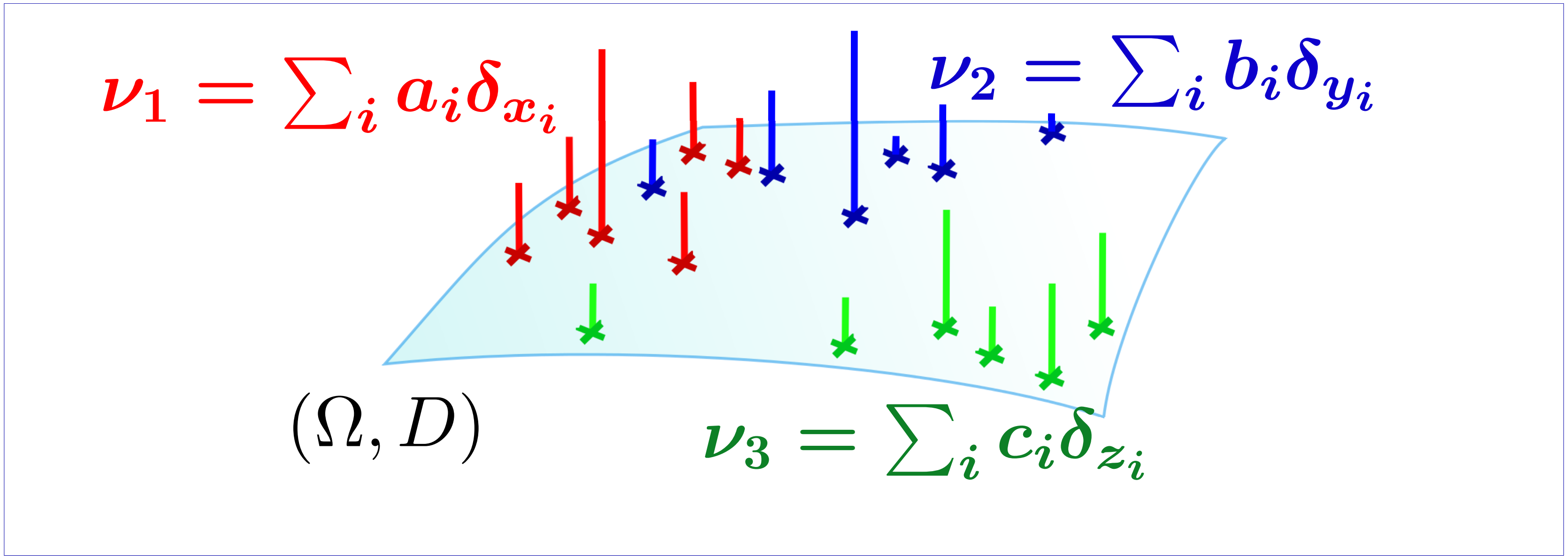
ICML 2014

Fast Computation of Wasserstein Barycenters

M. Cuturi[1] **A. Doucet**[2]

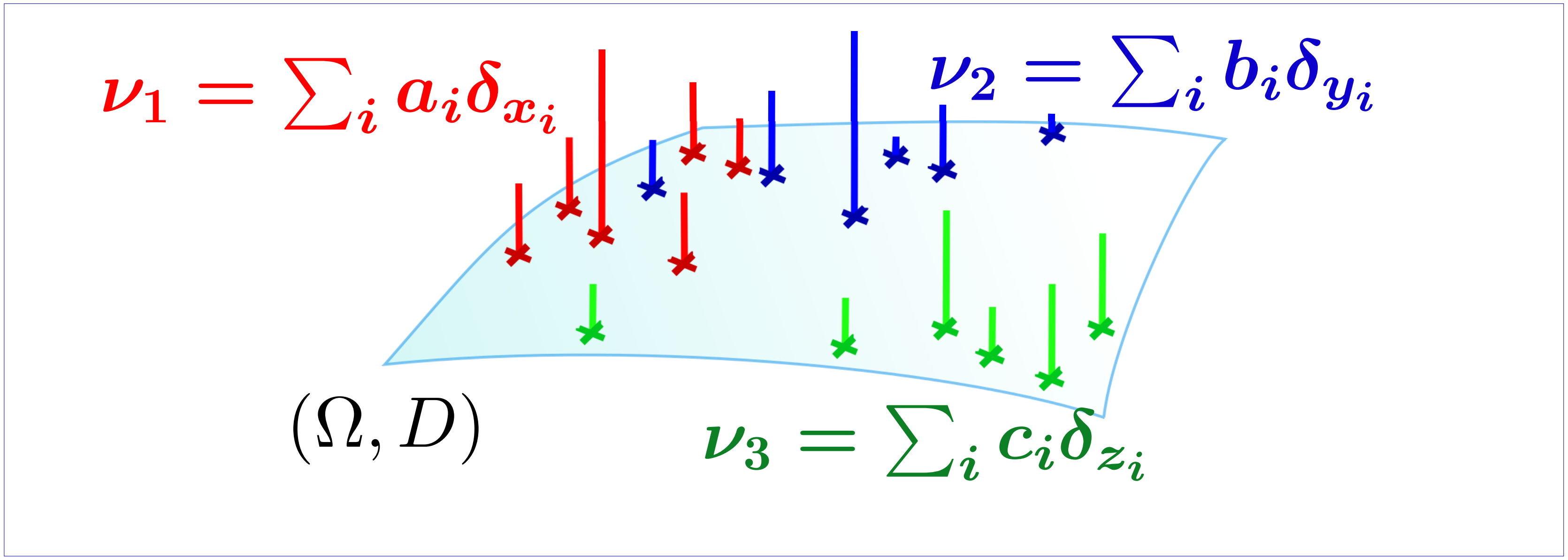
[1]Kyoto University [2]University of Oxford

Problem: Average N Probability Measures



- $\{\Omega, D\}$ a metric space
- $\{\nu_1, \cdots, \nu_N\}$ family of empirical probability measures.

Problem: Average N Probability Measures



Can we **summarize** the $\{\nu_i\}$ as an "**average**" or a "**barycentric**" single empirical probability measure?

interest in ML: empirical measure = dataset,

histogram/bags-of-features, single observation with uncertainty

Euclidean Means for Vectors

- For **vectors** $\{x_1, \cdots, x_N\}$ in a Hilbert, their average is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Euclidean Means for Vectors

- For **vectors** $\{x_1, \cdots, x_N\}$ in a Hilbert, their average is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- behind this formula lies a **variational** problem

$$\bar{x} = \operatorname*{argmin}_{u \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} \|u - x_i\|_2^2$$

Euclidean Means for Measures

- For **probability measures** $\{\nu_i\}_{i=1..N}$, we can also use:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \nu_i,$$

- as well as, using a **smoothing kernel** $k = e^{-D^2/\sigma}$,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{k} * \nu_i)$$

(a.k.a *RKHS* mean map **[Gretton'07]**)

Other Means for Probabilities

- Other means can be defined using other **metrics** or **divergences**:

$$\operatorname*{argmin}_{\mu \in P(\Omega)} \sum_{i=1}^{N} \boldsymbol{\Delta}(\mu, \nu_i).$$

  - **KL, Symmetrized KL** **[Nielsen'12]**
  - **Bregman Divergence** **[Bhanerjee'05]**
  - **Wasserstein Distance** (a.k.a *EMD*) **[Agueh'11]**

Wasserstein Barycenter Problem

- **[Agueh'11]** defined

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

  provided theoretical analysis, unicity of solution.

- Simple cases ($N = 2$, multivariate Gaussians) covered.
- **very challenging computational problem**.

Our Contribution

- **First computational approach** to solve efficiently **variational Wasserstein problems**,

- including the Wasserstein barycenter problem,

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

- that is applicable for arbitrary (Ω, D) and $p > 0$, using **entropy-smoothed optimal transport [Cuturi'13]**.

Our Contribution

- **First computational approach** to solve efficiently **variational Wasserstein problems**,
- including the Wasserstein barycenter problem,

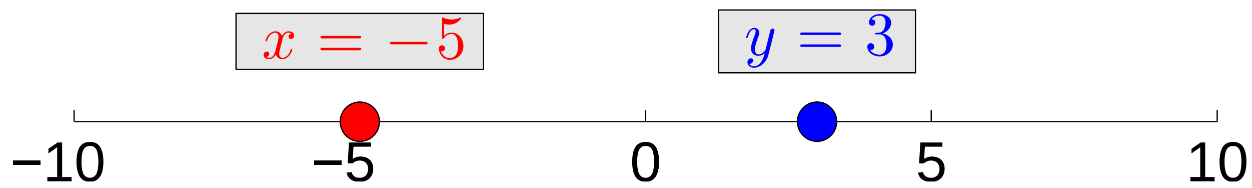
$$\underset{\mu \in P(\Omega)}{\operatorname{argmin}} \sum_{i=1}^{N} \boldsymbol{W_p^p}(\mu, \nu_i),$$

- that is applicable for arbitrary (Ω, D) and $p > 0$, using **entropy-smoothed optimal transport [Cuturi'13]**.

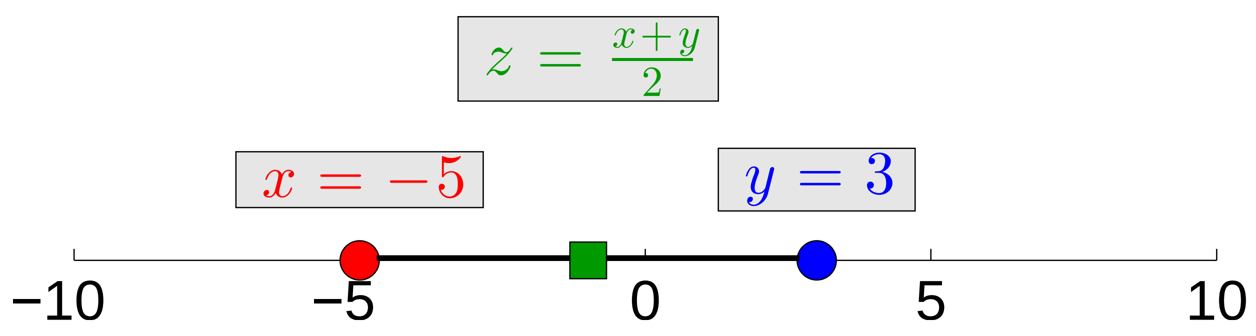
(**[Rabin'12,Bonneel'14]** studied case $\Omega = \mathbb{R}^2$)

Motivating Examples

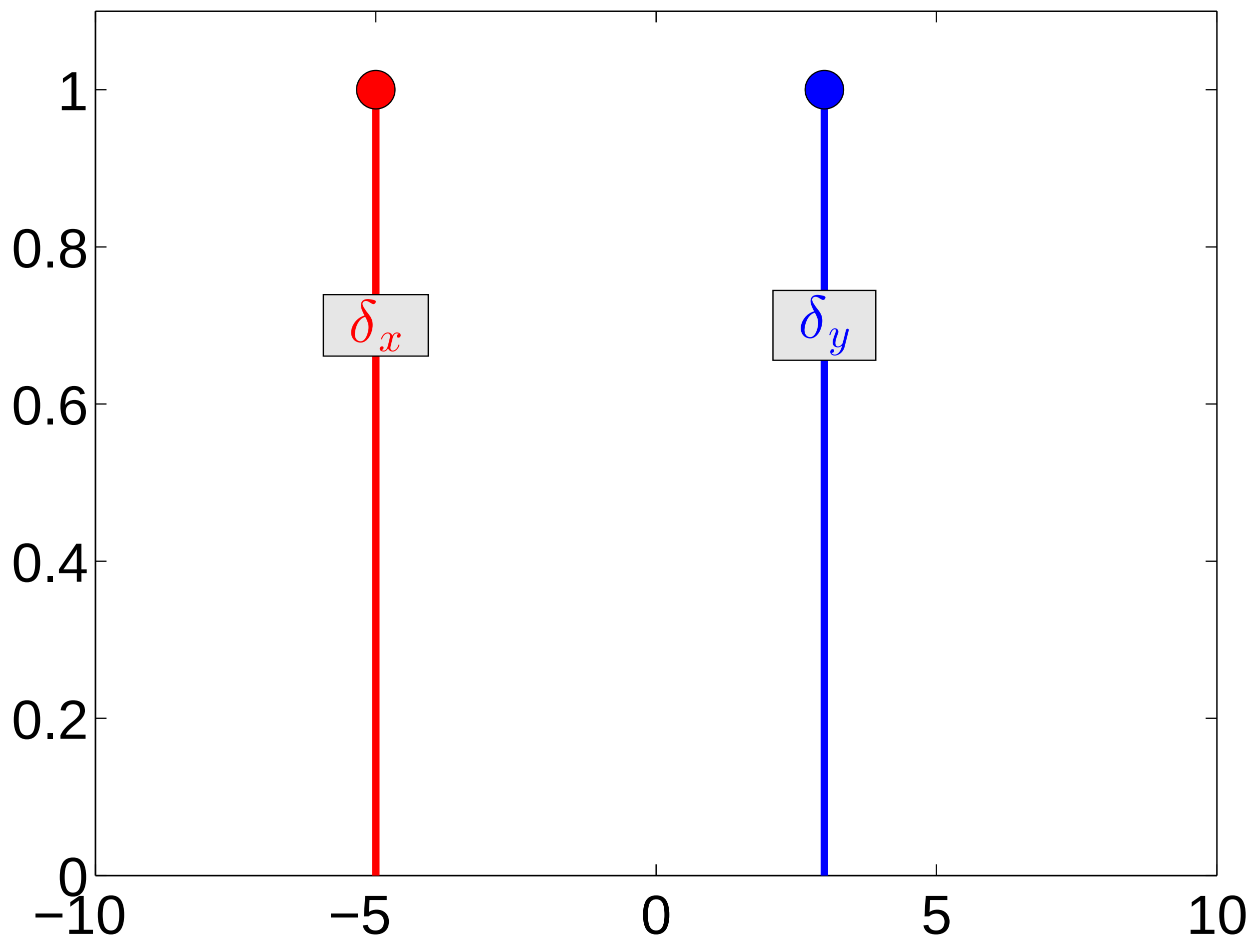
2 Points on the Real Line



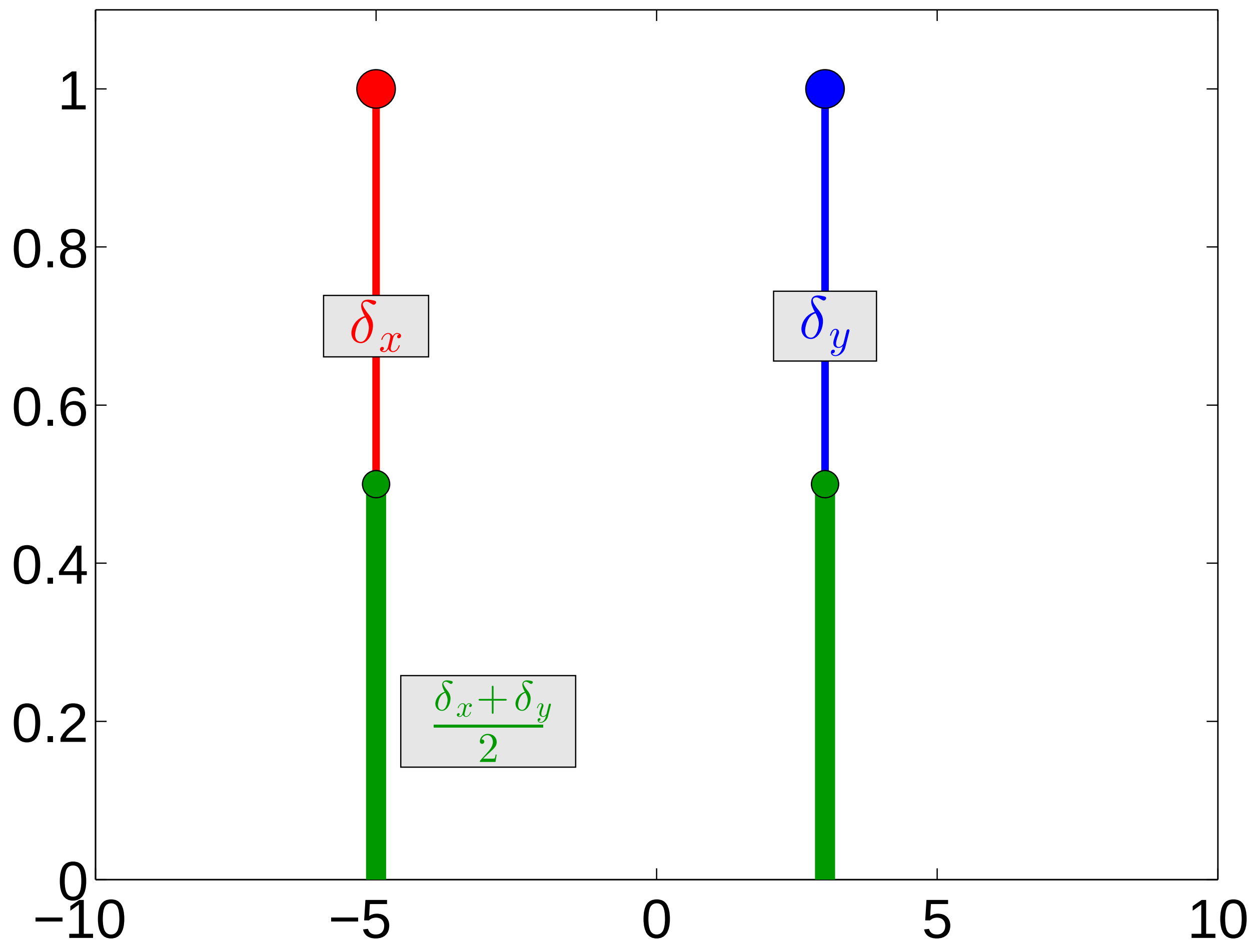
Their Average



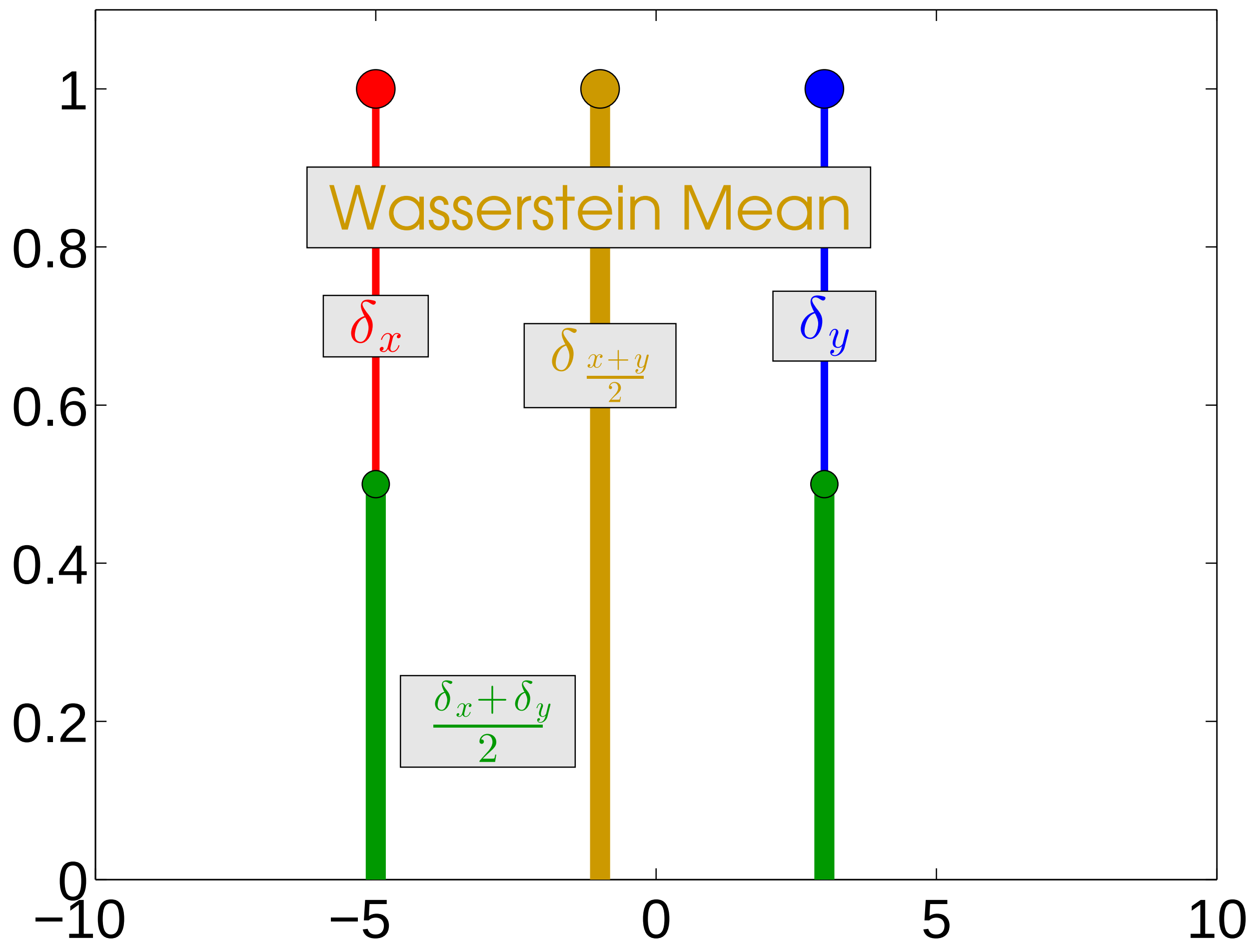
2 Points as Diracs



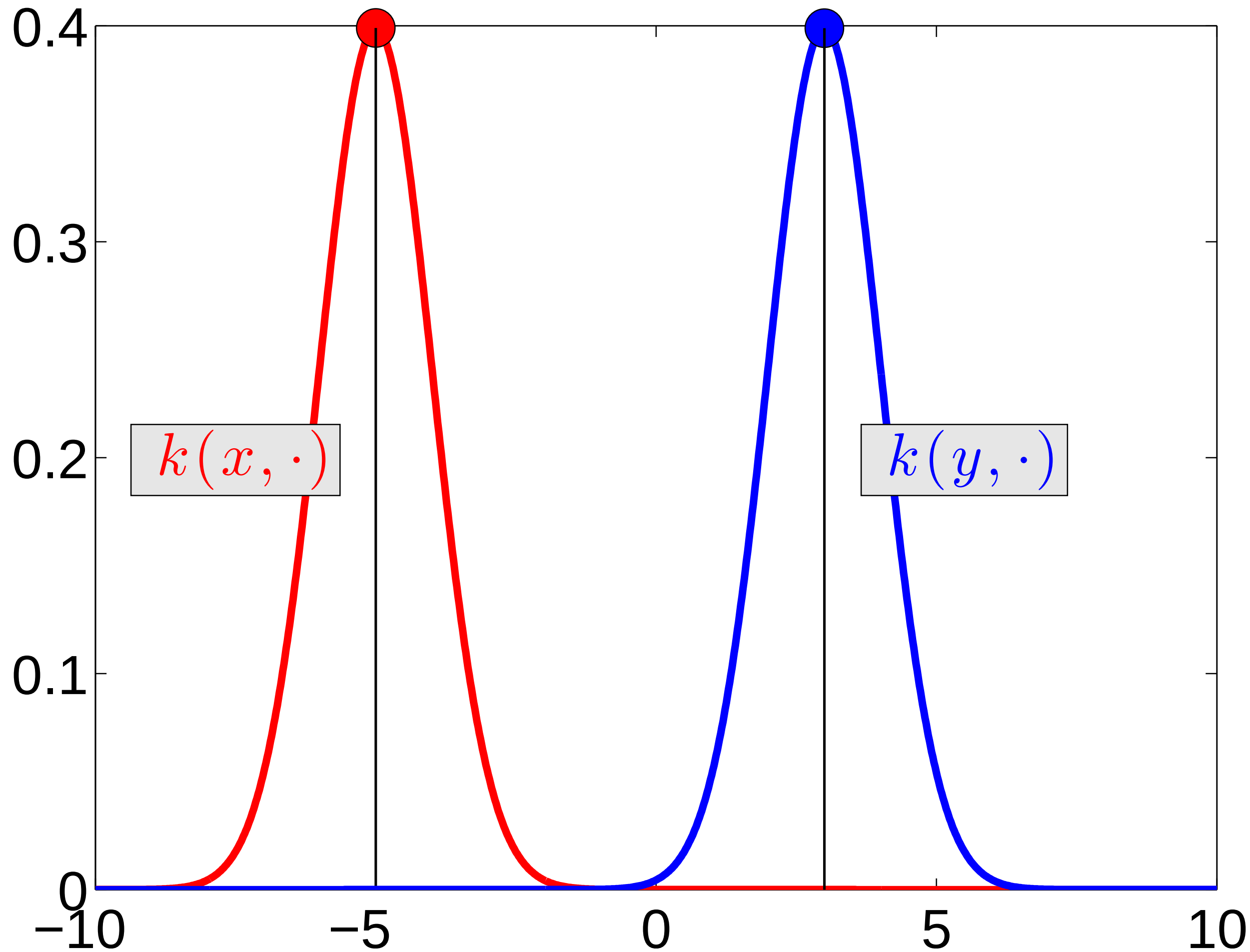
Euclidean Mean of Diracs



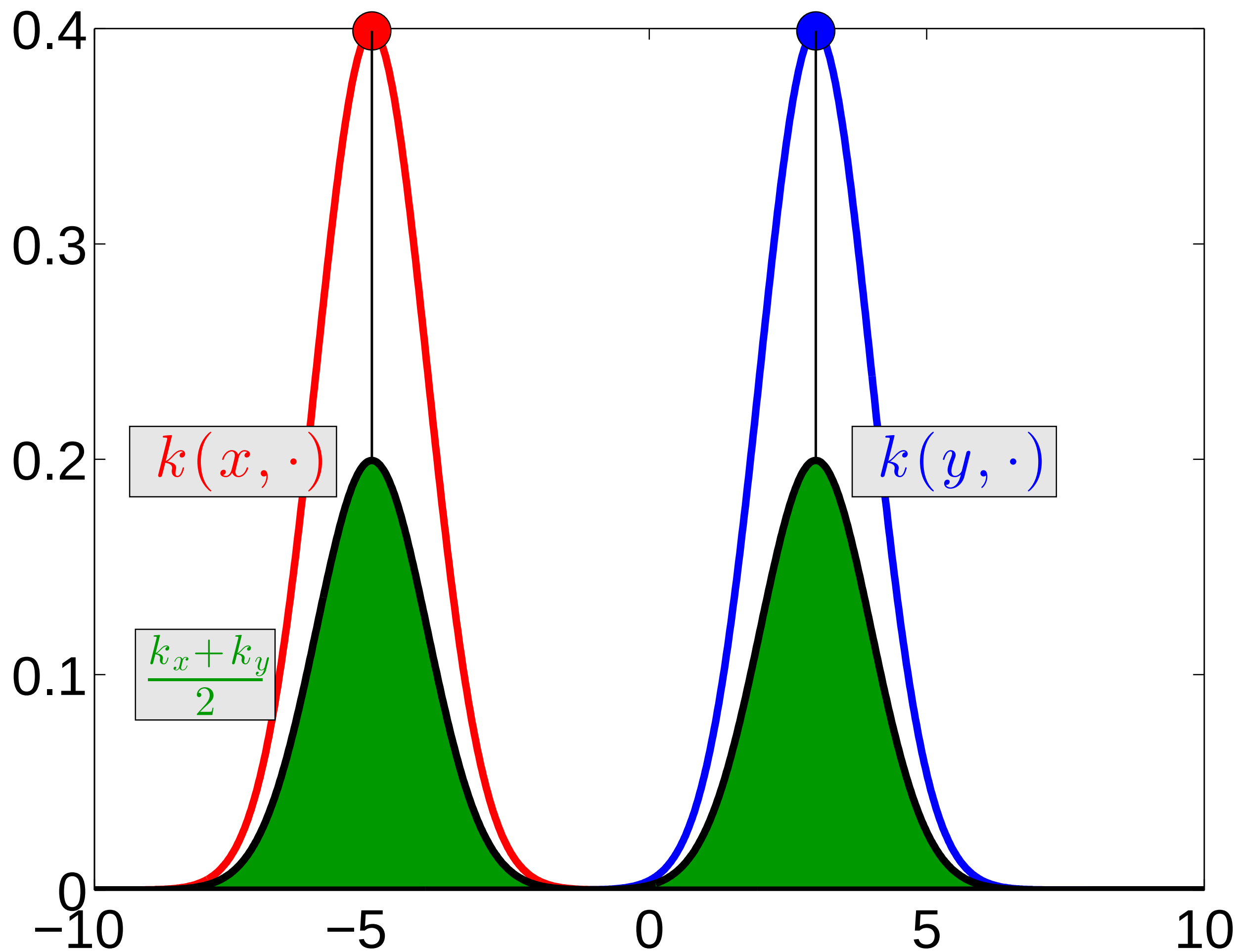
Wasserstein Mean of Diracs



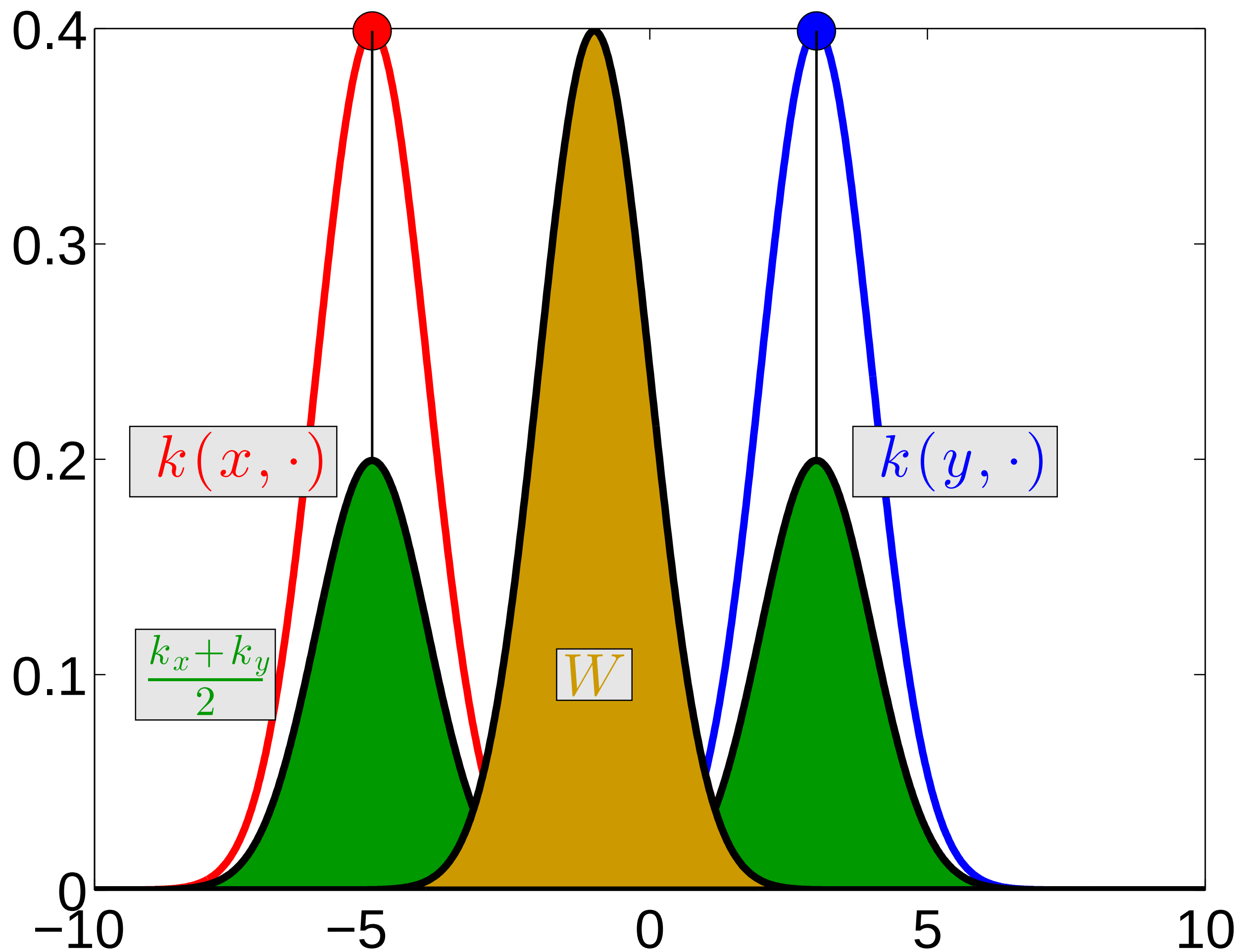
Smoothed Measures (RKHS mean map)



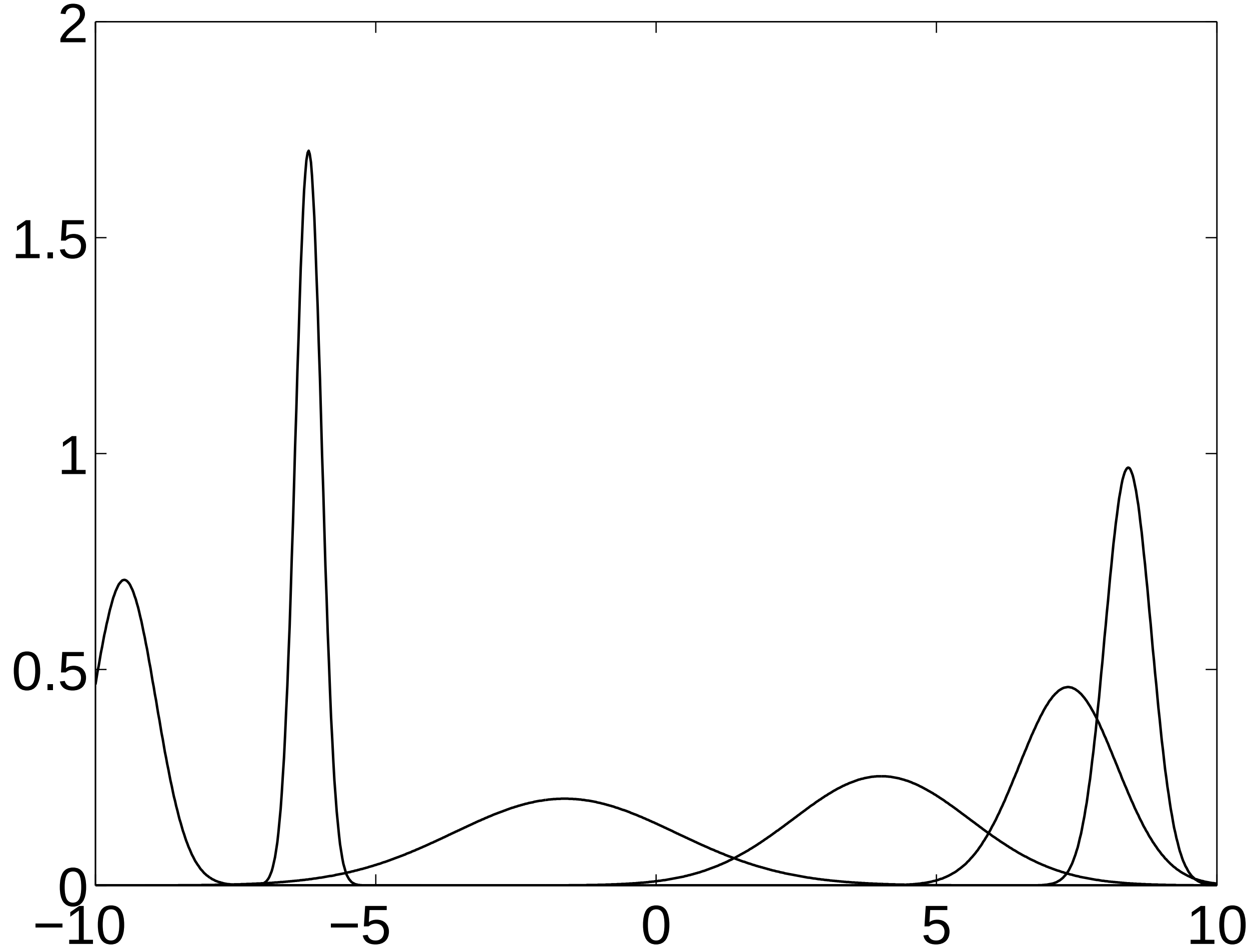
Euclidean Mean of 2 Gaussians



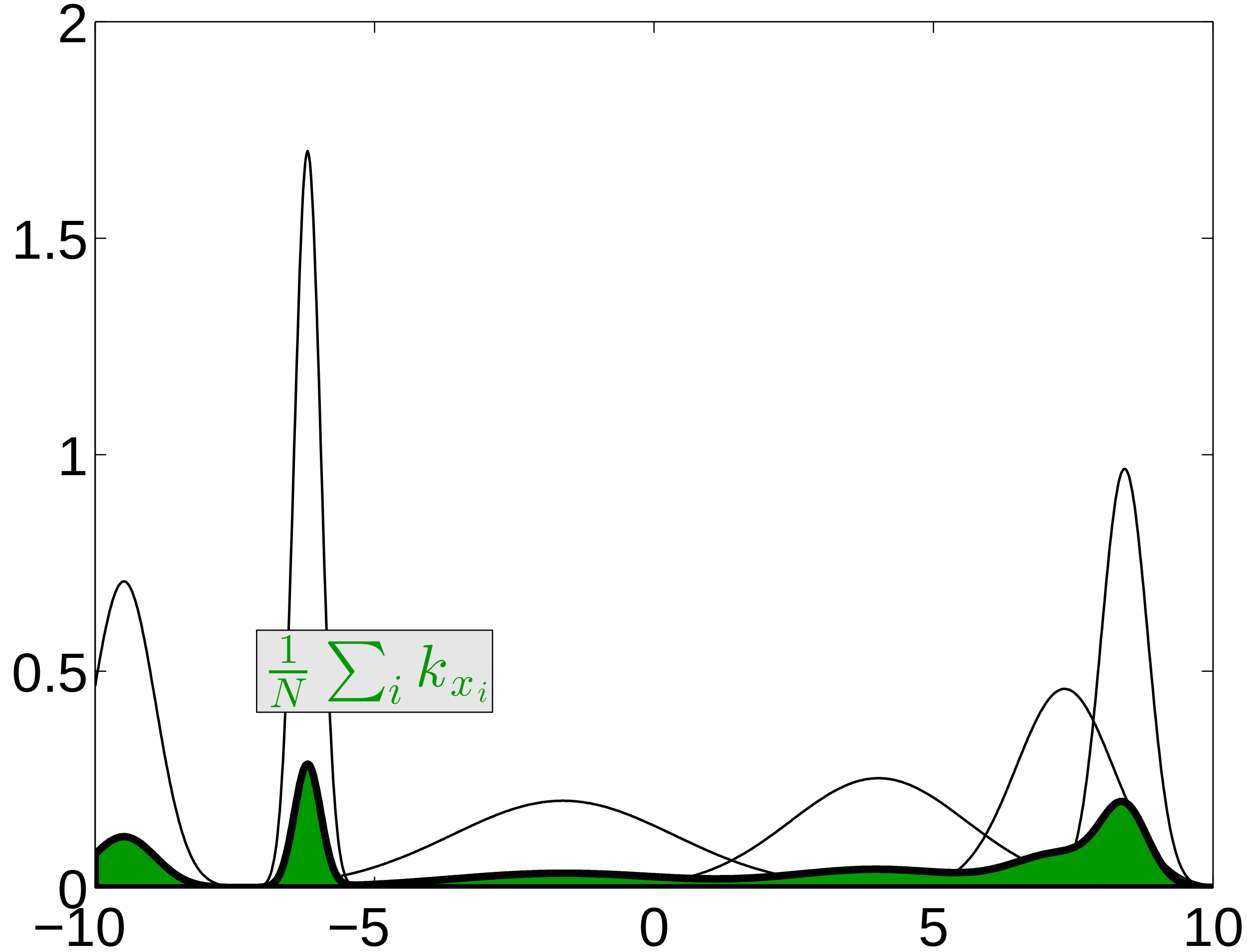
Wasserstein Mean of 2 Gaussians



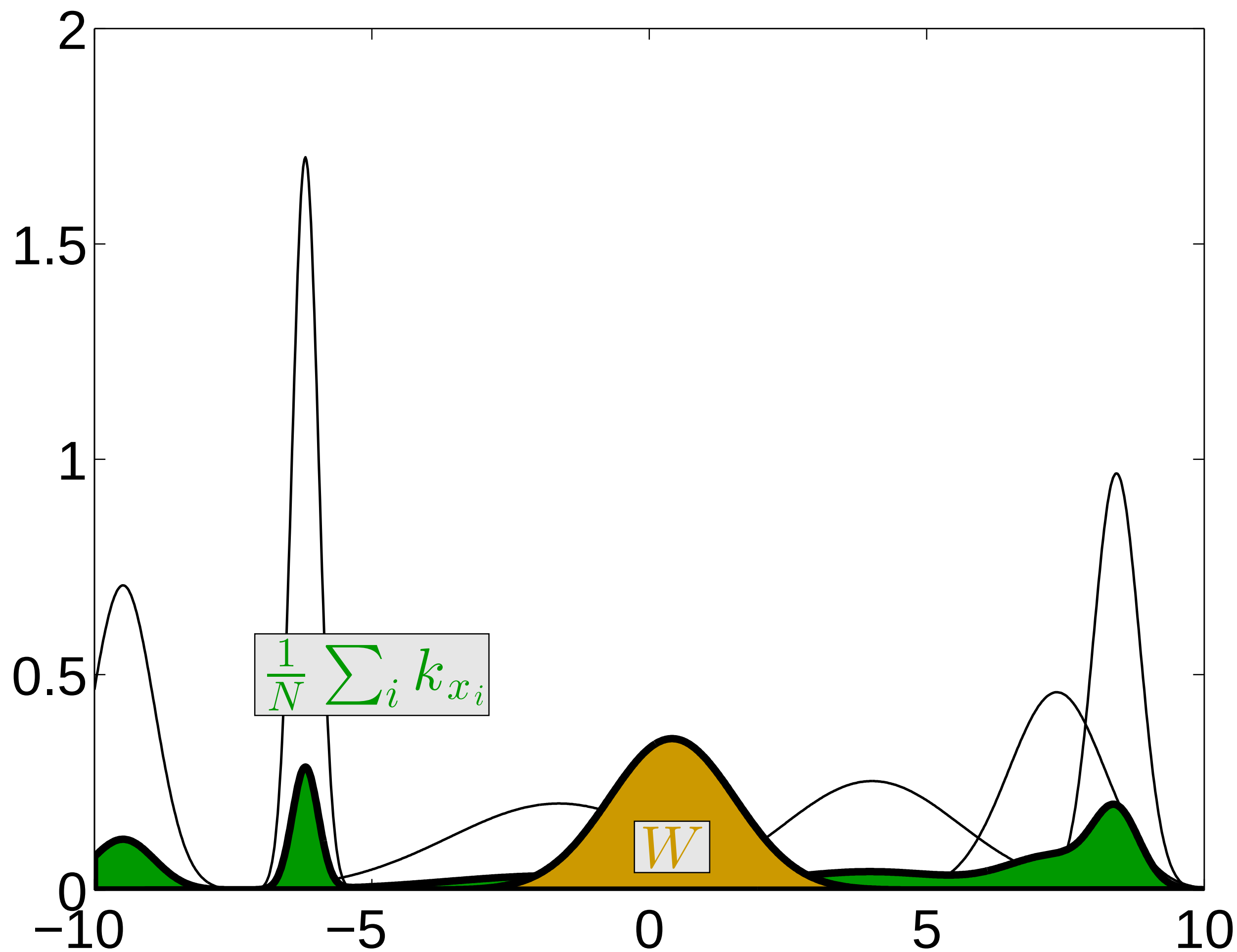
6 Gaussians



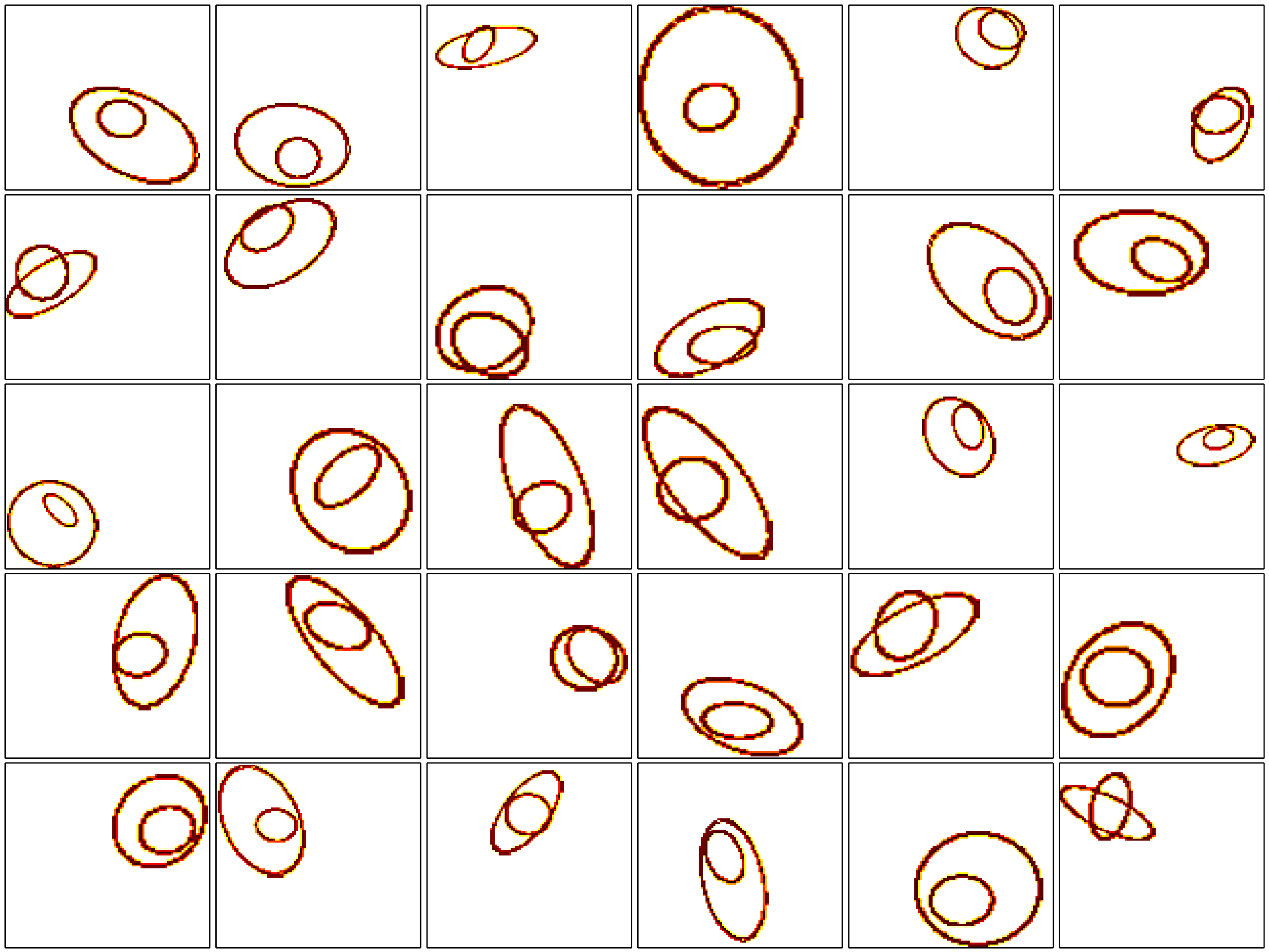
Euclidean Mean



Wasserstein Mean

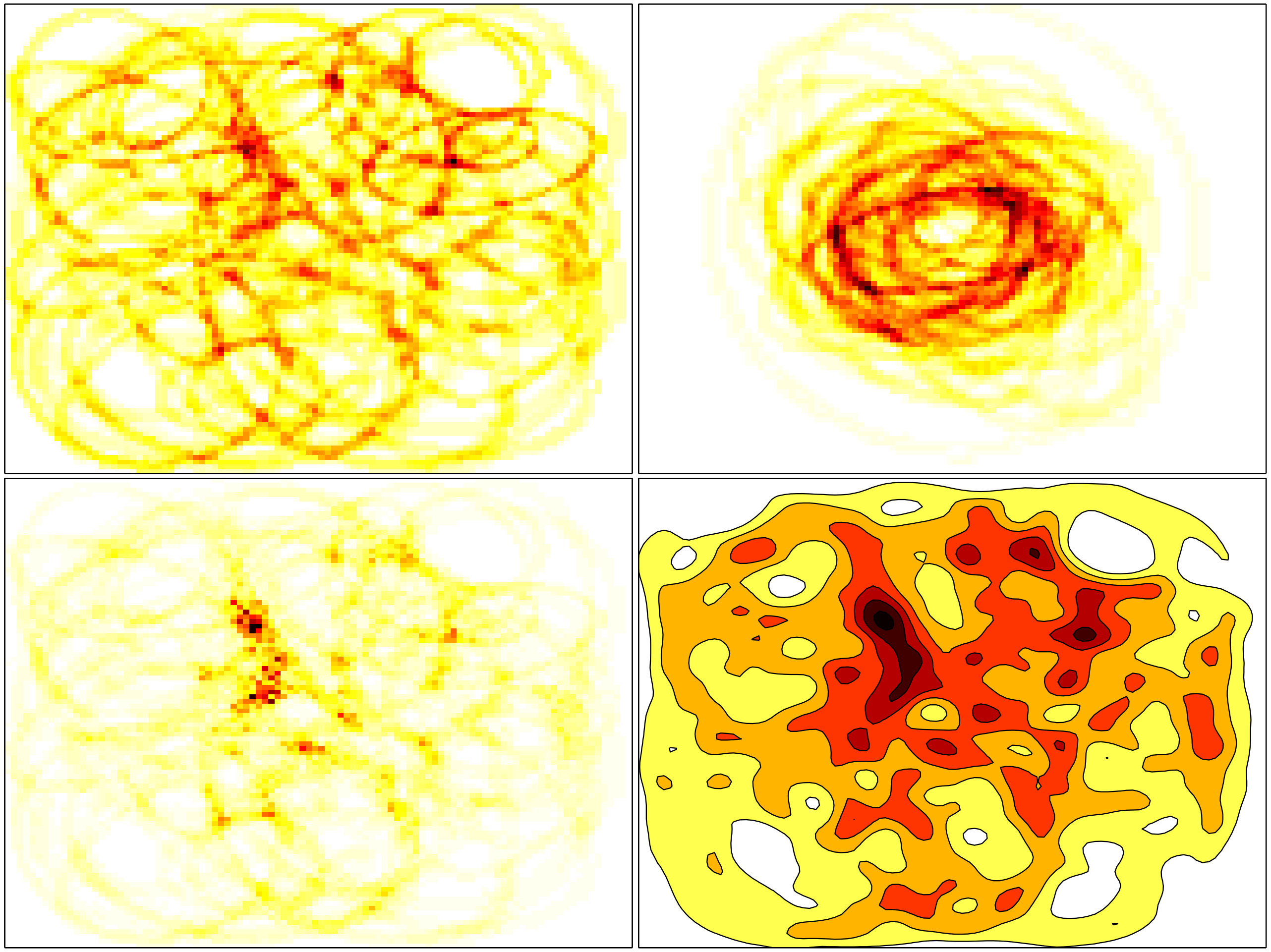


Motivation in 2D



$$\{\nu_1, \cdots, \nu_{30}\} \in P([0,1]^2).$$

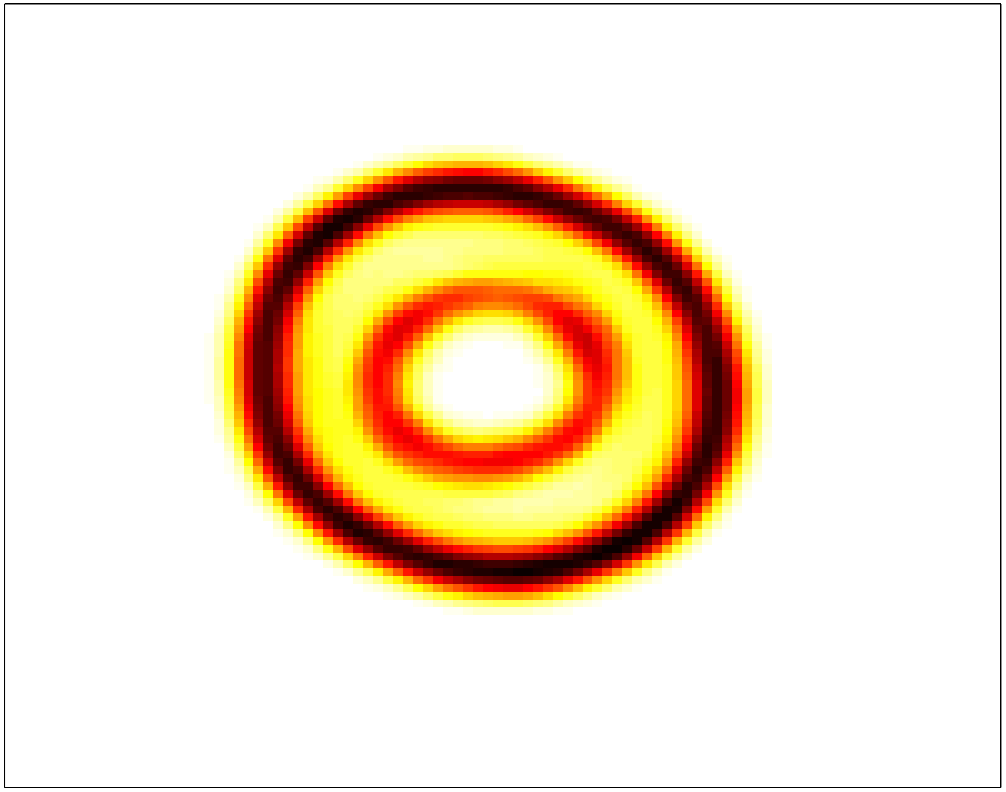
Euclidean / Centered / Jeffrey / RKHS



Euclidean distance / recentered,

Sym. Kullback / RKHS Mean Map

Wasserstein Barycenter

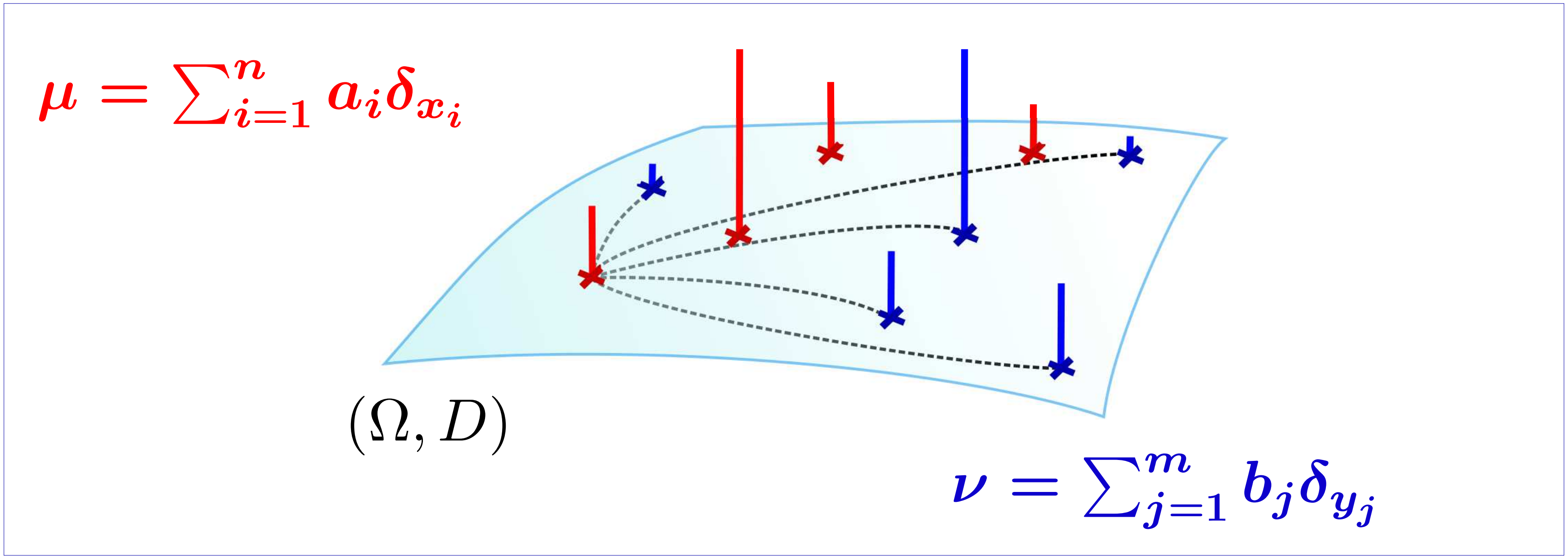


2-Wasserstein barycenter

(computed with our method)

Variational Perspective on the Wasserstein Distance

Wasserstein for Empirical Measures



- (Ω, D) metric. $p \geq 1$.
- Two empirical measures μ, ν.

p-Wasserstein distance $W_p(\mu, \nu)$?

Computing p Wasserstein Distances

$$\mu = \sum_{i=1}^{n} \boldsymbol{a_i} \delta_{\boldsymbol{x_i}}, \quad \nu = \sum_{j=1}^{m} \boldsymbol{b_j} \delta_{\boldsymbol{y_j}},$$

$W_p(\boldsymbol{\mu}, \boldsymbol{\nu})$ is the solution of a linear program involving:

1. $M_{\boldsymbol{XY}} \stackrel{\text{def}}{=} [D(\boldsymbol{x_i}, \boldsymbol{y_j})^p]_{ij} \in \mathbb{R}^{n \times m}$
2. $U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{T \in \mathbb{R}_+^{n \times m} \mid T\mathbb{1}_m = \boldsymbol{a},\ T^T\mathbb{1}_n = \boldsymbol{b}\}.$

Computing the OT Distance

- p-Wasserstein is the solution (primal or dual LP):

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

Computing the OT Distance

- p-Wasserstein is the solution (primal or dual LP):

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

$$\boxed{\text{Changes in } f(\boldsymbol{a},\boldsymbol{X}) \stackrel{\text{def}}{=} W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) \text{ as } \boldsymbol{a} \text{ \& } \boldsymbol{X} \text{ change?}}$$

Wasserstein (Sub)differentiability

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

- $\partial f|_a = \boldsymbol{\alpha}^\star$: the *dual optimum* $\boldsymbol{\alpha}^\star$ is a subgradient.

$$\boxed{f(a, \boldsymbol{X}) = \min_{T\in U(\boldsymbol{a},\boldsymbol{b})} \langle T, M_{\boldsymbol{XY}} \rangle}$$

- $\partial f|_X = Y\boldsymbol{T}^{\star\boldsymbol{T}}\,\mathbf{diag}(a^{-1})$: *primal optimum* $\boldsymbol{T}^{\star\boldsymbol{T}}$ yields a subgradient (when D =Euclidean, $p = 2$).

Average of Wasserstein Distances

$$g(\boldsymbol{a}, \boldsymbol{X}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$
$$= \frac{1}{N}\sum_{i=1}^{N} \mathbf{primal}(\boldsymbol{a}, \boldsymbol{b_i}, M_{\boldsymbol{X}\mathbf{Y}_{\boldsymbol{i}}})$$

- $a \to g(a, X)$ is **convex**, **non-smooth**
- $X \to g(a, X)$ is **not convex**, **non-smooth**

Wasserstein Barycenter Problem

$$\min_{\boldsymbol{a}} g(\boldsymbol{a}, X) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{primal}(\boldsymbol{a}, \boldsymbol{b_i}, M_{X\mathbf{Y}_{\boldsymbol{i}}})$$

- $\boldsymbol{a} \rightarrow g(\boldsymbol{a}, X)$ is **convex**
  - subgradient method works (in theory).
  - Great if X is fixed (b-o-w or discretized Ω)!
  - **Need to solve $\{\alpha_i^\star\}$ at each subgradient step.**

Wasserstein Barycenter Problem

$$\min_{\boldsymbol{X}} g(a, \boldsymbol{X}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{primal}(a, \boldsymbol{b_i}, M_{\boldsymbol{X}\mathbf{Y}_{\boldsymbol{i}}})$$

- $\boldsymbol{X} \to g(a, \boldsymbol{X})$ is **not convex**
  - (and so far only applicable when Ω is $\mathbb{R}^d$.)
  - local minimum with subgradient method
  - **Need to compute $\{T_i^\star\}$ at each subgradient step**

To recapitulate...

$$\min_{\boldsymbol{a},\boldsymbol{X}} g(\boldsymbol{a}, \boldsymbol{X})$$

- **convex** w.r.t weights a, not locations X.
- **only subgradients** (g is usually very degenerate).
- **computationally intractable** (cost of OT $\approx n^3 \log n$)
- **computationally inefficient** (hard to parallelize)

Solution: Entropic Smoothing

Original primal problem gives us $\boldsymbol{T}^{\star}$:

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

Original dual problem gives us $\boldsymbol{\alpha}^{\star}$:

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

Solution: Entropic Smoothing

Smoothed ($\boldsymbol{\lambda > 0}$) primal problem gives us $\boldsymbol{T^\star_\lambda}$:

$$\mathbf{primal}_{\boldsymbol{\lambda}}(a, b, M_{XY}) = \min_{\boldsymbol{T} \in U(a,b)} \langle \boldsymbol{T}, M_{XY} \rangle - \frac{\mathbf{1}}{\boldsymbol{\lambda}} \boldsymbol{h}(\boldsymbol{T})$$

Smoothed dual problem gives us $\boldsymbol{\alpha^\star_\lambda}$:

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

Benefits of Smoothing [Cuturi'13]

- Objective now **strongly convex** vs. **piecewise linear**: infinitely more efficient in practice **[Nesterov'05]**.
- Primal/dual smoothed optima $\boldsymbol{\alpha}_\lambda^\star, \boldsymbol{T}_\lambda^\star$ can be solved
  - **In** $O(n^2)$ with **Sinkhorn's algorithm**,
  - in **parallel on GPGPUs** for **any metric** on finite Ω,
  - **millions of time faster** than simplex,
  - can deal with **large dimensions** (≈ 20.000).

To conclude...

- our approach also **generalizes k-means**
  - can consider weight constraints (see paper),
  - can quantize simultaneously different datasets
- Versatile and scalable approach for **other variational Wasserstein problems** ($e.g.$ Wasserstein propagation **[Solomon'14]**)
- Future applications to visualization of measures on Riemanian manifolds, data-fusion, inference...