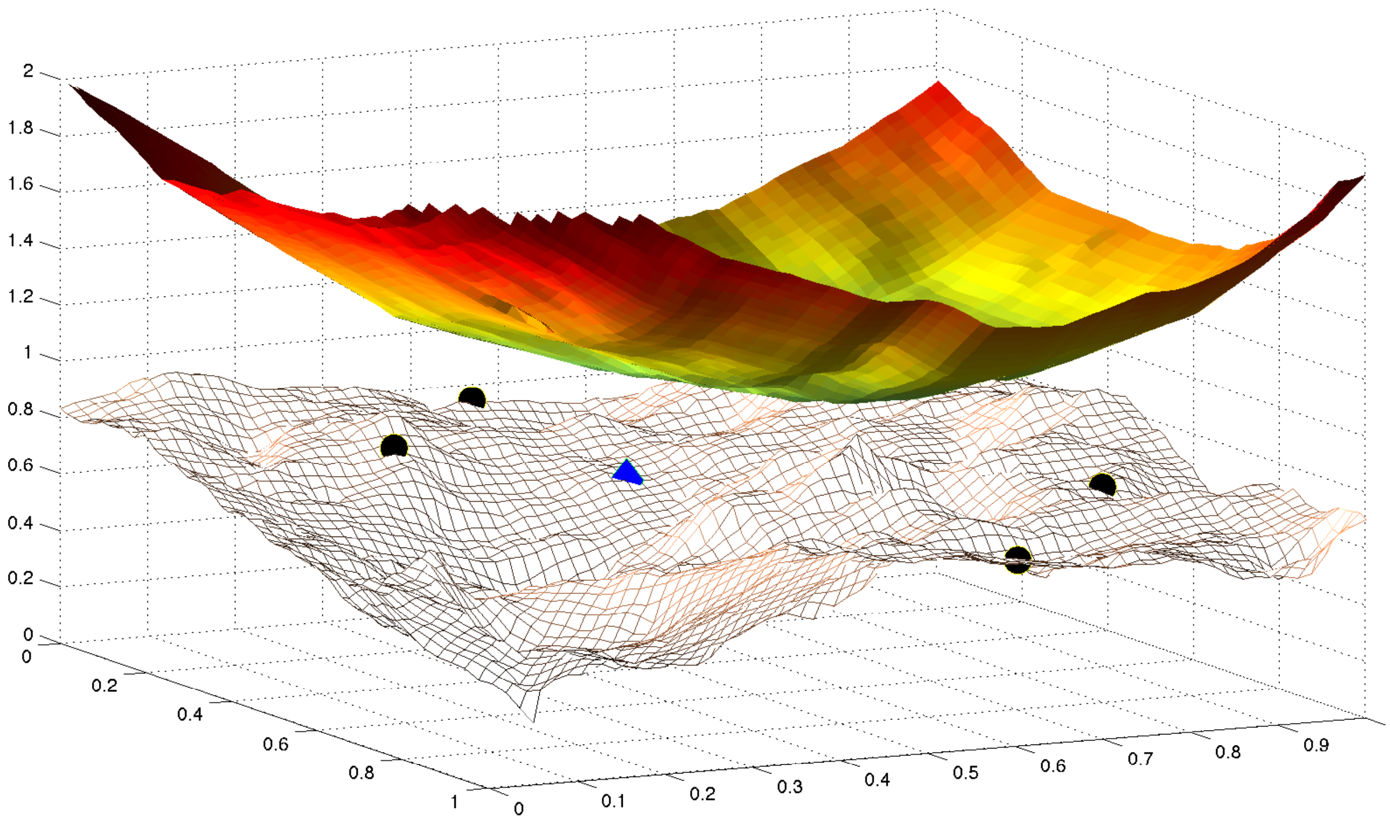
Using *e.g.* geodesic distances. Here $\Delta(\bullet, \bullet) = 0.994$

Means in Metric Spaces



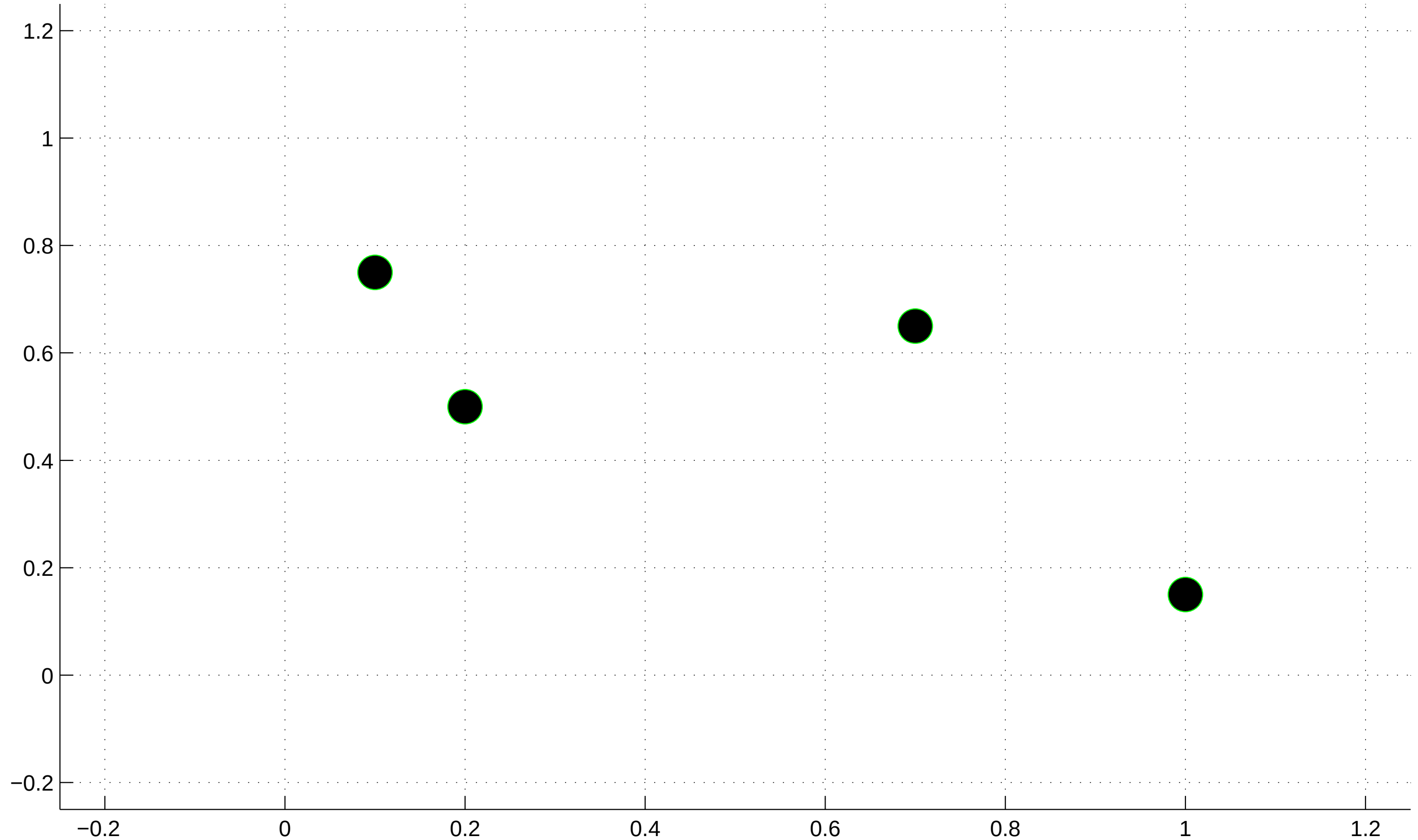
Consider the distance functions $\Delta(\cdot, x_i), i = 1, 2, 3, 4.$

Means in Metric Spaces

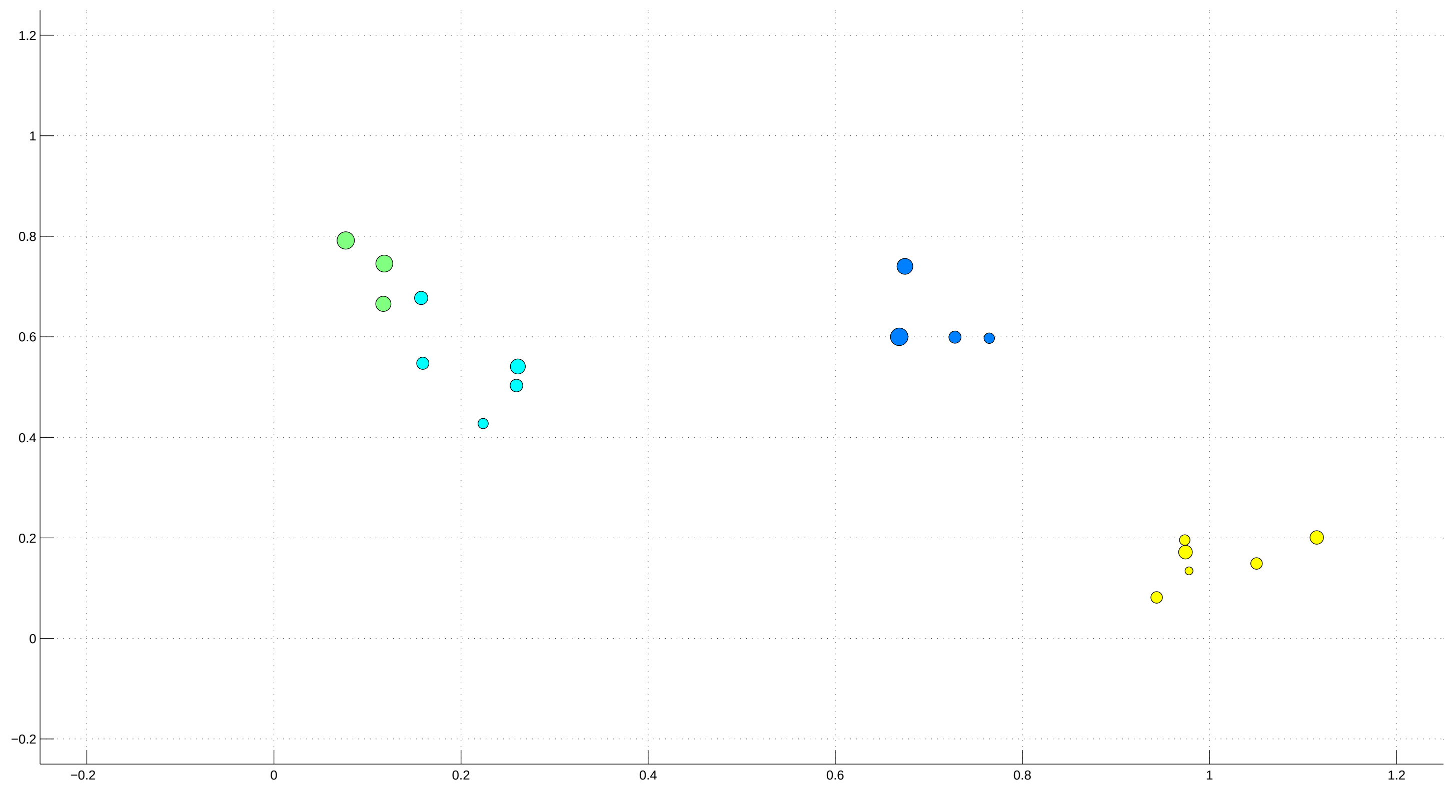


$$\blacklozenge = \operatorname{argmin} \frac{1}{N} \sum_{i=1}^{N} \Delta(\cdot, x_i).$$

From points

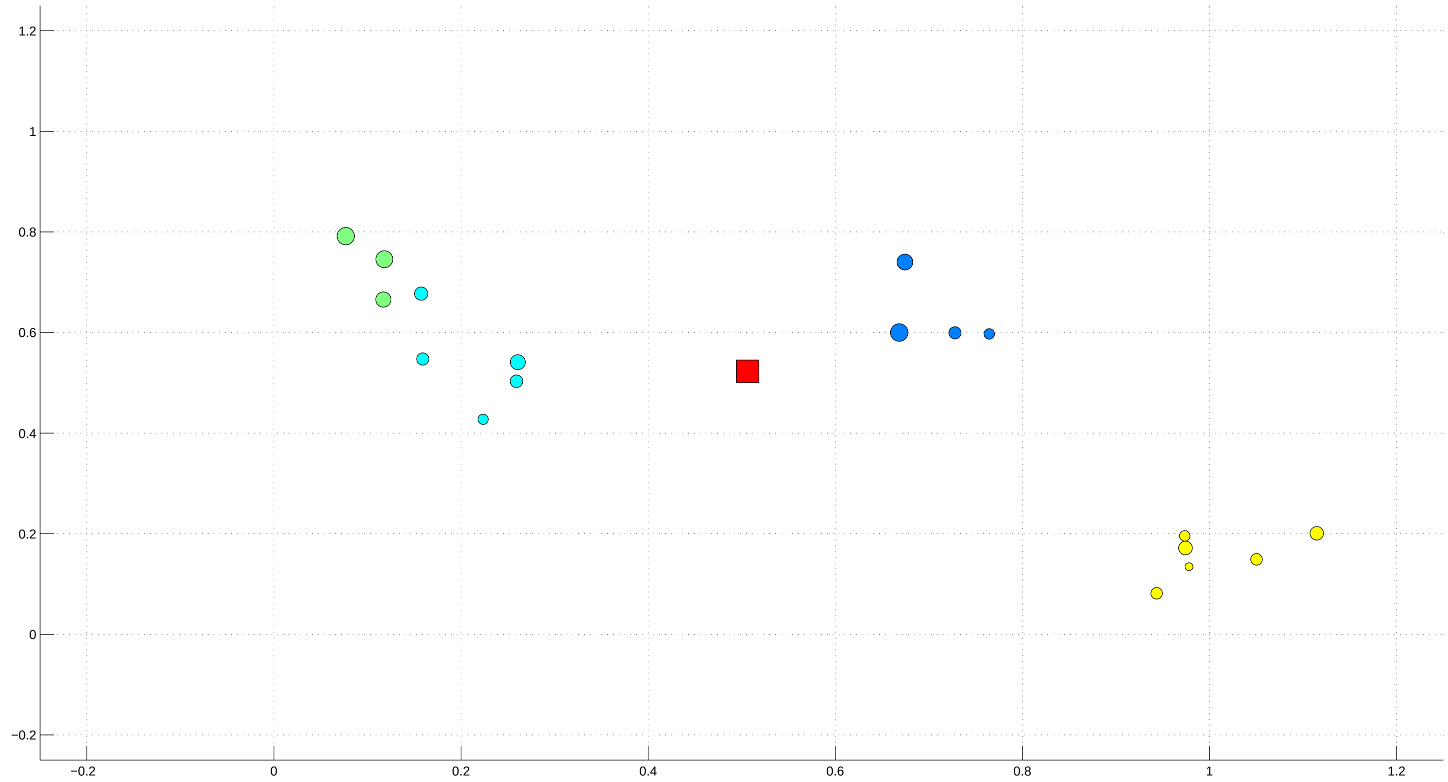


to Probability Measures



Assume that each datum is now an **empirical measure**.
What could be the mean of these 4 measures?

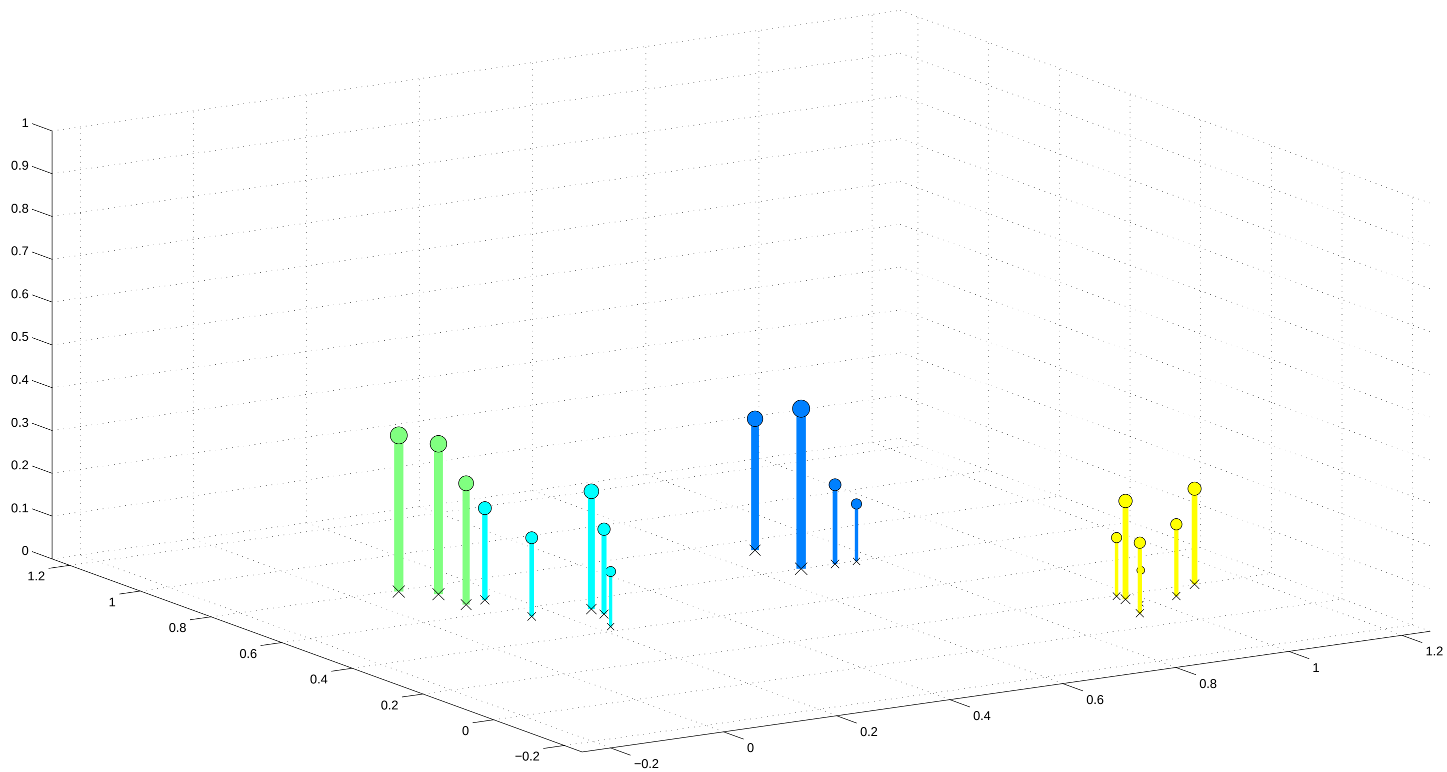
1. Naive Averaging



■ = naive mean of *all* observations.

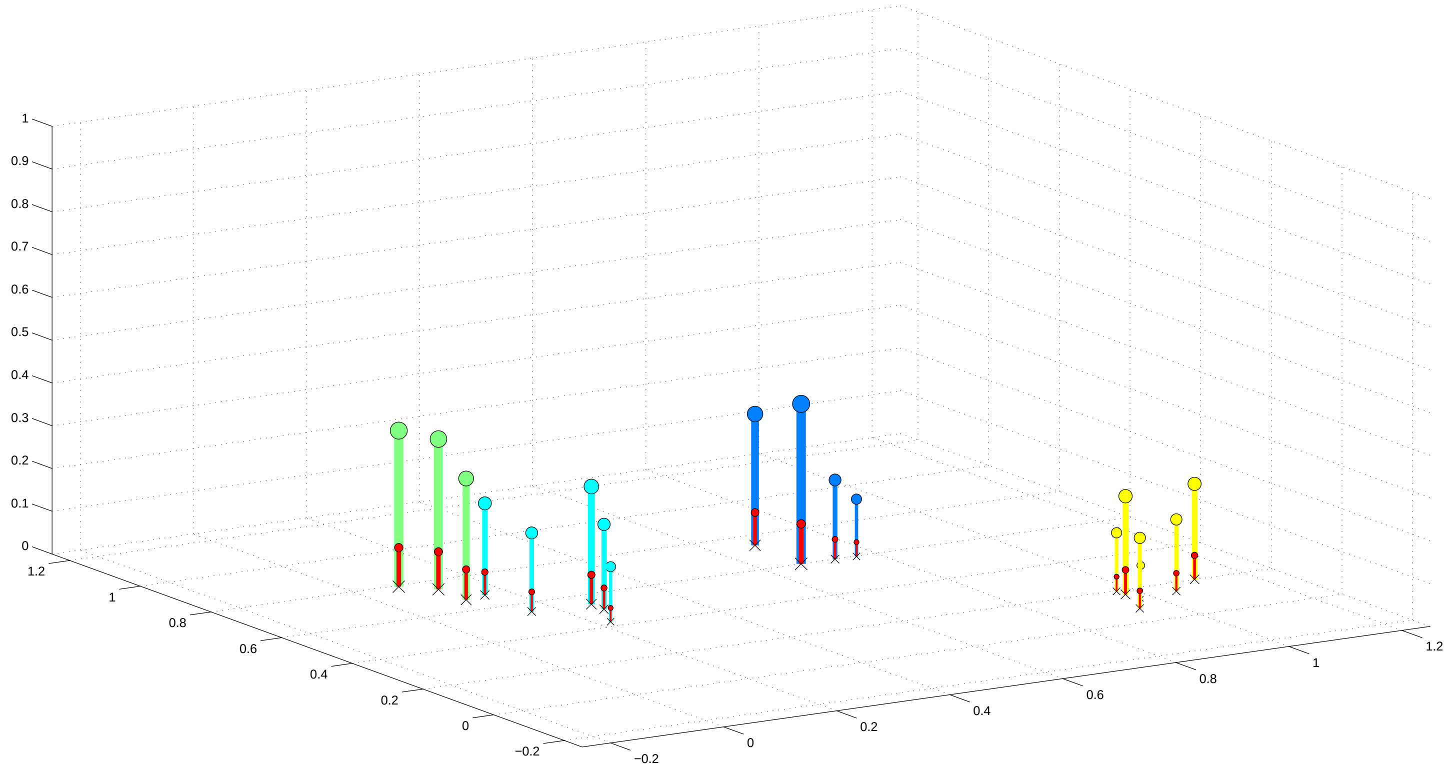
Mean of 4 measures = a point?

Averaging Probability Measures



Same measures, in a 3D perspective.

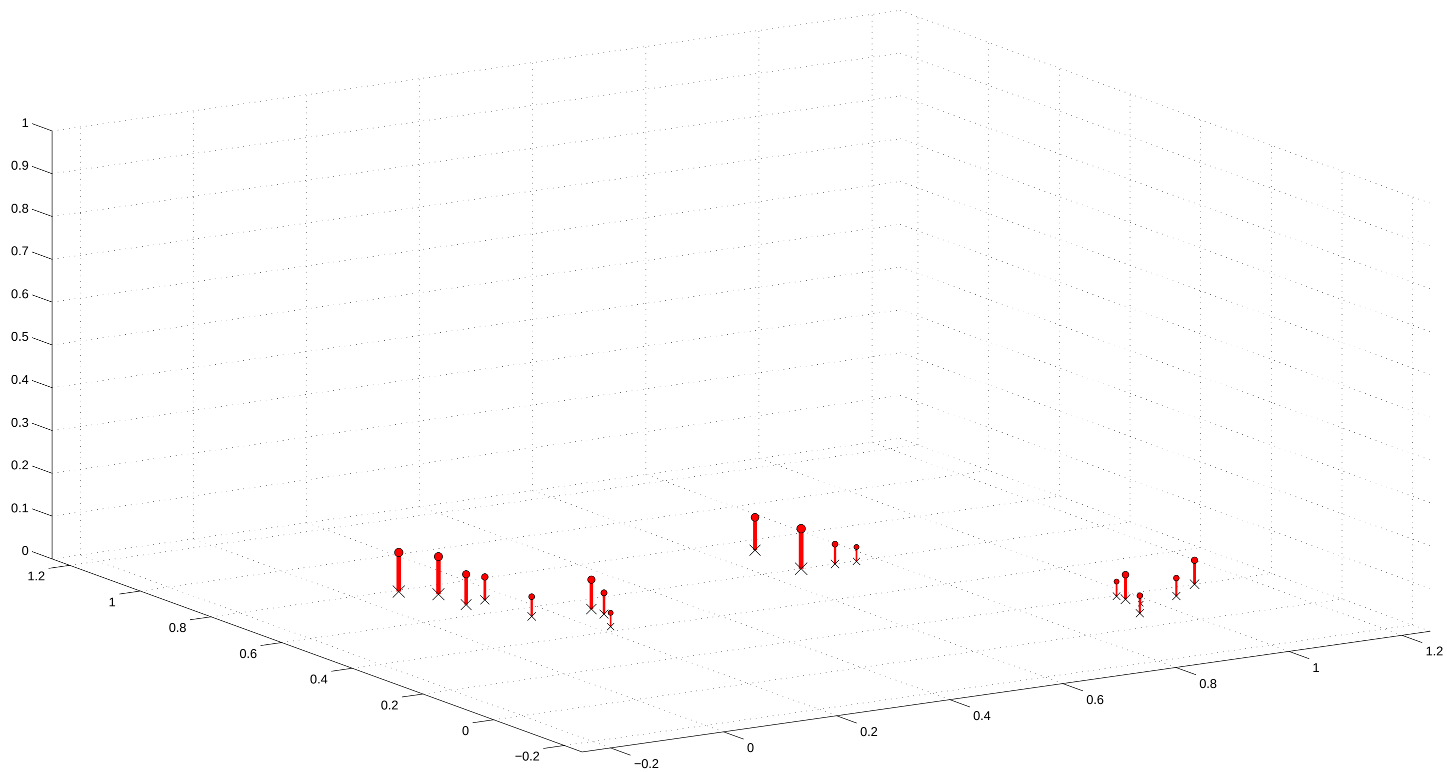
2. Naive Averaging



Euclidean mean of measures is their sum / N.

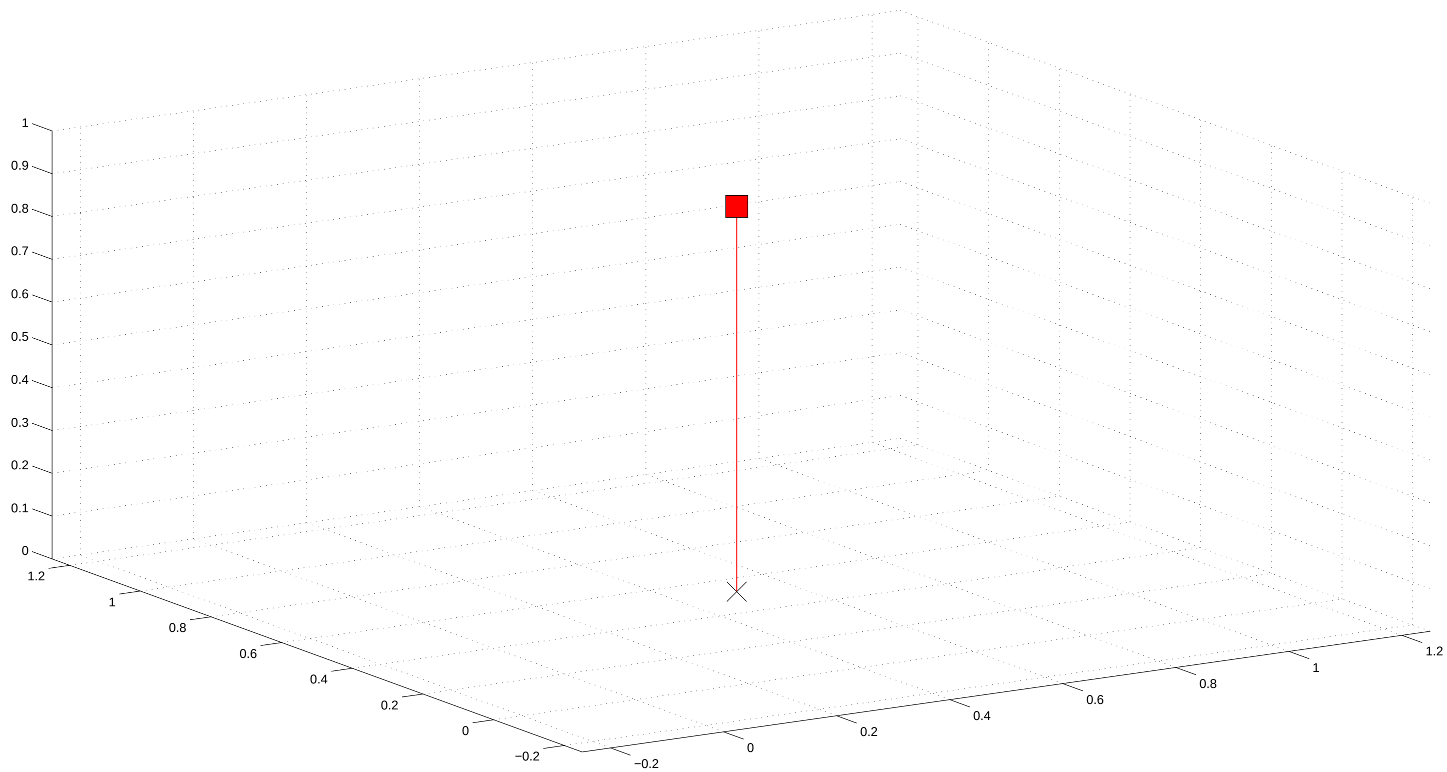
Here, $\Delta(\mu, \nu) = \int_{\mathbb{R}^2}[d\mu - d\nu]^2$.

Focus on uncertainty



...but geometric knowledge ignored.

Focus on geometry



...but uncertainty is lost.

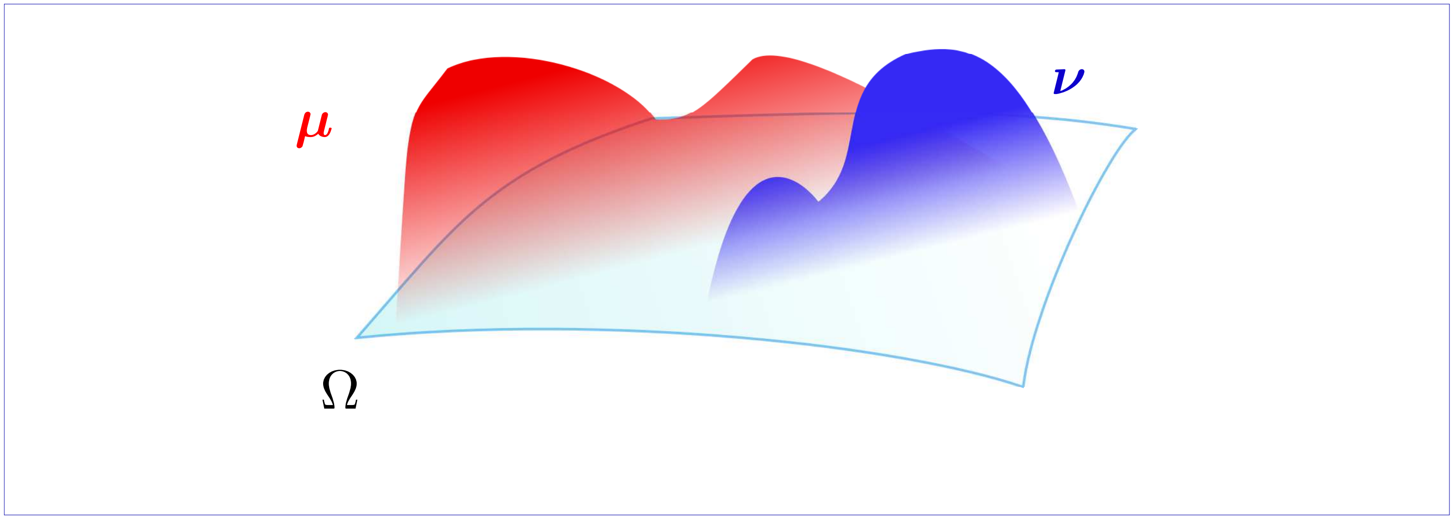
Problem of interest

Given a discrepancy function Δ between probabilities, compute **their mean:** $\text{argmin} \sum_i \Delta(\cdot, \nu_i)$

- The idea is useful, sometimes tractable & appears in
  - Bregman clustering for histograms **[Banerjee'05]..**
  - Topic modeling **[Blei & al.'03]..**
  - Clustering problems (k-means).
- Our goal in this talk: study the case $\Delta =$ Wasserstein

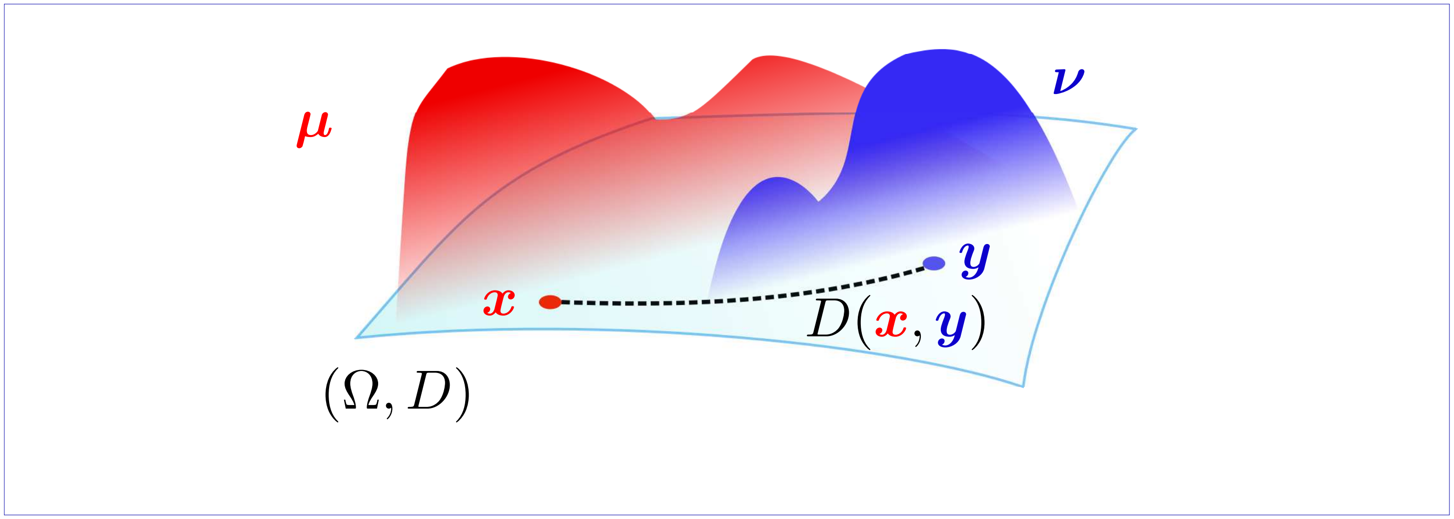
Wasserstein Distances

Comparing Two Measures



Two measures $\mu, \nu \in P(\Omega)$.

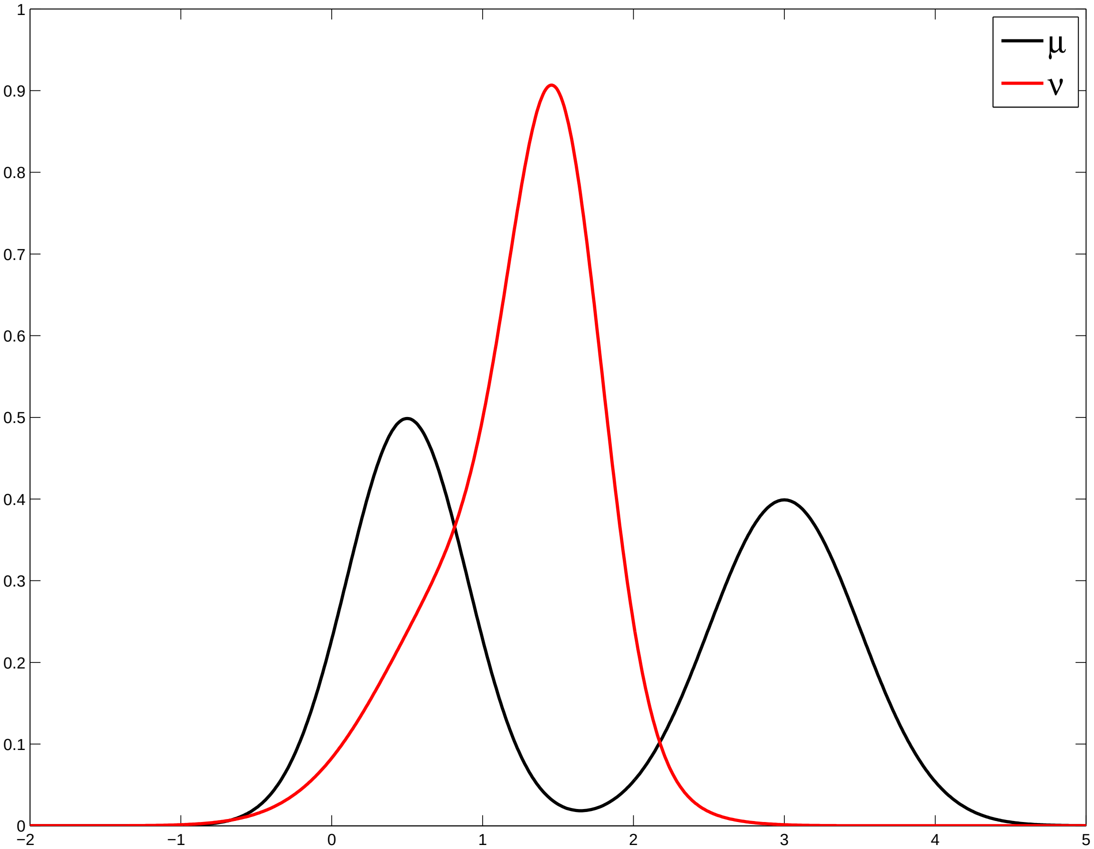
The Optimal Transport Approach



Optimal Transport distances rely **on 2 key concepts**:

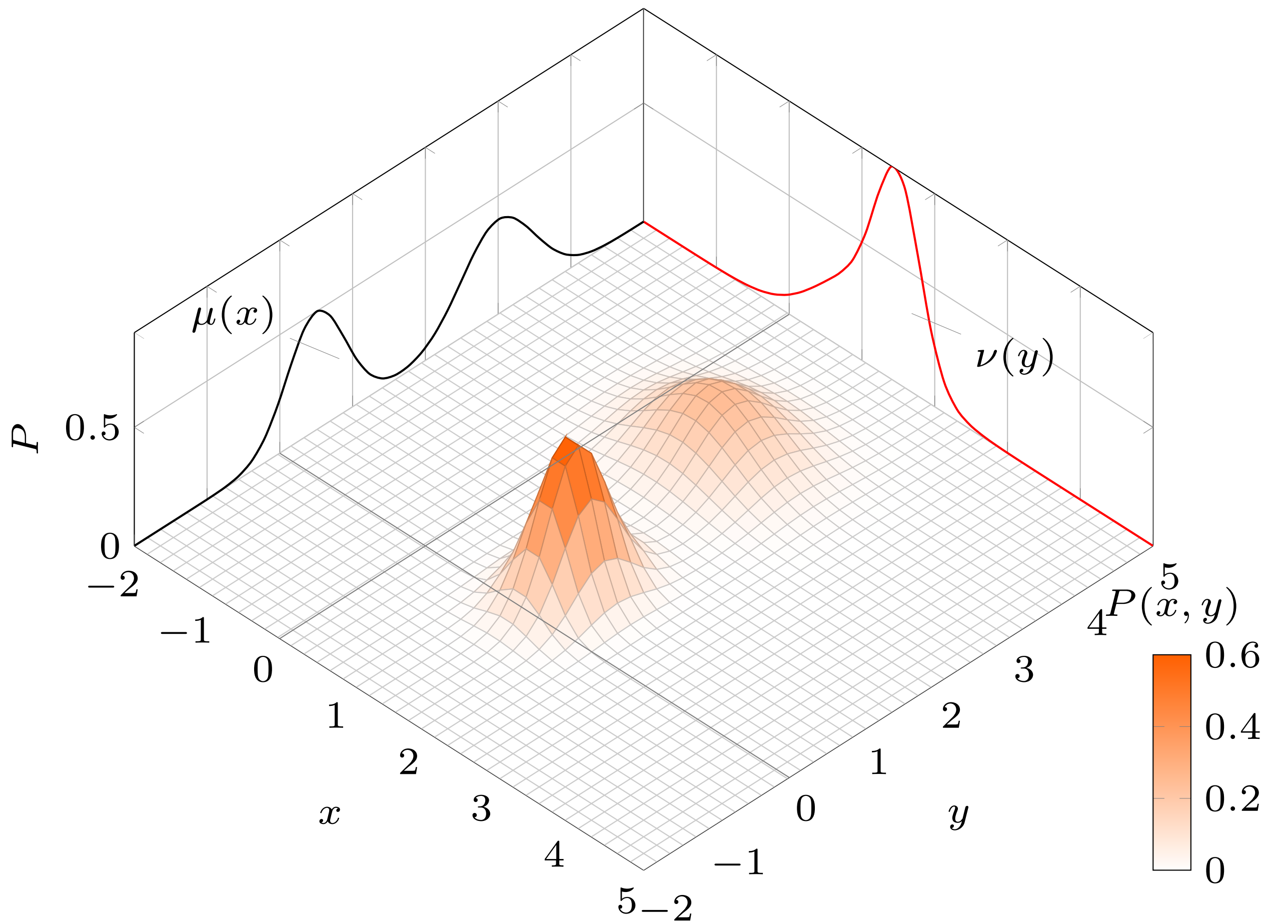
- A **metric** $D : \Omega \times \Omega \to \mathbb{R}_+$;
- $\Pi(\mu, \nu)$: **joint probabilities** with marginals μ, ν.

Joint Probabilities of (μ, ν)



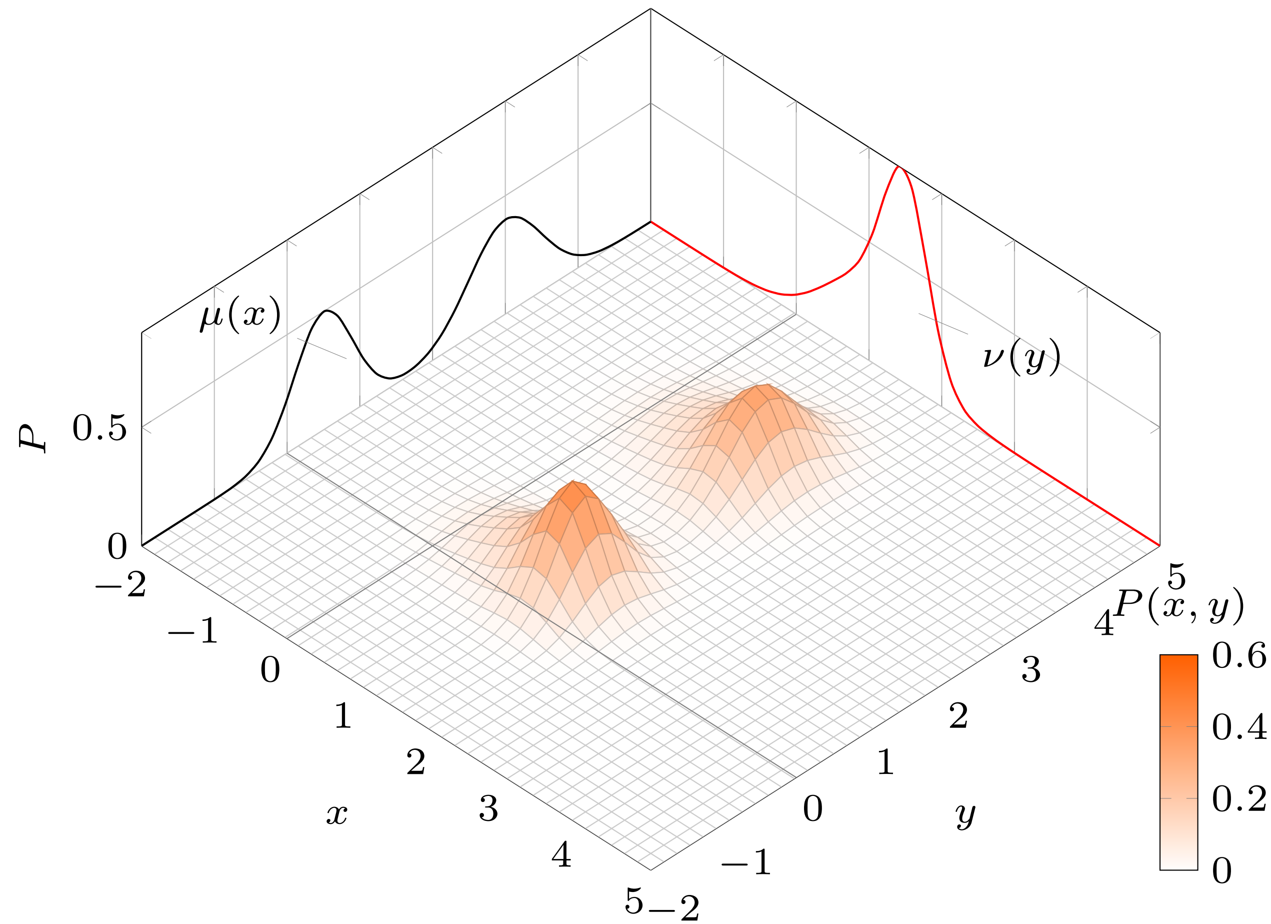
Consider μ, ν two measures on the real line.

Joint Probabilities of (μ, ν)



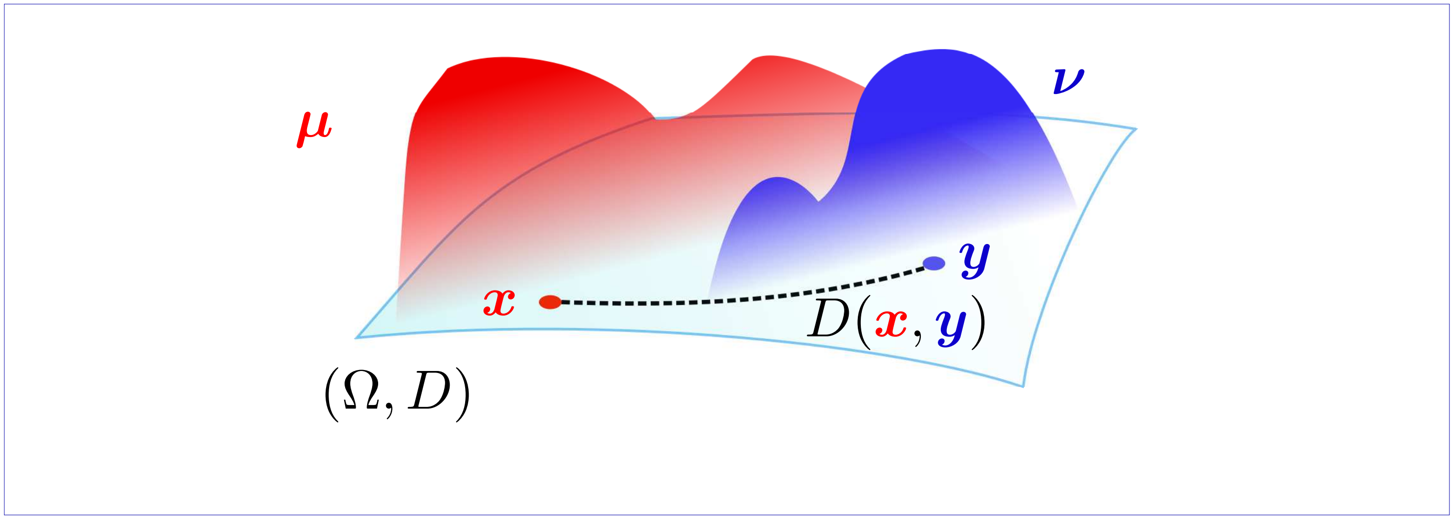
$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ = probability measures on Ω^2
with marginals $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

Joint Probabilities of (μ, ν)



$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \text{probability measures on } \Omega^2$$
with marginals $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

Optimal Transport Distance



p-Wasserstein (or OT) distance, assuming $p \geq 1$, is:

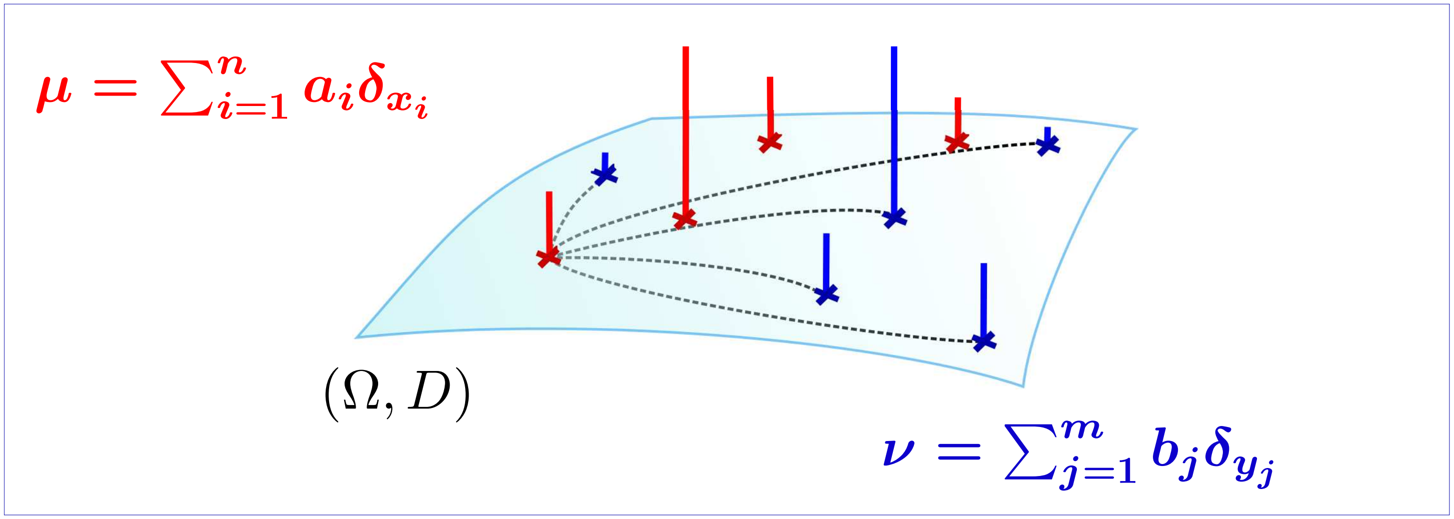
$$W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left(\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathbb{E}_{\boldsymbol{P}}[D(X, Y)^p] \right)^{1/p}.$$

(Historical Parenthesis)

Monge-Kantorovich, Kantorovich-Rubinstein, Wasserstein, Earth Mover's Distance, Mallows

- Monge 1781 *Mémoire sur la théorie des déblais et des remblais*
- **Optimization & Operations Research**
  - Kantorovich'42, Dantzig'47, Ford Fulkerson'55, $etc.$
- **Probability & Statistical Physics**
  - Rachev'92, Talagrand'96, Villani'09
- **Computer Vision**: Rubner et al'98

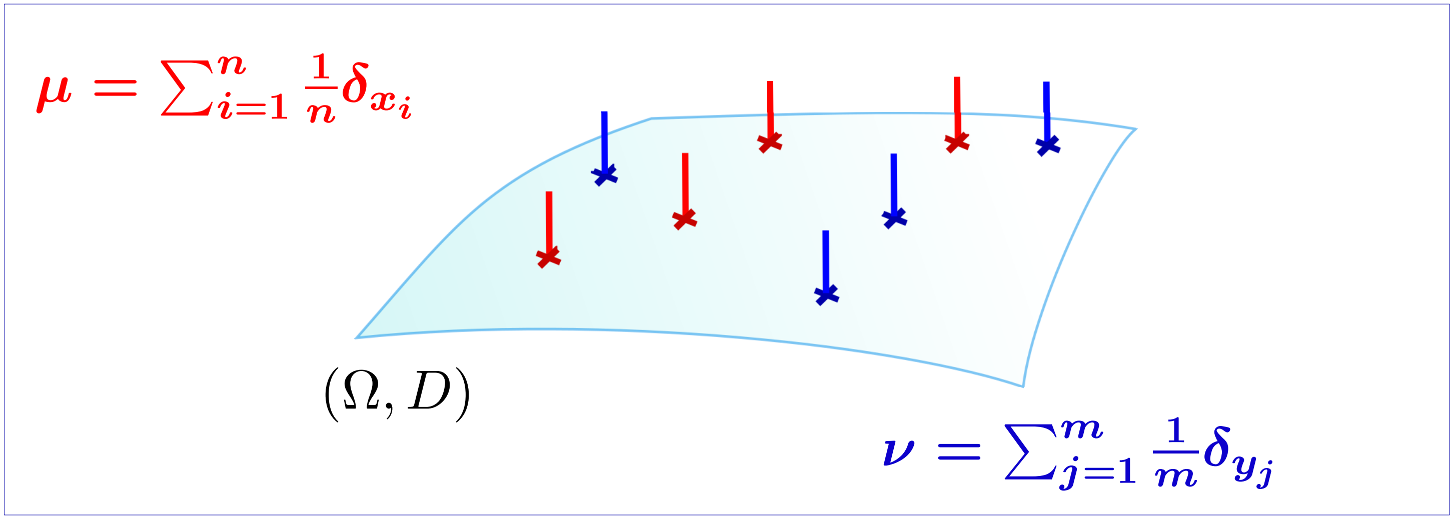
OT Distance for Empirical Measures



$$W_p(\mu, \nu) = \left(\inf_{P \in \Pi(\mu,\nu)} \mathbb{E}_P[D(X,Y)^p] \right)^{1/p}.$$

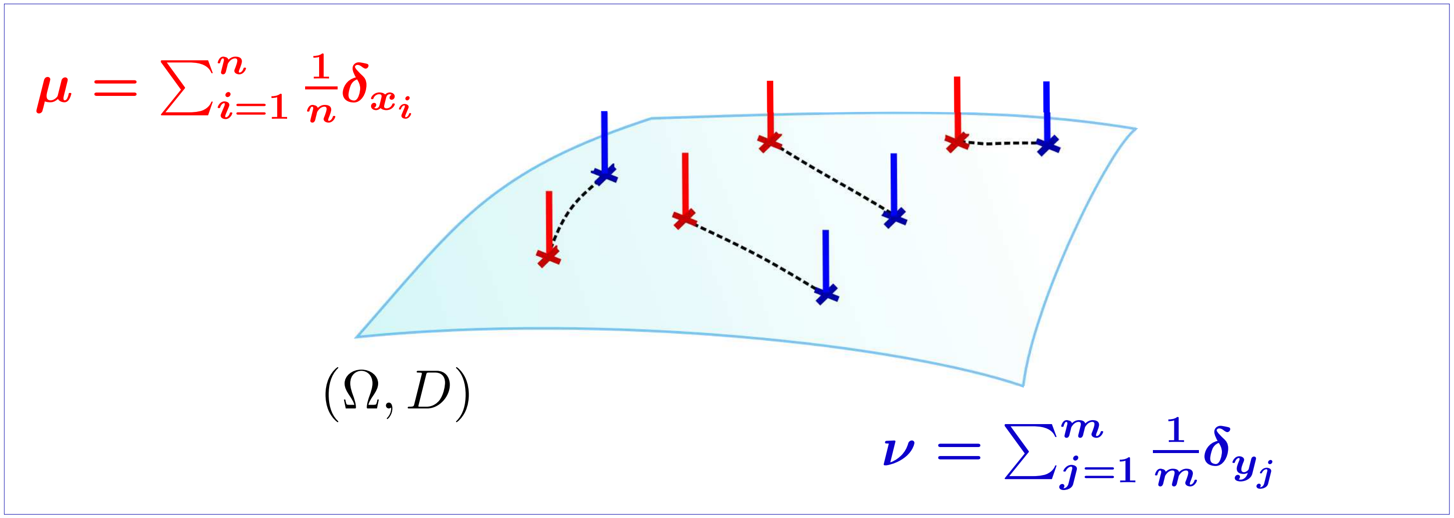
Algorithmically?

OT Distance for Empirical Measures



Suppose $n = m$ and all weights are uniform

OT Distance for Empirical Measures

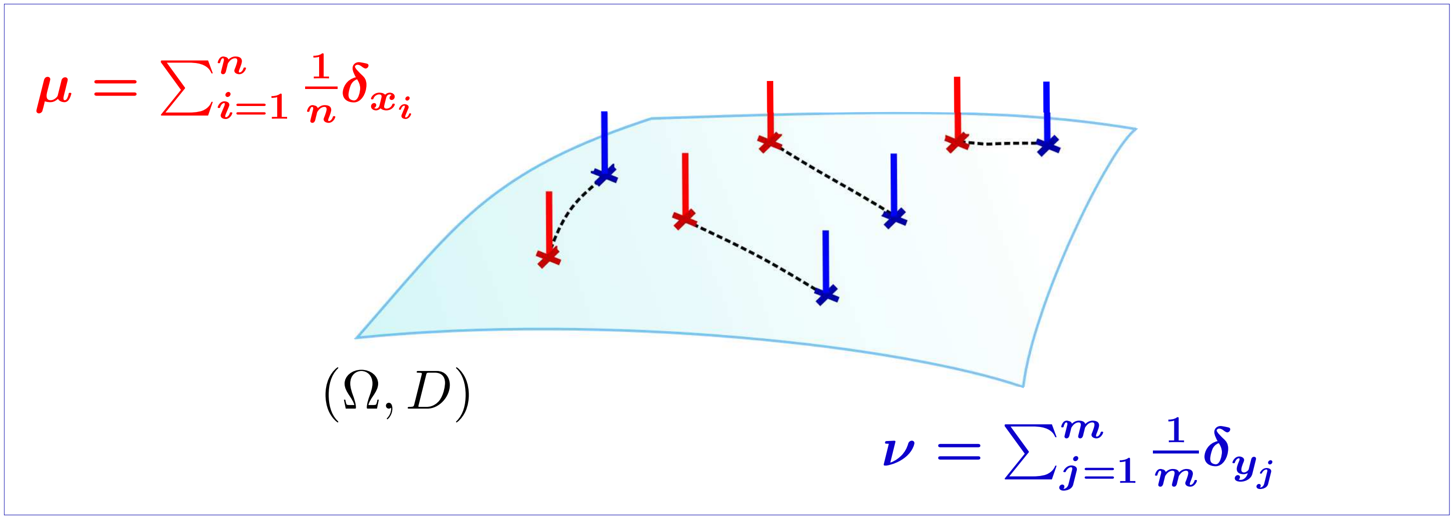


Then W_p^p = *optimal matching cost*

(solved for instance with Hungarian algorithm)

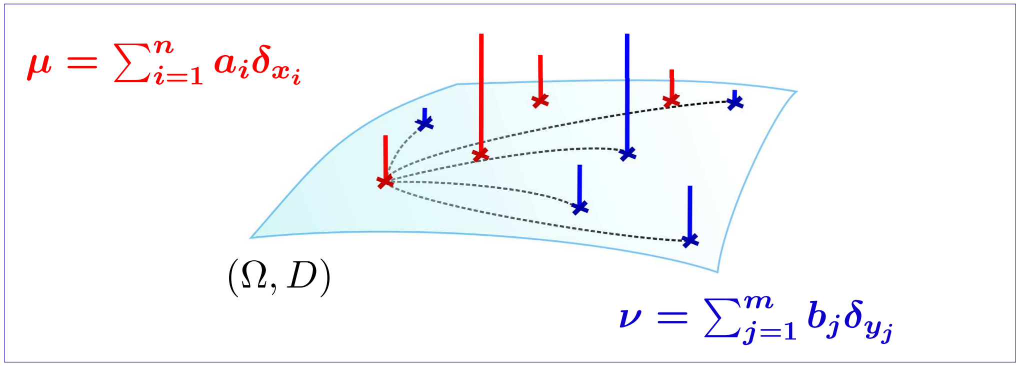
$$\left(\min_{\sigma \in S_n} \frac{1}{n} \sum_{i=1}^{n} D(x_i, y_{\sigma_i})^p\right)^{1/p}$$

OT Distance for Empirical Measures



As soon as $n \neq m$ or weights are non uniform, optimal matching does not make sense.

Computing the OT Distance



$W_p^p(\mu, \nu)$ can be cast as a linear program in $\mathbb{R}^{n \times m}$:

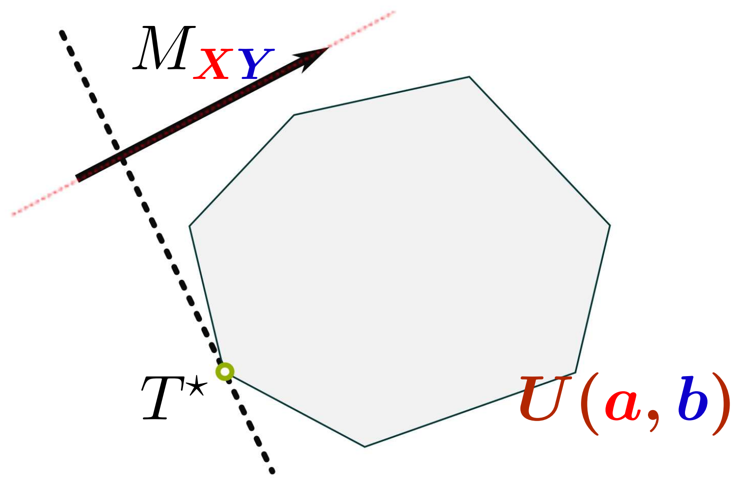
1. $M_{XY} \stackrel{\text{def}}{=} [D(x_i, y_j)^p]_{ij} \in \mathbb{R}^{n \times m}$ ***(metric information)***

2. Transportation Polytope ***(joint probabilities)***

$$U(a, b) = \{P \in \mathbb{R}_+^{n \times m} \mid P\mathbb{1}_m = a, P^T\mathbb{1}_n = b\}$$

Computing p-Wasserstein Distances

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \mathbf{primal}(\boldsymbol{a},\boldsymbol{b},M_{\mathbf{XY}}) \stackrel{\text{def}}{=} \min_{T\in U(\boldsymbol{a},\boldsymbol{b})} \langle T, M_{\mathbf{XY}} \rangle$$



$$\begin{aligned} W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) &= \langle T^\star, M_{\mathbf{XY}}\rangle \\ &= \min_{T\in U(\boldsymbol{a},\boldsymbol{b})} \langle T, M_{\mathbf{XY}}\rangle \end{aligned}$$

[Kantorovich'42] Duality

- This primal problem has an equivalent, dual LP:

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \begin{cases} \textbf{primal}(\boldsymbol{a}, \boldsymbol{b}, M_{\boldsymbol{XY}}) \overset{\text{def}}{=} \min\limits_{T \in U(\boldsymbol{a}, \boldsymbol{b})} \langle T, M_{\boldsymbol{XY}} \rangle \\ \text{or} \\ \textbf{dual}(\boldsymbol{a}, \boldsymbol{b}, M_{\boldsymbol{XY}}) \overset{\text{def}}{=} \max\limits_{(\alpha,\beta) \in C_{M_{\boldsymbol{XY}}}} \alpha^T \boldsymbol{a} + \beta^T \boldsymbol{b}\,, \\ \text{where } C_M = \{(\alpha, \beta) \in \mathbb{R}^{n+m} \,|\, \alpha_i + \beta_j \leq M_{ij}\} \end{cases}$$

[Kantorovich'42] Duality

- This primal problem has an equivalent, dual LP:

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \begin{cases} \mathbf{primal}(\boldsymbol{a},\boldsymbol{b},M_{\mathbf{XY}}) \stackrel{\text{def}}{=} \min_{T\in U(\boldsymbol{a},\boldsymbol{b})} \langle T, M_{\mathbf{XY}} \rangle \\ \text{or} \\ \mathbf{dual}(\boldsymbol{a},\boldsymbol{b},M_{\mathbf{XY}}) \stackrel{\text{def}}{=} \max_{(\alpha,\beta)\in C_{M_{\mathbf{XY}}}} \alpha^T\boldsymbol{a} + \beta^T\boldsymbol{b}\,, \\ \text{where } C_M = \{(\alpha,\beta)\in\mathbb{R}^{n+m}\,|\,\alpha_i+\beta_j\leq M_{ij}\} \end{cases}$$

⚠ **Both problems require $O(n^3\log(n))$ operations. Typically solved using the network simplex.**

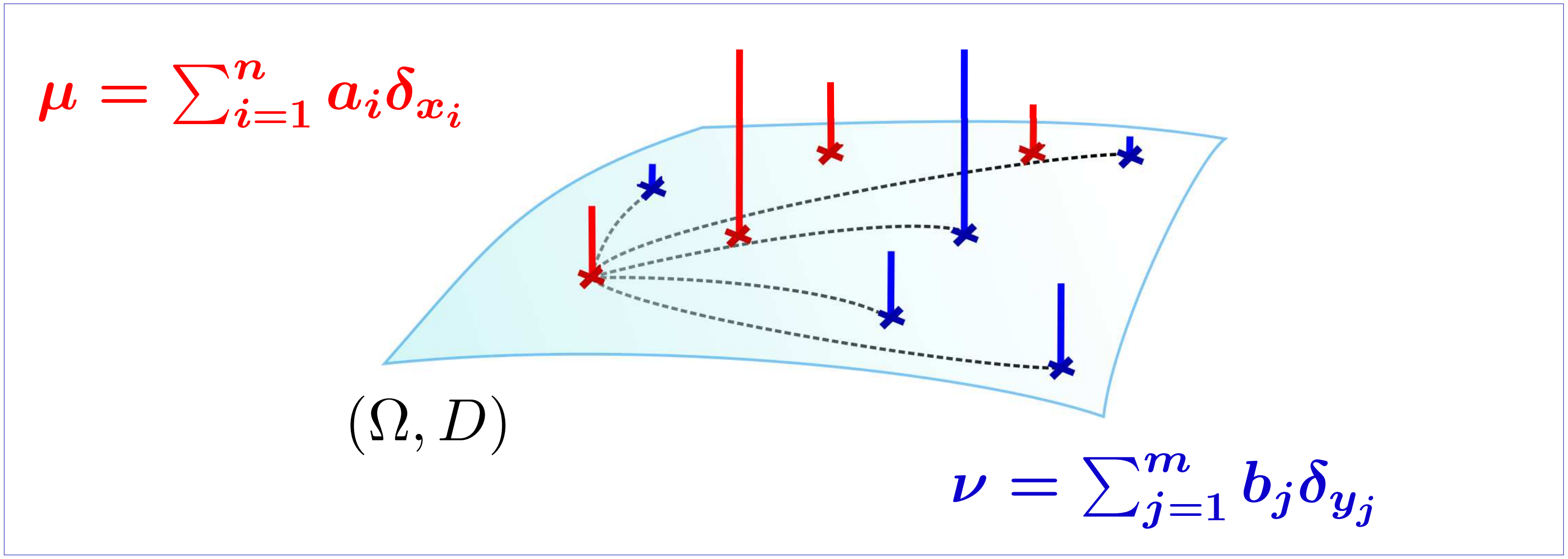
Wasserstein Barycenter Problem (WBP)

- **[Agueh'11]** introduced the WBP:

$$\operatorname*{argmin}_{\mu \in P(\Omega)} C(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \sum_{i=1}^{N} W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu_i}),$$

- Can be solved with a **multi-marginal** OT problem.
- **Intractable**: LP of $\prod_i \operatorname{card}(\operatorname{supp}(\nu_i))$ variables.

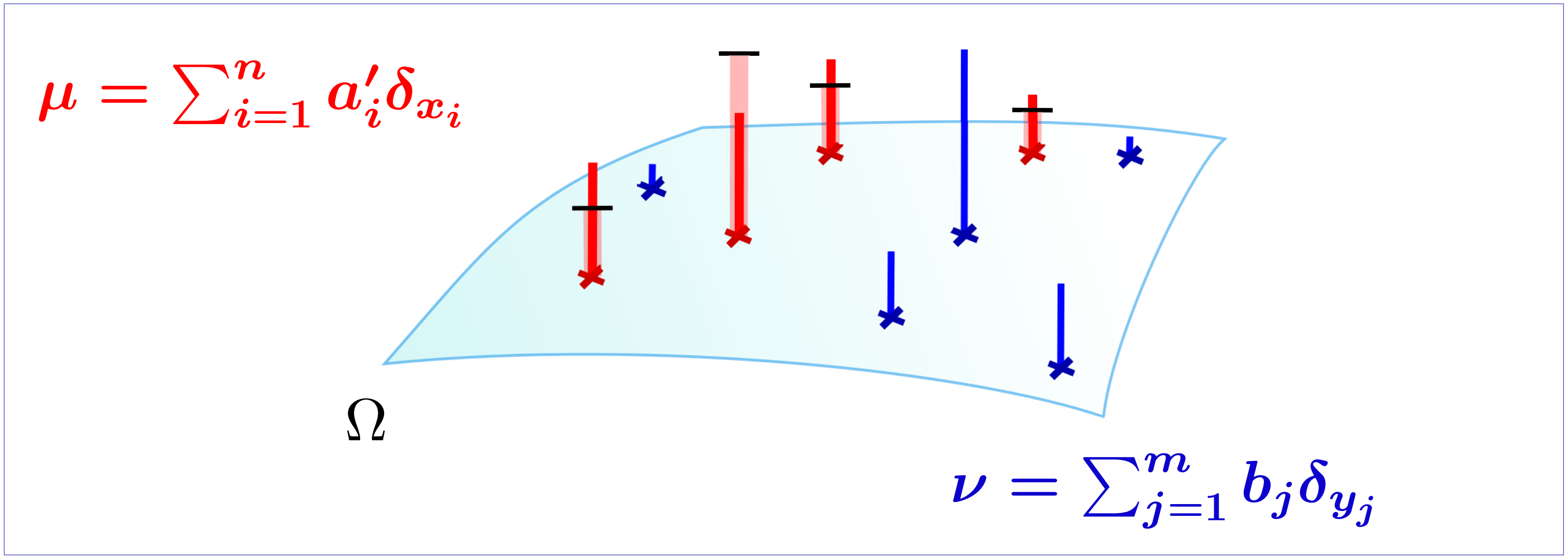
Differentiability w.r.t. X or a



To solve it **numerically**, we must understand how

$f_{\nu}(a, X) \stackrel{\text{def}}{=} W_p^p(\mu, \nu)$ varies when a & X varies.

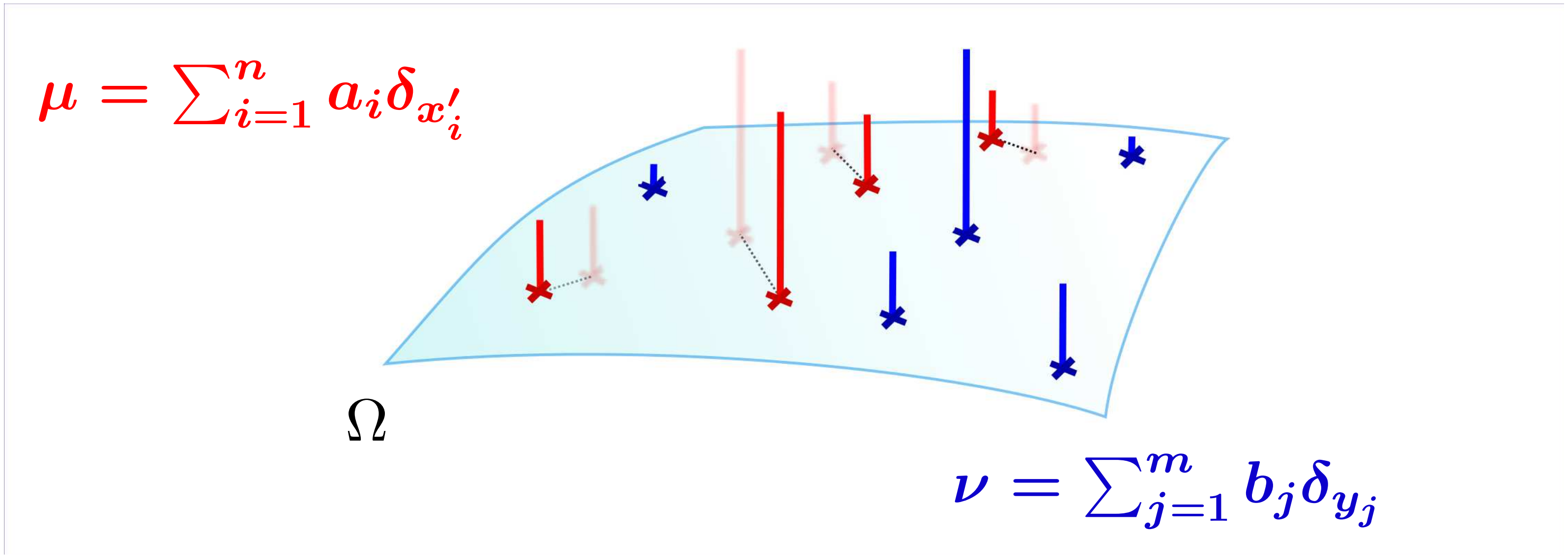
Differentiability w.r.t. X or a



1. Infinitesimal Variation in Weights

$$f_{\nu}(\boldsymbol{a'}, \boldsymbol{X})?, \quad \text{if } a' \approx a$$

Differentiability w.r.t. X or a

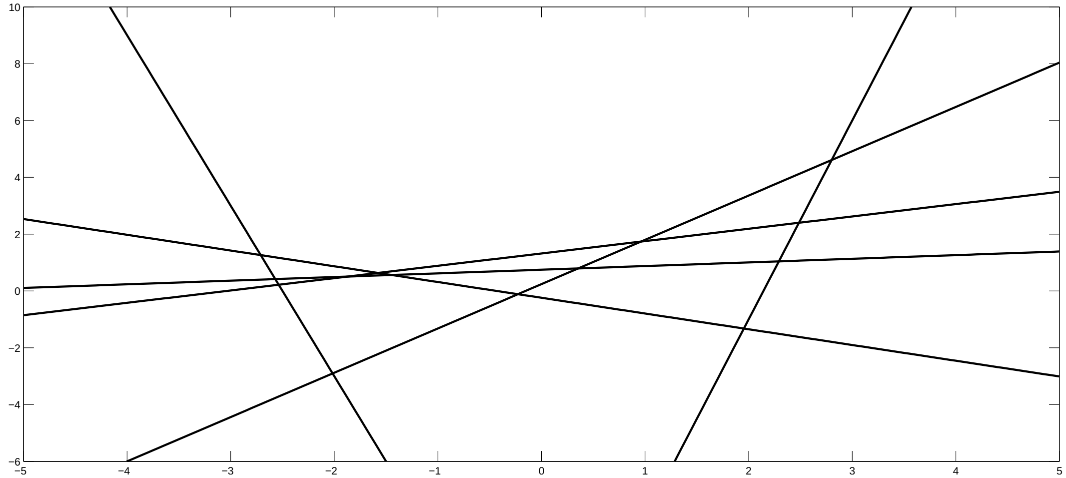


2. Infinitesimal Variation in Locations

$$f_{\nu}(a, X')?, \quad \text{if } X' \approx X$$

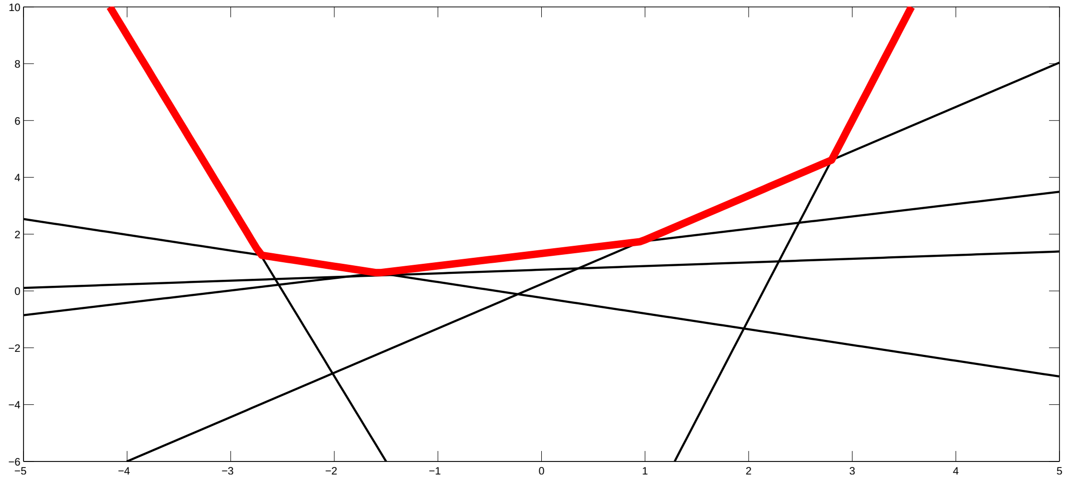
Using the dual, $\partial|_{\boldsymbol{a}}$

$$f_{\boldsymbol{\nu}}(\boldsymbol{a}, X) = \max_{(\alpha,\beta)\in C_{M_{\boldsymbol{XY}}}} \alpha^T \boldsymbol{a} + \beta^T \boldsymbol{b}$$



Using the dual, $\partial|_{\boldsymbol{a}}$

$$f_{\boldsymbol{\nu}}(\boldsymbol{a}, X) = \max_{(\alpha,\beta)\in C_{M_{\boldsymbol{XY}}}} \alpha^T\boldsymbol{a} + \beta^T\boldsymbol{b}$$



$\boldsymbol{a} \mapsto f_{\boldsymbol{\nu}}(\boldsymbol{a}, X)$ is a **convex** **non-smooth** map.
The *dual optimum* $\alpha^\star$ is a subgradient $f_{\boldsymbol{\nu}}(\boldsymbol{a}, X)$.

Using the primal $\partial|_{\boldsymbol{X}}$

$$f_{\boldsymbol{\nu}}(a, \boldsymbol{X}) = \min_{T \in U(\boldsymbol{a},\boldsymbol{b})} \langle T, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

- More involved computations. Tractable when $\boldsymbol{D}$ =Euclidean, $p = 2$.

- Convex quadratic + piecewise linear concave of $\boldsymbol{X}$

- $\partial f_{\boldsymbol{\nu}}|_X = \boldsymbol{Y}\boldsymbol{T}^{\star \boldsymbol{T}}\,\mathbf{diag}(a^{-1})$: *optimal transport* $\boldsymbol{T}^{\star \boldsymbol{T}}$ yields a subgradient.

To sum up: (1) the WBP is challenging

$$C(\boldsymbol{a},\boldsymbol{X}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} W_p^p(\boldsymbol{\mu},\boldsymbol{\nu_i}) = \frac{1}{N}\sum_{i=1}^{N} f_{\boldsymbol{\nu_i}}(\boldsymbol{a},\boldsymbol{X})$$

- $\boldsymbol{a} \to C(\boldsymbol{a}, X)$ is **convex**, **non-smooth**, **computing one subgradient requires solving N OT problems!**
- $\boldsymbol{X} \to C(a, \boldsymbol{X})$ is **not convex**, **non-smooth**

(2) the WBP is unstable

- Assume $X = Y_1 = \cdots = Y_N$ (fixed grid).

$$C(\boldsymbol{a}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{primal}(\boldsymbol{a}, \boldsymbol{b_i}, M) = \frac{1}{N}\sum_{i=1}^{N} \min_{T_i \in U(\boldsymbol{a},\boldsymbol{b_i})} \langle T_i, M \rangle$$

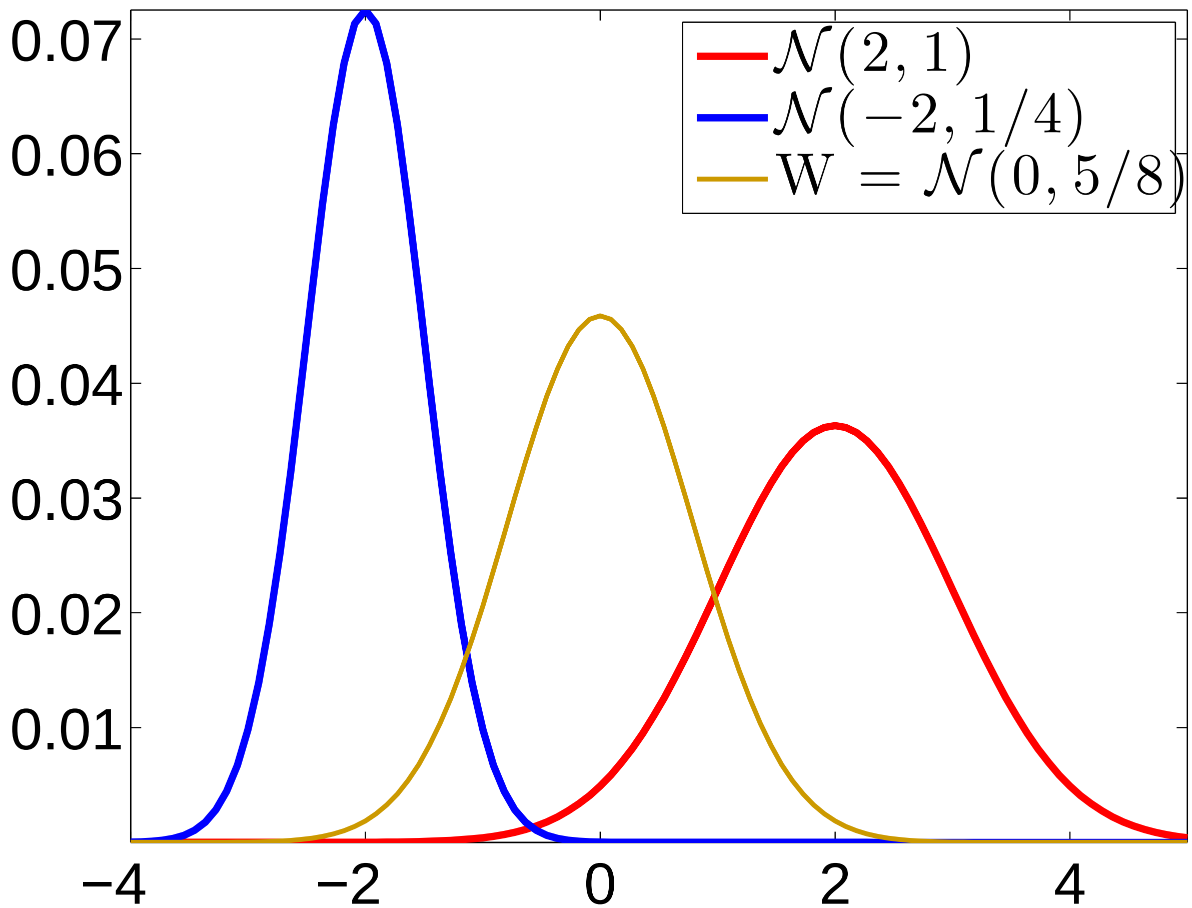
- In that case, the WBP can be solved as a large LP:

$$\min_{T_1,\cdots,T_N,\boldsymbol{a}} \sum_{i=1}^{N} \langle T_i, M \rangle$$

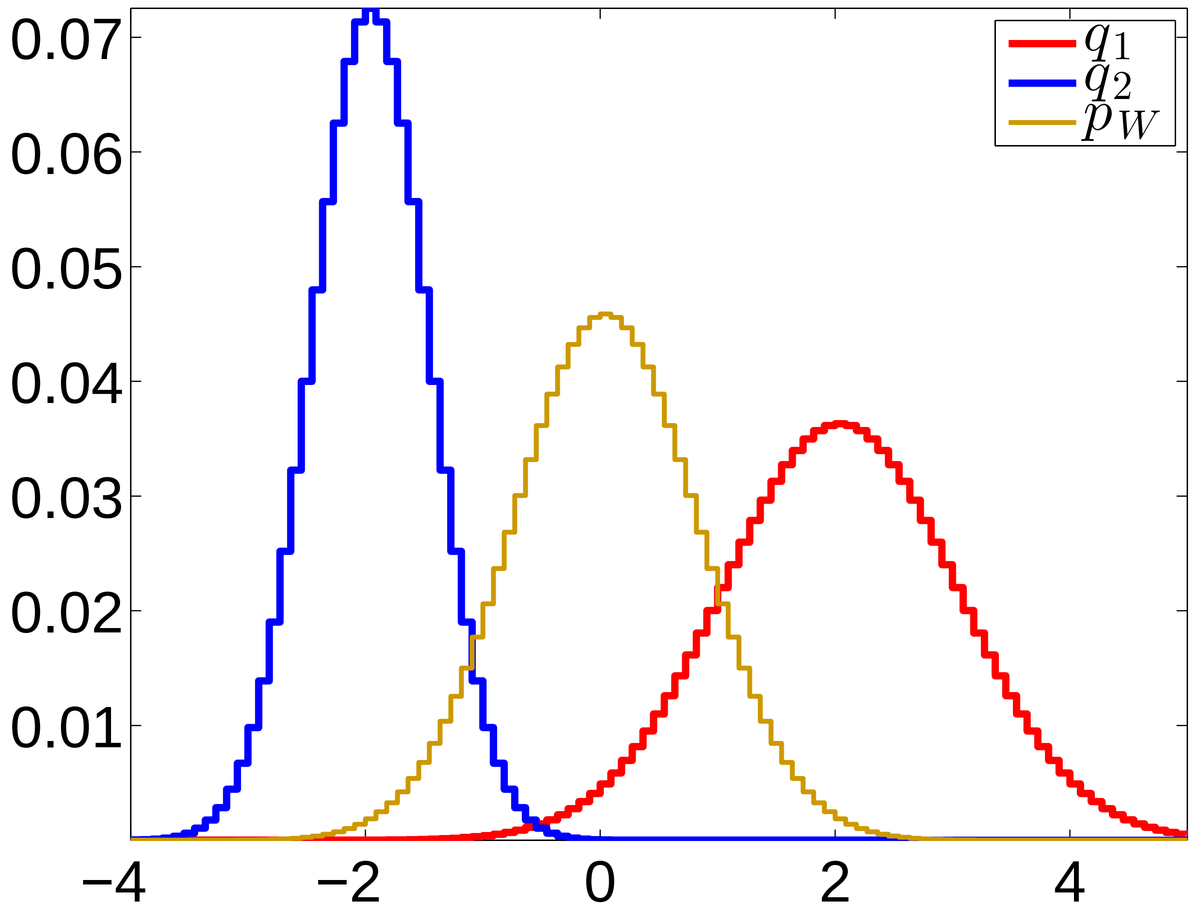
s.t. $T_i^T \mathbb{1}_d = \boldsymbol{b_i}, \forall i \leq N,$

$T_1\mathbb{1}_d = \cdots = T_N\mathbb{1}_d = \boldsymbol{a}.$

Averaging Two Gaussians

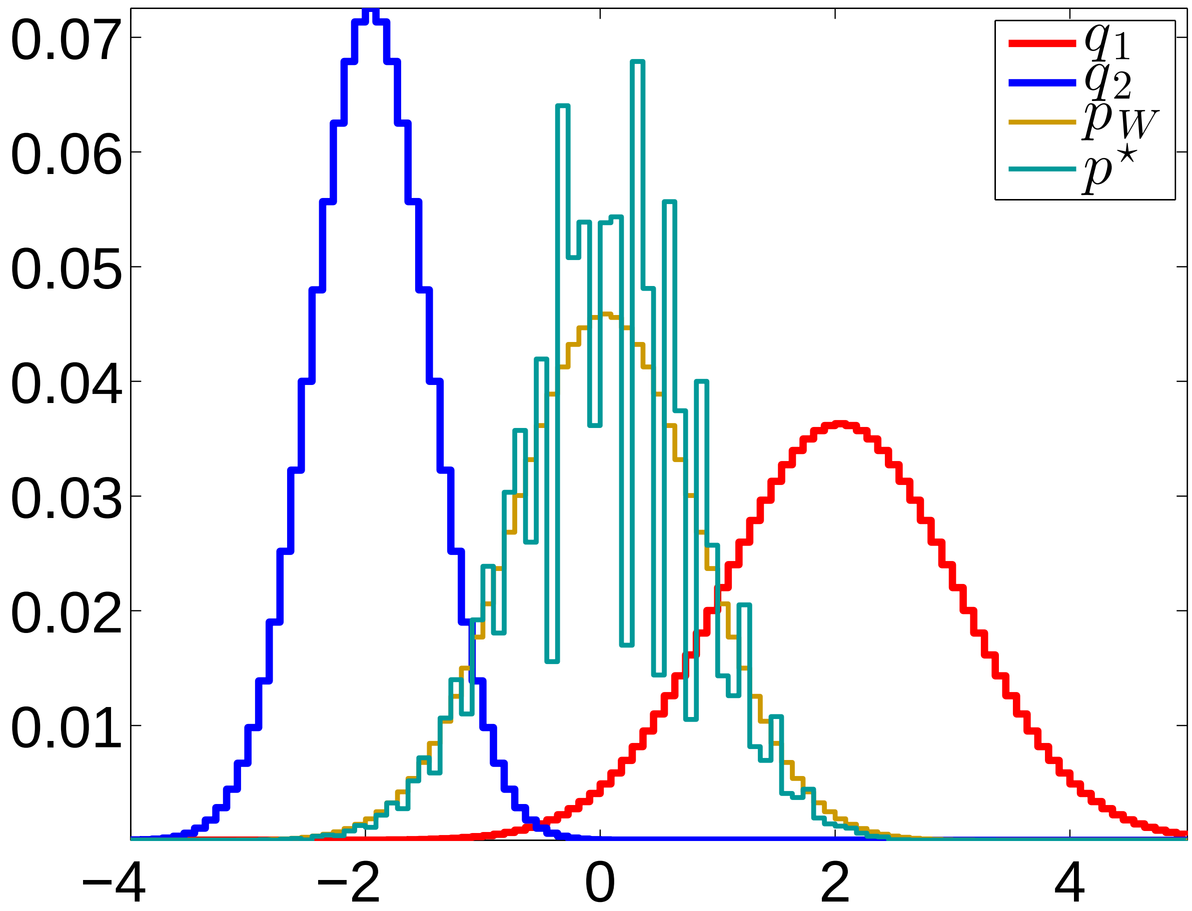


Discretized



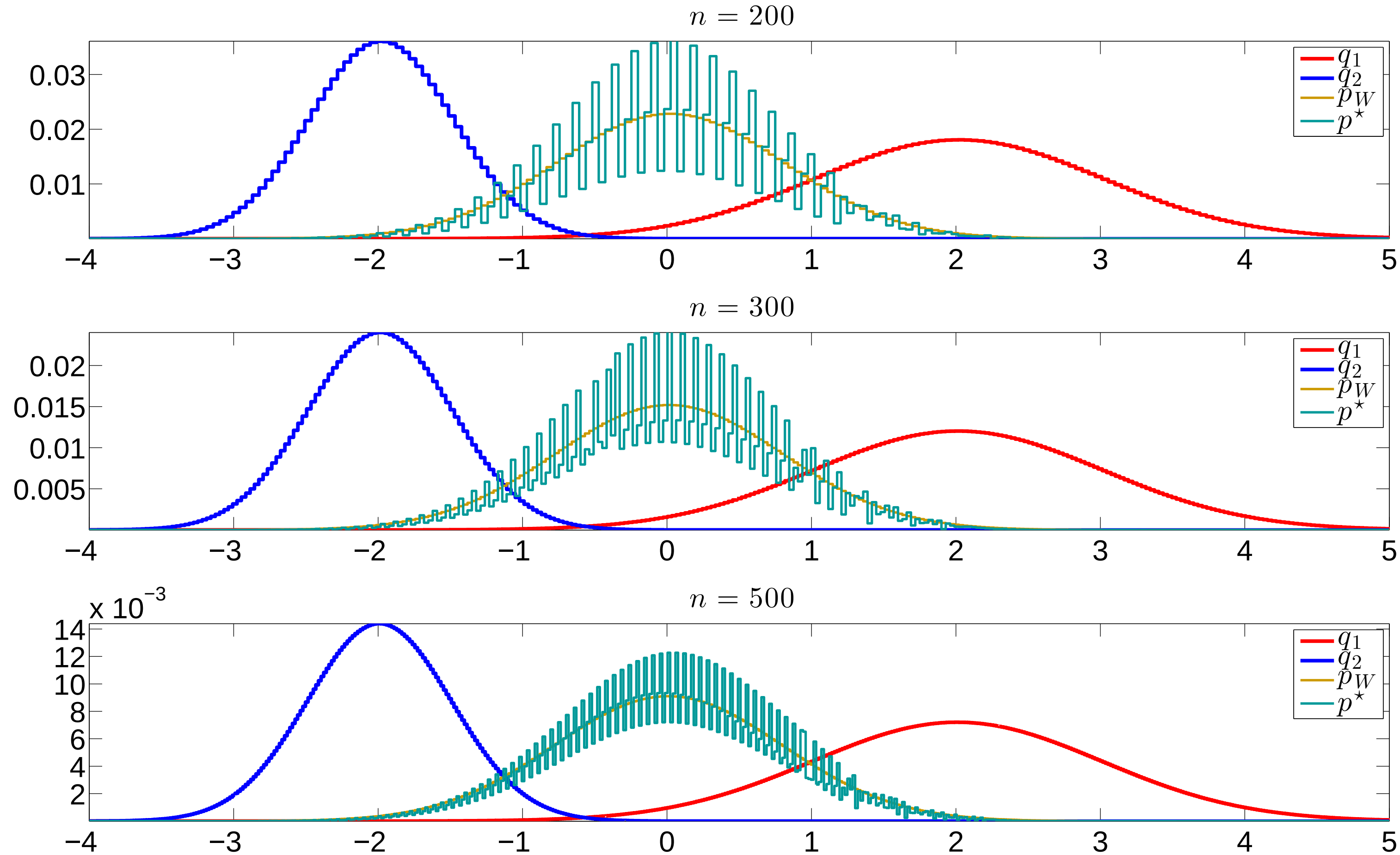
p_W is the discrete equivalent of the true barycenter.

Exact Solution



$p^\star$ is the solution to that LP

Does not get much better with large n...



Entropic Smoothing of OT

Smoothing solves (almost) everything

Original OT primal:

$$\mathbf{primal}(a, b, M_{XY}) = \min_{\boldsymbol{T} \in U(a,b)} \langle \boldsymbol{T}, M_{XY} \rangle$$

Original OT Kantorovich dual:

$$\mathbf{dual}(a, b, M_{XY}) = \max_{(\boldsymbol{\alpha},\boldsymbol{\beta}), \boldsymbol{\alpha_i} + \boldsymbol{\beta_j} \leq M_{ij}} \boldsymbol{\alpha}^T a + \boldsymbol{\beta}^T b$$

Smoothing solves (almost) everything

Entropy-smoothed ($\boldsymbol{\gamma > 0}$) primal problem:

$$\mathbf{primal}_{\gamma}(a, b, M_{XY}) = \min_{\boldsymbol{T} \in U(a,b)} \langle \boldsymbol{T}, M_{XY} \rangle - \boldsymbol{\gamma} H(\boldsymbol{T})$$

Smoothed dual problem:

$$\mathbf{dual}_{\gamma}(a, b, M_{XY}) = \max_{(\boldsymbol{\alpha},\boldsymbol{\beta})} \boldsymbol{\alpha}^T a + \boldsymbol{\beta}^T b - \boldsymbol{\gamma} \sum_{i \leq n, j \leq m} e^{-\left(M_{ij} - \boldsymbol{\alpha_i} - \boldsymbol{\beta_j}\right)/\gamma}$$

Smoothing solves (almost) everything

Entropy-smoothed ($\gamma > \mathbf{0}$) primal problem:

$$\mathbf{primal}_{\gamma}(a,b,M_{XY}) = \min_{\boldsymbol{T}\in U(a,b)} \mathbf{KL}(\boldsymbol{T} \| e^{-\boldsymbol{M_{XY}}/\gamma})$$

Smoothed dual problem:

$$\mathbf{dual}_{\gamma}(a,b,M_{XY}) = \max_{(\boldsymbol{\alpha},\boldsymbol{\beta})} \boldsymbol{\alpha}^T a + \boldsymbol{\beta}^T b - \boldsymbol{\gamma} \sum_{i\leq n, j\leq m} e^{-\left(M_{ij}-\boldsymbol{\alpha_i}-\boldsymbol{\beta_j}\right)/\gamma}$$

Why is entropy a good regularizer for OT?

The penalized problem

$$T_\gamma = \underset{T \in U(r,c)}{\operatorname{argmin}} \langle P, M \rangle - \boldsymbol{\gamma H(T)}$$

implies that T_γ has the form: (first order cond.)

$$\exists \boldsymbol{u} \in \mathbb{R}_n^+, \boldsymbol{v} \in \mathbb{R}_m^+ \,|\, T_\gamma = \mathbf{diag}(\boldsymbol{u}) e^{-M/\gamma} \mathbf{diag}(\boldsymbol{v}).$$

Gravity Model in Transportation[Wilson'69]
Schrödinger Problem['32]

Sinkhorn - Matrix Scaling

Theorem 1 (Sinkhorn'62). *For* **any** $n \times m$ *matrix* A *with positive entries, any* $\boldsymbol{r}$ *and* $\boldsymbol{c}$ *in the simplex,* $\exists! \boldsymbol{u} \in \mathbb{R}^+_n, \boldsymbol{v} \in \mathbb{R}^+_m$ *such that*

$$\begin{bmatrix} \boldsymbol{u_1} & 0 & \dots & 0 \\ 0 & \boldsymbol{u_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{u_n} \end{bmatrix} \begin{bmatrix} & A & \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} & 0 & \dots & 0 \\ 0 & \boldsymbol{v_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{v_m} \end{bmatrix} \in U(\boldsymbol{r}, \boldsymbol{c})$$

$\boldsymbol{u}, \boldsymbol{v}$ can be computed in $O(nm)$ time using the Sinkhorn fixed-point iteration.

Sinkhorn Algorithm

1. Set $K = \exp(-M/\gamma)$ (note: if M is Euclidean metric, this is a Gaussian convolution...)

2. Seed initial random values for $\boldsymbol{u}$.

3. Loop until convergence

   (a) Set $\boldsymbol{v} \leftarrow (K'\boldsymbol{u}^{-1})./c$

   (b) Set $\boldsymbol{u} \leftarrow (K\boldsymbol{v}^{-1})./r$

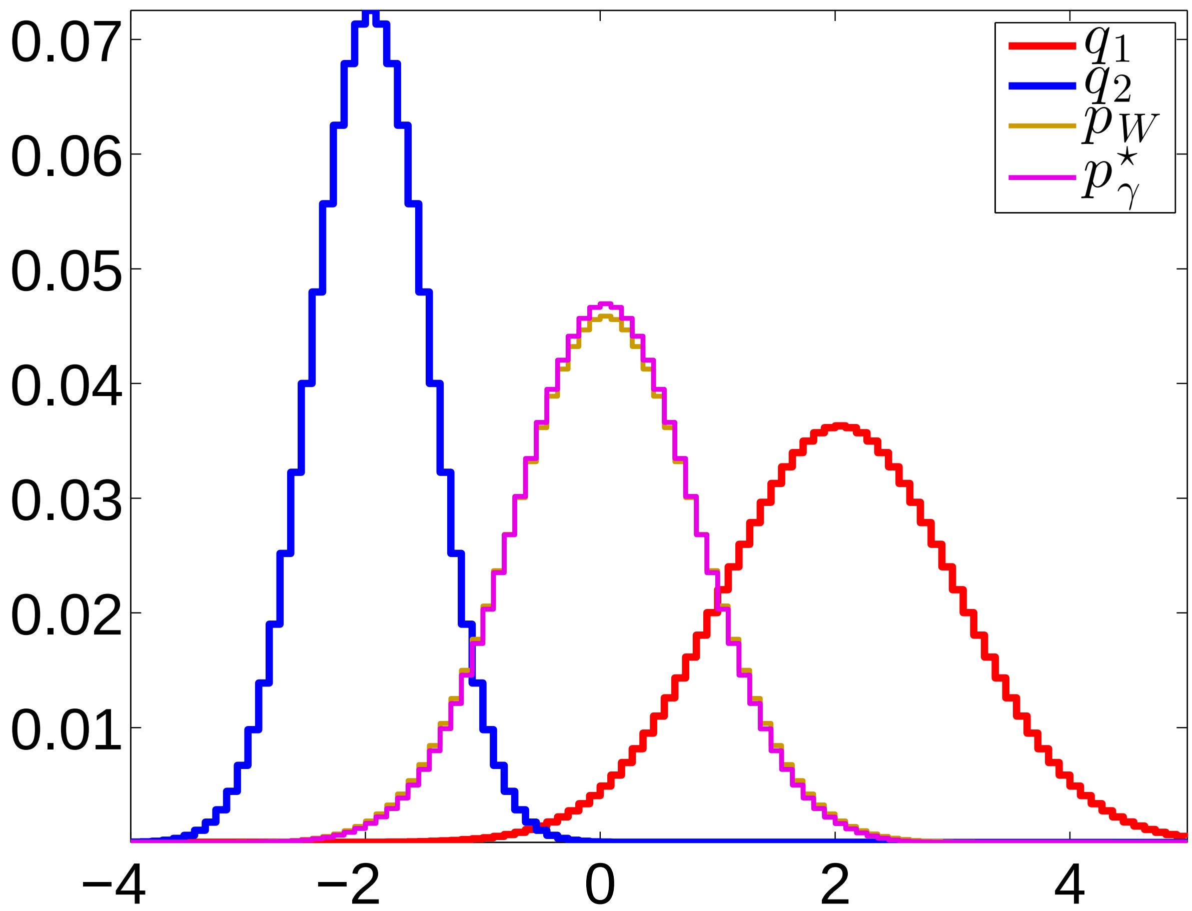
- $T_\gamma^\star = \mathbf{diag}(\boldsymbol{u}^\star)K\,\mathbf{diag}(\boldsymbol{v}^\star), \alpha_\gamma^\star = \log(u^\star)/\gamma.$

- $W_\gamma(\mu,\nu) = \langle T_\gamma, M\,\rangle = \boldsymbol{u}^{\star T}\,(K.*M)\,\boldsymbol{v}^\star$

Benefits of Smoothing [C.'13]

- These OT problems are **strongly convex** vs. **LPs**. **Unicity** of solutions, **differentiable**.
- Considerably more efficient in practice **[Nesterov'05]**.
- Primal/dual smoothed optima $\boldsymbol{\alpha}^\star_\gamma, \boldsymbol{T}^\star_\gamma$ can be solved
  - **In** $O(n^2)$ with **Sinkhorn's (IPFP) algorithm**,
  - in **parallel on GPGPUs** for **any metric** on finite Ω,
  - **millions of time faster** than simplex,
  - can deal with **large dimensions** (≈ 50.000 so far).

Our Solution (using regularization)



1. Smoothed primal [C.Doucet'14]

$$C(\boldsymbol{a}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{primal}_\gamma(\boldsymbol{a}, \boldsymbol{b_i}, M)$$

- (Projected) gradient descent:
  - Solve N smoothed (dual) OT problems $\boldsymbol{\alpha}^\star_{i,\gamma}$
  - Update $\boldsymbol{a}$ using gradient $\frac{1}{N}\sum_i \boldsymbol{\alpha}^\star_{i,\gamma}$
- ⚠ each step requires **computing** $\alpha^\star_{i,\gamma}$.

2. Dual appproach [C. Peyré'14]

- The Fenchel-Legendre conjugate of

$$f_{\boldsymbol{b}}(\boldsymbol{a}) = \mathbf{primal}_\gamma(\boldsymbol{a}, \boldsymbol{b}, M),$$

namely

$$f_{\boldsymbol{b}}^*(\boldsymbol{g}) = \max_{\boldsymbol{p}\in\Sigma_n} \langle \boldsymbol{g}, \boldsymbol{p} \rangle - f_{\boldsymbol{b}}(\boldsymbol{p}).$$

has a **closed form**

$$\boxed{f_{\boldsymbol{b}}^*(\boldsymbol{g}) = \gamma \left(H(\boldsymbol{b}) + \langle \boldsymbol{b}, \log e^{-M/\gamma} e^{\boldsymbol{g}/\gamma} \rangle \right)}$$

2. Dual appproach [C. Peyré'14]

- The original problem in splitted form:

$$\min_{a_1,\cdots,a_N\in\Sigma_n} \sum_{i=1}^N f_{b_i}(a_i) \text{ subj. to } a_1 = \cdots = a_N$$

- can be replaced with an easier problem:

$$\min_{g_1,\cdots,g_N\in\mathbb{R}^n} \sum_{i=1}^N f^*_{b_i}(g_i) \text{ subj. to } \sum_{i=1}^N g_i = 0.$$

gradient/Hessian explicit, equality constraint $\rightarrow$ **truncated Newton**.

at convergence, all $\nabla f^*_{b_i}(g_i)$ are equal to solution $a^\star$.

3. Generalized KL Projections [NBCCP'14]

- Idea: generalize $\mathbf{KL}$ projection for two marginals

$$\underset{T \in U(\boldsymbol{a},\boldsymbol{b_i})}{\operatorname{argmin}} \mathbf{KL}(T|e^{-M/\gamma})$$

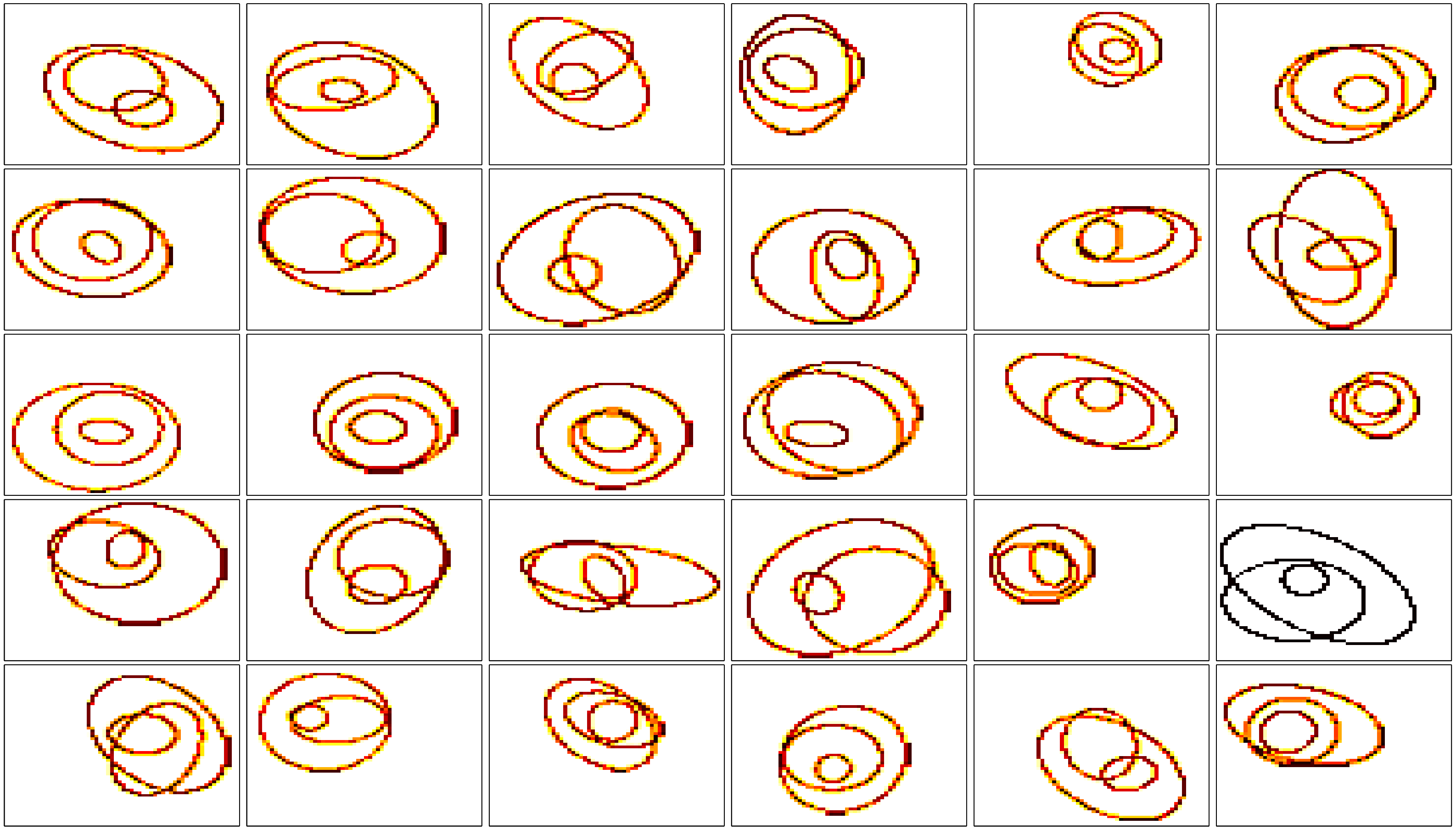
- to alternated $\mathbf{KL}$ projections for N + common (unknown) one.

$$\underset{T_i^T\mathbf{1}=\boldsymbol{b_i},\, T_i\mathbf{1}=T_{i+1}\mathbf{1}}{\operatorname{argmin}} \quad \sum_i \mathbf{KL}(T_i|e^{-M/\gamma})$$

- 2 lines of matlab code.

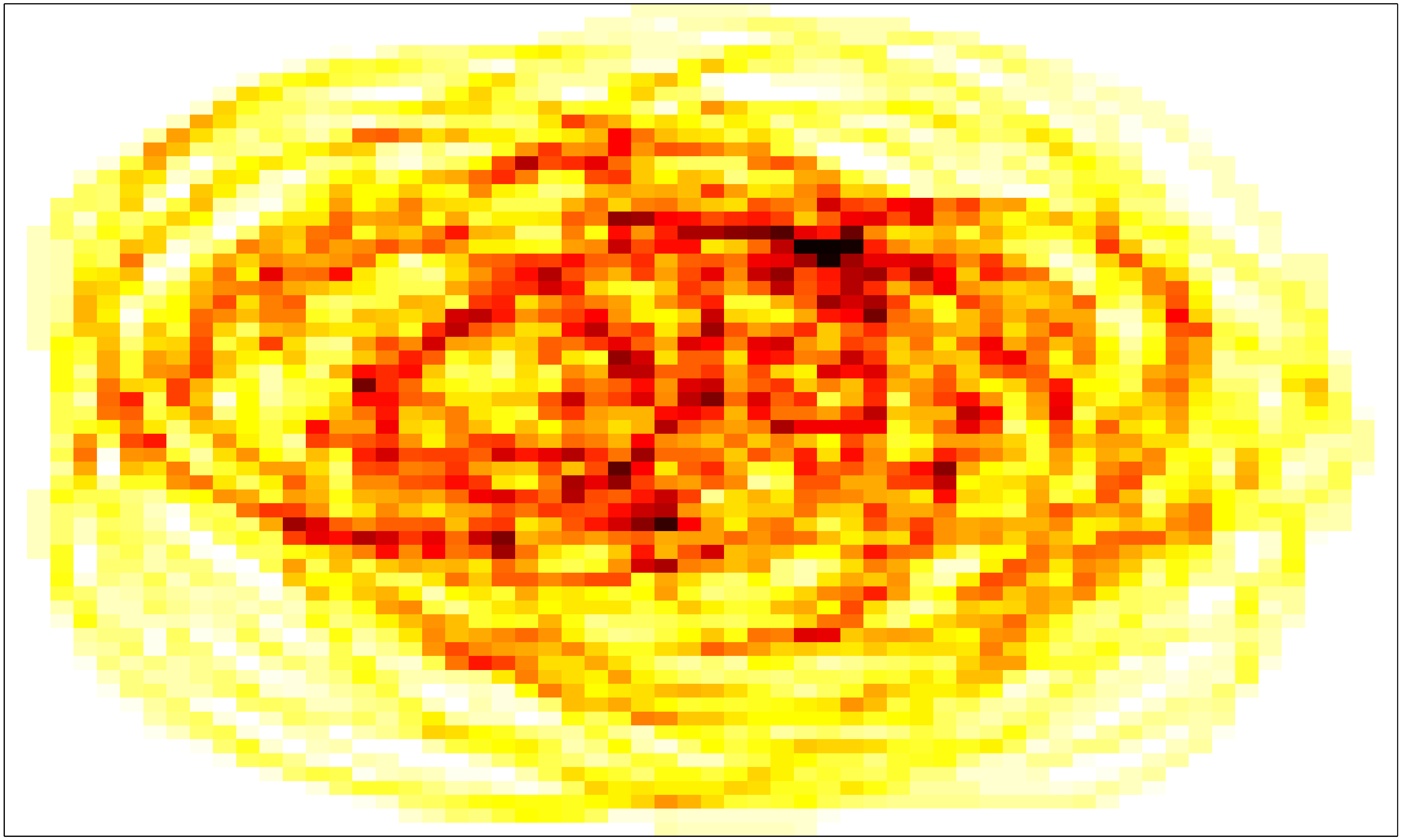
Applications

Averaging 30 Measures

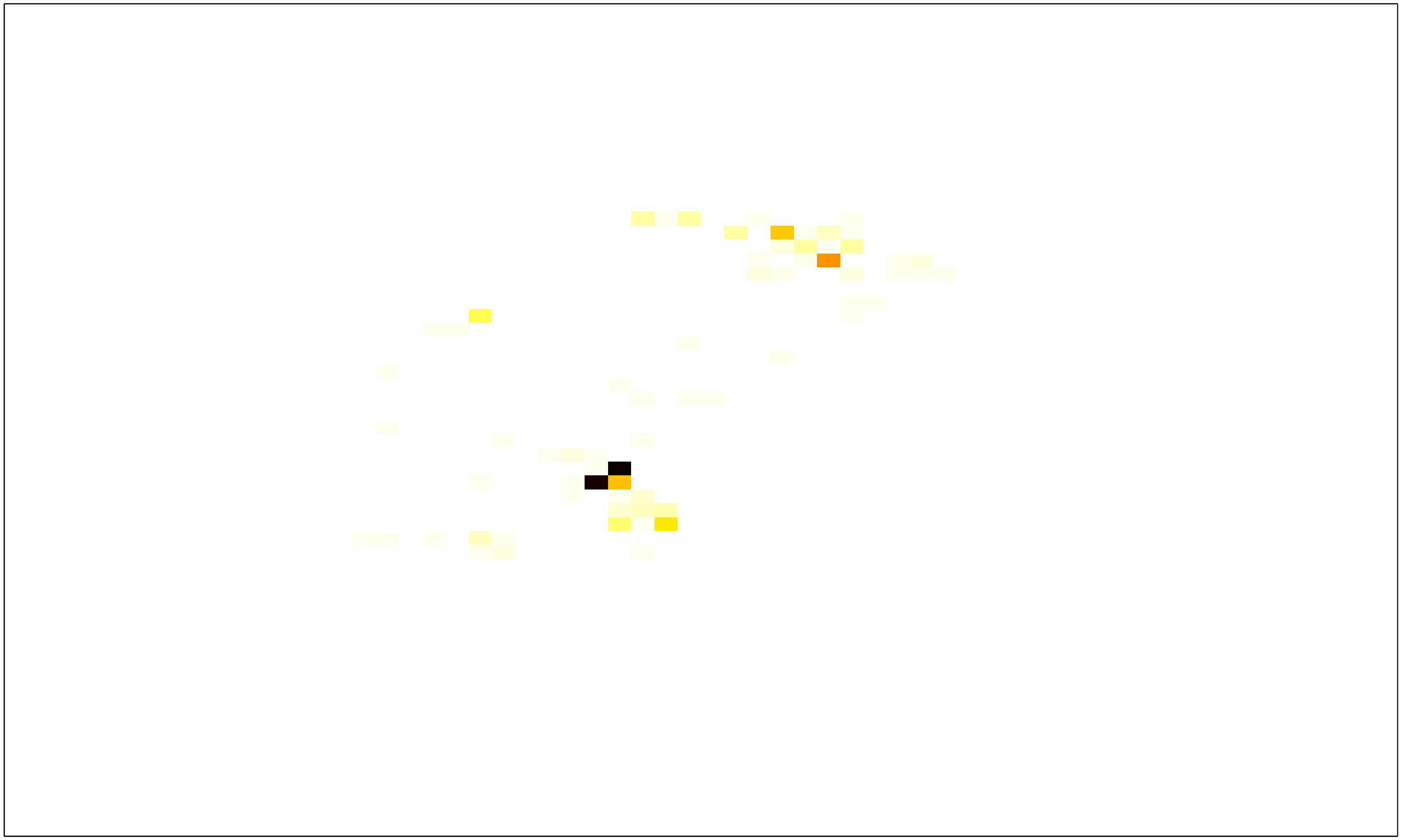


30 measures on $\mathbb{R}^2$.

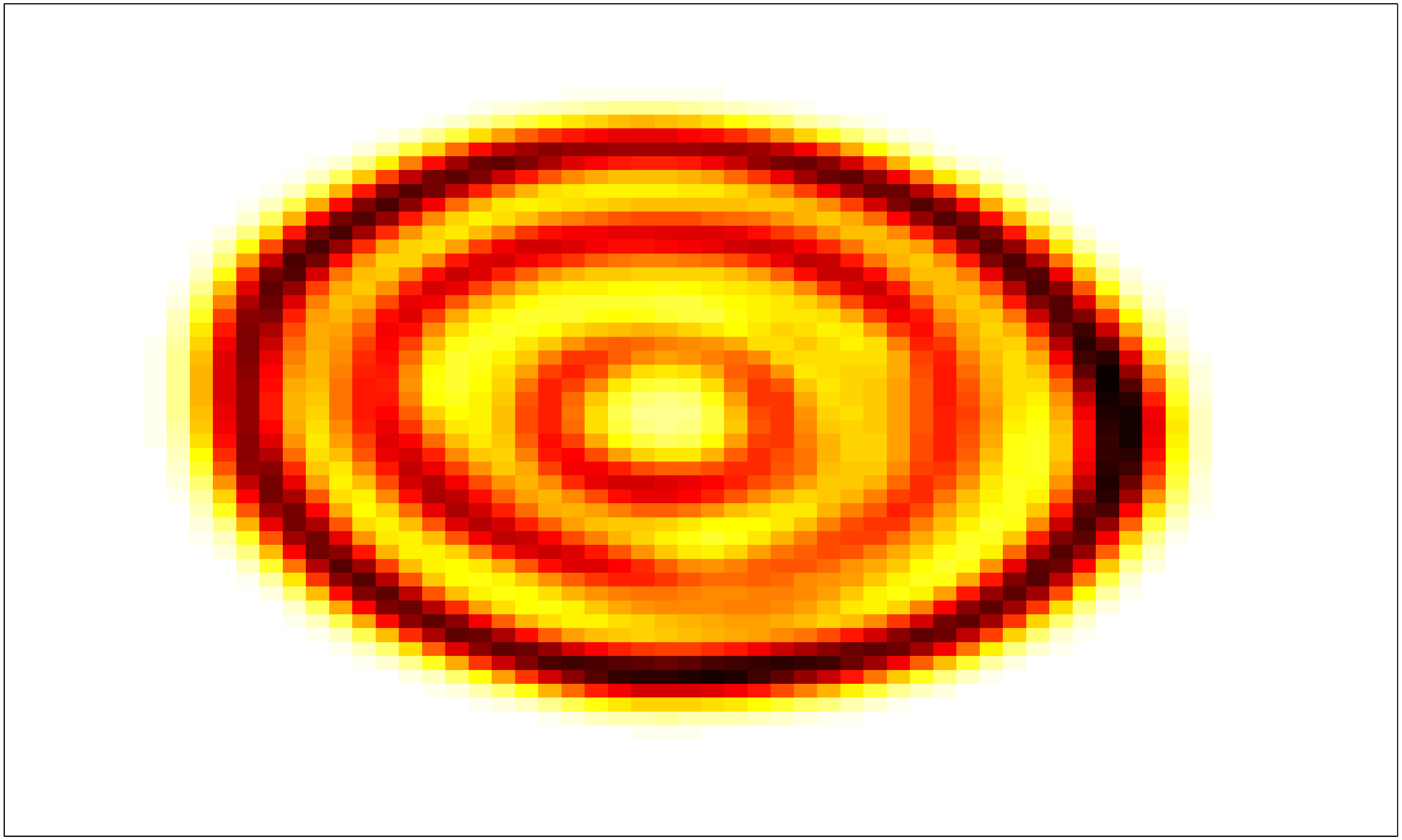
Euclidean Mean



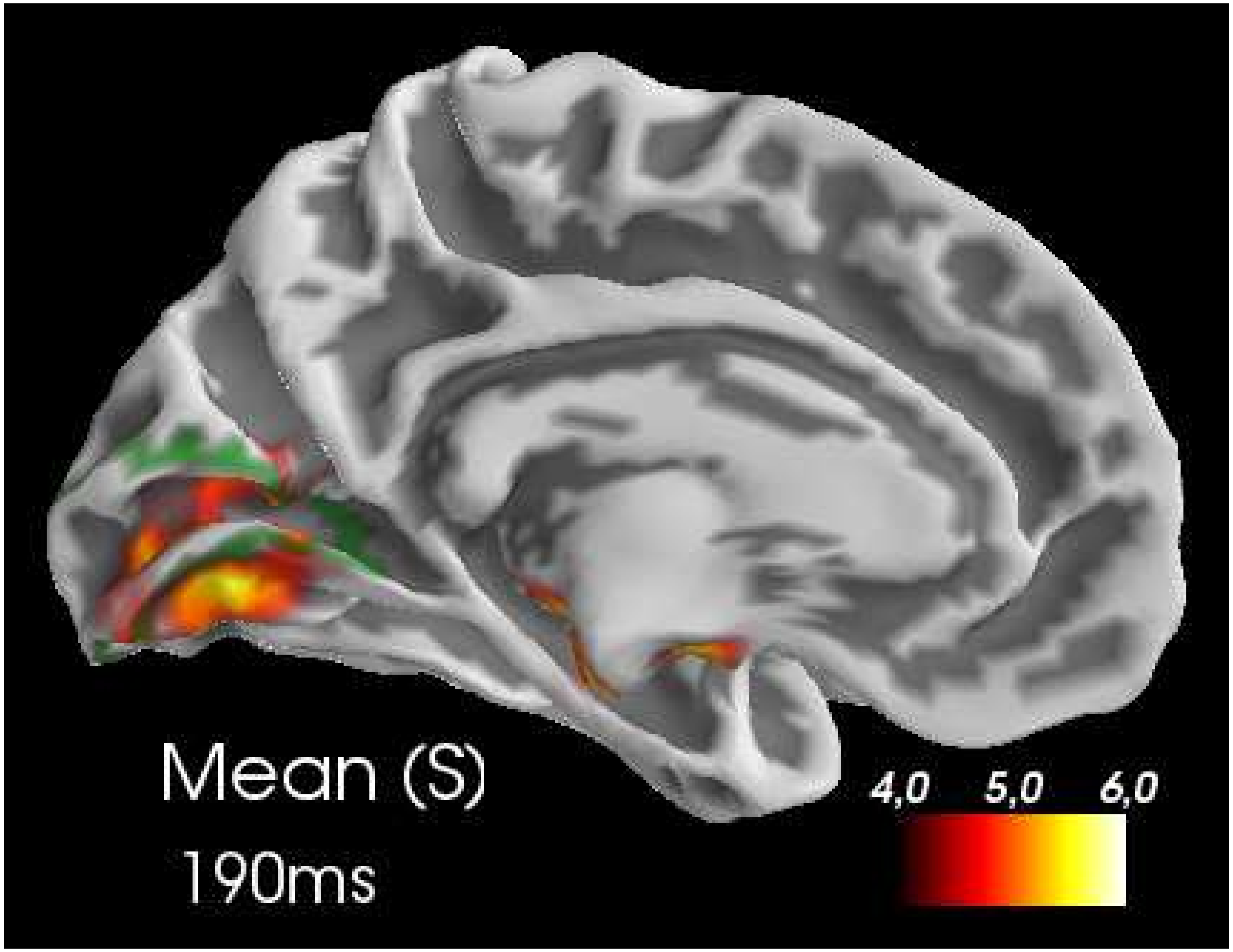
Symmetric KL Mean



2-Wasserstein

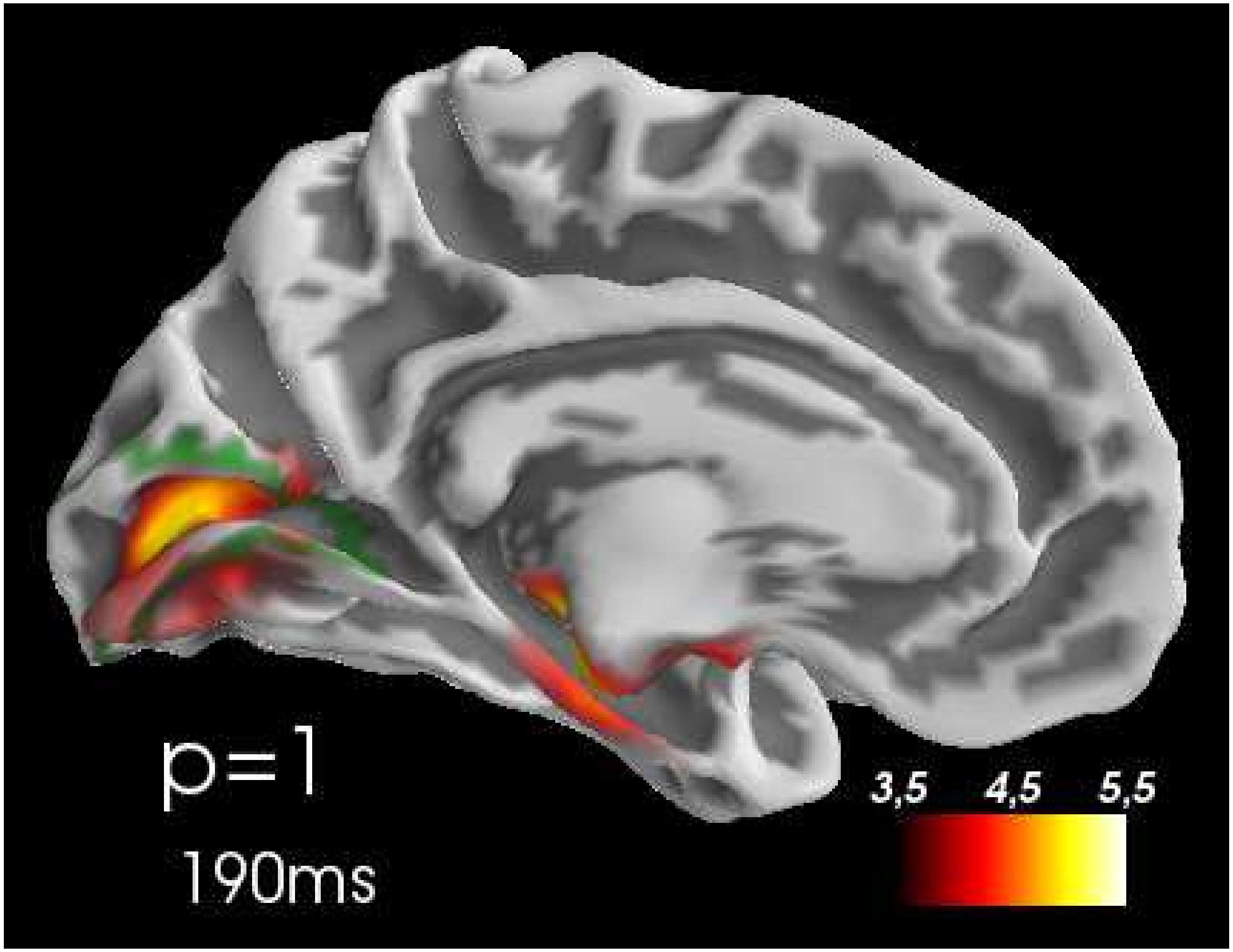


Averaging Brain Activations real

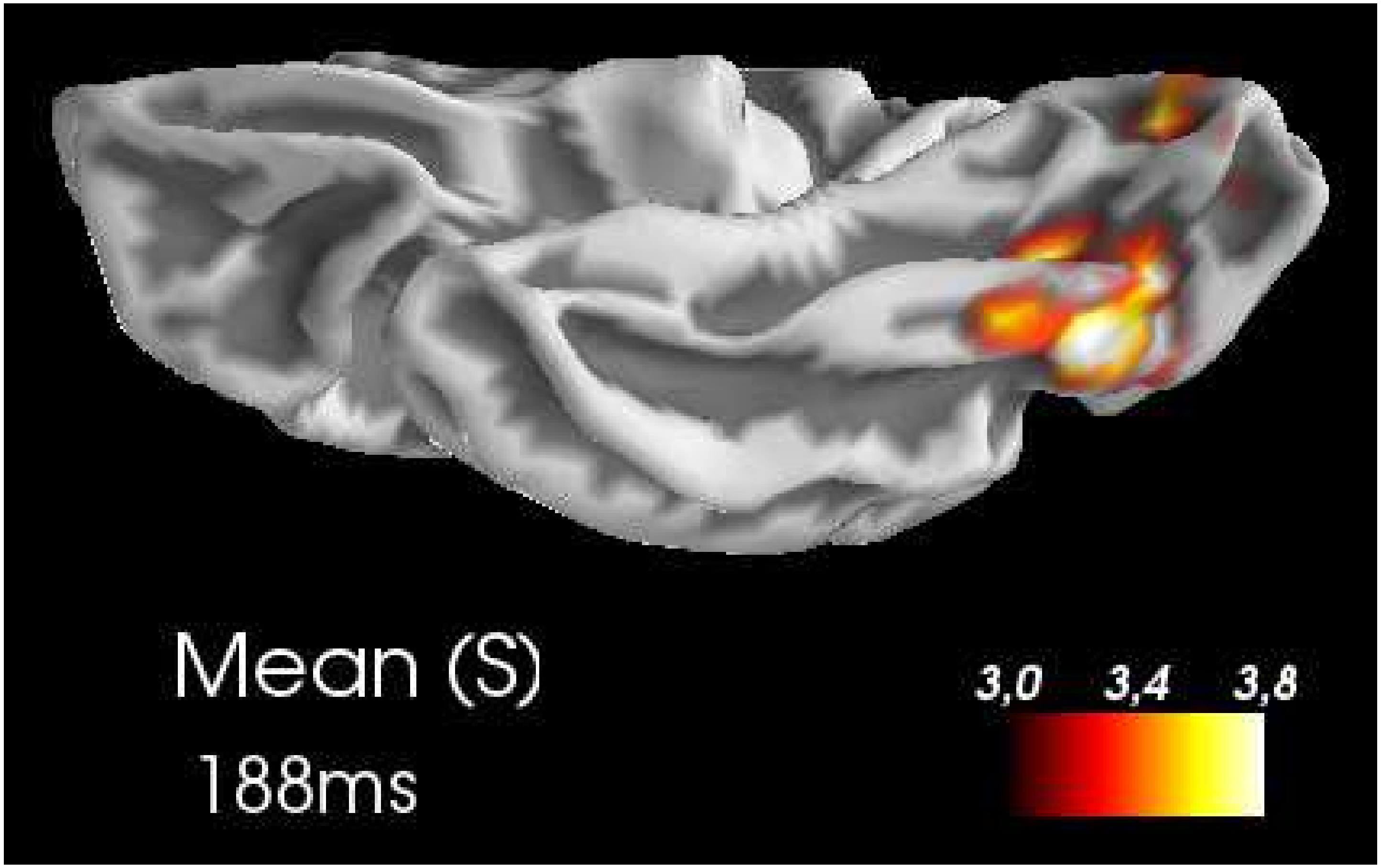


MEG ERF data, N=16. Left, medial view. **border** of (V1)

1-Wasserstein

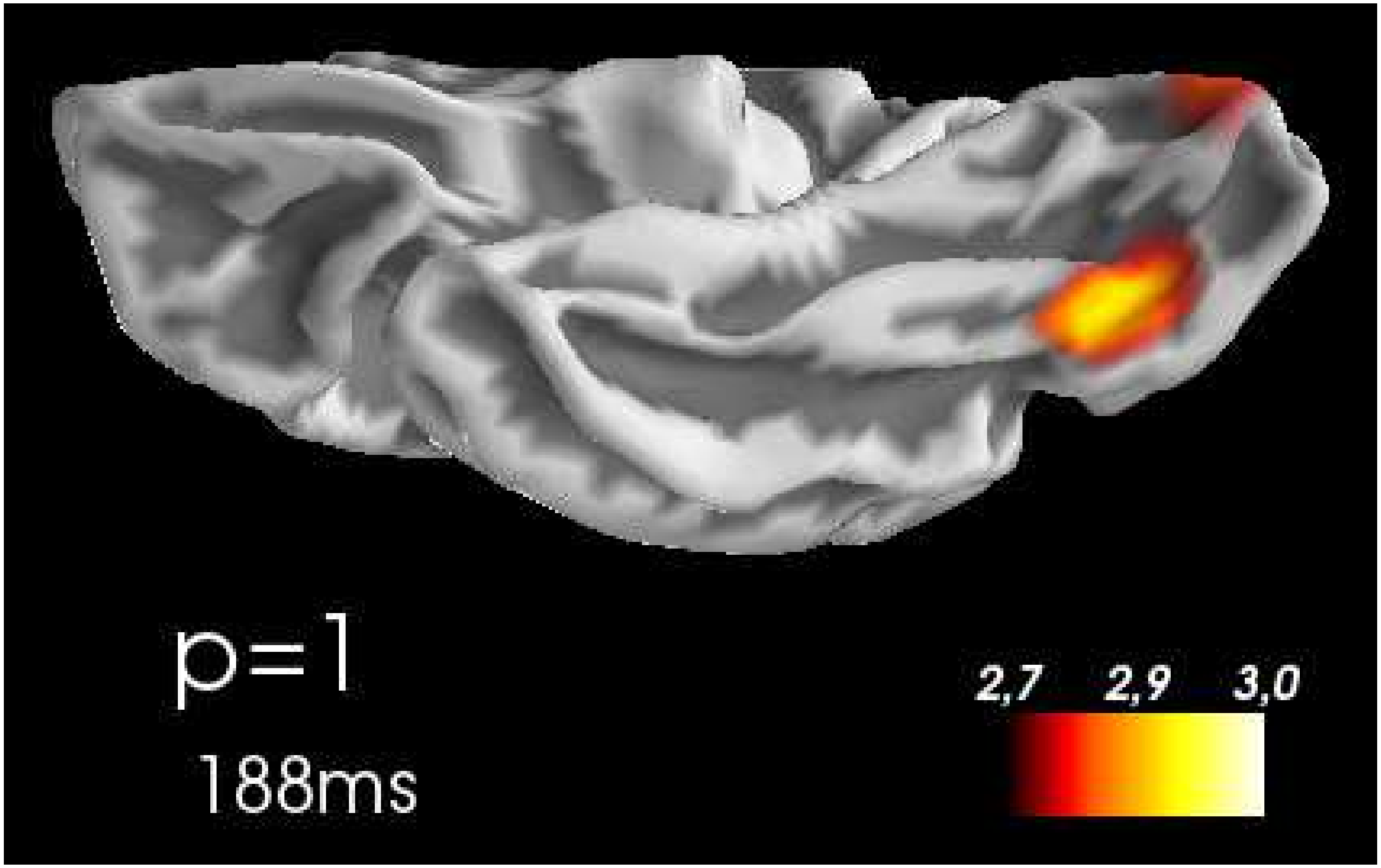


Averaging Brain Activations real



Right, ventral view

Averaging Brain Activations real



Centered on the Fusiform gyrus

Averaging Text Histograms

- Using GLOVE embeddings for words, 2-Wasserstein.

Graphics

Graphics

Graphics

Graphics

Graphics

Graphics

End