Fast Computation of Wasserstein Distances and Applications to Parameter Estimation

Marco Cuturi
Kyoto University

京都大学
KYOTO UNIVERSITY

科研費
KAKENHI

Joint work with G. Peyré, G. Carlier (Dauphine), J. Solomon (Princeton/MIT), F. de Goes (Pixar), A. Gramfort (ParisTech), J.D. Benamou (INRIA), V. Seguy, A. Rolet (Kyoto), G. Montavon and K.R. Müller (TU Berlin) and more...

What is Optimal Transport?

A geometric toolbox to
compare probability measures
supported on a metric space.

Monge

Kantorovich

Dantzig

Wasserstein

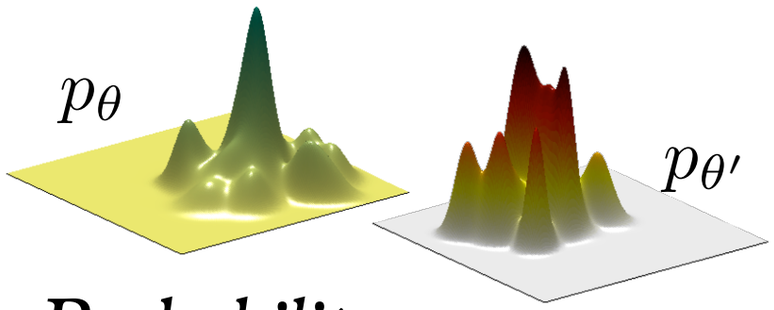
Brenier

McCann

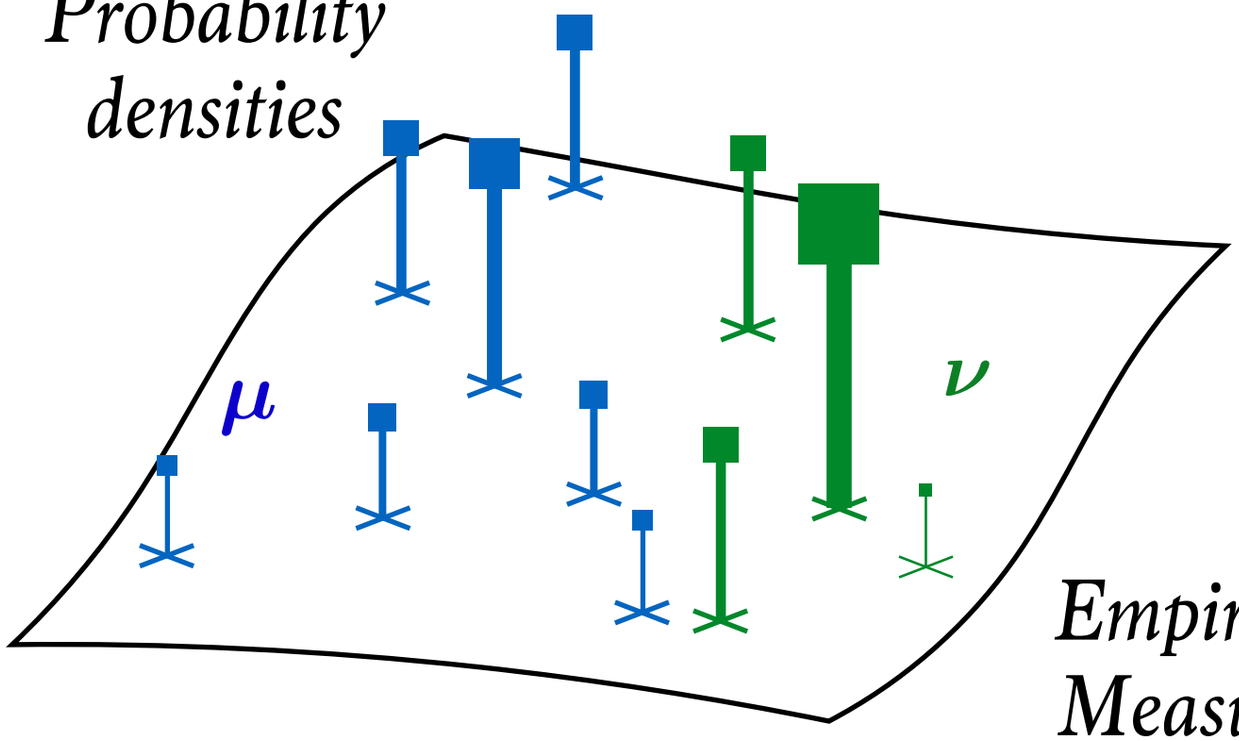
Villani

What is Optimal Transport?

A geometric toolbox to
compare **probability measures**
supported on a metric space.

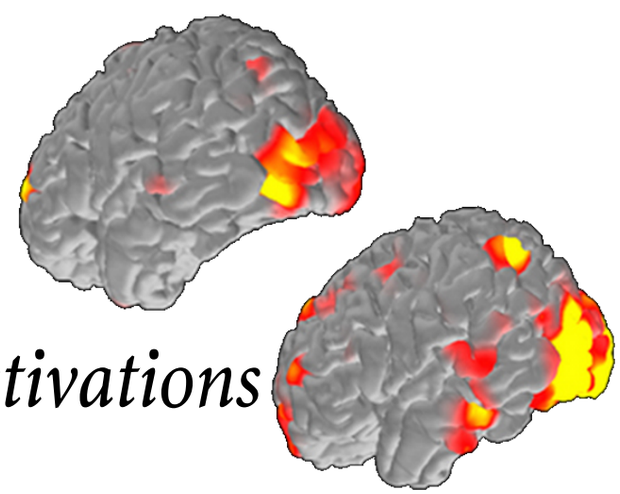


Probability densities



Empirical Measures

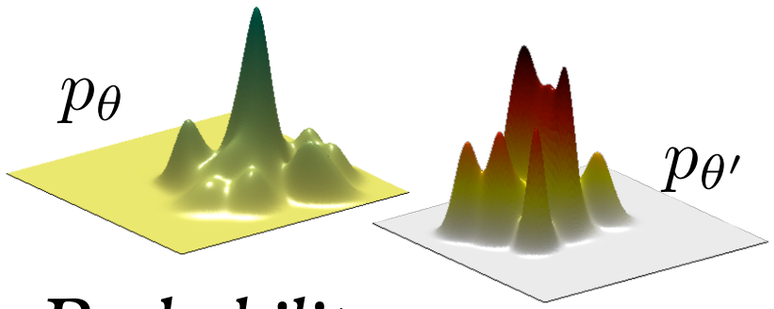
Bags of Features



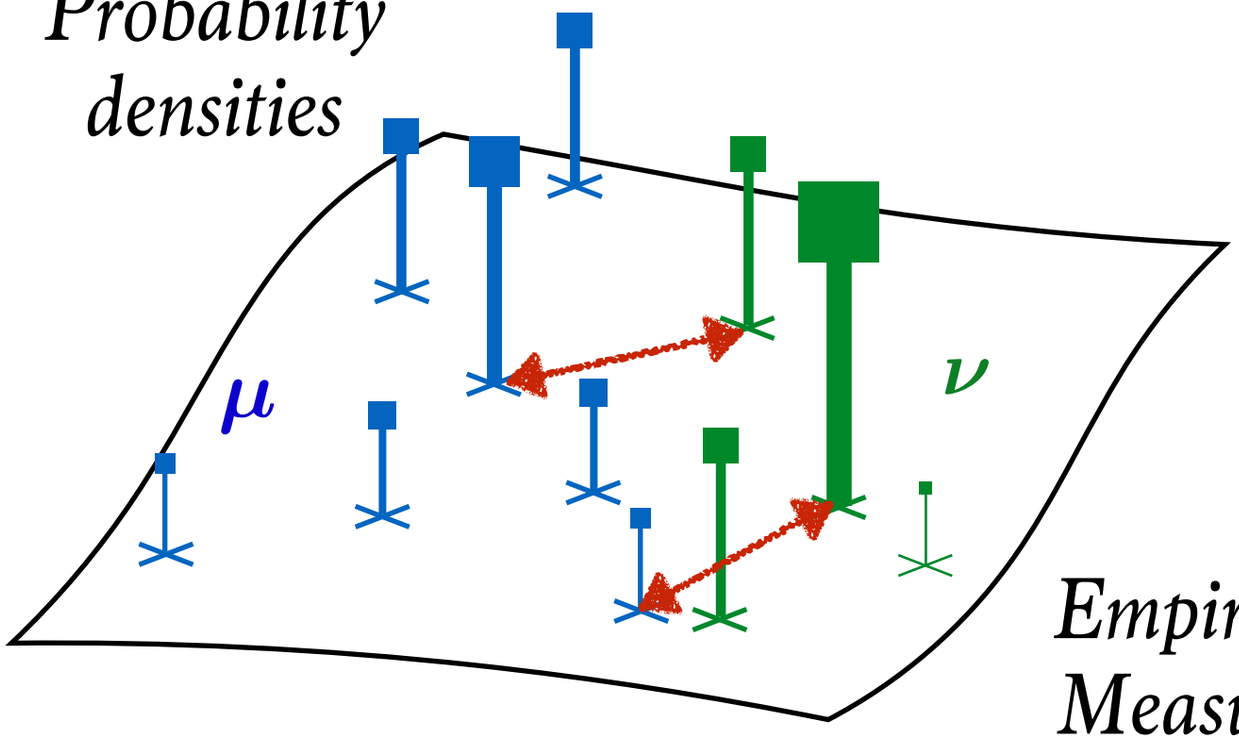
Brain Activations

What is Optimal Transport?

A geometric toolbox to compare **probability measures supported on a metric space.**

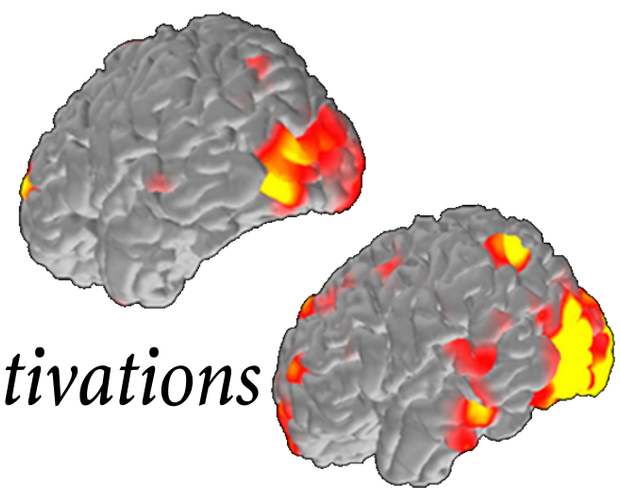


Probability densities



Empirical Measures

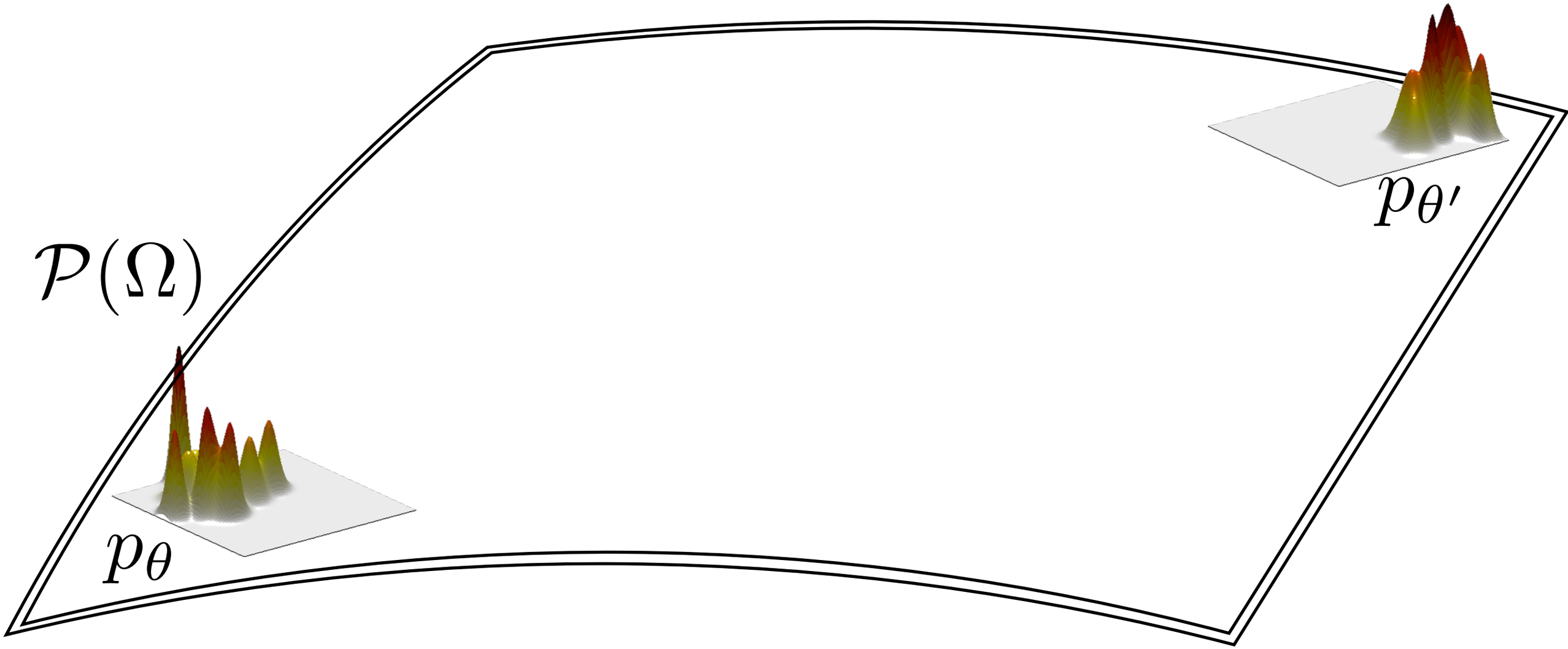
Bags of Features



Brain Activations

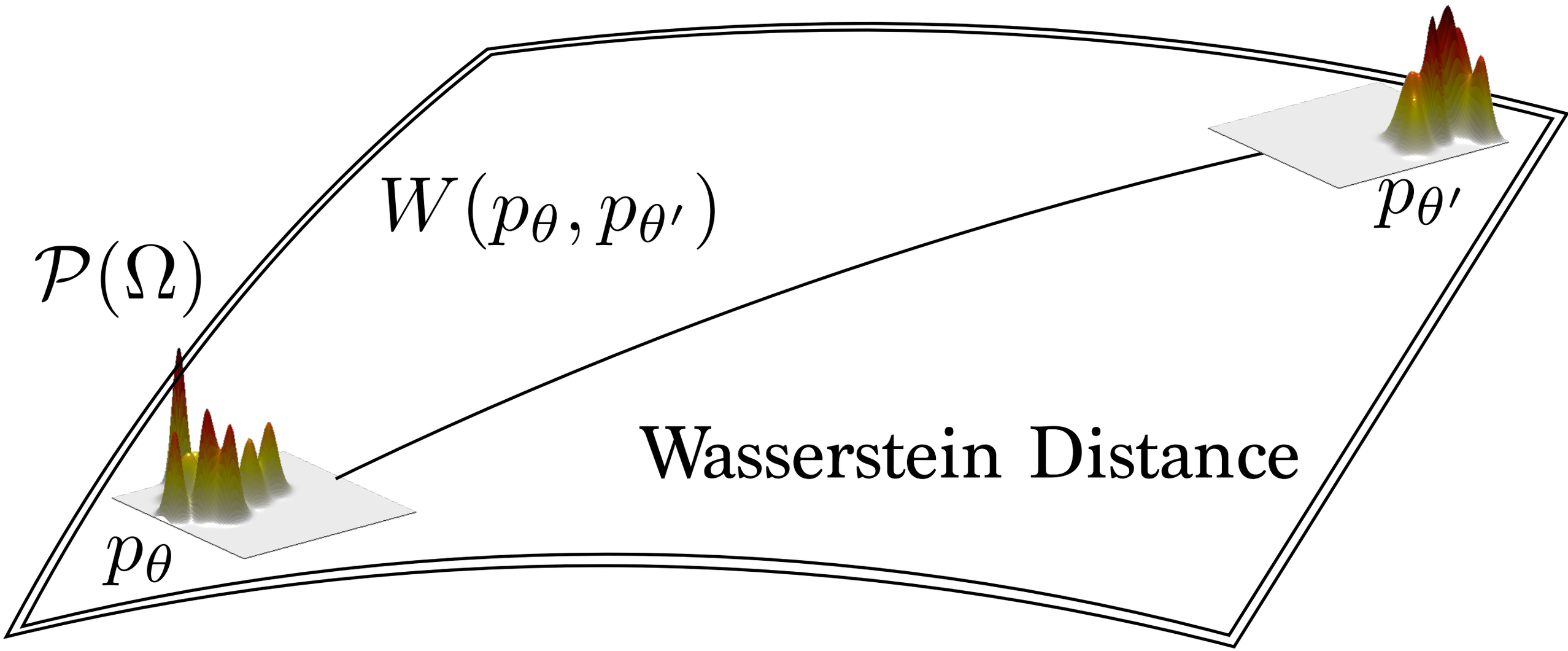
What is Optimal Transport?

A **geometric toolbox** to
compare probability measures
supported on a metric space.



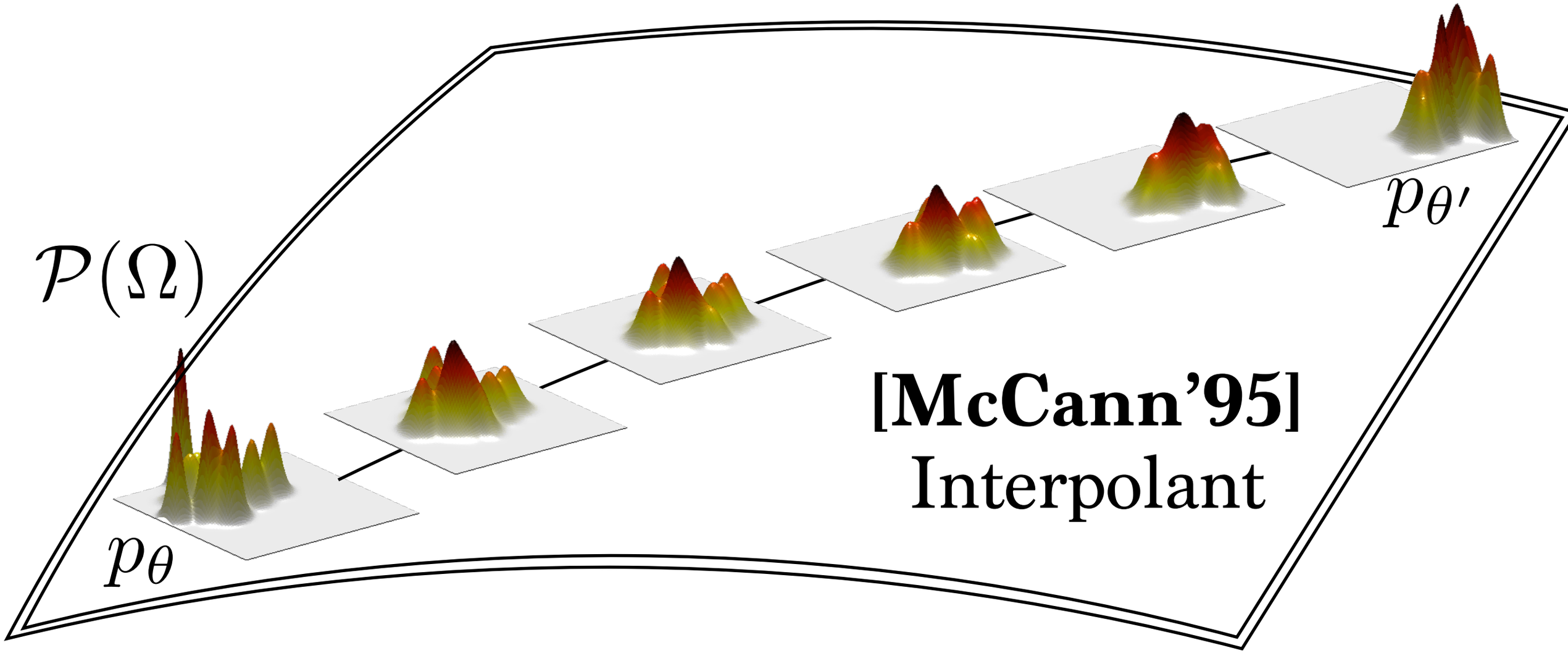
What is Optimal Transport?

A **geometric toolbox** to
compare probability measures
supported on a metric space.



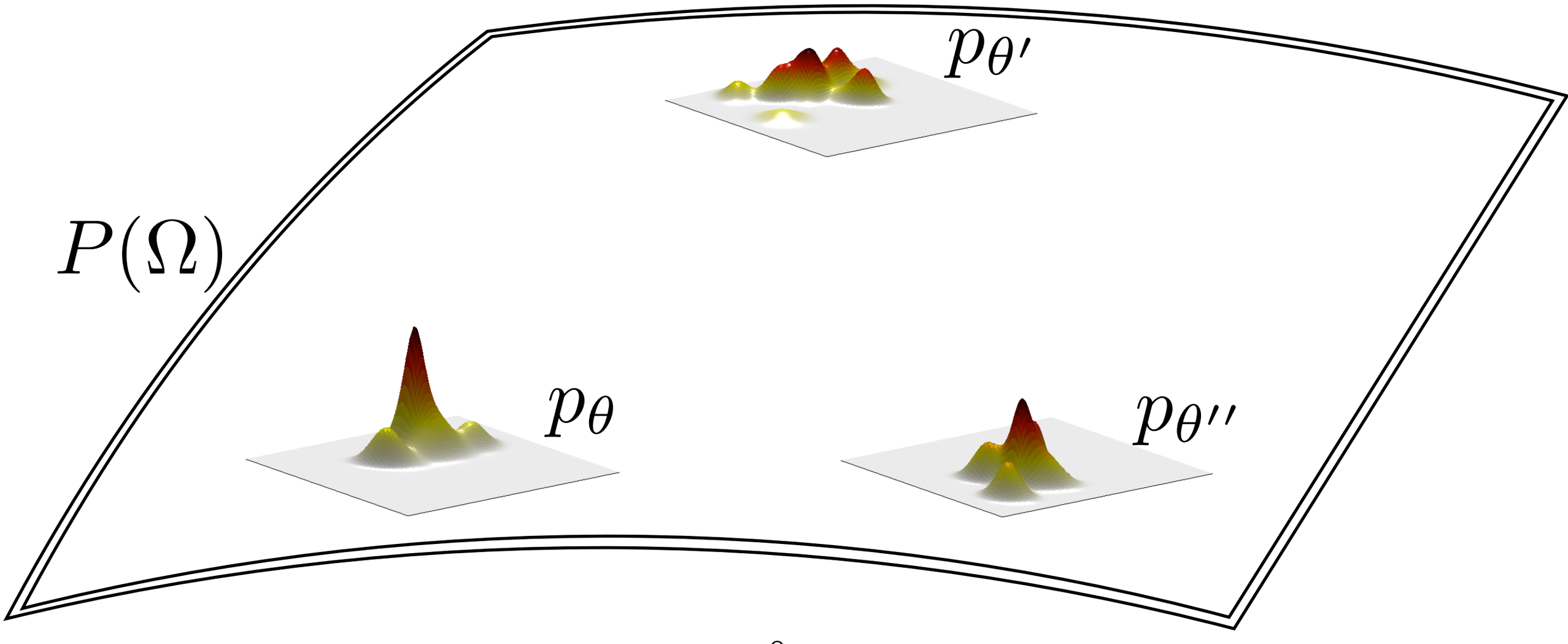
What is Optimal Transport?

A **geometric toolbox** to
compare probability measures
supported on a metric space.



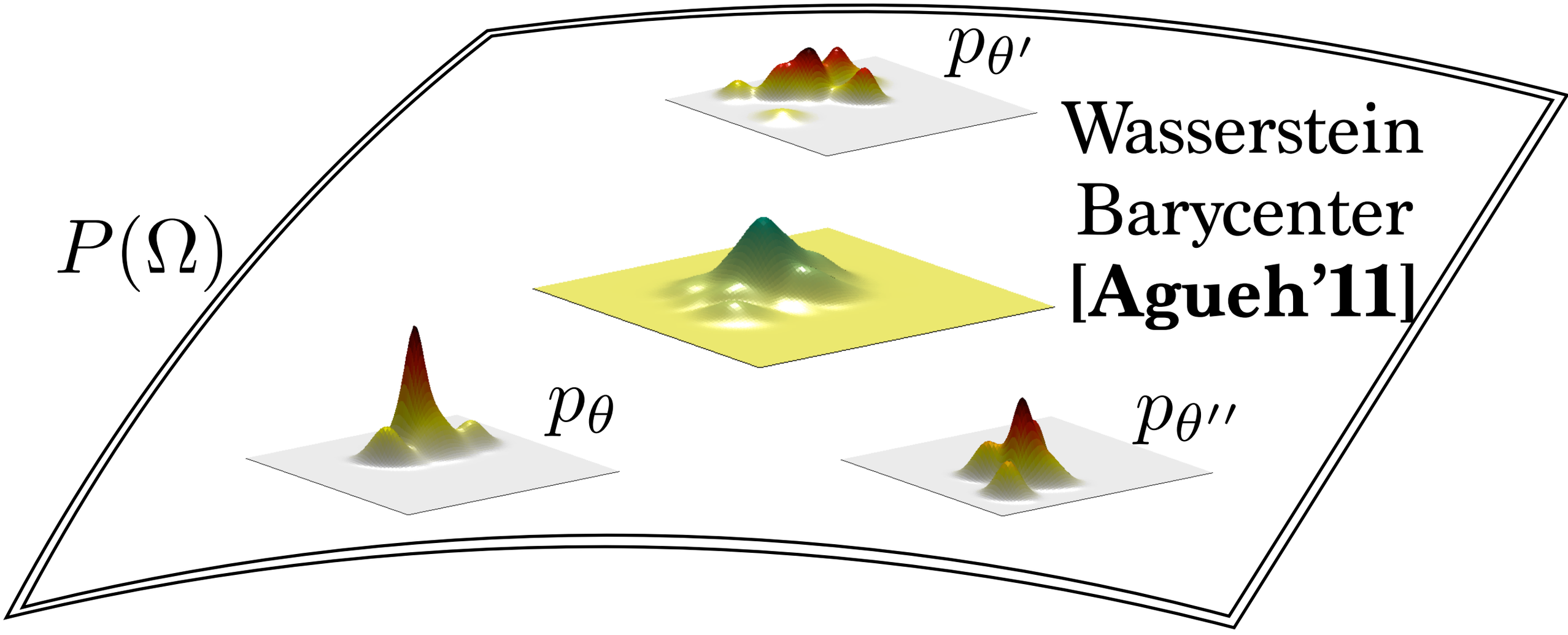
What is Optimal Transport?

A **geometric toolbox** to
compare probability measures
supported on a metric space.



What is Optimal Transport?

A **geometric toolbox** to
compare probability measures
supported on a metric space.



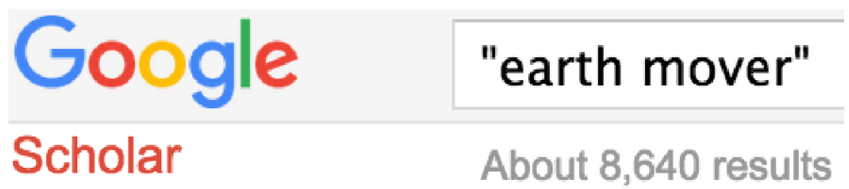
Why is it relevant to ML?

- New geometry for statistical modeling
  - Information geometry is crucial in stats.
  - That geometry is often KL (*e.g.* MLE).

- New algorithms to study histogram data
  - Bags-of-features are everywhere.
  - Knowledge on these features is often known but not used.

Why now?

- Key results in maths since ’95~
  - **[McCann’95], [JKO’98], [Benamou’98], [Ambrosio’06], [Villani’03/’09]**
- More work in CV/TCS/Graphics since ’98~
  - Earth Mover’s Distance **[Rubner’98]**, Embeddings **[Indyk’03]**



- Longstanding roadblock: **computation**
  - Regularization **[C.’13]** can provide the key

Outline

- Definitions: The Wasserstein Distances
- Fast computations with regularization
- Wasserstein variational problems
  - barycenters
  - dictionary learning
  - PCA
  - minimum Kantorovich estimation

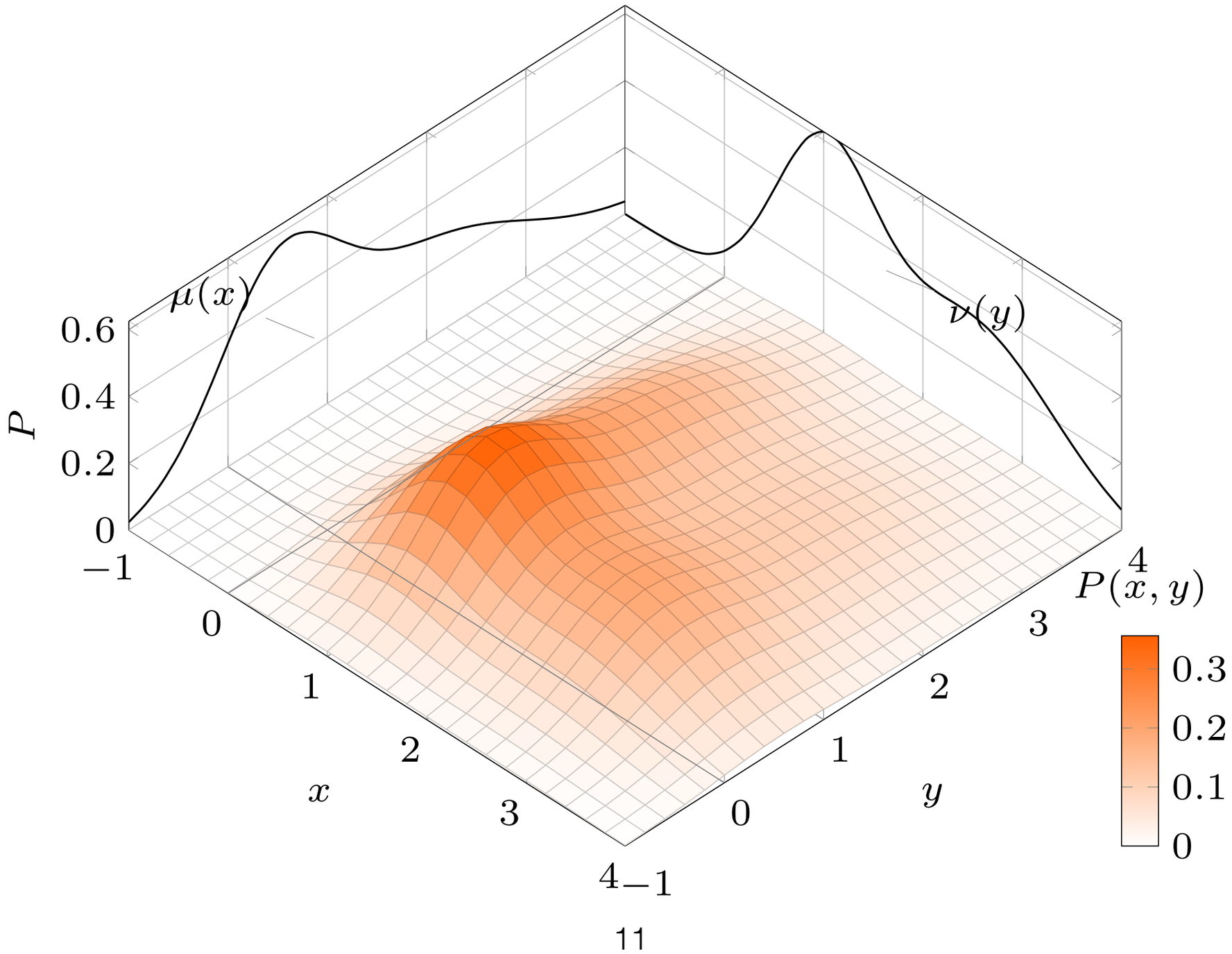
Definitions: Couplings & Wasserstein

Assume $(\Omega, \boldsymbol{D})$ is a probability space endowed with a metric.

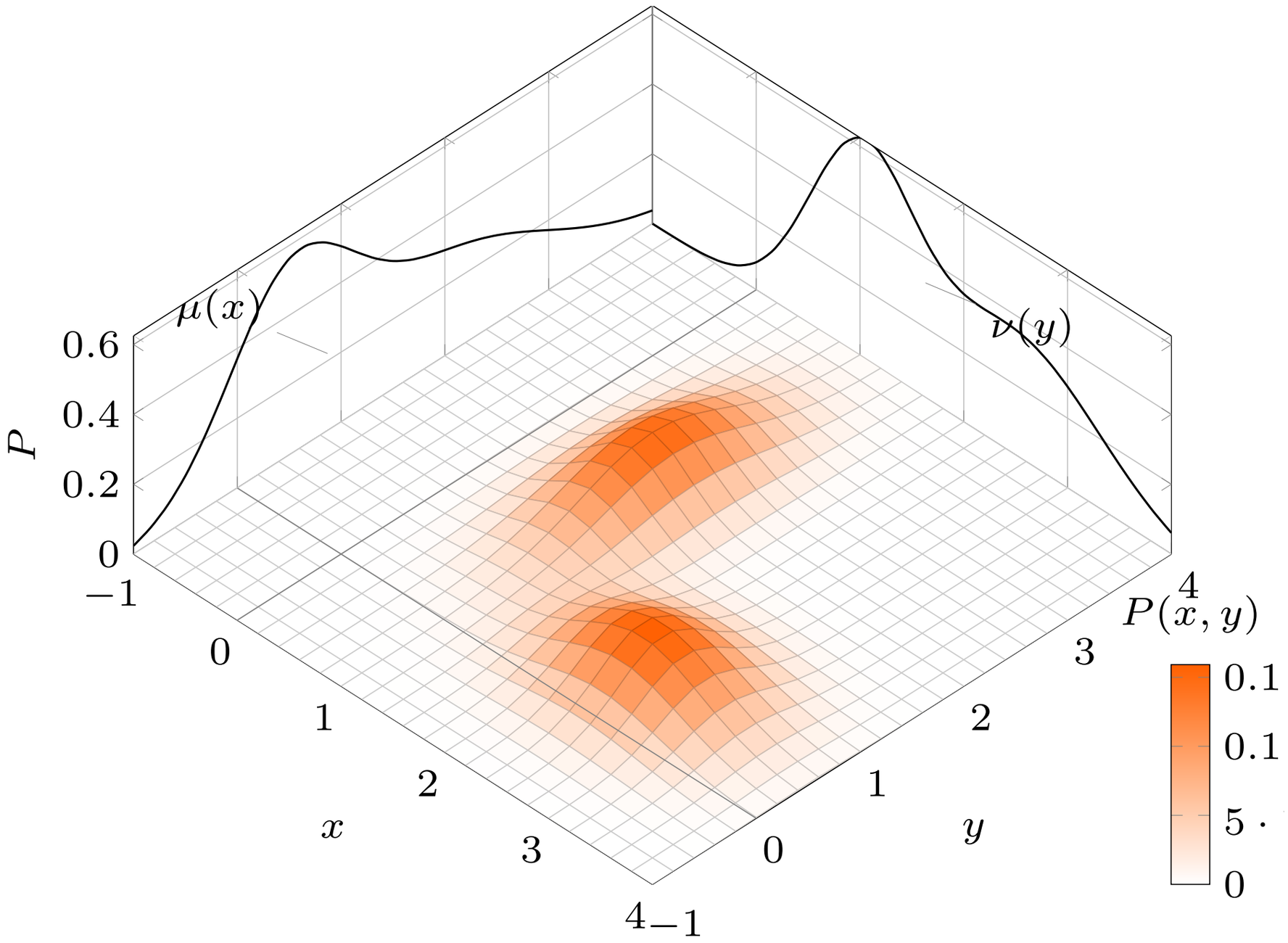
For $\boldsymbol{\mu}, \boldsymbol{\nu}$ probability measures in $\mathcal{P}(\Omega)$,

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) | \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega,$$
$$\boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}),$$
$$\boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$

Couplings



Couplings

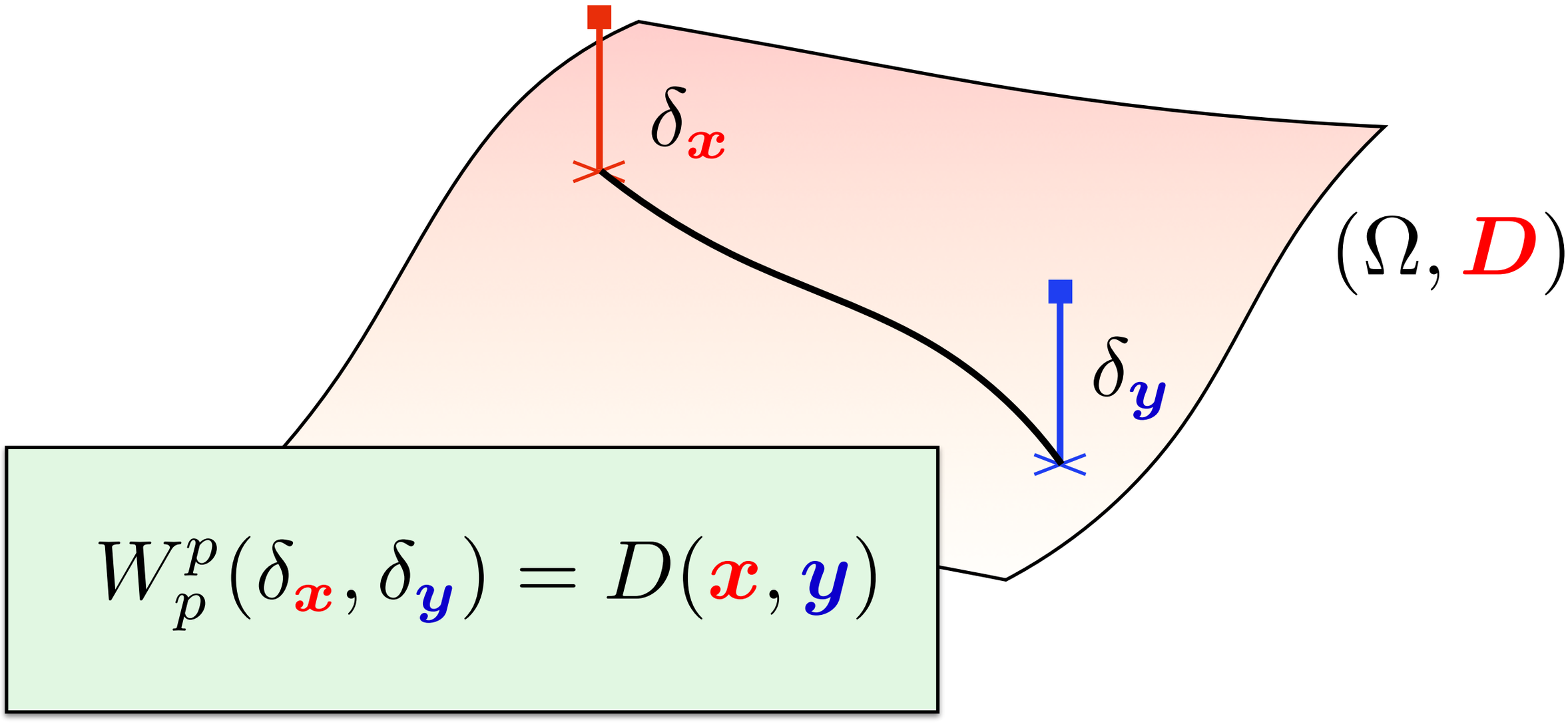


Wasserstein Distance

Def. For $p \geq 1$, the p-Wasserstein distance between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$ is

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

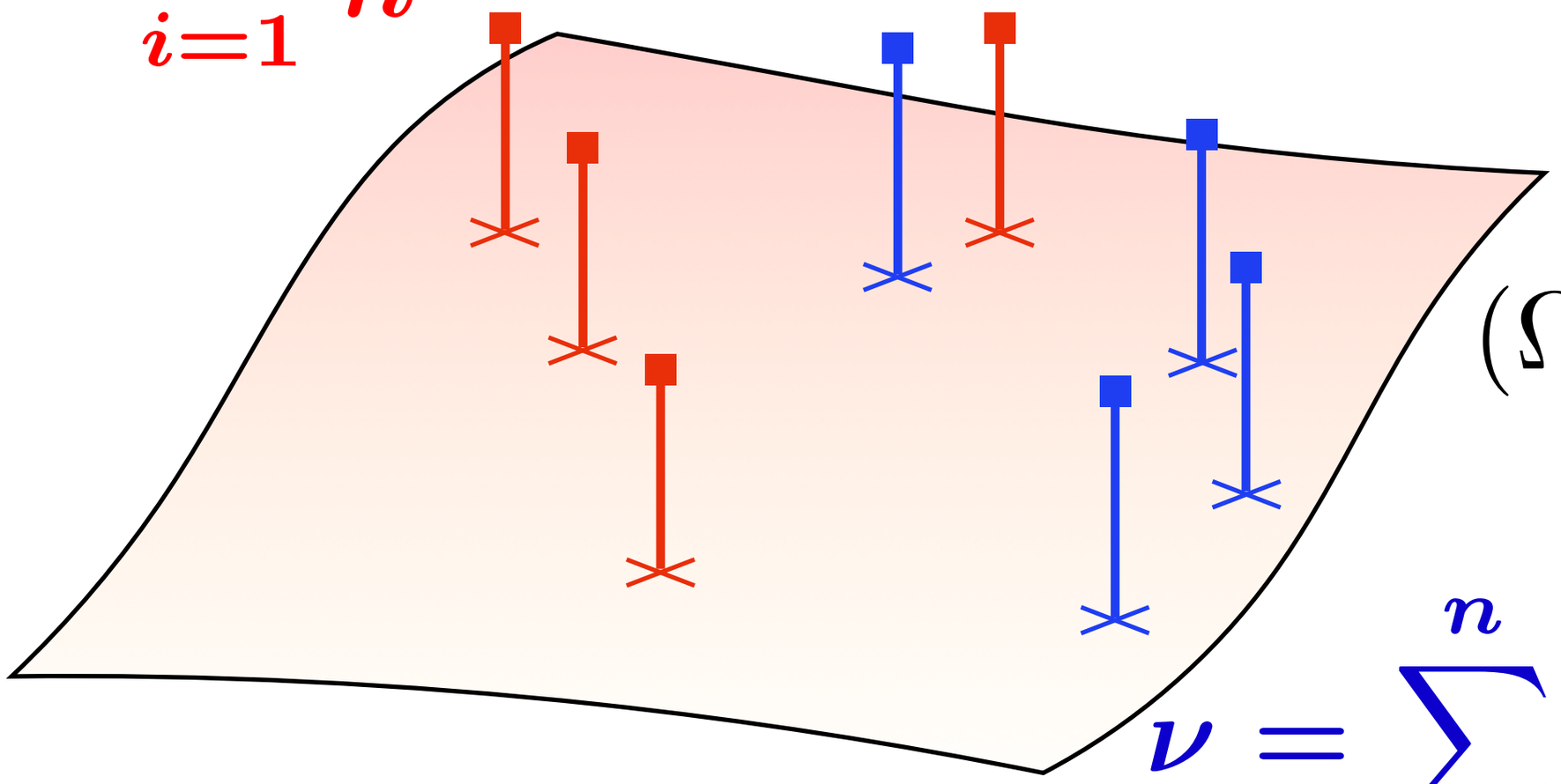
Wasserstein on 2 Diracs



$$W_p^p(\delta_{\boldsymbol{x}}, \delta_{\boldsymbol{y}}) = D(\boldsymbol{x}, \boldsymbol{y})$$

Wasserstein on Uniform Measures

$$\mu = \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}$$

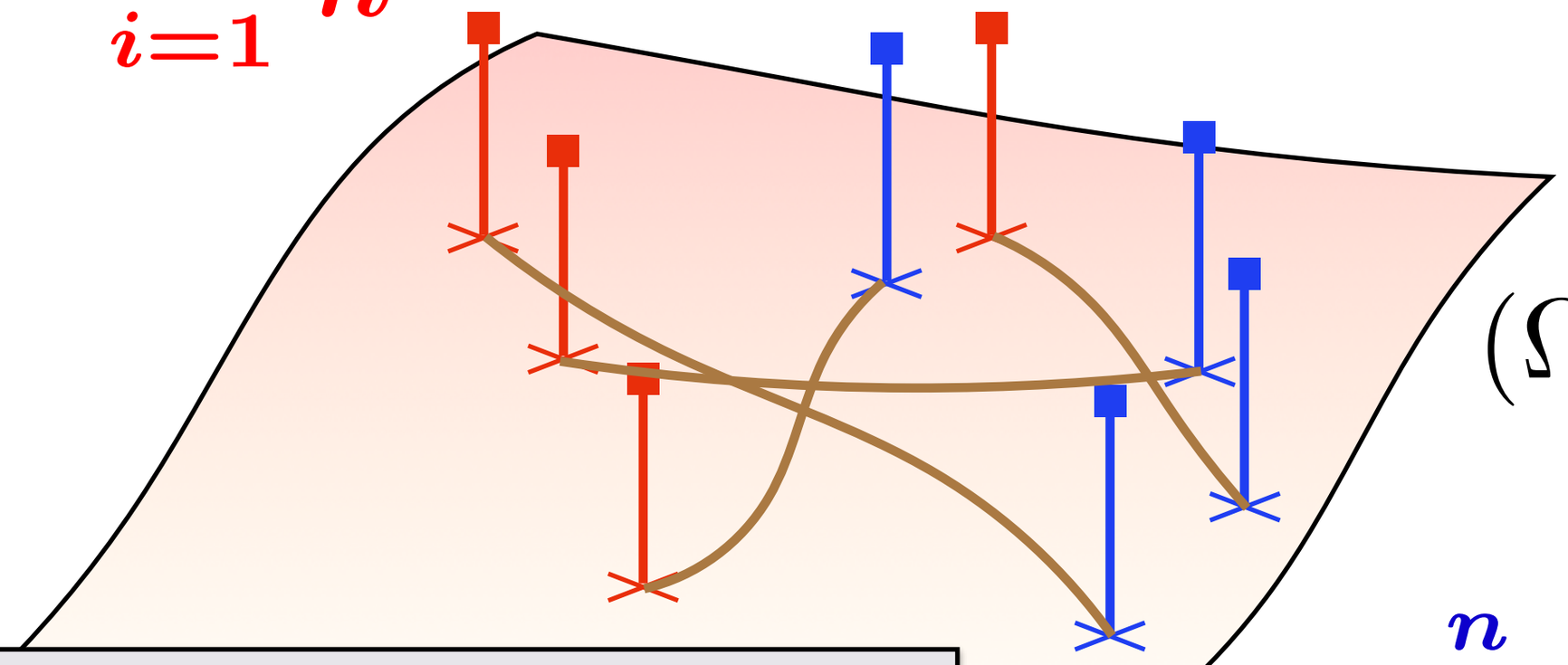


(Ω, D)

$$\nu = \sum_{j=1}^{n} \frac{1}{n} \delta_{y_j}$$

Wasserstein on Uniform Measures

$$\mu = \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}$$



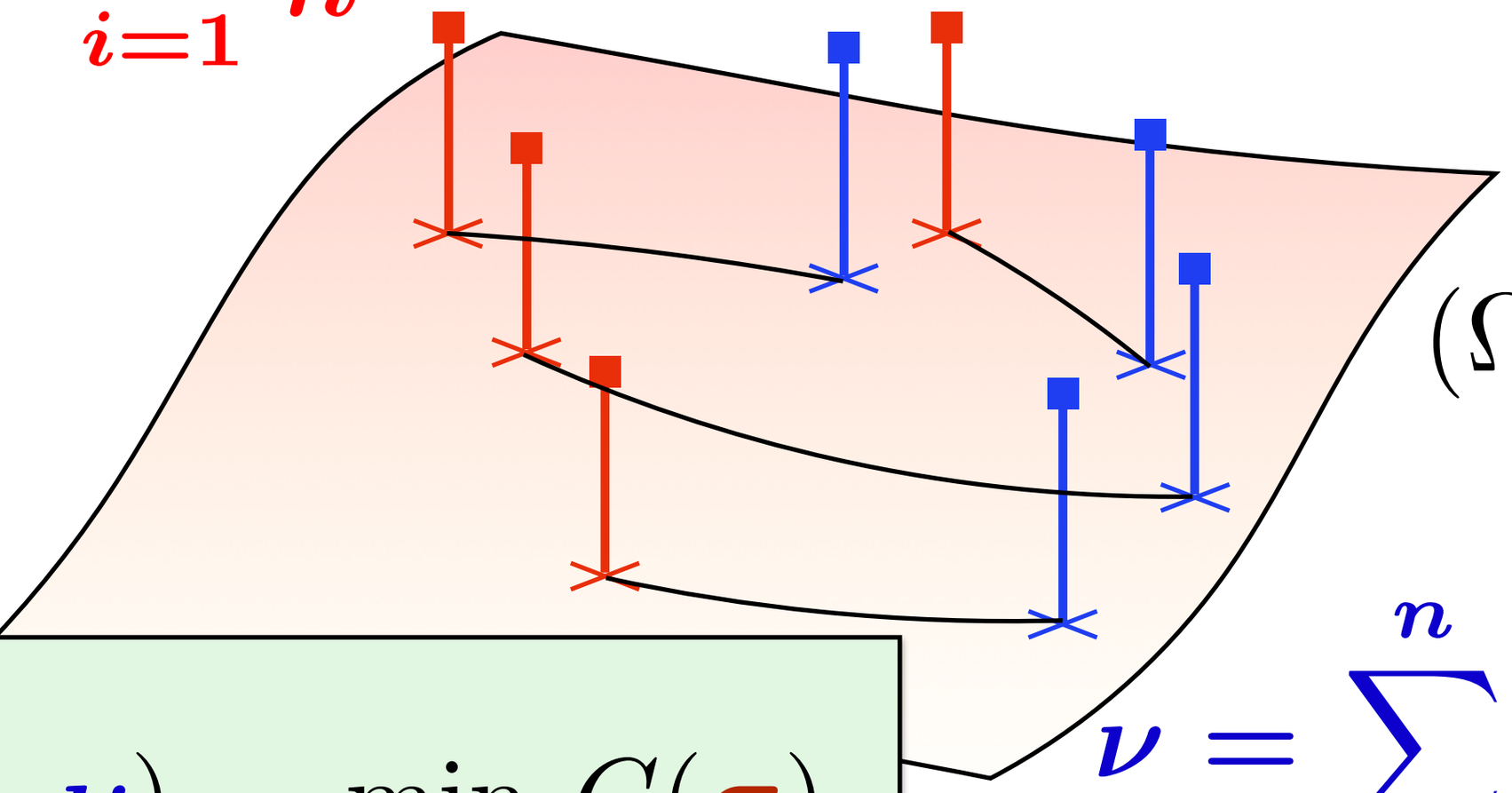
(Ω, D)

$$\nu = \sum_{j=1}^{n} \frac{1}{n} \delta_{y_j}$$

$$C(\sigma) = \sum_{i=1}^{n} D(x_i, y_{\sigma_i})^p$$

Optimal Assignment $\subset$ Wasserstein

$$\mu = \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}$$



(Ω, D)

$$\nu = \sum_{j=1}^{n} \frac{1}{n} \delta_{y_j}$$

$$W_p^p(\mu, \nu) = \min_{\sigma \in S_n} C(\sigma)$$

Wasserstein on Empirical Measures

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

$$M_{\boldsymbol{XY}} \stackrel{\text{def}}{=} [D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij}$$

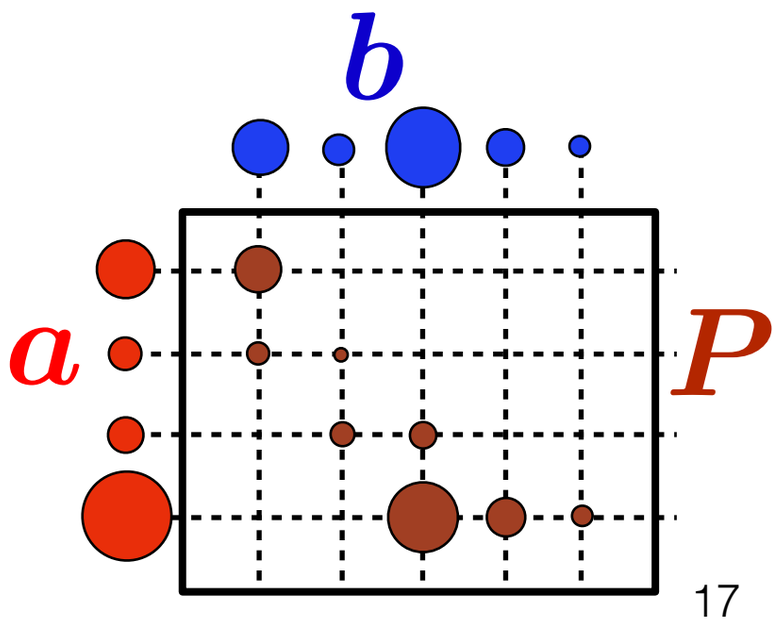
$$U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

Wasserstein on Empirical Measures

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

$$M_{\boldsymbol{XY}} \stackrel{\text{def}}{=} [D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij}$$

$$U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$



Wasserstein on Empirical Measures

Assume $\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{\delta}_{\boldsymbol{x}_i}$ and $\boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{\delta}_{\boldsymbol{y}_j}$.

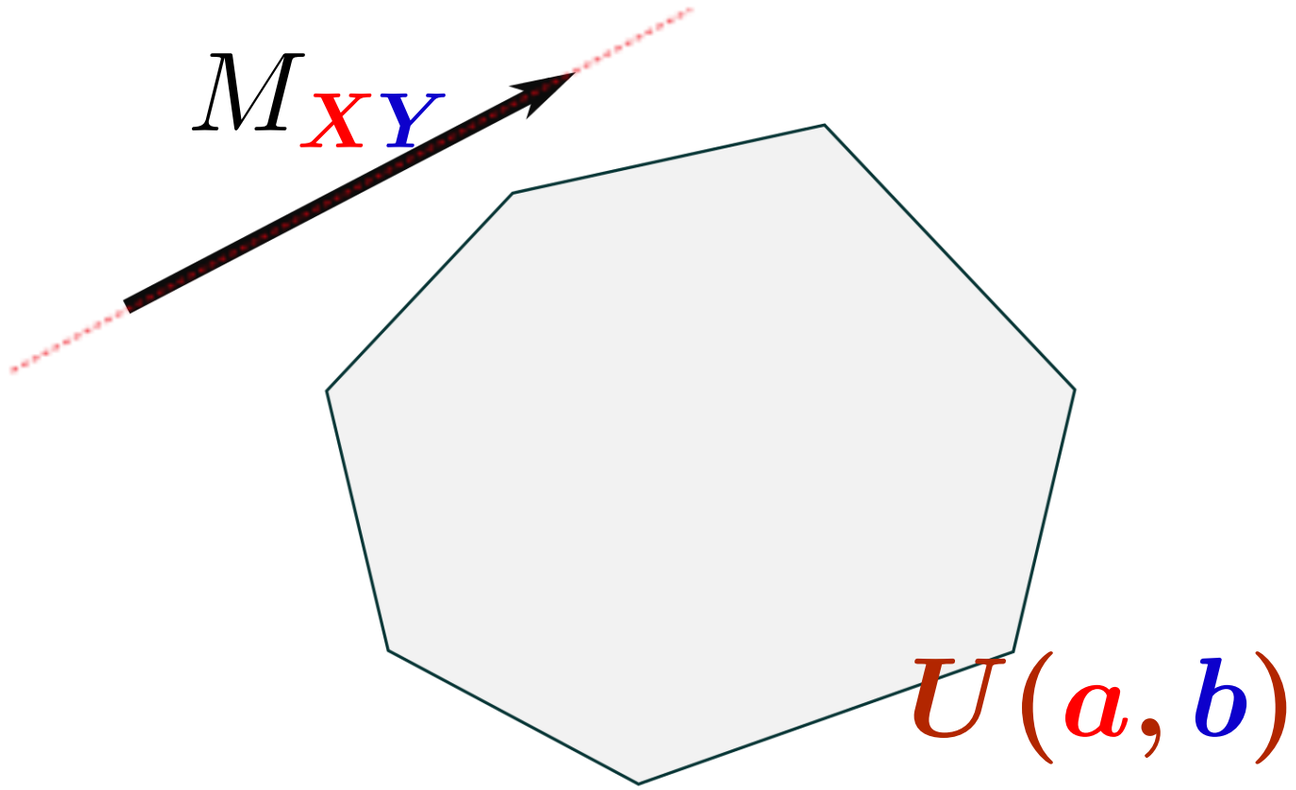
$$M_{\boldsymbol{X}\boldsymbol{Y}} \stackrel{\text{def}}{=} [D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij}$$

$$U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

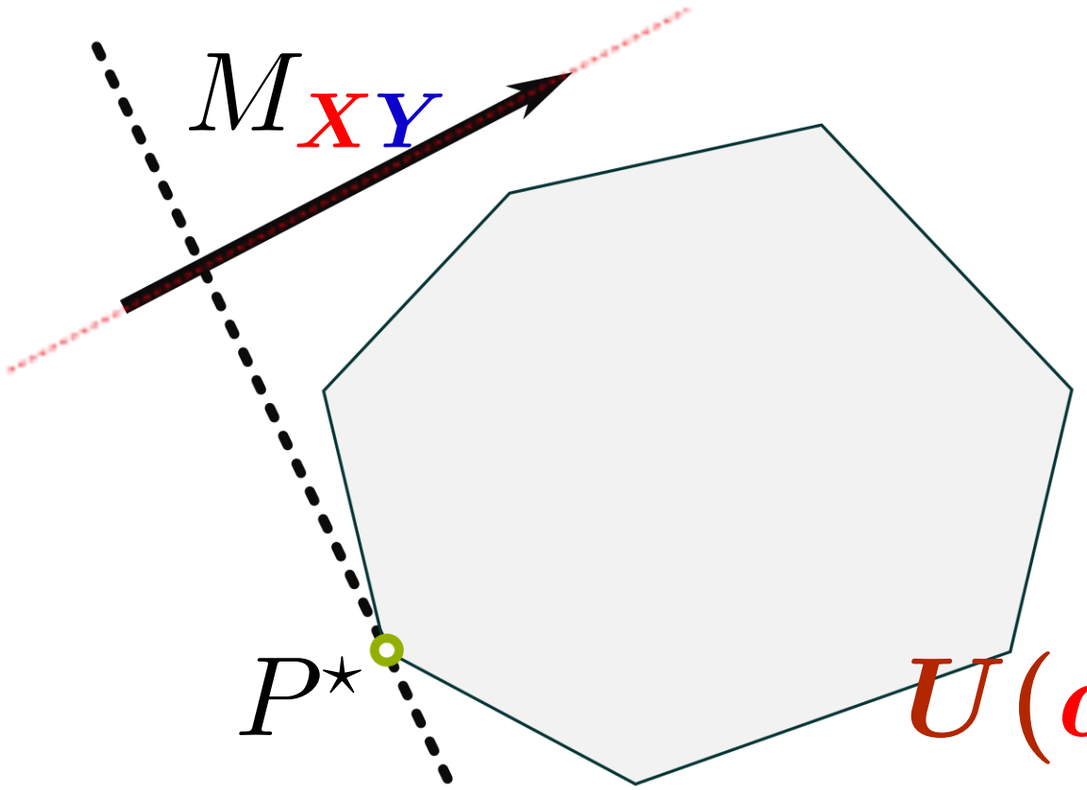
Def. Optimal Transport Problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

Wasserstein on Empirical Measures



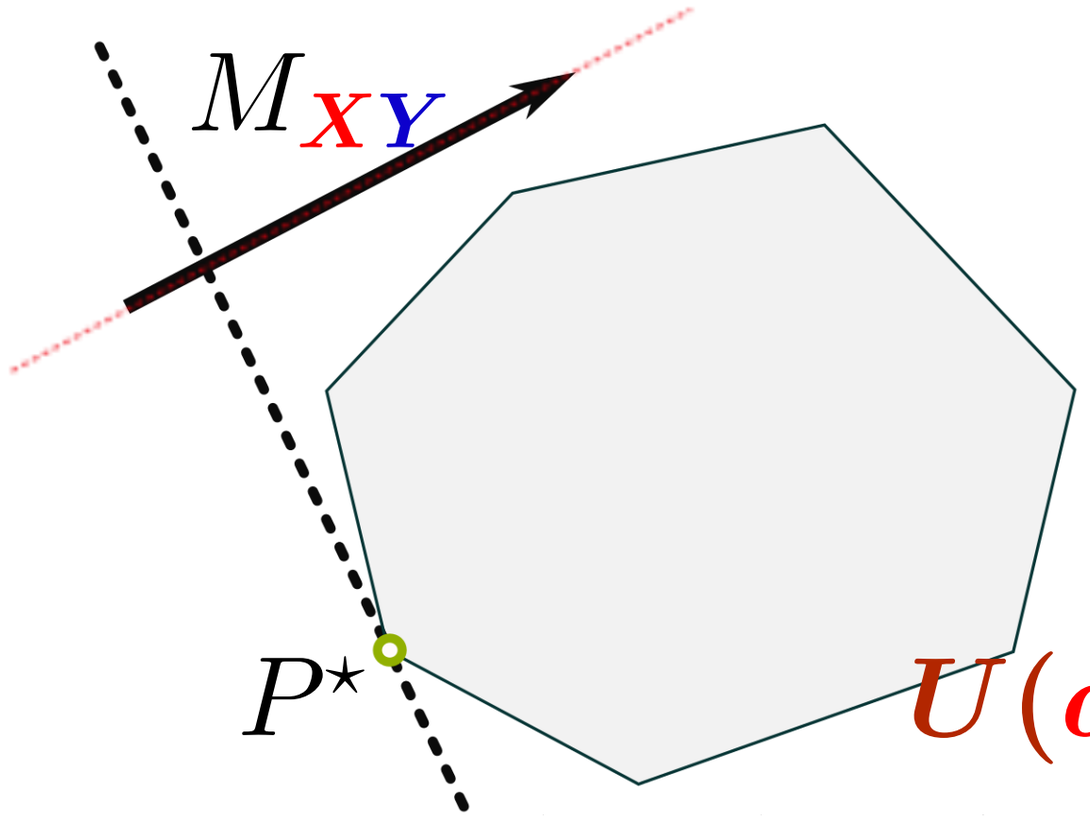
Wasserstein on Empirical Measures



$$W_p^p(\mu, \nu) = \langle P^\star, M_{XY} \rangle$$

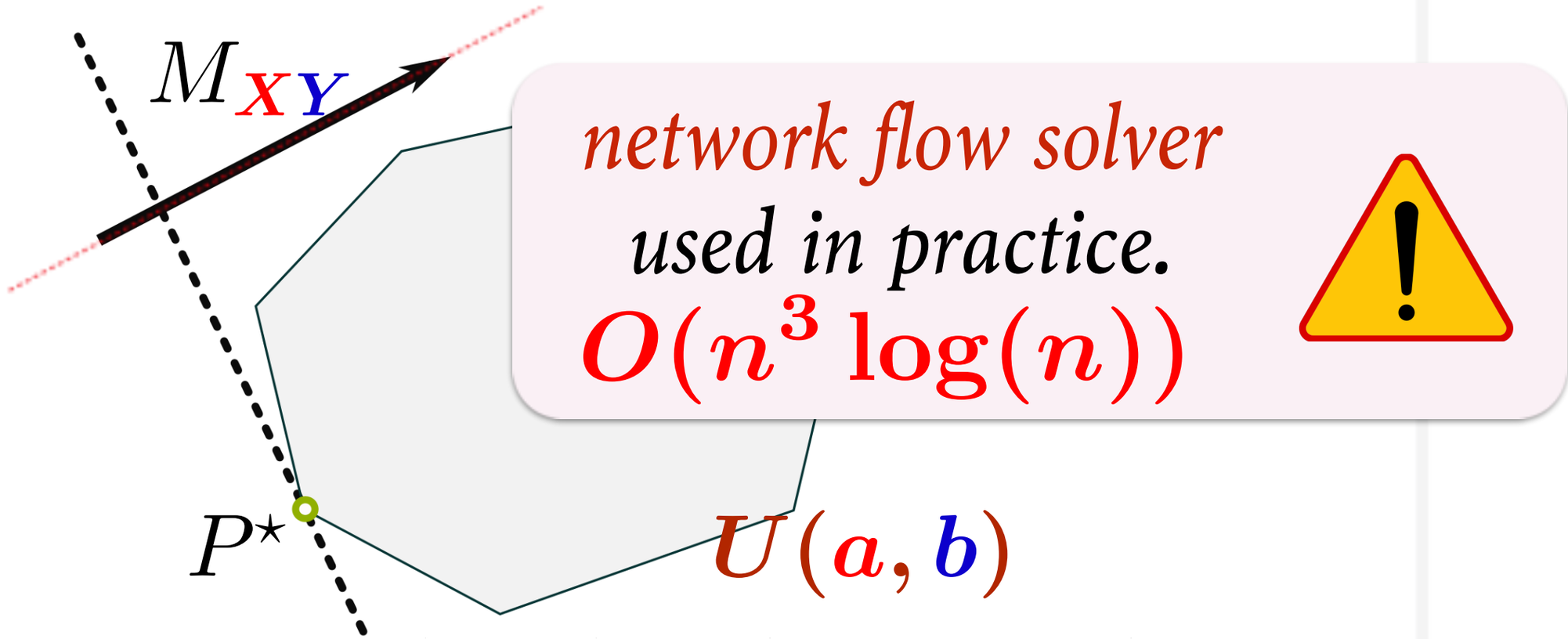
$$= \min_{P \in U(a,b)} \langle P, M_{XY} \rangle$$

Wasserstein on Empirical Measures



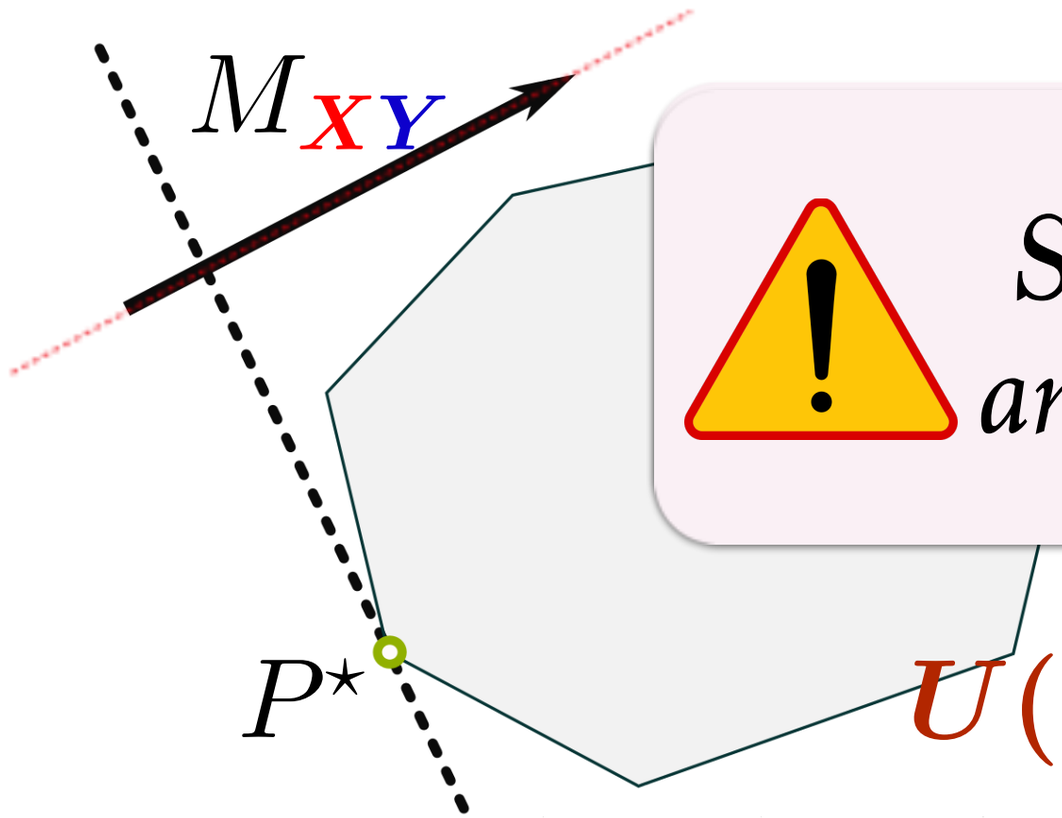
$$W_p^p(\mu, \nu) = \max_{\substack{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m \\ \alpha_i + \beta_j \leq D(x_i, y_j)}} \alpha^T a + \beta^T b$$

Wasserstein on Empirical Measures



$$W_p^p(\mu, \nu) = \max_{\substack{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m \\ \alpha_i + \beta_j \leq D(x_i, y_j)}} \alpha^T a + \beta^T b$$

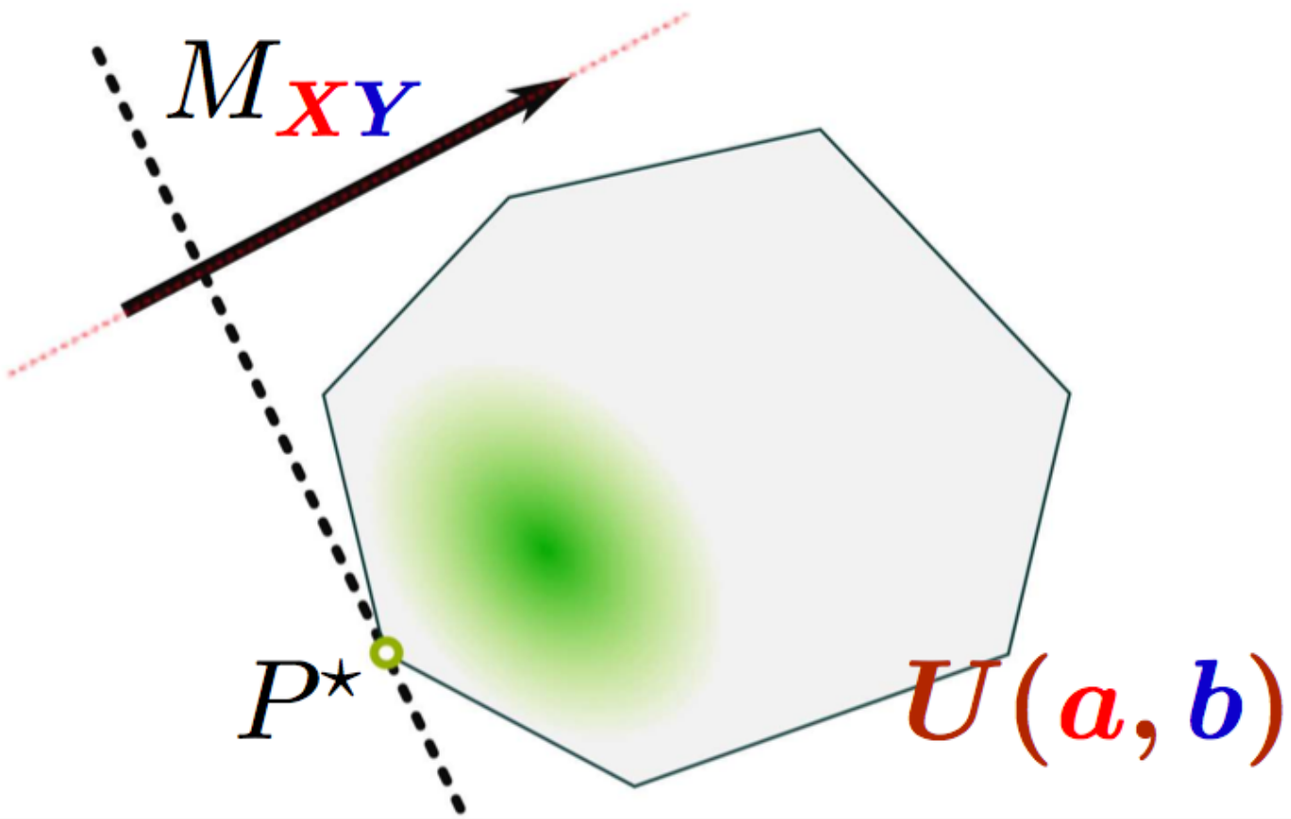
Wasserstein on Empirical Measures



Solution $P^\star$ unstable and not always unique.

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n, \boldsymbol{\beta}\in\mathbb{R}^m \\ \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j \leq D(\boldsymbol{x}_i, \boldsymbol{y}_j)}} \alpha^T \boldsymbol{a} + \beta^T \boldsymbol{b}$$

Regularized Optimal Transport



Wishlist:
faster & *scalable*, more *stable*.
approximate optimality is OK.

Entropic Regularization **[Wilson'62]**

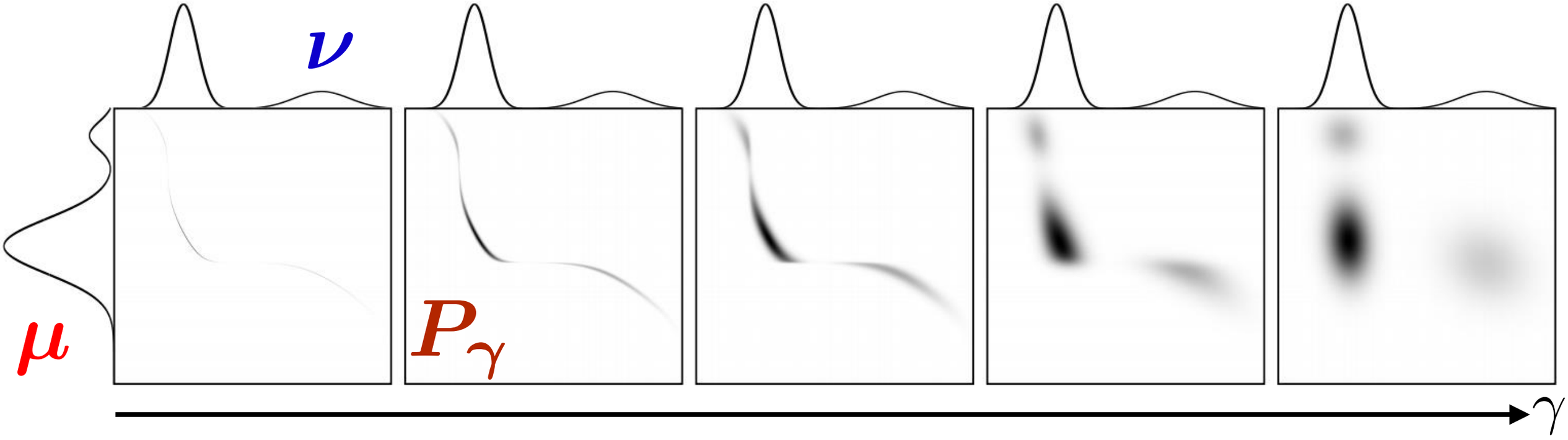
$$E(P) \stackrel{\text{def}}{=} \sum_{i,j=1}^{nm} P_{ij}(\log P_{ij} - 1) + \iota_{\mathbb{R}_+}(P_{ij})$$

Def. Regularized Wasserstein, $\gamma \geq 0$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$

Note. **Unique** optimal solution because of strong concavity of Entropy

Entropic Regularization [Wilson'62]



Def. Regularized Wasserstein, $\gamma \geq 0$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$

Note. **Unique** optimal solution because of strong concavity of Entropy

Fast & Scalable Algorithm

Prop. If $P_\gamma \stackrel{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists ! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

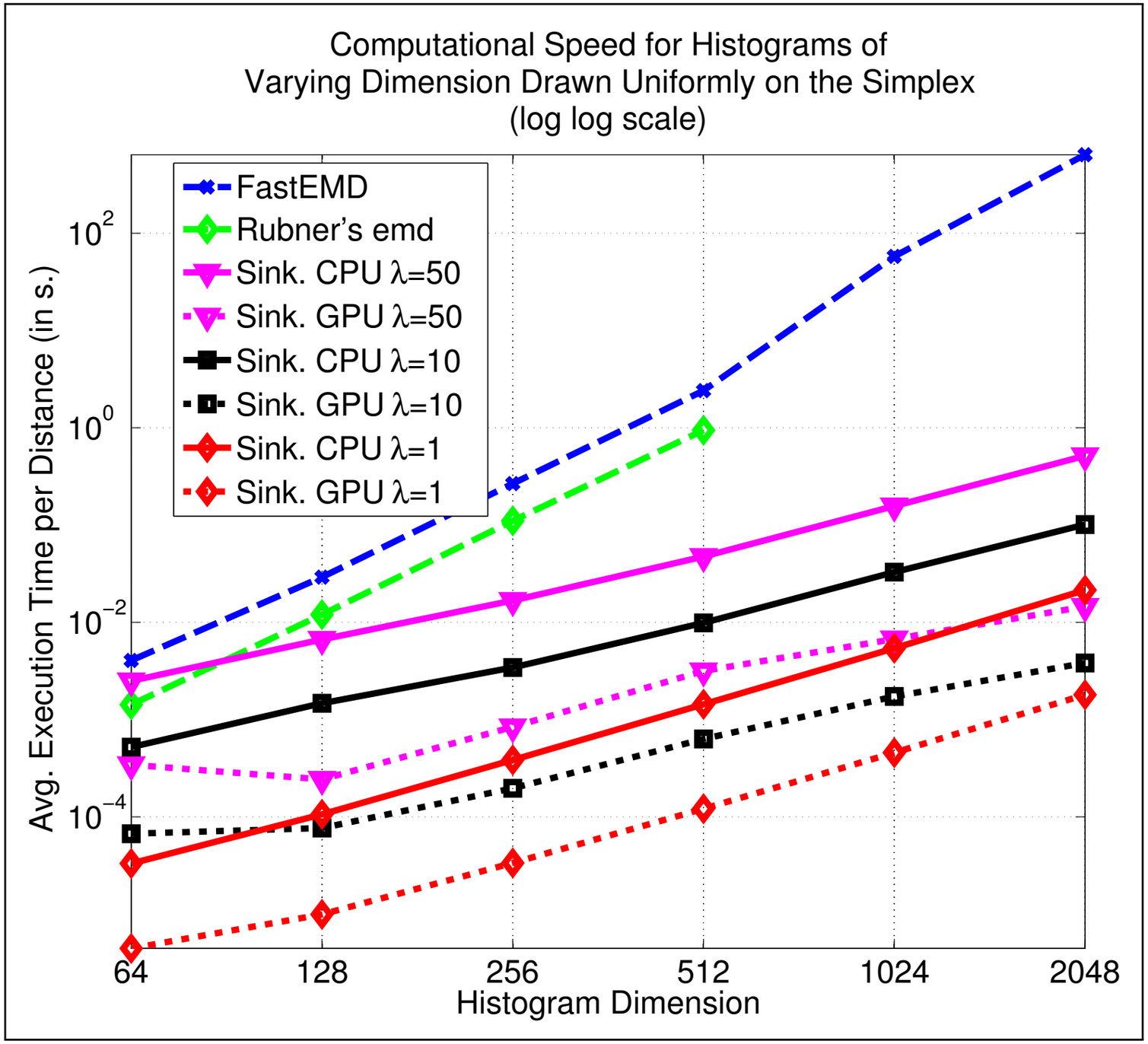
$$P_\gamma = \mathbf{D}(\boldsymbol{u})\ K \mathbf{D}(\boldsymbol{v}), \qquad K \stackrel{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

- **[Sinkhorn'64]** fixed-point iterations for $(\boldsymbol{u}, \boldsymbol{v})$

$$\boldsymbol{u} \leftarrow \boldsymbol{a} / K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b} / K^T \boldsymbol{u}$$

- Fast, $O(nm)$ **or less,** GPU parallel **[C'13]** .

Regularized Transport: *Fast*



Regularized Transport: *differentiable*

Prop. Gradients w.r.t $\boldsymbol{a}, \boldsymbol{X}$: **[CD'14]**

1. $W_\gamma = \max_{\alpha,\beta} \alpha^T \boldsymbol{a} + \beta^T \boldsymbol{b} - \frac{1}{\gamma}(e^{\alpha/\gamma})^T \ \boldsymbol{K} e^{\beta/\gamma}$

2. W_γ is convex w.r.t. $\boldsymbol{a}$; $\nabla_{\boldsymbol{a}} W_\gamma = \gamma \log(\boldsymbol{u})$.

3. If $p = 2, \Omega = \mathbb{R}^d$,

$$\nabla_{\boldsymbol{X}} W_\gamma = \boldsymbol{X}\mathbf{D}(\boldsymbol{a}^{\frac{1}{2}}) - \boldsymbol{Y} P_\gamma^T \mathbf{D}(\boldsymbol{a}^{-\frac{1}{2}})$$

Regularized Transport: *duality*

Prop. Writing $H_{\boldsymbol{\nu}} : \boldsymbol{a} \mapsto W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$, **[CP'15]**

1. The Legendre transform of $H_{\boldsymbol{\nu}}$ has a **closed form**:

$$H^*_{\boldsymbol{\nu}} : \boldsymbol{g} \in \mathbb{R}^n \mapsto \gamma \left(E(\boldsymbol{b}) + \boldsymbol{b}^T \log(\boldsymbol{K} e^{\boldsymbol{g}/\gamma}) \right)$$

2. By Fenchel duality, if f concave on Σ_n,

$$\min_{\boldsymbol{a} \in \Sigma_n} W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) - f(\boldsymbol{a}) = \max_{\boldsymbol{g} \in \mathbb{R}^n} f_*(\boldsymbol{g}) - H^*_{\boldsymbol{\nu}}(\boldsymbol{g})$$

Regularized Transport: *duality*

Prop. Writing $H_{\boldsymbol{\nu}} : \boldsymbol{a} \mapsto W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$, **[CP'15]**

1. The Legendre transform of $H_{\boldsymbol{\nu}}$ has a closed form:

$$H^*_{\boldsymbol{\nu}} : \boldsymbol{g} \in \mathbb{R}^n \mapsto \gamma\left(E(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \log\left(K e^{\frac{\boldsymbol{g}}{\gamma}}\right)\right)$$

2. By Fenchel duality, if f concave on Σ_n,

$$\min_{\boldsymbol{a} \in \Sigma_n} W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) - f(\boldsymbol{a}) = \max_{\boldsymbol{g} \in \mathbb{R}^n} f_*(\boldsymbol{g}) - H^*_{\boldsymbol{\nu}}(\boldsymbol{g})$$

Relevance?

Optimizing over measures with the Wasserstein metric is crucial to use OT in statistics / machine learning.

Variational Wasserstein Problems

$$\min_{\boldsymbol{\mu} \in Q \subset \mathcal{P}(\Omega)} F\left(\boldsymbol{\mu}, W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu_1}), W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu_2}), \cdots, W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu_N})\right)$$

- *k*-means Algorithm **[Lloyd'82]**

$$\min_{\substack{\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d) \\ |\operatorname{supp} \boldsymbol{\mu}| = k}} W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}_{\text{data}})$$

- **[McCann'95]** Interpolant

$$\min_{\boldsymbol{\mu} \in \mathcal{P}(\Omega)} (1-t) W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu_1}) + t W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu_2})$$

Variational Wasserstein Problems

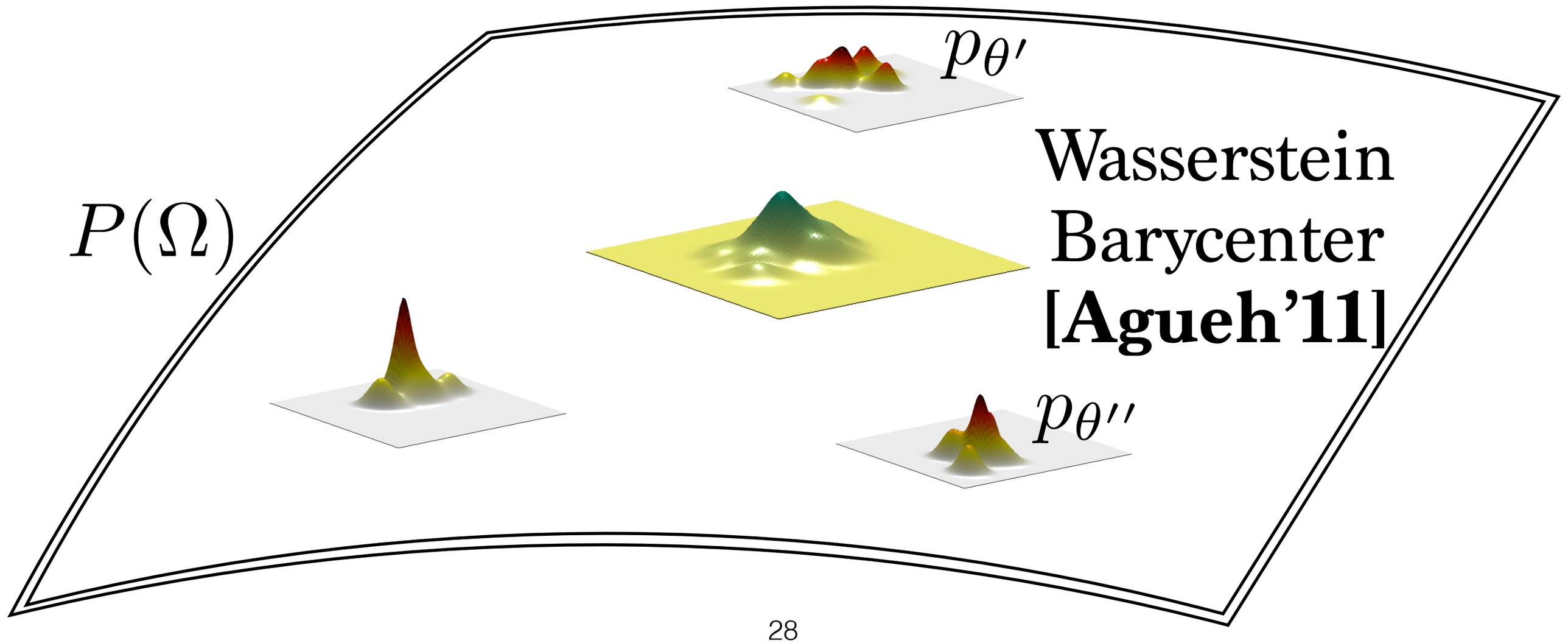
- **[JKO'98]** gradient flow

$$\mu_{t+1} = \underset{\boldsymbol{\mu} \in \mathcal{P}(\Omega)}{\operatorname{argmin}} J(\boldsymbol{\mu}) + \lambda_t W_p^p(\boldsymbol{\mu}, \mu_t)$$

1. **[Agueh'11]** Wasserstein barycenters
2. Wasserstein Dictionary Learning **[RCP'15]**
3. **[Bigot'15]** *Wasserstein PCA* **[SC'15]**
4. **[Bassetti'06]** *Min. Kantorovich Estimation* **[MMC'15]**

1. *Wasserstein Barycenters*

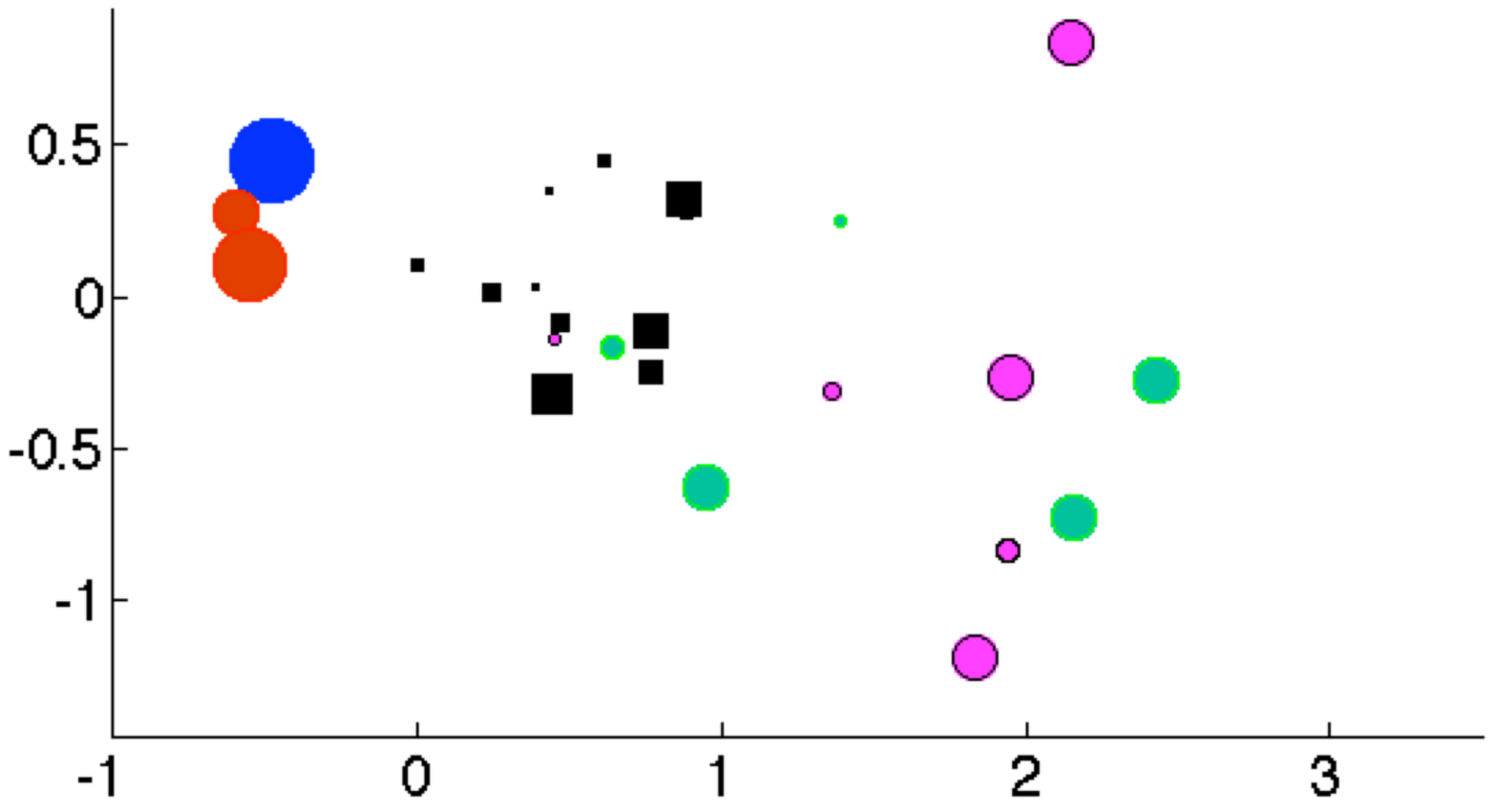
$$\min_{\boldsymbol{\mu} \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$



LP Formulations

- Can solve exactly this problem with empirical measures in 2-Wasserstein case, MM-OT:

If $|\operatorname{supp} \boldsymbol{\nu_i}| = \boldsymbol{n_i}$, LP of size $(\prod_i \boldsymbol{n_i}, \sum_i \boldsymbol{n_i})$



LP Formulations

- If solving on a grid (all locations fixed), LP:

$$\min_{P_1,\cdots,P_N,\boldsymbol{a}} \sum_{i=1}^{N} \lambda_i \langle P_i, M \rangle$$

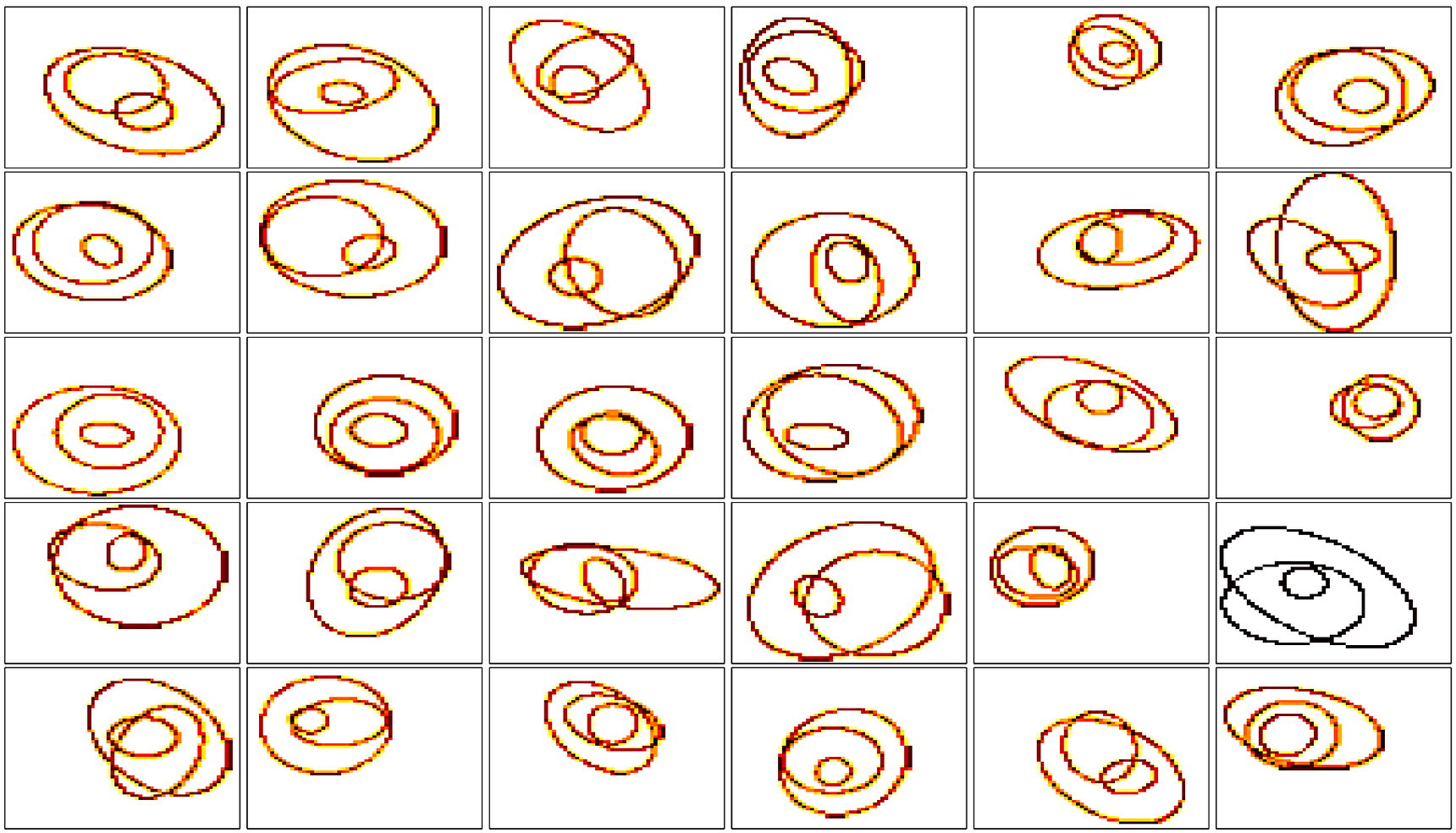
$$\text{s.t. } P_i^T \mathbf{1}_d = \boldsymbol{b_i}, \forall i \leq N,$$

$$P_1 \mathbf{1}_d = \cdots = P_N \mathbf{1}_d = \boldsymbol{a}.$$

Primal Descent on Regularized W

[CD'14]

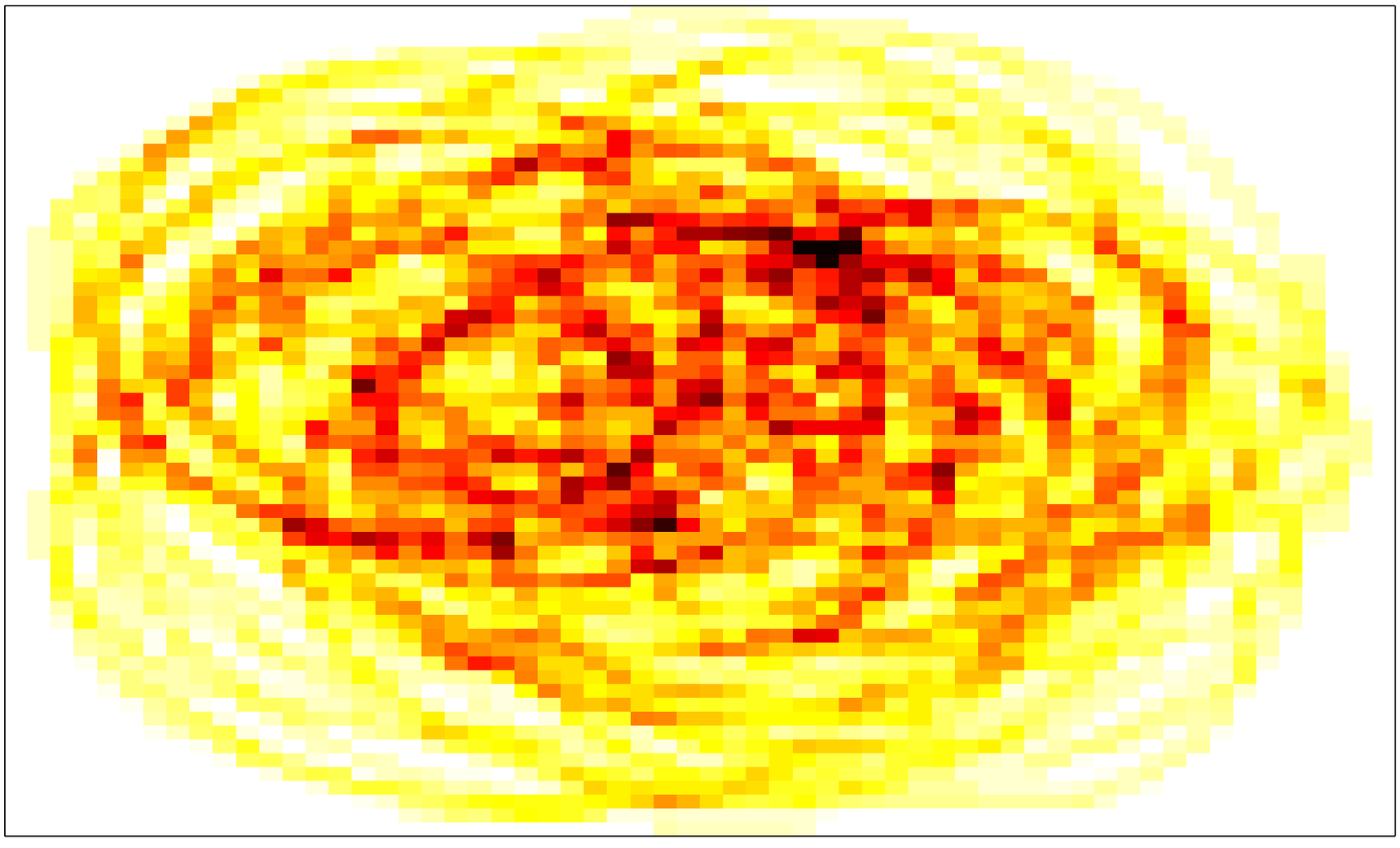
$$\min_{\boldsymbol{\mu} \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} W_{\gamma}(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$



Primal Descent on Regularized W

[CD'14]

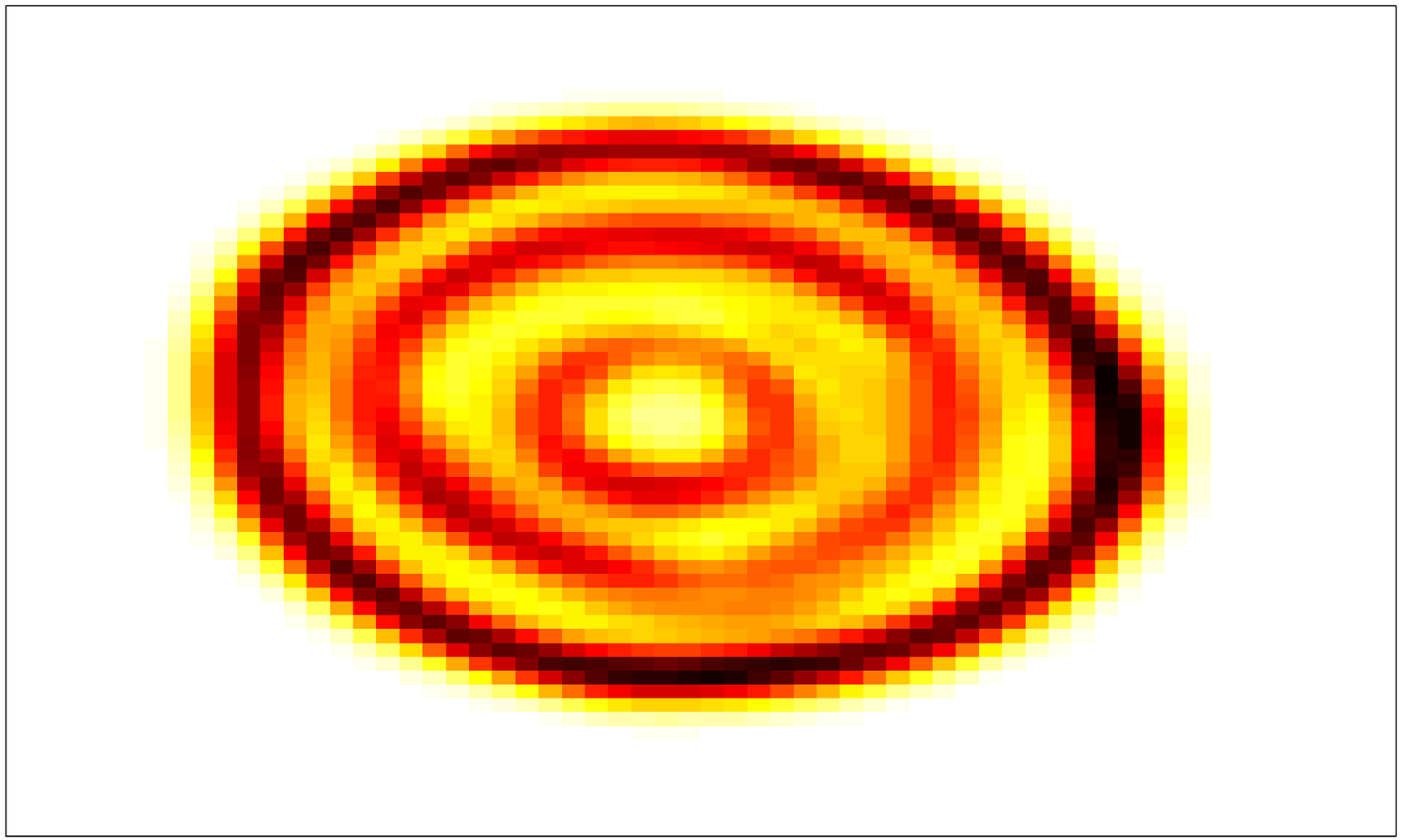
$$\min_{\boldsymbol{\mu} \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} W_{\gamma}(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$



Primal Descent on Regularized W

[CD'14]

$$\min_{\boldsymbol{\mu} \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} W_{\gamma}(\boldsymbol{\mu}, \boldsymbol{\nu_i})$$



Regularized OT as KL Projection

$$\mathbf{KL}(P|\ K) = \sum_{ij} P_{ij} \log\left(P_{ij}/\ K_{ij}\right)$$

$$\langle P, M_{\boldsymbol{XY}} \rangle - \gamma E(P) = \gamma \mathbf{KL}(P|\ K)$$

Prop. $P_\gamma = \mathrm{Proj}_{C_{\boldsymbol{a}} \cap C'_{\boldsymbol{b}}}(K)$

$$C_{\boldsymbol{a}} = \{P | P\mathbf{1}_m = \boldsymbol{a}\},\ C'_{\boldsymbol{b}} = \{P | P^T\mathbf{1}_n = \boldsymbol{b}\}$$

Regularized OT = KL Projections

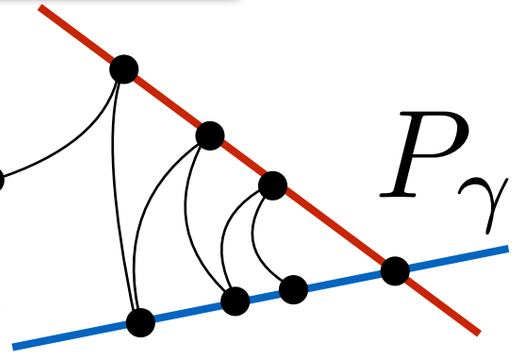
$$\textbf{Prop. } P_\gamma = \mathrm{Proj}_{C_{\boldsymbol{a}} \cap C'_{\boldsymbol{b}}}(K)$$

$$C_{\boldsymbol{a}} = \left\{P \middle| P\mathbf{1}_m = \boldsymbol{a}\right\},\ C'_{\boldsymbol{b}} = \left\{P \middle| P^T\mathbf{1}_n = \boldsymbol{b}\right\}$$

$$\mathrm{Proj}_{C_{\boldsymbol{a}}}(P) = \mathbf{D}\left(\frac{\boldsymbol{a}}{P\mathbf{1}_m}\right) P,$$

$$\mathrm{Proj}_{C'_{\boldsymbol{b}}}(P) = P\,\mathbf{D}\left(\frac{\boldsymbol{b}}{P^T\mathbf{1}_n}\right).$$

1. Sinkhorn = Dykstra's alternate projection
2. Only need to store & update diagonal multipliers



Wasserstein Barycenter = KL Projections

$$\langle P, M_{\boldsymbol{XY}} \rangle - \gamma E(P) = \gamma \mathbf{KL}(P|\ K)$$

$$\min_{\boldsymbol{a}} \sum_{i=1}^{N} W_\gamma(\boldsymbol{a}, \boldsymbol{b_i}) = \min_{\substack{\mathbf{P}=[\boldsymbol{P_1},\ldots,\boldsymbol{P_N}] \\ \mathbf{P}\in \boldsymbol{C_1}\cap\boldsymbol{C_2}}} \sum_{i=1}^{N} \lambda_i \mathbf{KL}(\boldsymbol{P_i}|K)$$

$$\boldsymbol{C_1} = \{\mathbf{P}|\exists \boldsymbol{a}, \forall i, P_i \mathbf{1}_m = \boldsymbol{a}\}$$

$$\boldsymbol{C_2} = \{\mathbf{P}|\forall i, P_i^T \mathbf{1}_n = \boldsymbol{b_i}\}$$

Wasserstein Barycenter = KL Projections

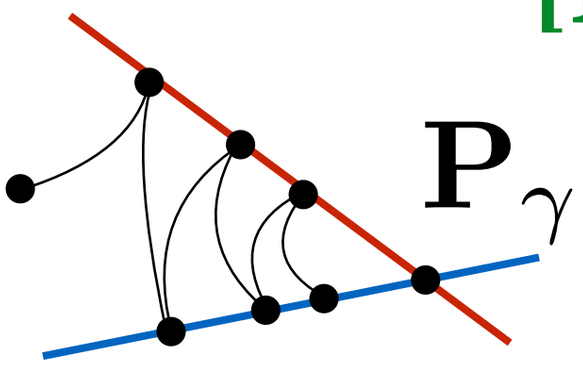
$$\min_{\boldsymbol{a}} \sum_{i=1}^{N} W_\gamma(\boldsymbol{a}, \boldsymbol{b_i}) = \min_{\substack{\mathbf{P}=[\boldsymbol{P_1},\ldots,\boldsymbol{P_N}] \\ \mathbf{P}\in \boldsymbol{C_1}\cap\boldsymbol{C_2}}} \sum_{i=1}^{N} \lambda_i \mathbf{KL}(\boldsymbol{P_i}|\boldsymbol{K})$$

$$\boldsymbol{C_1} = \{\mathbf{P}|\exists \boldsymbol{a}, \forall i, P_i \mathbf{1}_m = \boldsymbol{a}\}$$

$$\boldsymbol{C_2} = \{\mathbf{P}|\forall i, P_i^T \mathbf{1}_n = \boldsymbol{b_i}\}$$

[BCCNP'15]

$$[\ \boldsymbol{K} \cdots\ \boldsymbol{K}]$$



$$\mathbf{P}_\gamma$$

Wasserstein Barycenter = KL Projections

$$\min_{\boldsymbol{a}} \sum_{i=1}^{N} W_\gamma(\boldsymbol{a}, \boldsymbol{b_i}) = \min_{\substack{\mathbf{P}=[\boldsymbol{P_1},\ldots,\boldsymbol{P_N}] \\ \mathbf{P}\in \boldsymbol{C_1}\cap\boldsymbol{C_2}}} \sum_{i=1}^{N} \lambda_i \mathbf{KL}(\boldsymbol{P_i}|\boldsymbol{K})$$

$$\boldsymbol{C_1} = \{\mathbf{P}|\exists \boldsymbol{a}, \forall i, P_i \mathbf{1}_m = \boldsymbol{a}\}$$

$$\boldsymbol{C_2} = \{\mathbf{P}|\forall i, P_i^T \mathbf{1}_n = \boldsymbol{b_i}\}$$

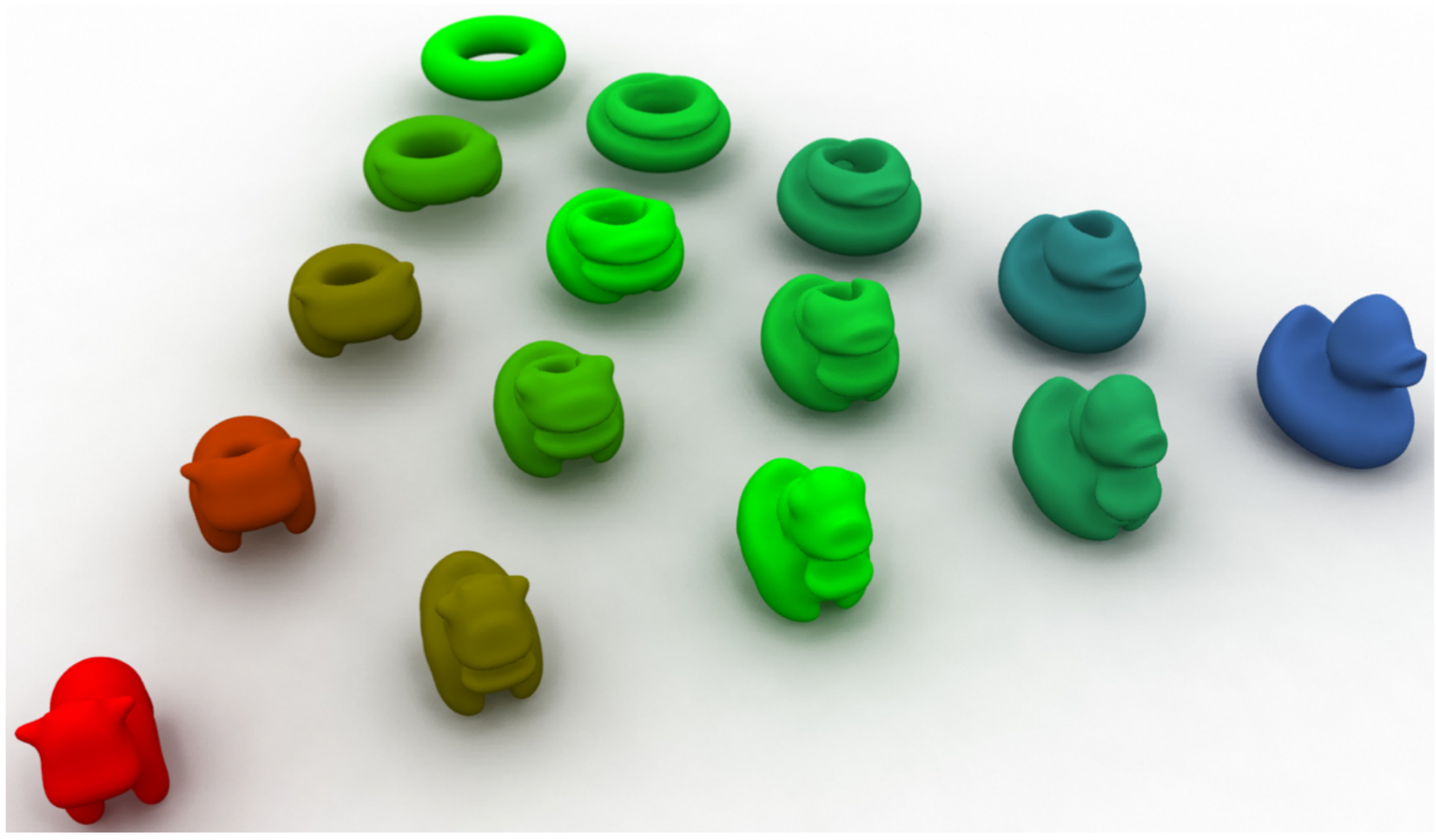
[BCCNP'15]

```
u=ones(size(B)); % d x N matrix
while not converged
   v=u.*(K'*(B./(K*u))); % 2(Nd^2) cost
   u=bsxfun(@times,u,exp(log(v)*weights))./v;
end
a=mean(v,2);
```

Application: Graphics

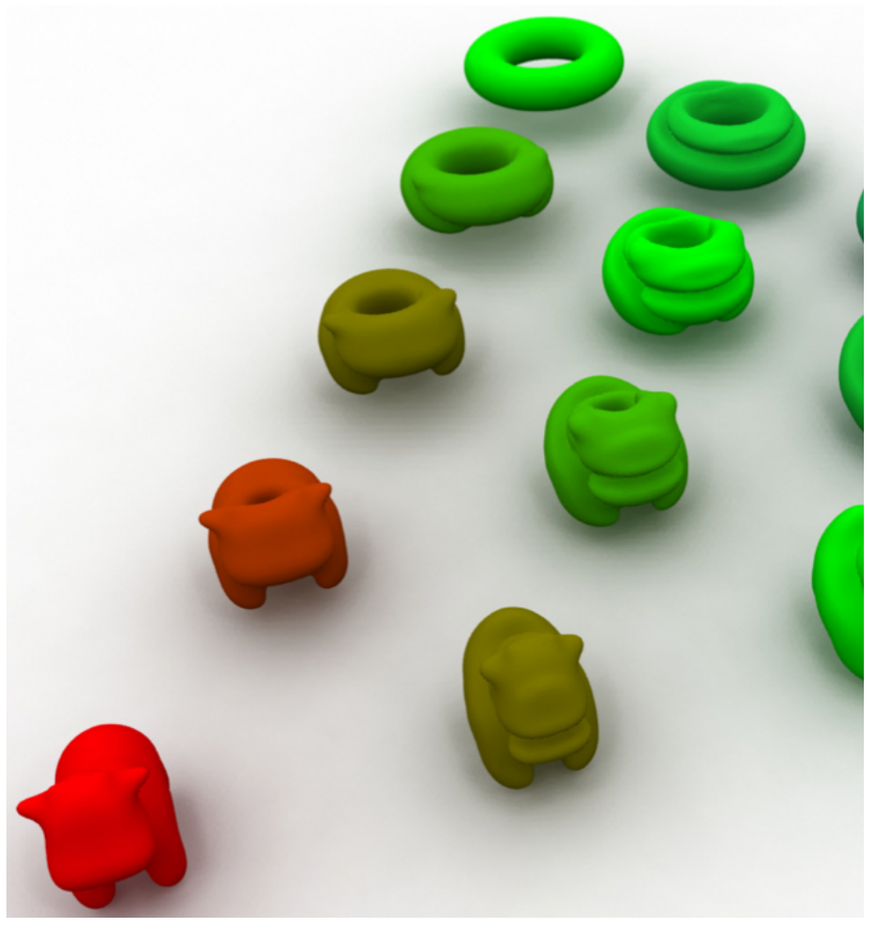
Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15

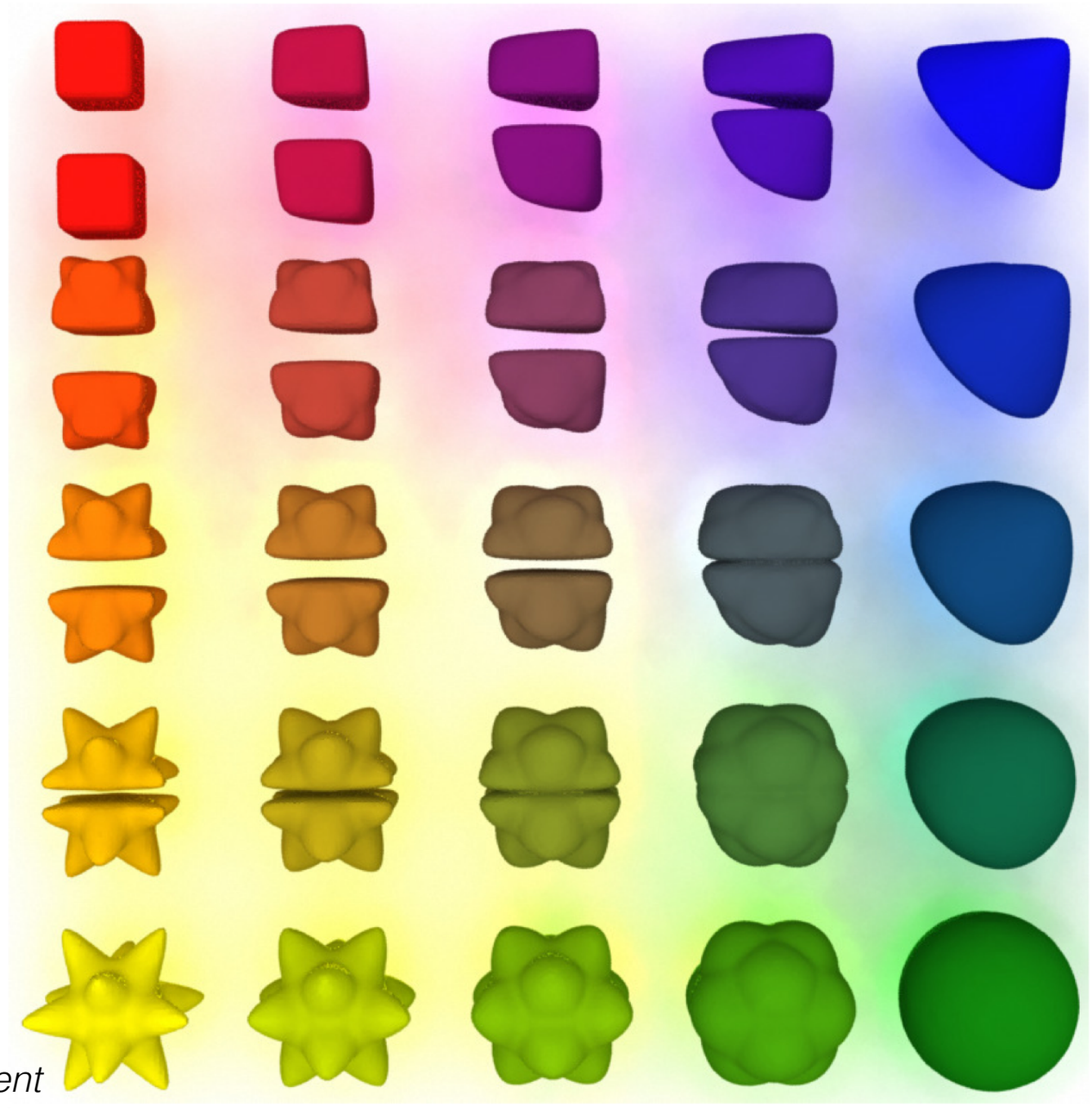
Application: Graphics



Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15

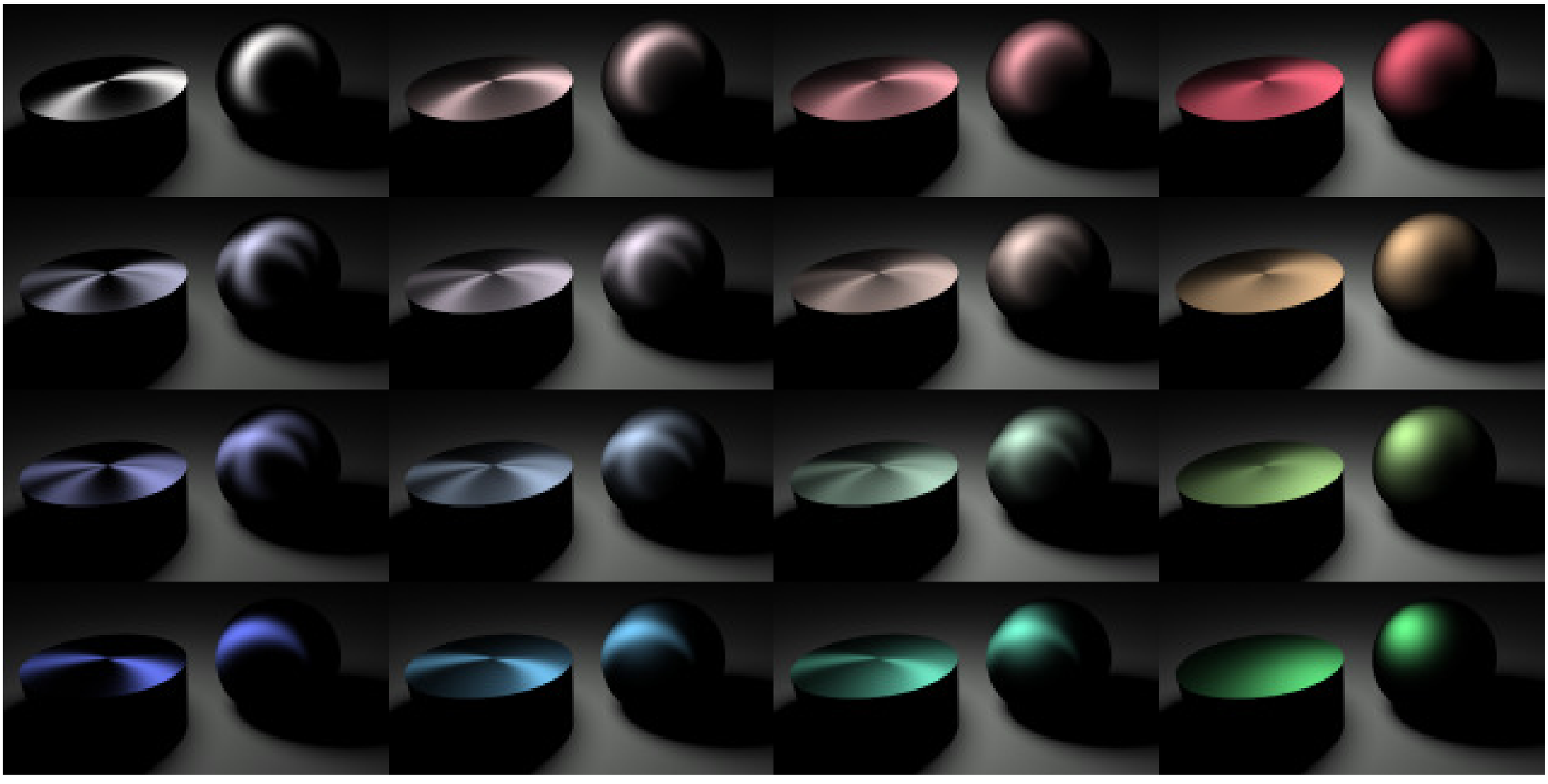
Application: Graphics





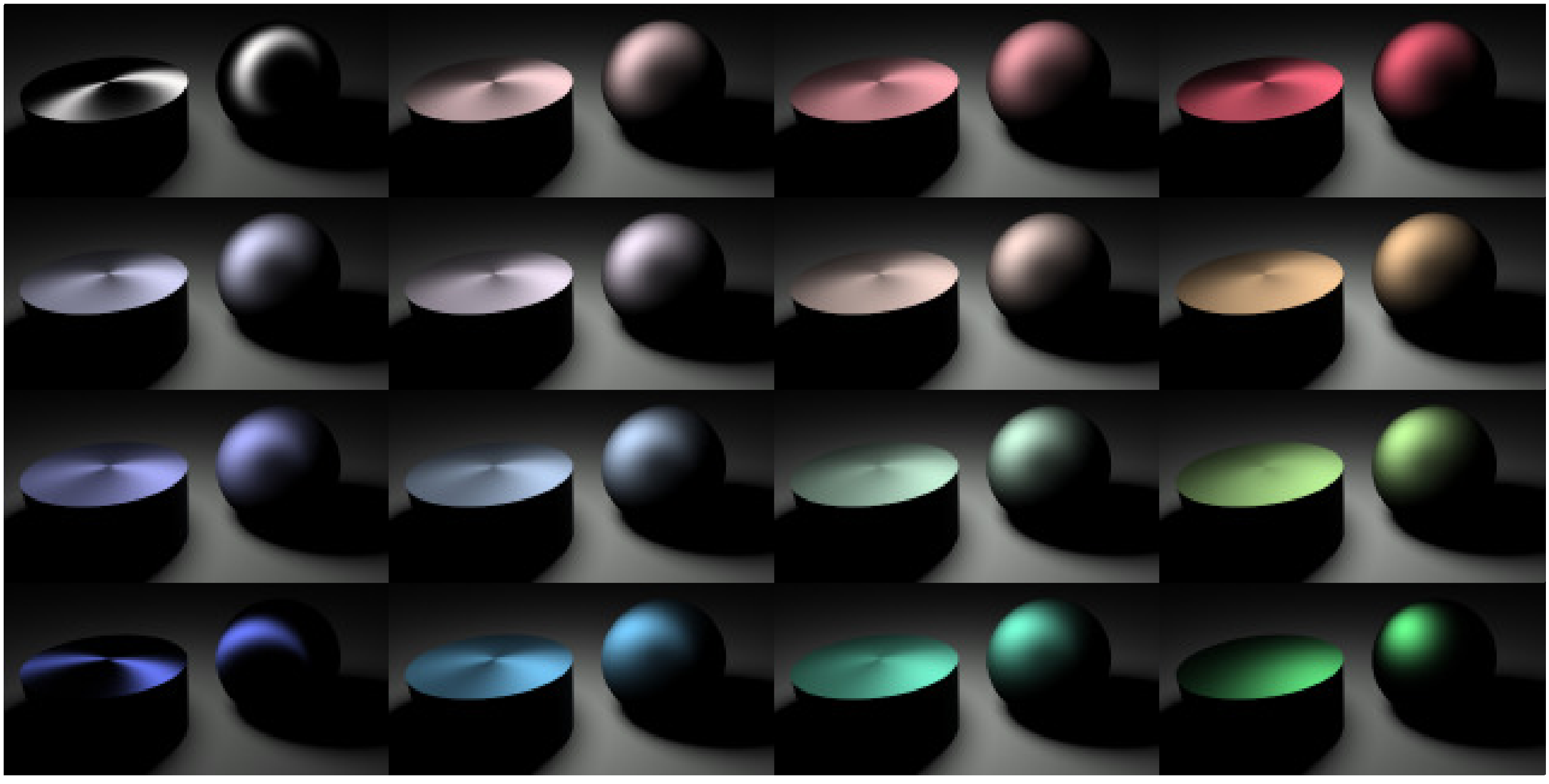
Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15

Application: Graphics



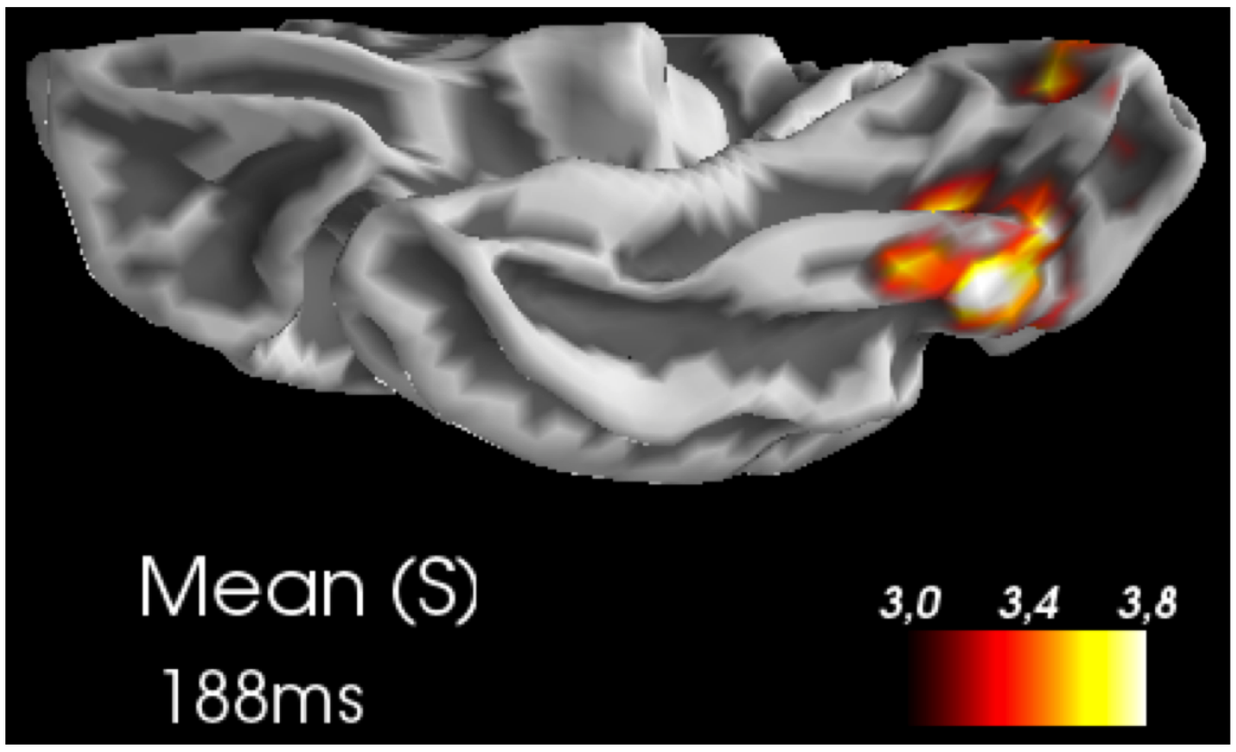
Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15

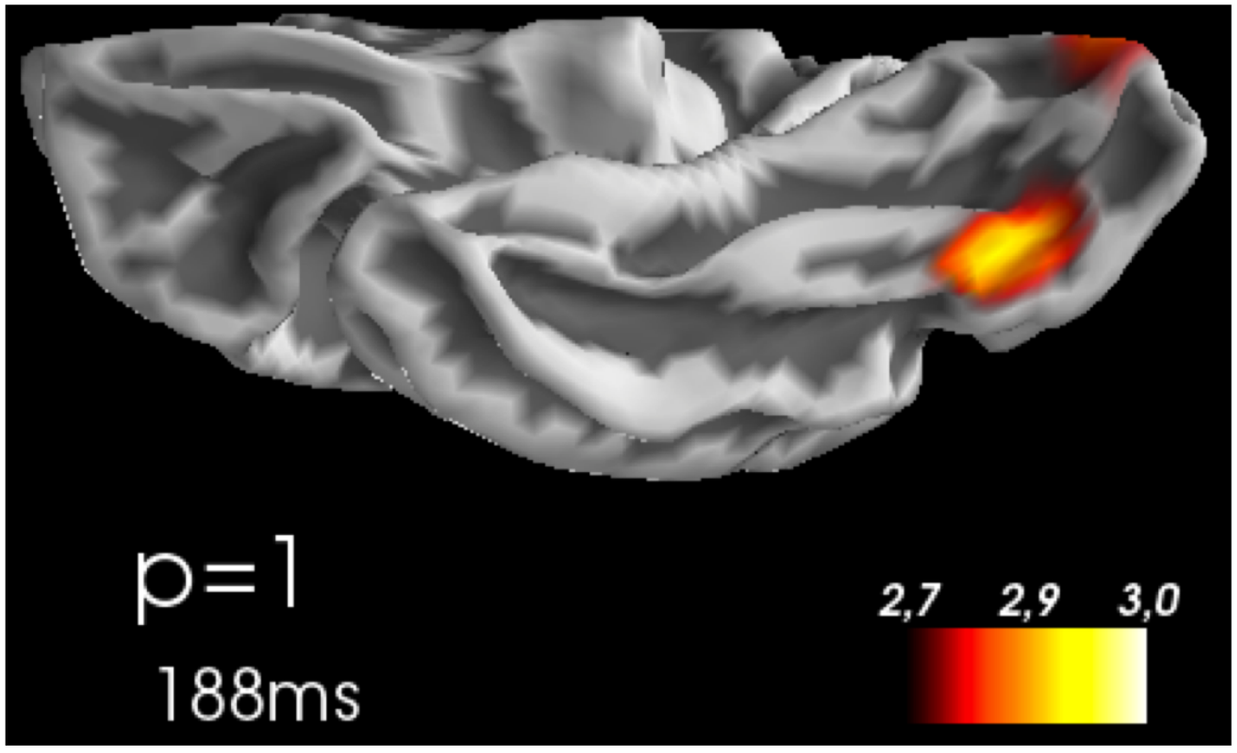
Application: Graphics



Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains,
SIGGRAPH'15

Applications: Brain Imaging

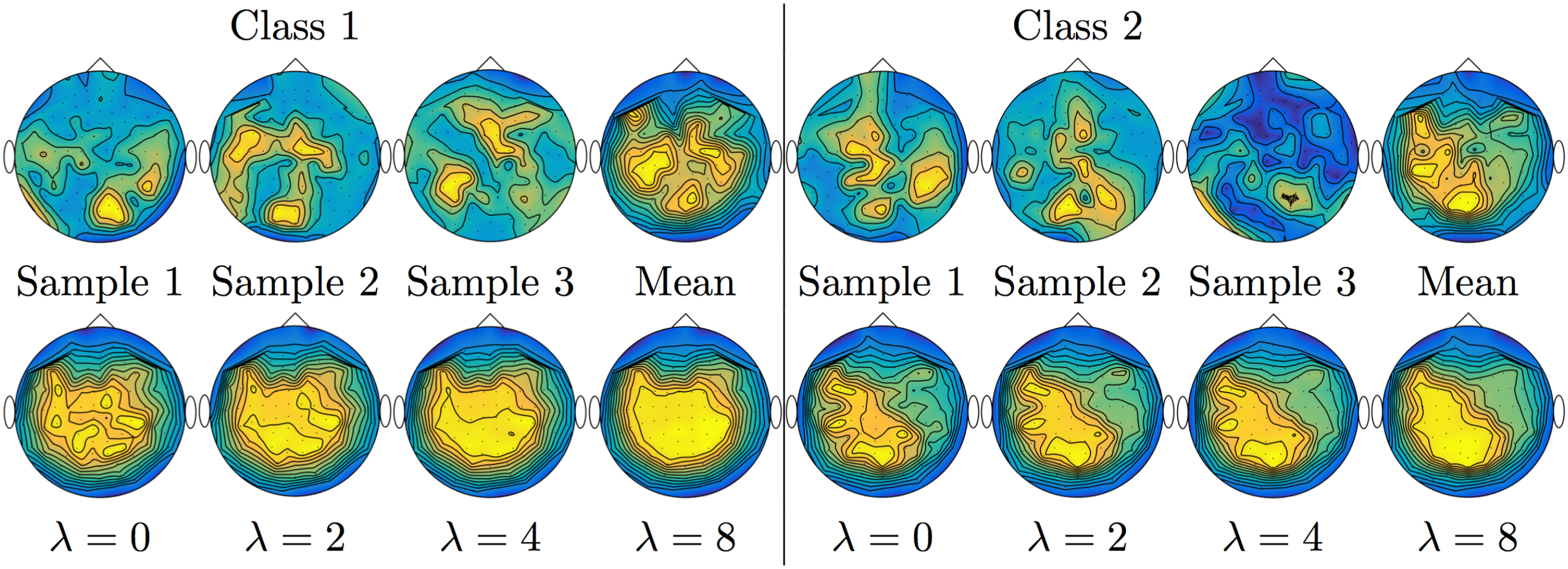




Fast Optimal Transport Averaging of Neuroimaging Data
Information Processing in Medical Imaging 2015

Applications: Brain Imaging

$$\min_{\boldsymbol{\mu} \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} W_{\boldsymbol{\gamma}}(\boldsymbol{\mu}, \boldsymbol{\nu_i}) + \lambda TV(\boldsymbol{\mu})$$

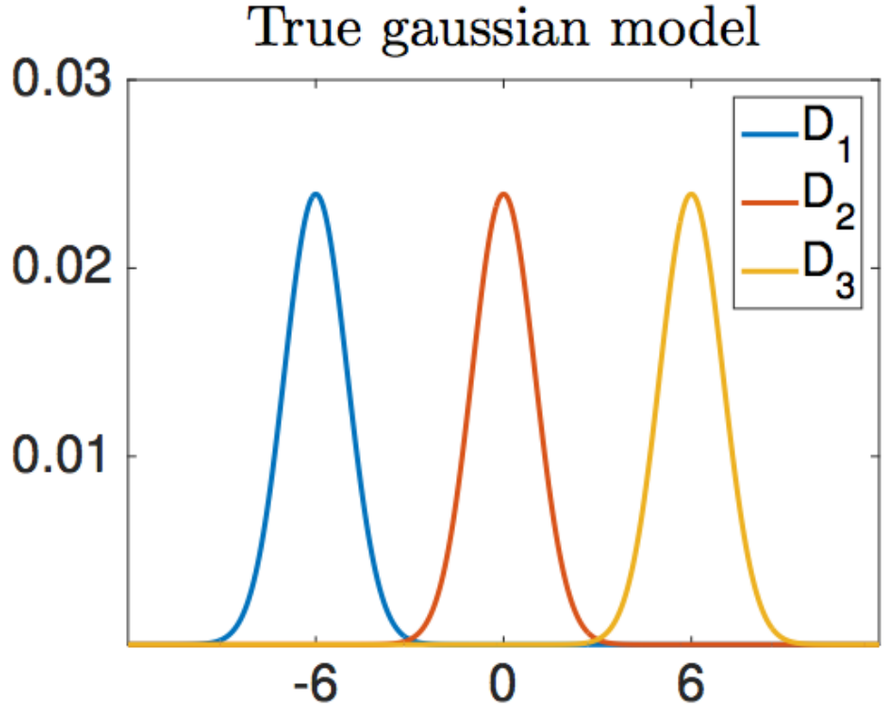


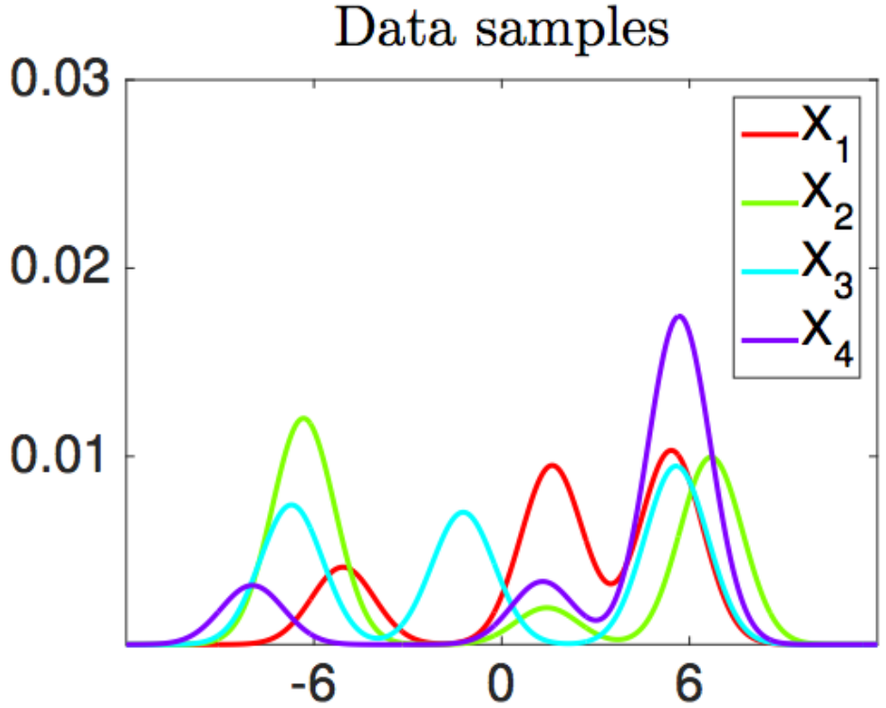
A Smoothed Dual Approach for Variational Wasserstein Problems

to appear in SIAM Imaging Sciences

2. *Dictionary Learning*

$$\min_{\boldsymbol{A}\in(\Sigma_n)^K,\boldsymbol{\Lambda}\in(\Sigma_K)^N} \sum_{i=1}^{N} W_\gamma\left(\boldsymbol{b_i}, \sum_{k=1}^{K} \boldsymbol{\Lambda_k^i a_k}\right)$$

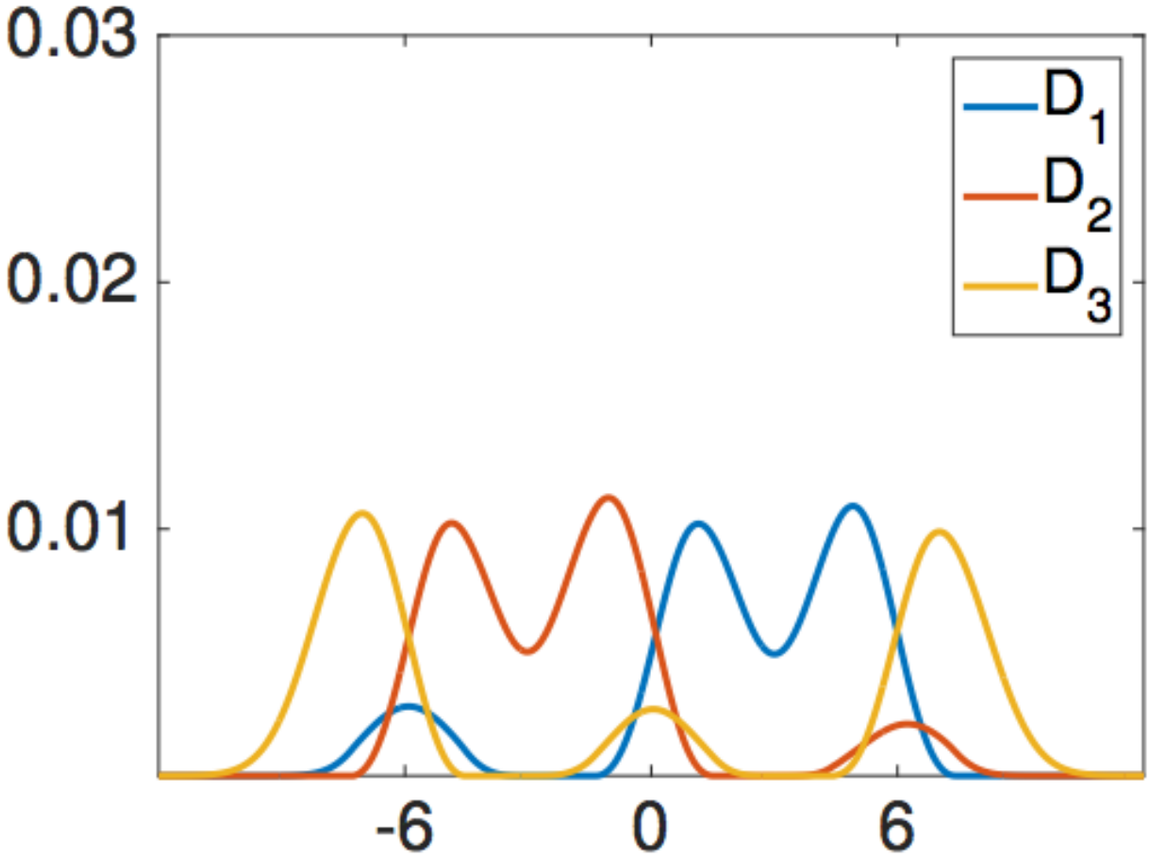




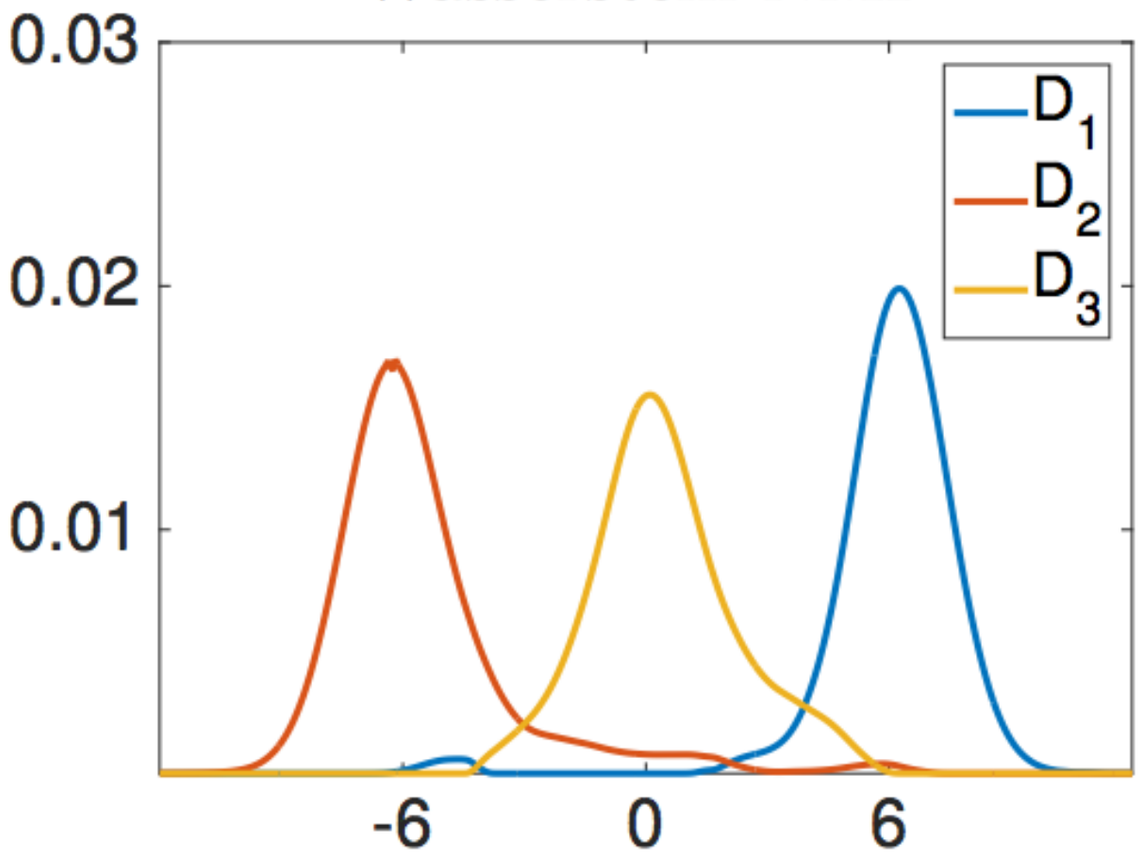
2. *Dictionary Learning*

$$\min_{\boldsymbol{A}\in(\Sigma_n)^K,\boldsymbol{\Lambda}\in(\Sigma_K)^N} \sum_{i=1}^N W_\gamma\left(\boldsymbol{b_i}, \sum_{k=1}^K \boldsymbol{\Lambda}_{\boldsymbol{k}}^{\boldsymbol{i}} \boldsymbol{a_k}\right)$$

KL-NMF

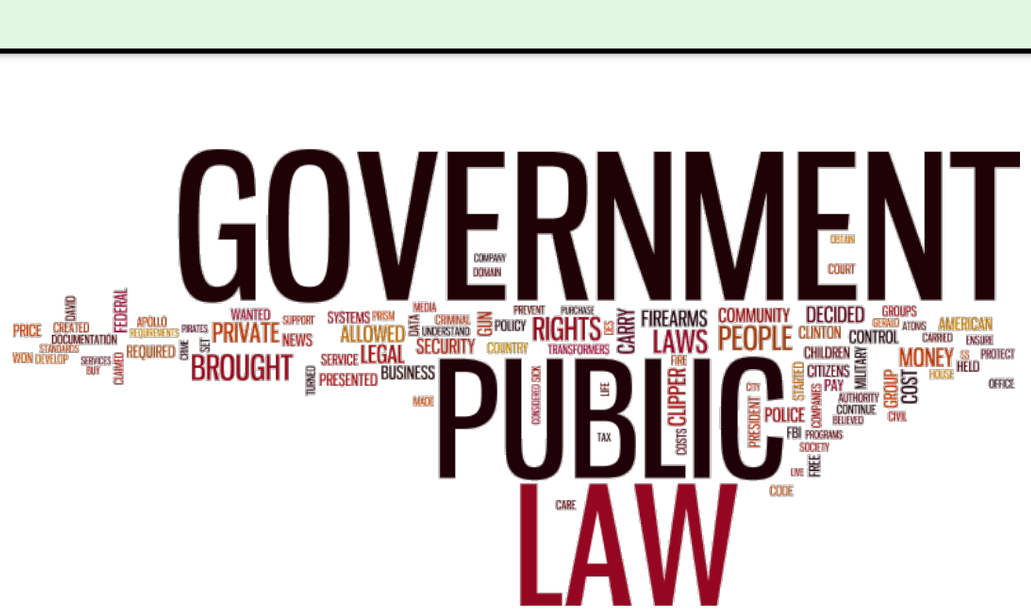


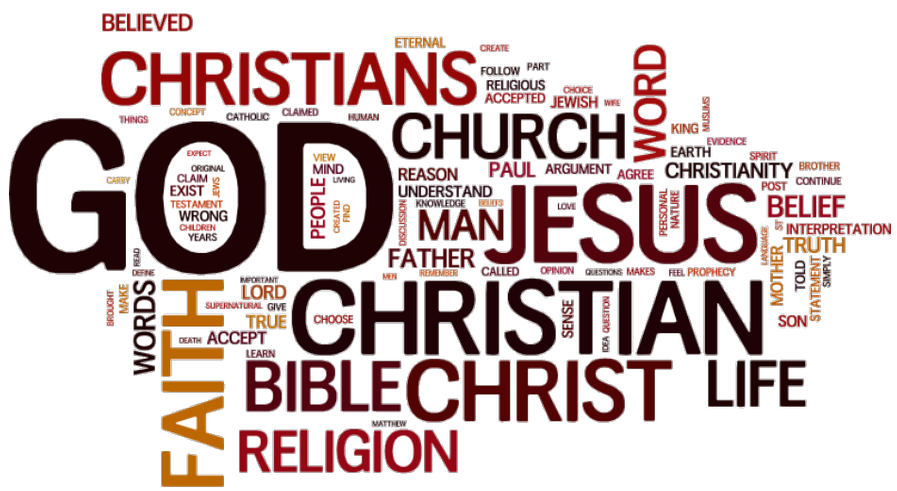
Wasserstein NMF



2. *Dictionary Learning*

$$\min_{\boldsymbol{A}\in(\Sigma_n)^K,\boldsymbol{\Lambda}\in(\Sigma_K)^N} \sum_{i=1}^{N} W_\gamma\left(\boldsymbol{b_i}, \sum_{k=1}^{K} \boldsymbol{\Lambda_k^i a_k}\right)$$

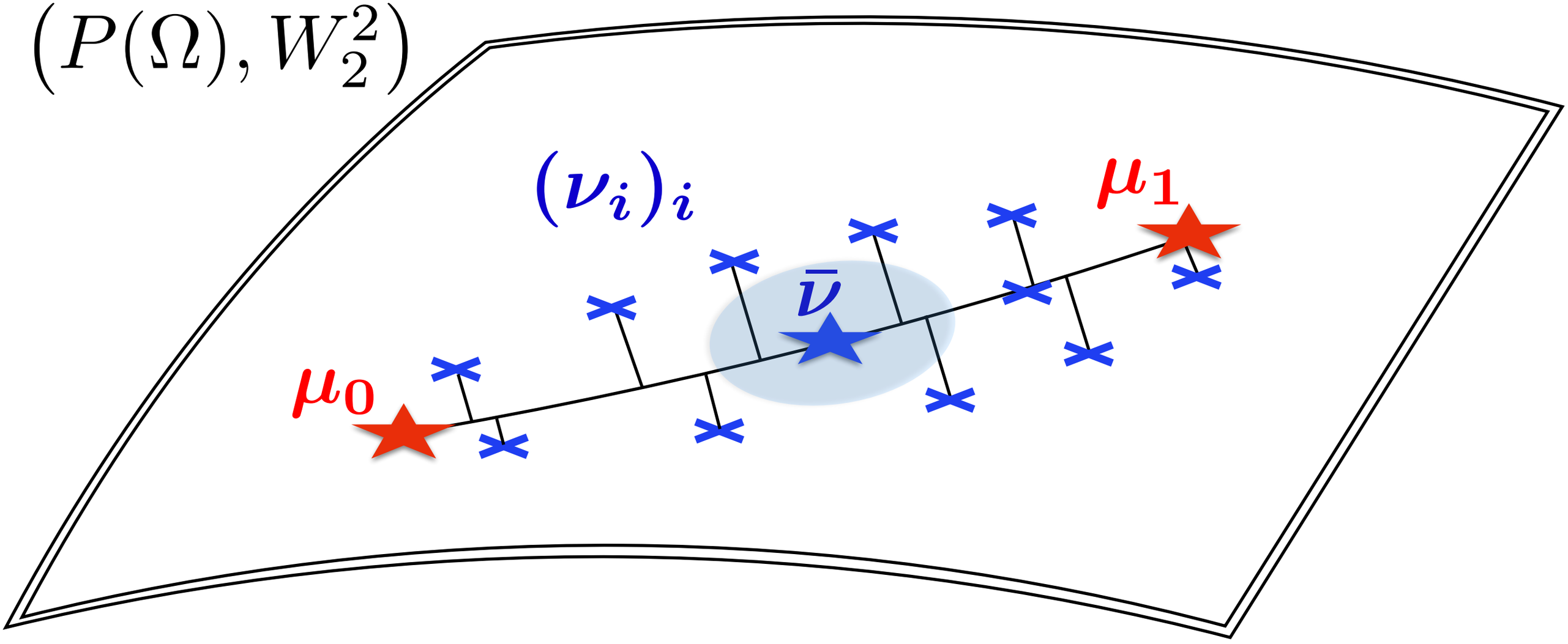




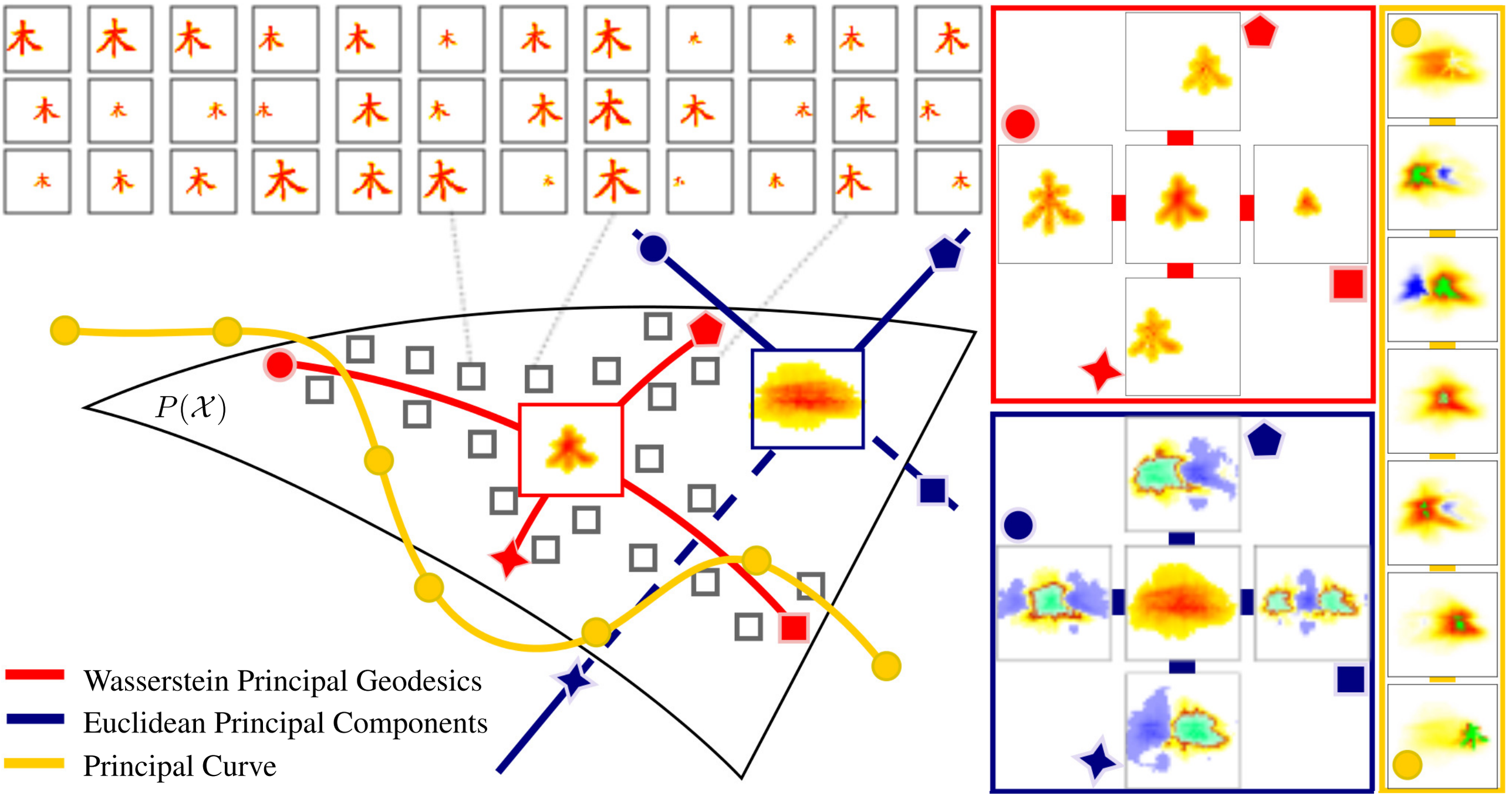
3. *Wasserstein PCA*

$$\min_{\boldsymbol{\mu_0},\boldsymbol{\mu_1}} \sum_{i=1}^{N} \min_{t} W_2^2(\rho^t_{\boldsymbol{\mu_0}\to\boldsymbol{\mu_1}}, \boldsymbol{\nu_i})$$

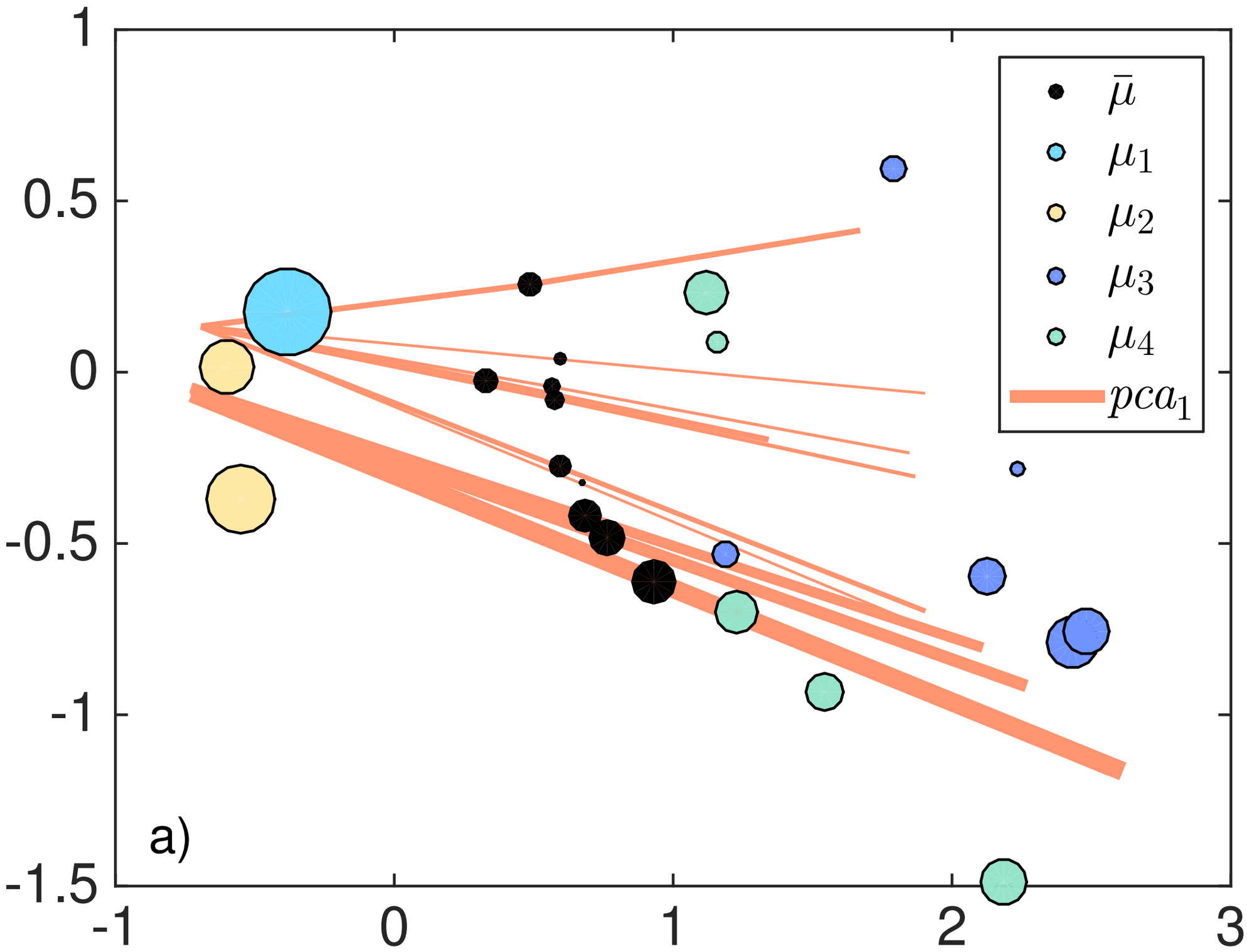
$\left(P(\Omega), W_2^2\right)$



Wasserstein PCA *vs.* Euclidean PCA



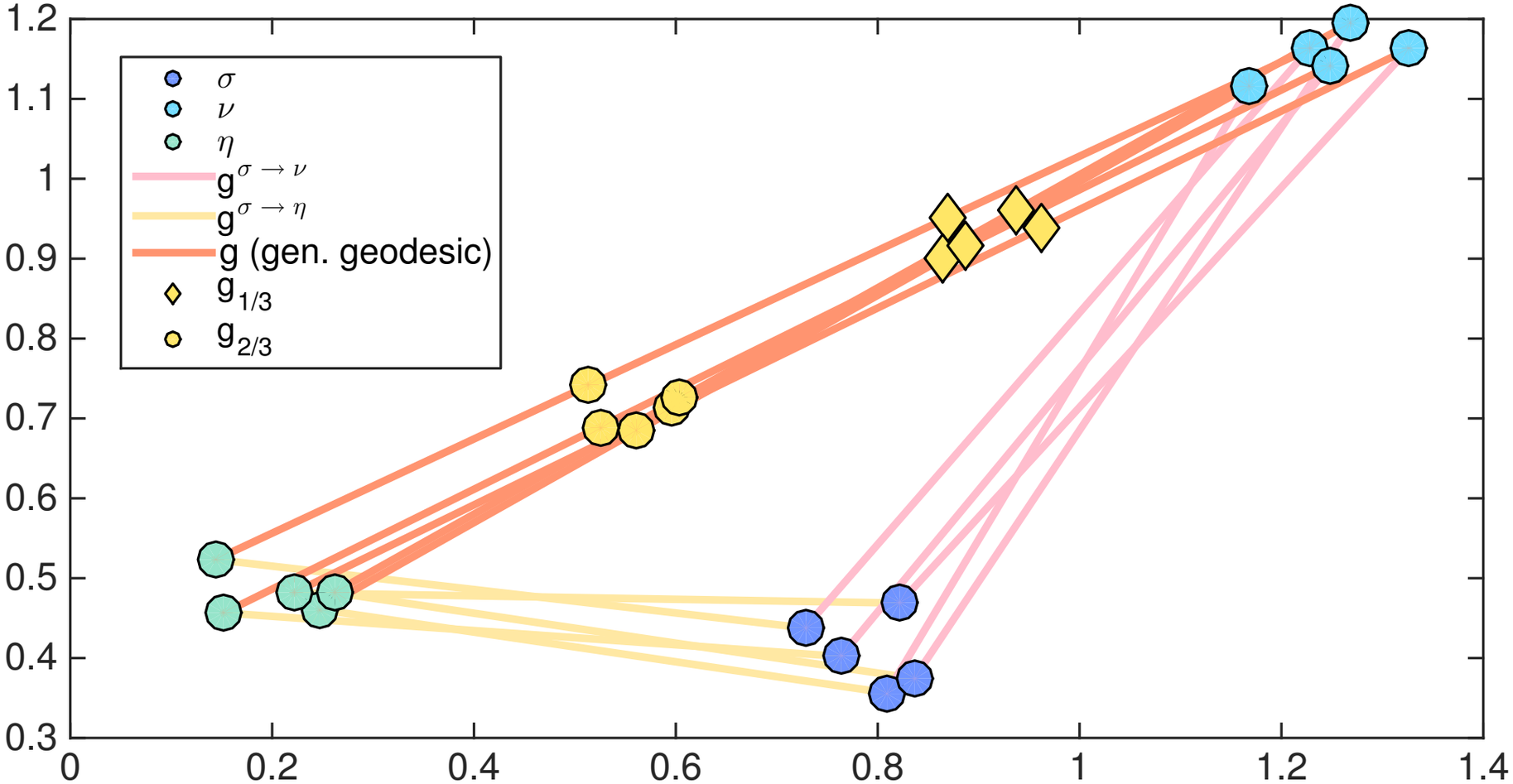
Generalized Principal Geodesics



[Ambrosio'06] Generalized Geodesics

$$\min_{\boldsymbol{v_1},\boldsymbol{v_2}\in L^2(\bar{\boldsymbol{\nu}},\Omega)} \sum_{i=1}^{N} \min_{t\in[0,1]} W_2^2\left(g_t(\boldsymbol{v_1},\boldsymbol{v_2}),\boldsymbol{\nu_i}\right) + \lambda R(\boldsymbol{v_1},\boldsymbol{v_2}),$$

$$\text{subject to} \begin{cases} g_t(\boldsymbol{v_1},\boldsymbol{v_2}) = (\mathrm{Id} - \boldsymbol{v_1} + t(\boldsymbol{v_1}+\boldsymbol{v_2}))\,\#\bar{\boldsymbol{\nu}} \\ \mathrm{Id} - \boldsymbol{v_1} \text{ and } \mathrm{Id} + \boldsymbol{v_2} \text{ are Monge maps from } \bar{\boldsymbol{\nu}} \end{cases}$$



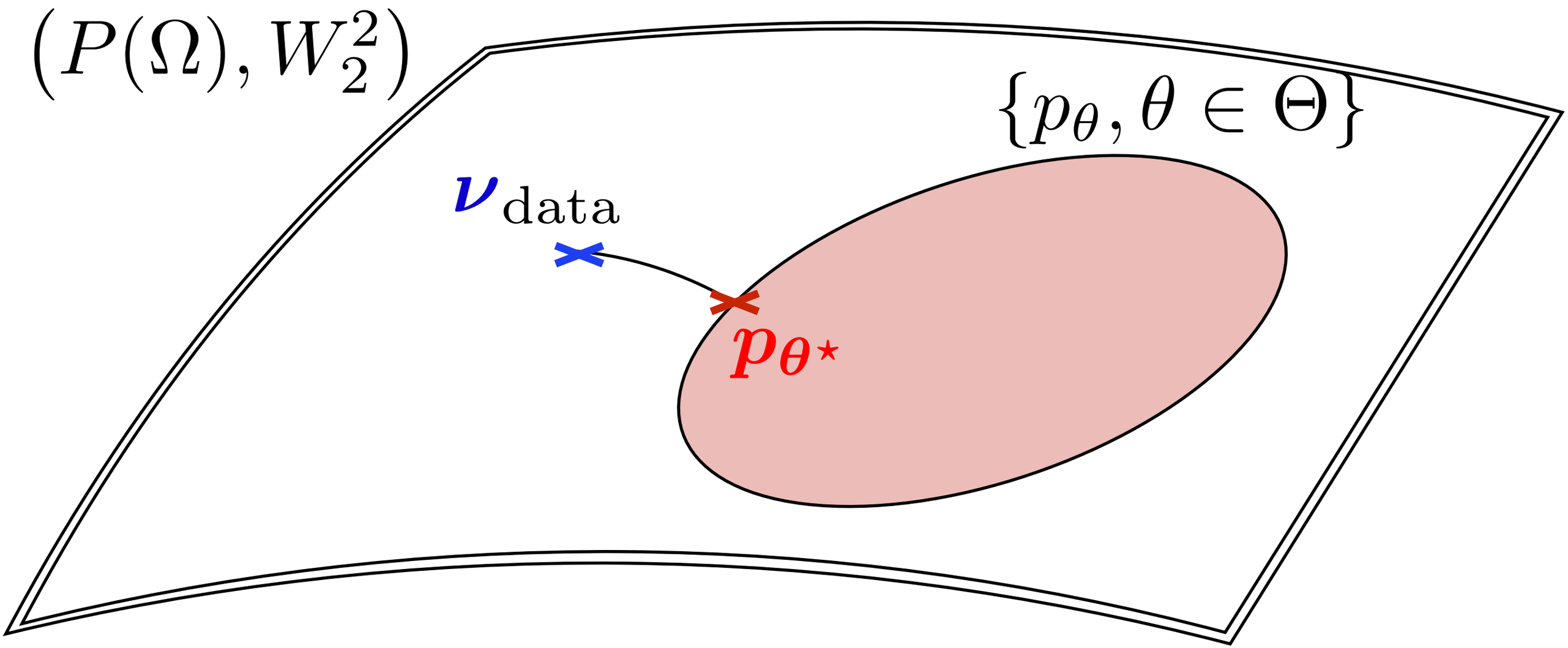
[SC'15]

Generalized Principal Geodesics

MNIST data

4. *Minimum Kantorovich Estimation*

$$\theta^\star = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \, W_p^p(\boldsymbol{p_\theta}, \boldsymbol{\nu}_{\text{data}})$$



4. *Minimum Kantorovich Estimation*

$$W_\gamma(p_{\boldsymbol{\theta}}, \boldsymbol{\nu}_{\text{data}}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \langle \boldsymbol{\alpha}, p_{\boldsymbol{\theta}} \rangle + \langle \boldsymbol{\beta}, \boldsymbol{\nu}_{\text{data}} \rangle - \gamma \langle e^{\boldsymbol{\alpha}/\gamma}, \; K e^{\boldsymbol{\beta}/\gamma} \rangle$$

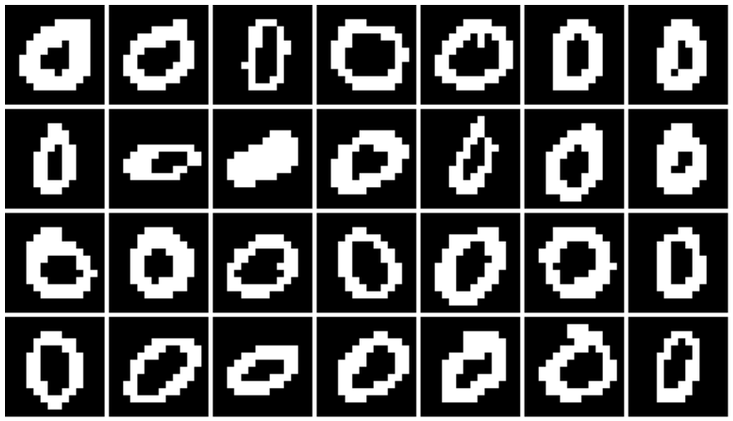
$$\nabla_\theta W_\gamma = \left(\frac{\partial p_\theta}{\partial \theta}\right)^T \boldsymbol{\alpha}^\star$$

- Application to Boltzmann machines **[MMC'15]** using contrastive divergence, better performance in denoising, data completion.

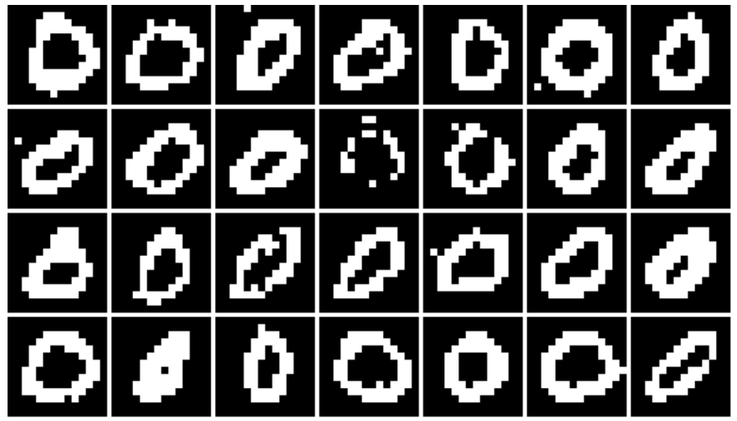
- Important statistical regularization / stochastic optimization problems.

Minimum Kantorovich Estimation

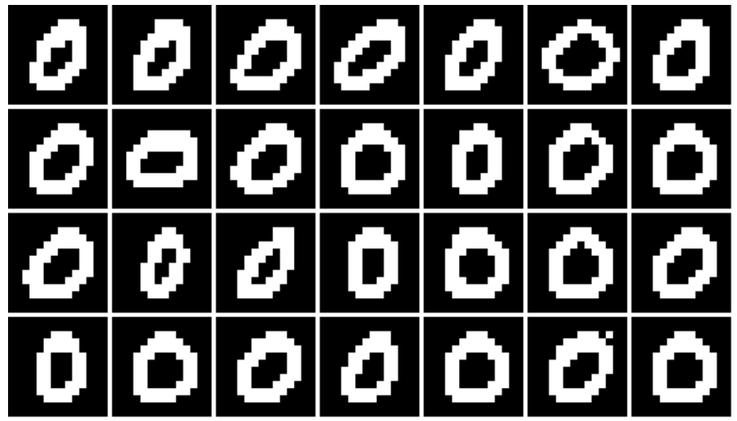
MNIST Data



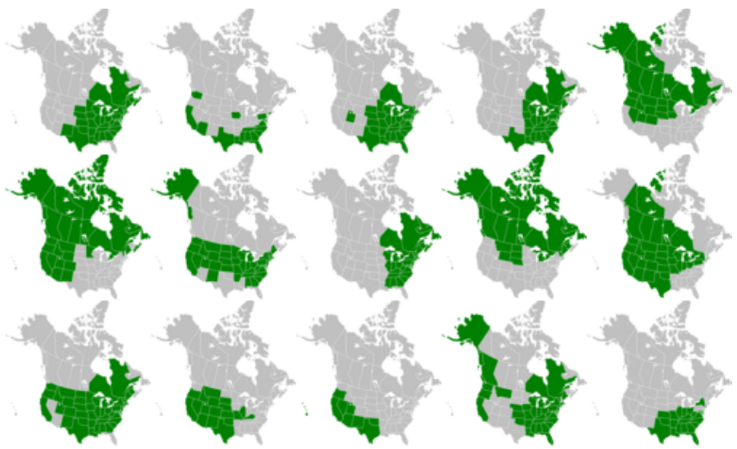
MNIST RBM



MNIST RBM-$\mathcal{W}$



PLANTS Data



PLANTS RBM

PLANTS RBM-$\mathcal{W}$

To conclude

- OT has deep/rich mathematical foundations.
- Regularized OT **[C.'13]** provides a convenient idea to import these ideas into stats/ML.
- Adopted in graphics/imaging, now in stats, ML, data fusion, Bayesian computation.