

Scaling Optimal Transport to High Dimensional Learning

Gabriel Peyré



Joint work with:



Shun'ichi
Amari



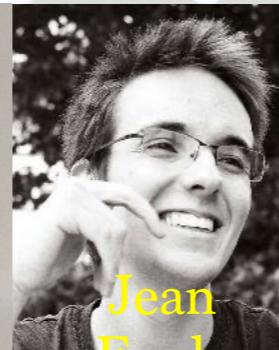
Francis
Bach



Lénaïc
Chizat



Marco
Cuturi



Jean
Feydy



Aude
Genevay



Thibault
Séjourné



Alain
Trouvé



François-Xavier
Vialard

<https://optimaltransport.github.io>

Home

Computational Optimal Transport

BOOK

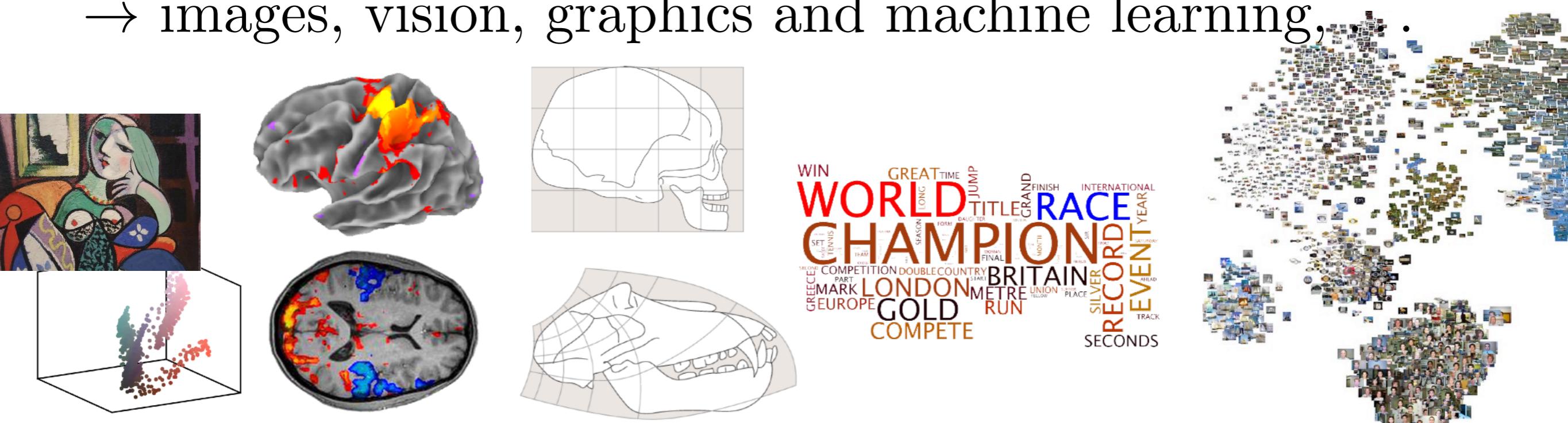
CODE

SLIDES

Probability Distributions in Data Sciences

Probability distributions and histograms

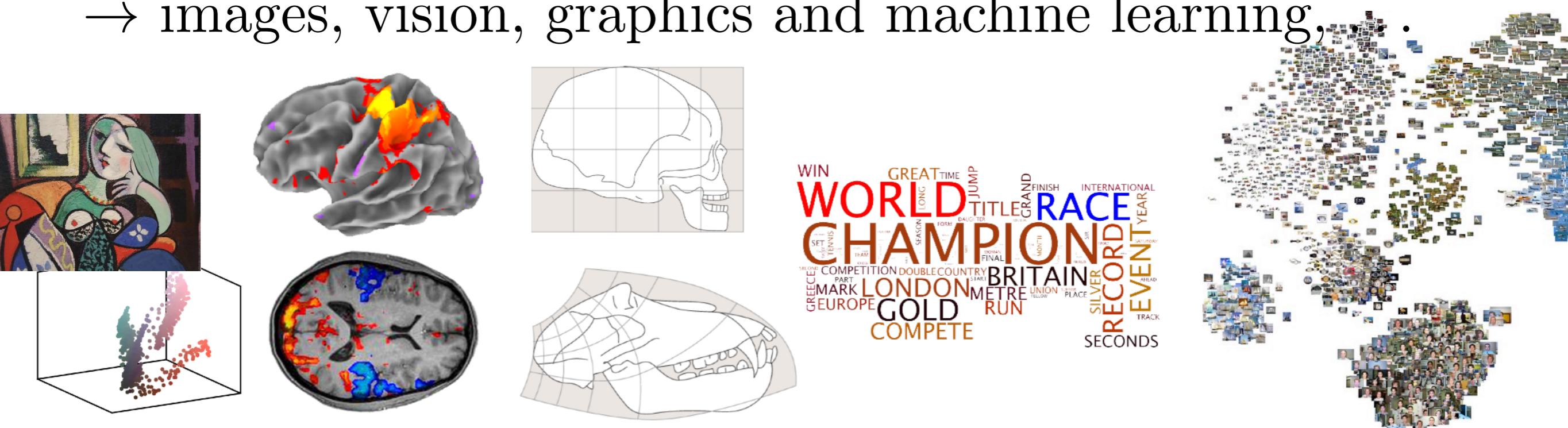
→ images, vision, graphics and machine learning,



Probability Distributions in Data Sciences

Probability distributions and histograms

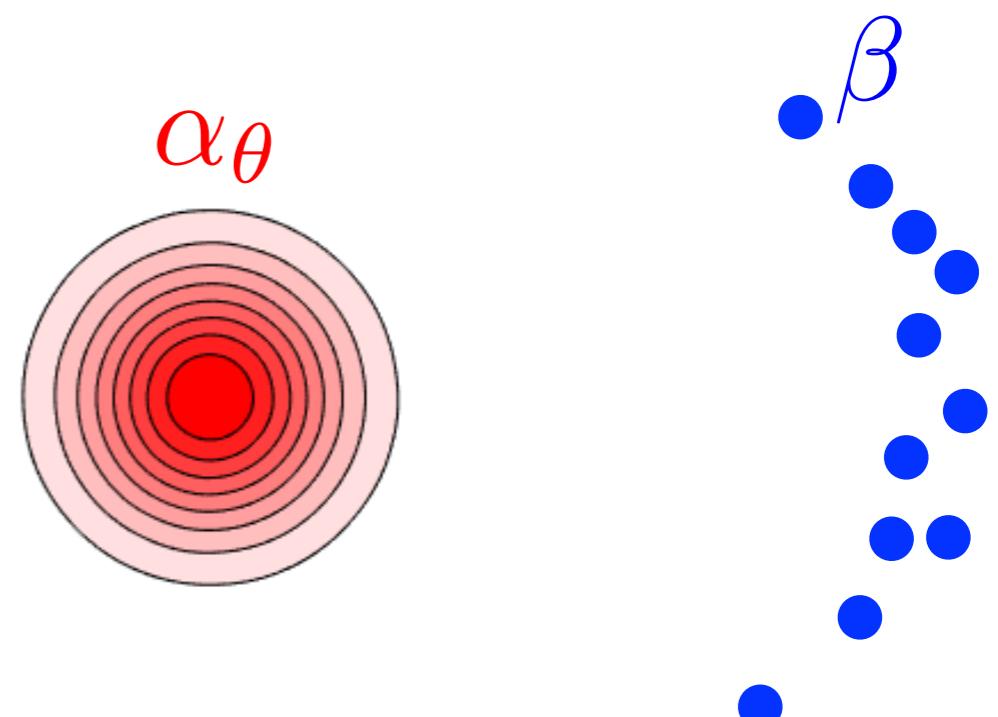
→ images, vision, graphics and machine learning,



Unsupervised learning

Observations: $\beta \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$

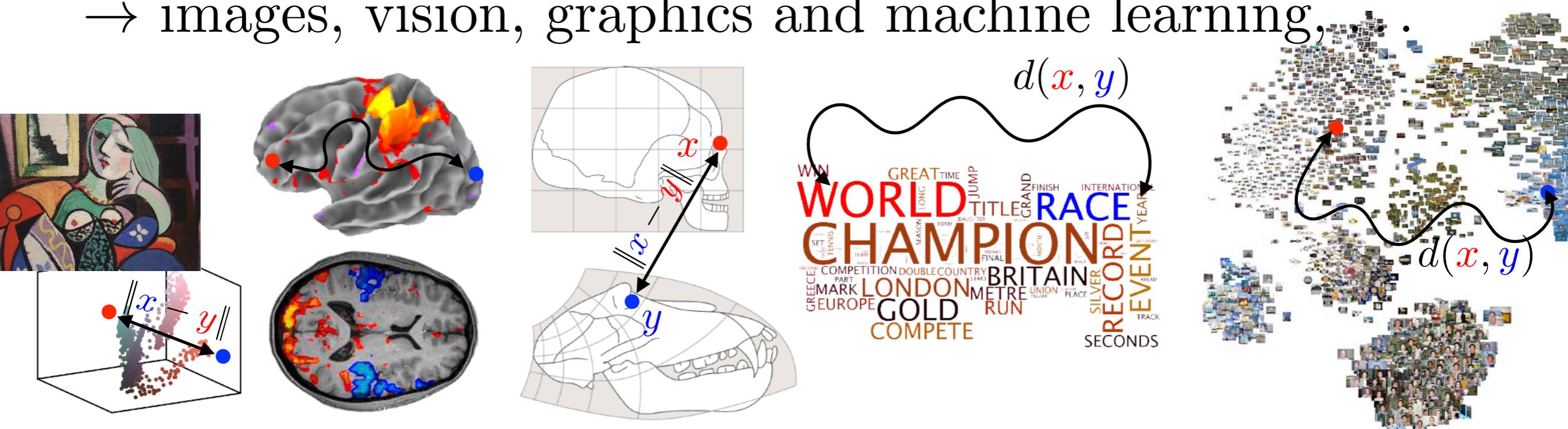
Parametric model: $\theta \mapsto \alpha_\theta$



Probability Distributions in Data Sciences

Probability distributions and histograms

→ images, vision, graphics and machine learning,

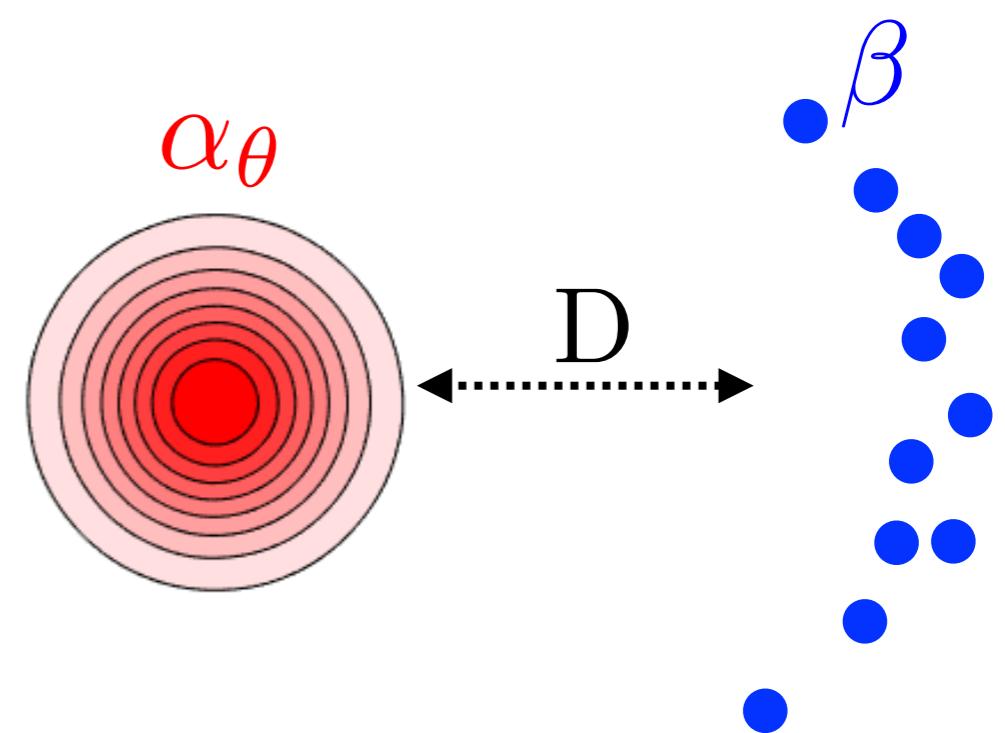


Unsupervised learning

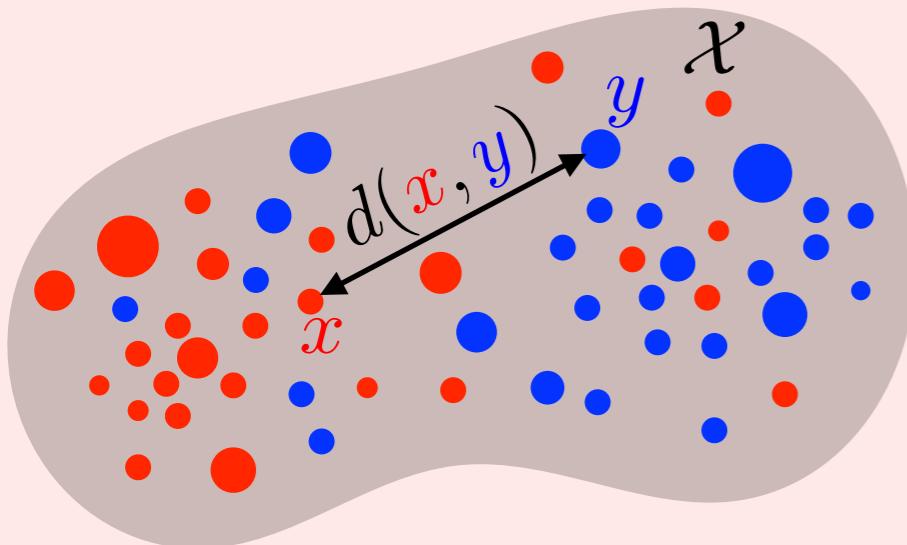
Observations: $\beta \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$

Parametric model: $\theta \mapsto \alpha_\theta$

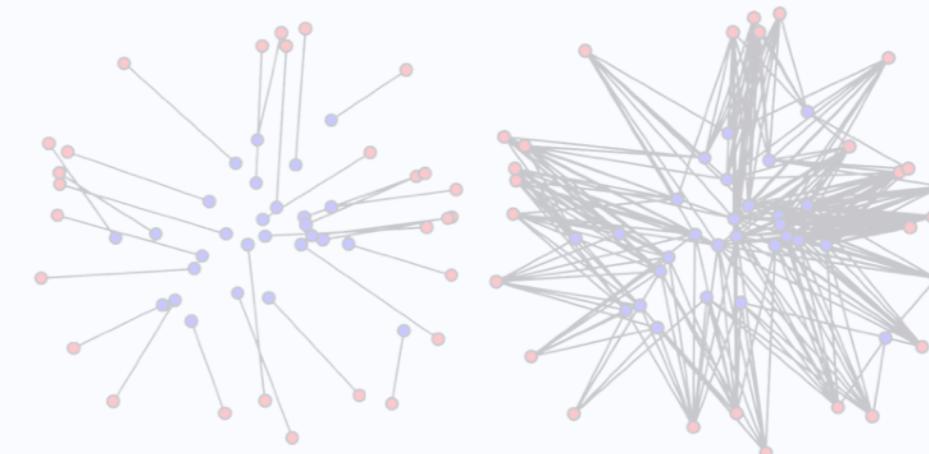
Density fitting: $\min_{\theta} D(\alpha_\theta, \beta)$
 → takes into account a metric d .



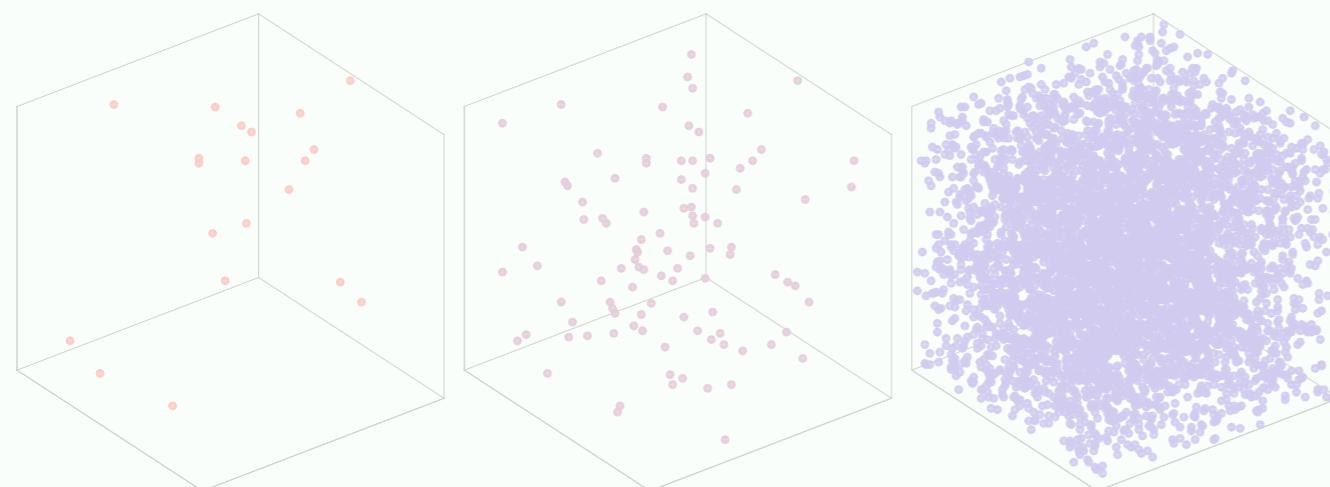
1. Optimal Transport



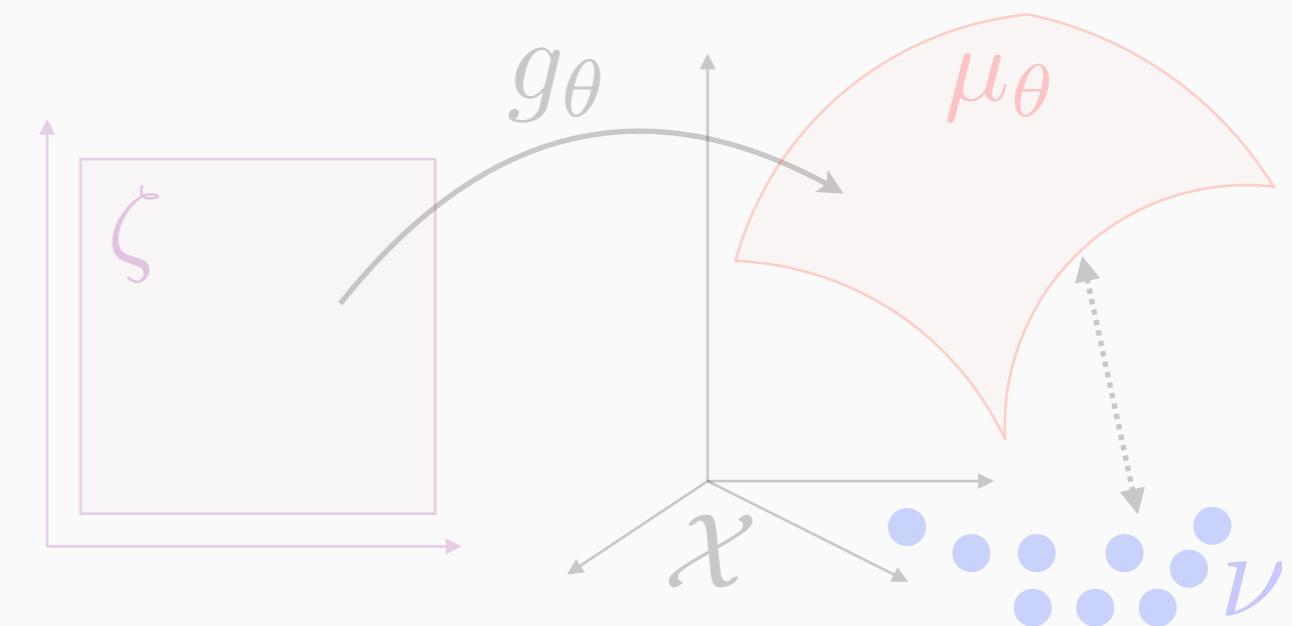
2. Entropic Regularization



3. Sinkhorn Divergences



4. Application to Generative Models

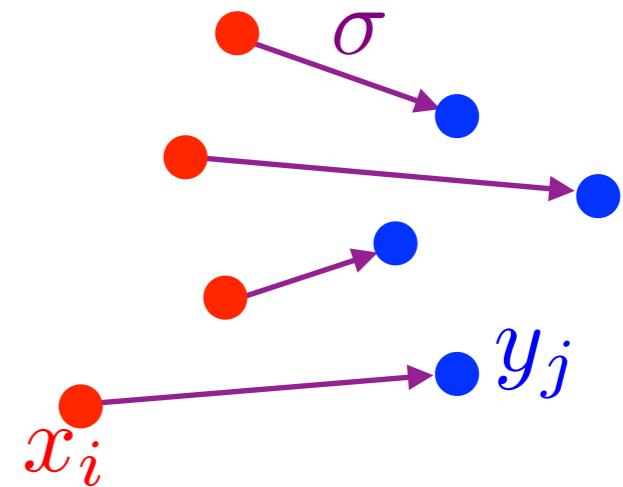


Monge's Problem

Points $(x_i)_i$, $(y_j)_j$

Permutation:

$$\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$$

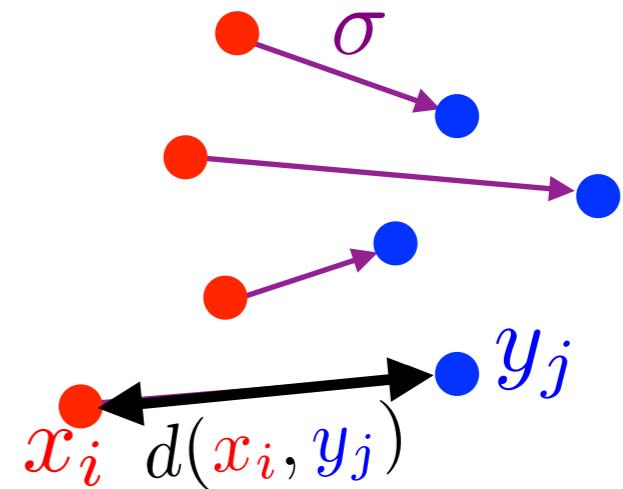


Monge's Problem

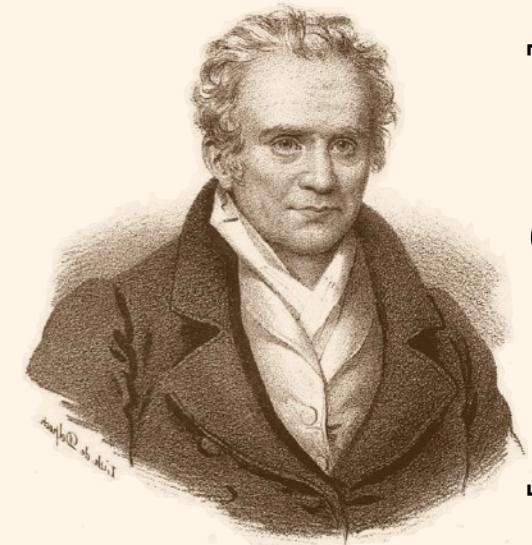
Points $(x_i)_i$, $(y_j)_j$

Permutation:

$$\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$$



Monge optimal matching: $\min_{\sigma} \sum_{i=1}^n d(\textcolor{red}{x}_i, \textcolor{blue}{y}_{\sigma(i)})$



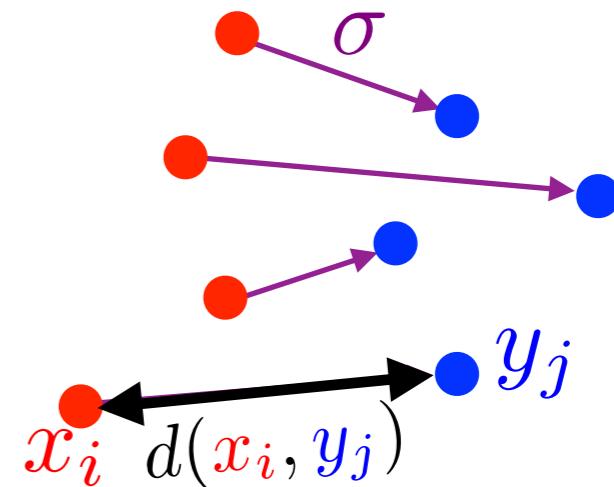
[Monge 1784]

Monge's Problem

Points $(x_i)_i$, $(y_j)_j$

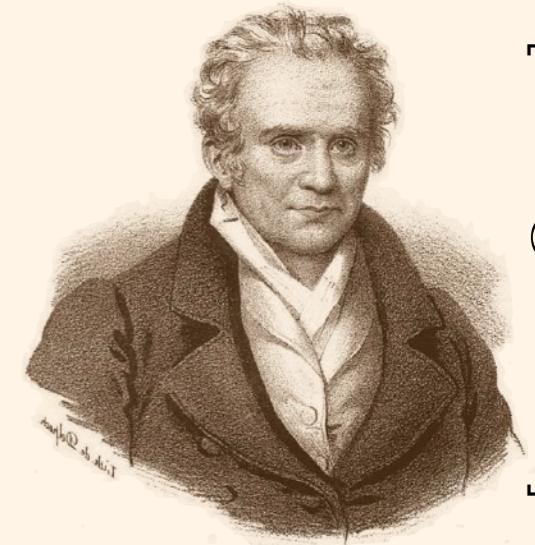
Permutation:

$$\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$$



Monge optimal matching: $\min_{\sigma} \sum_{i=1}^n d(\textcolor{red}{x}_i, \textcolor{blue}{y}_{\sigma(i)})$

→ Seems intractable: $n!$ possibilities.



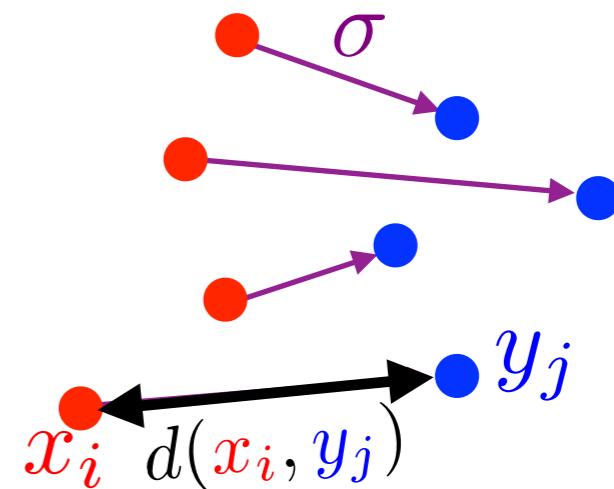
[Monge 1784]

Monge's Problem

Points $(x_i)_i$, $(y_j)_j$

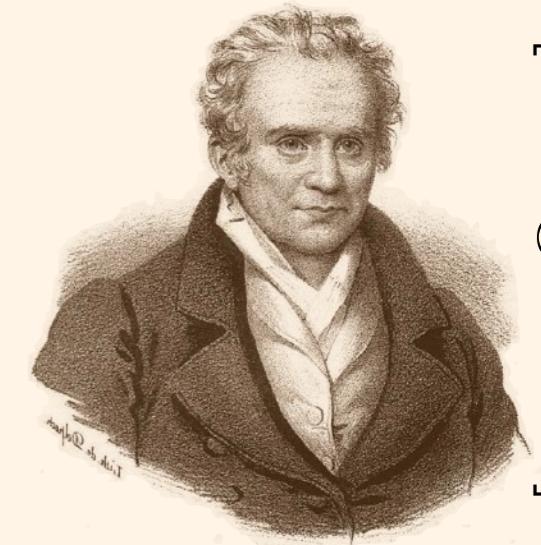
Permutation:

$$\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$$

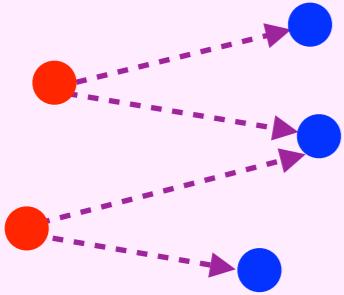


Monge optimal matching: $\min_{\sigma} \sum_{i=1}^n d(\textcolor{red}{x}_i, y_{\sigma(i)})$

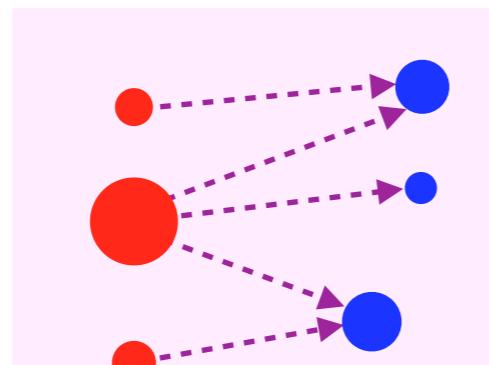
→ Seems intractable: $n!$ possibilities.



[Monge 1784]



Different
points?



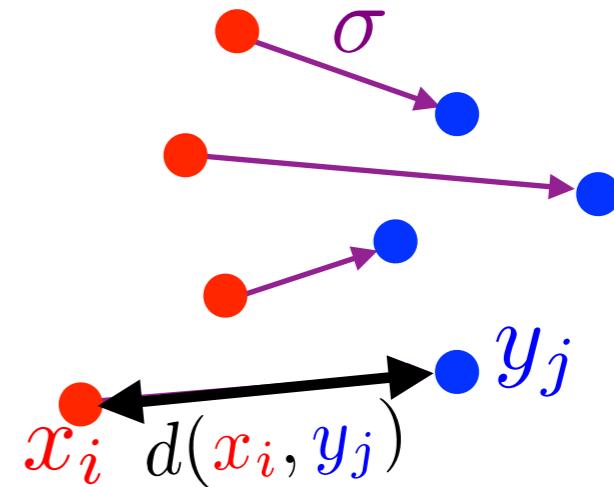
Weights?

Monge's Problem

Points $(x_i)_i, (y_j)_j$

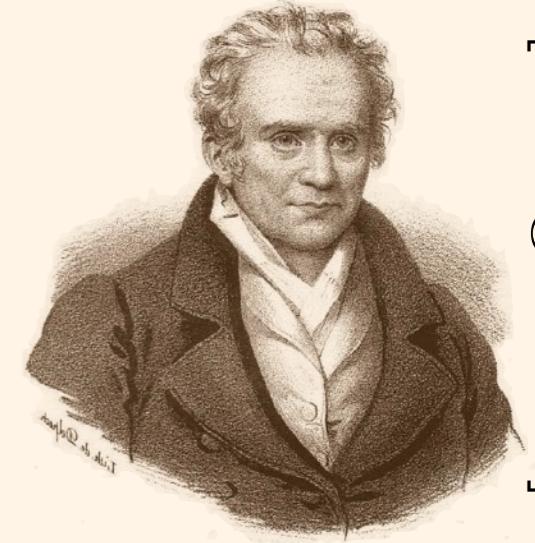
Permutation:

$$\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$$

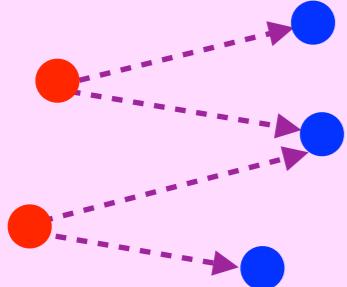


Monge optimal matching: $\min_{\sigma} \sum_{i=1}^n d(\textcolor{red}{x}_i, y_{\sigma(i)})$

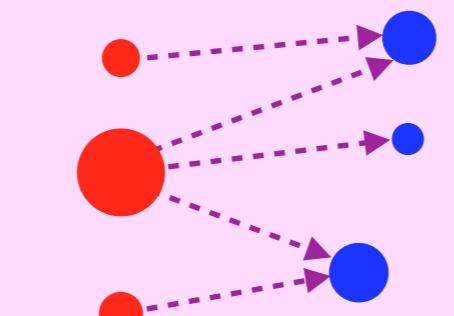
→ Seems intractable: $n!$ possibilities.



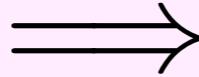
[Monge 1784]



Different
points?



Weights?



“Relax”
points → mass
permutation → coupling

Kantorovitch's Formulation

Discrete distributions:

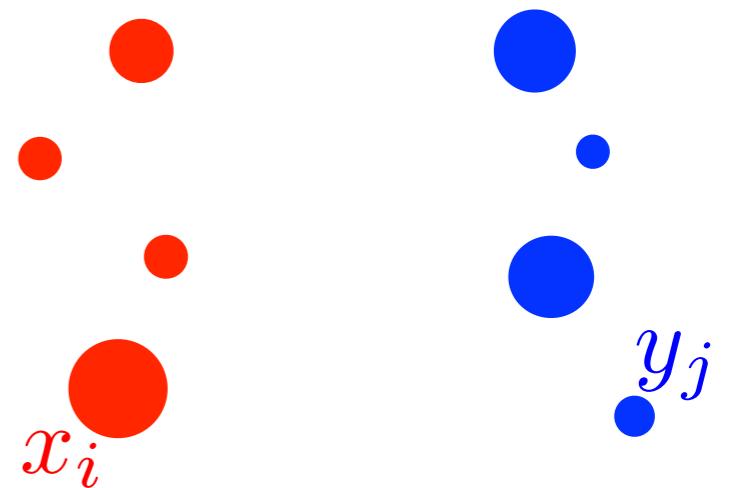
$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$$

$$\beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$$

Points $(x_i)_i, (y_j)_j$

Weights $\mathbf{a}_i \geq 0, \mathbf{b}_j \geq 0.$

$$\sum_{i=1}^n \mathbf{a}_i = \sum_{j=1}^m \mathbf{b}_j = 1$$



Kantorovitch's Formulation

Discrete distributions:

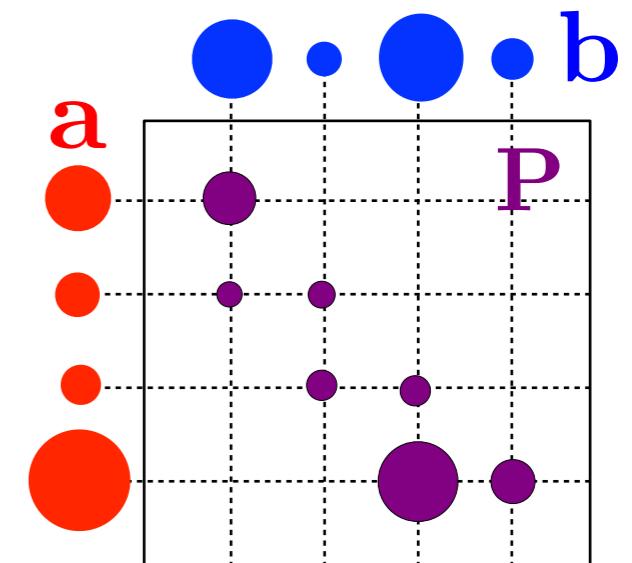
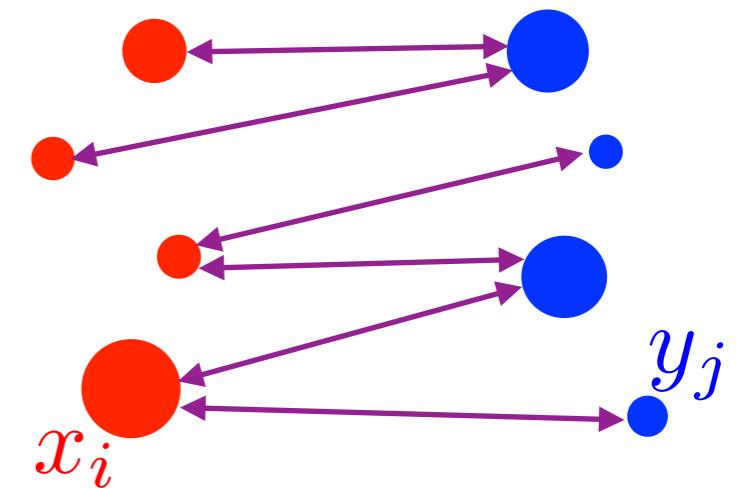
$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$$

$$\beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$$

Points $(x_i)_i, (y_j)_j$

Weights $\mathbf{a}_i \geq 0, \mathbf{b}_j \geq 0.$

$$\sum_{i=1}^n \mathbf{a}_i = \sum_{j=1}^m \mathbf{b}_j = 1$$



Couplings:

$$\sum_j \mathbf{P}_{i,j} = \mathbf{a}_i$$

$$\sum_i \mathbf{P}_{i,j} = \mathbf{b}_j$$

$$\mathbf{P} \geq 0, \mathbf{P}\mathbf{1}_m = \mathbf{a}, \mathbf{P}^\top \mathbf{1}_n = \mathbf{b}$$

Kantorovitch's Formulation

Discrete distributions:

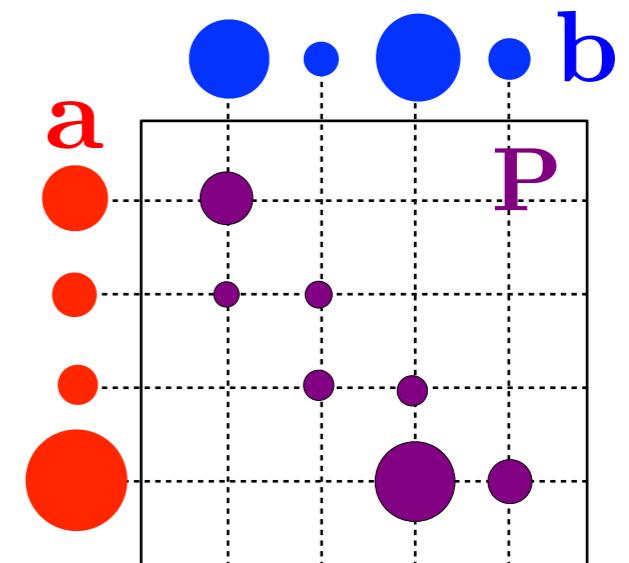
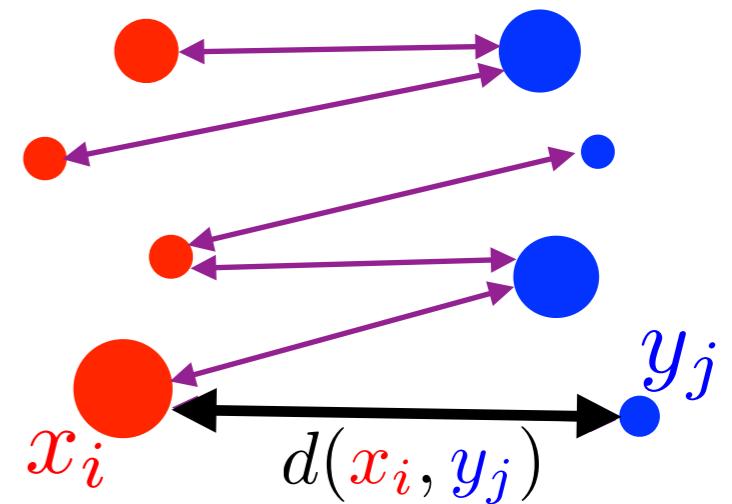
$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$$

$$\beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$$

Points $(x_i)_i, (y_j)_j$

Weights $\mathbf{a}_i \geq 0, \mathbf{b}_j \geq 0.$

$$\sum_{i=1}^n \mathbf{a}_i = \sum_{j=1}^m \mathbf{b}_j = 1$$



Couplings: $\sum_j \mathbf{P}_{i,j} = \mathbf{a}_i$

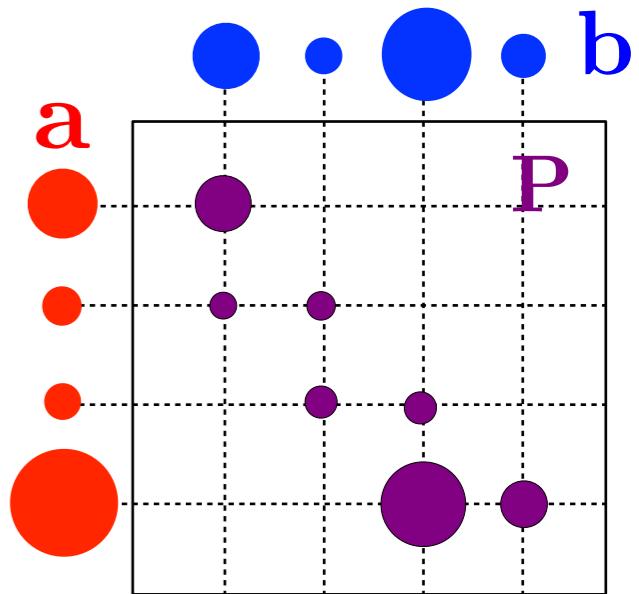
$\sum_i \mathbf{P}_{i,j} = \mathbf{b}_j$

[Kantorovich 1942]

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} ; \mathbf{P} \geq 0, \mathbf{P} \mathbf{1}_m = \mathbf{a}, \mathbf{P}^\top \mathbf{1}_n = \mathbf{b} \right\}$$

Optimal Transport Distances

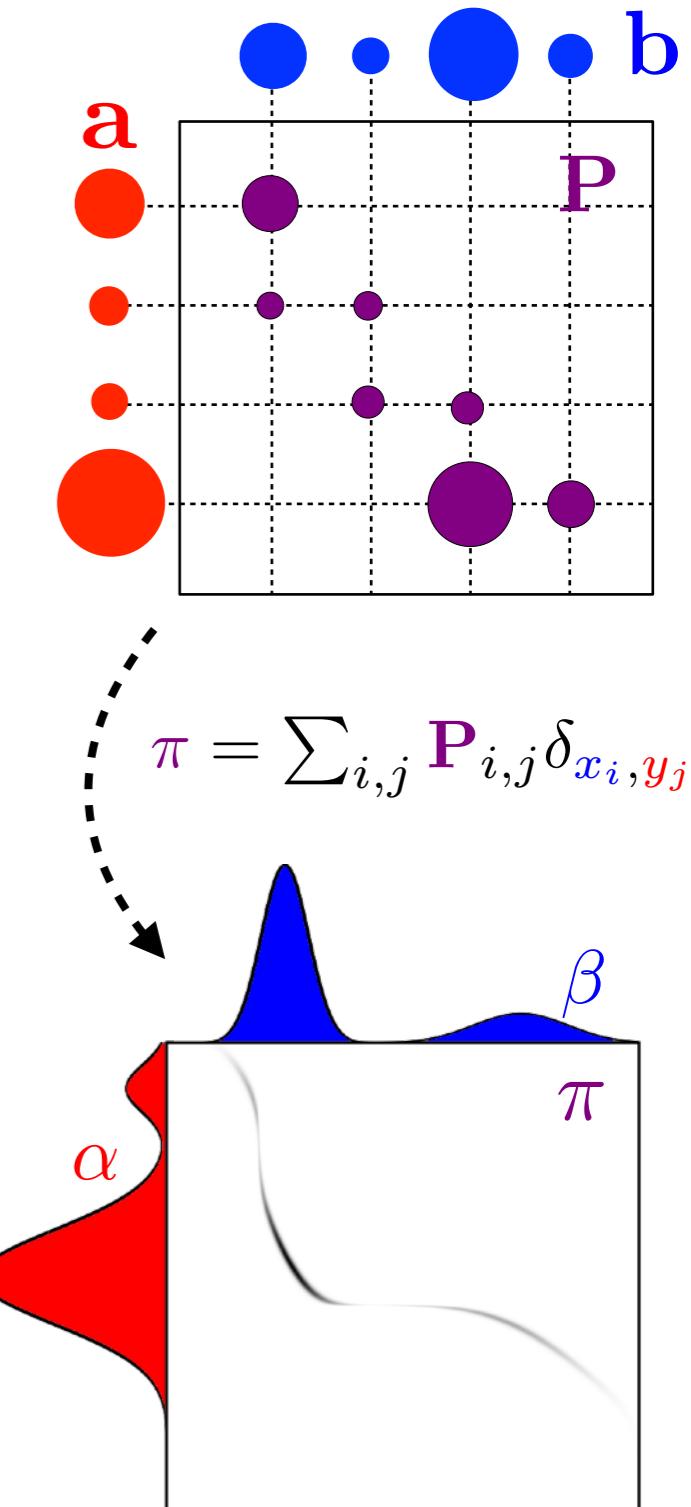
$$W_p(\alpha, \beta) \stackrel{\text{def.}}{=} \left(\min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} \right)^{\frac{1}{p}}$$



Optimal Transport Distances

$$W_p(\alpha, \beta) \stackrel{\text{def.}}{=} \left(\min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} \right)^{\frac{1}{p}}$$

$$\min_{\pi \in \mathcal{M}_+^1(\mathcal{X}^2)} \left\{ \int_{\mathcal{X}^2} d(x, y)^p d\pi(x, y) ; \pi_1 = \alpha, \pi_2 = \beta \right\}$$



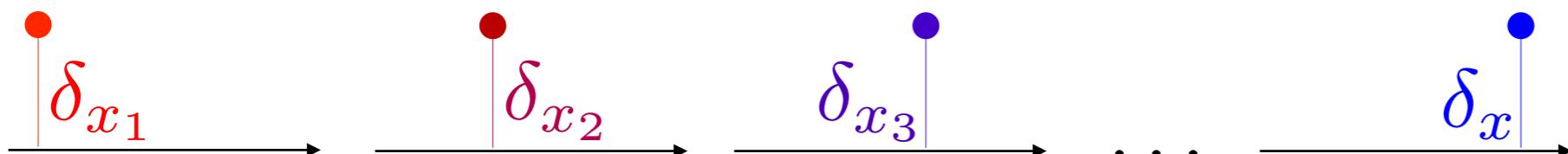
Optimal Transport Distances

$$W_p(\alpha, \beta) \stackrel{\text{def.}}{=} \left(\min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} \right)^{\frac{1}{p}}$$

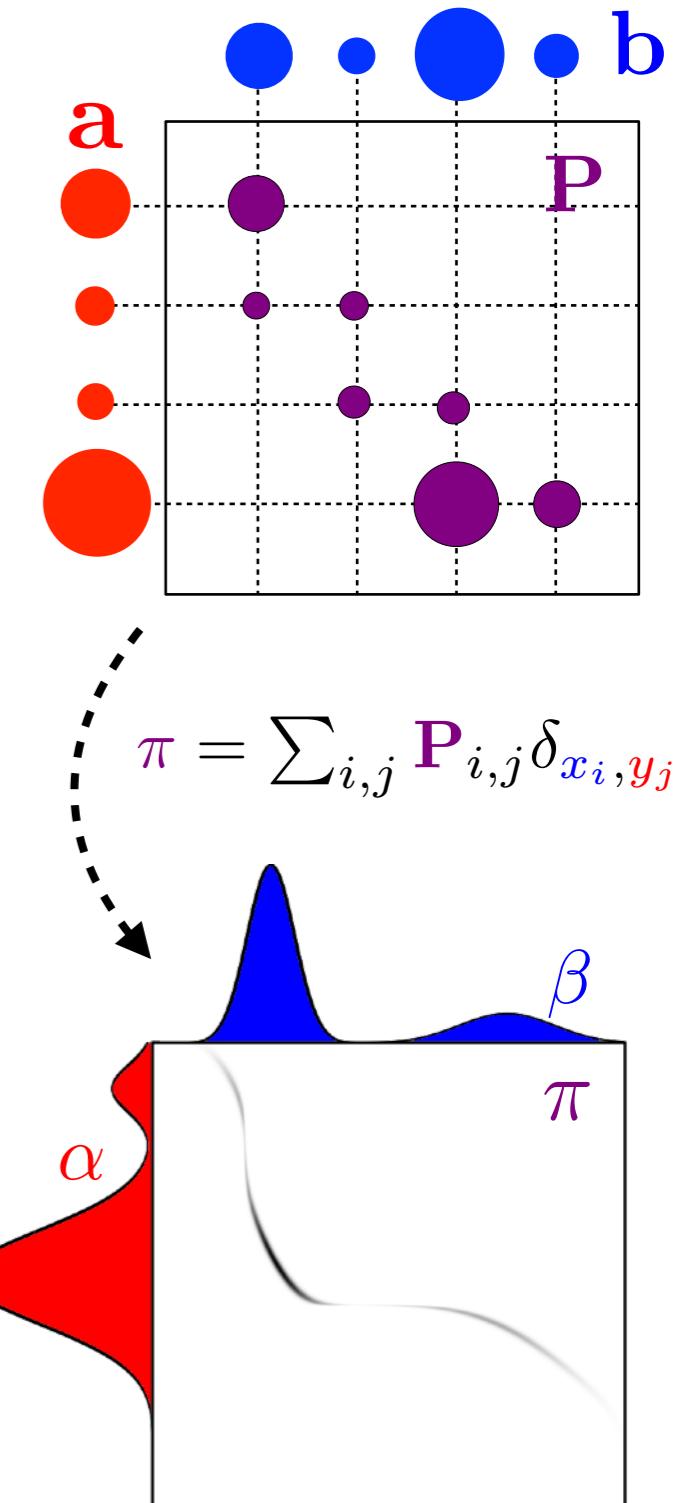
$$\min_{\pi \in \mathcal{M}_+^1(\mathcal{X}^2)} \left\{ \int_{\mathcal{X}^2} d(x, y)^p d\pi(x, y) ; \pi_1 = \alpha, \pi_2 = \beta \right\}$$

Convergence in law: $\alpha_n \rightarrow \beta$

$$\Leftrightarrow \forall f \in \mathcal{C}(\mathcal{X}), \int_{\mathcal{X}} f d\alpha_n \rightarrow \int_{\mathcal{X}} f d\beta$$



$$\|\delta_{x_n} - \delta_x\|_1 = 2 \quad \text{vs.} \quad W_p(\delta_{x_n}, \delta_x) = d(x_n, x)$$



Optimal Transport Distances

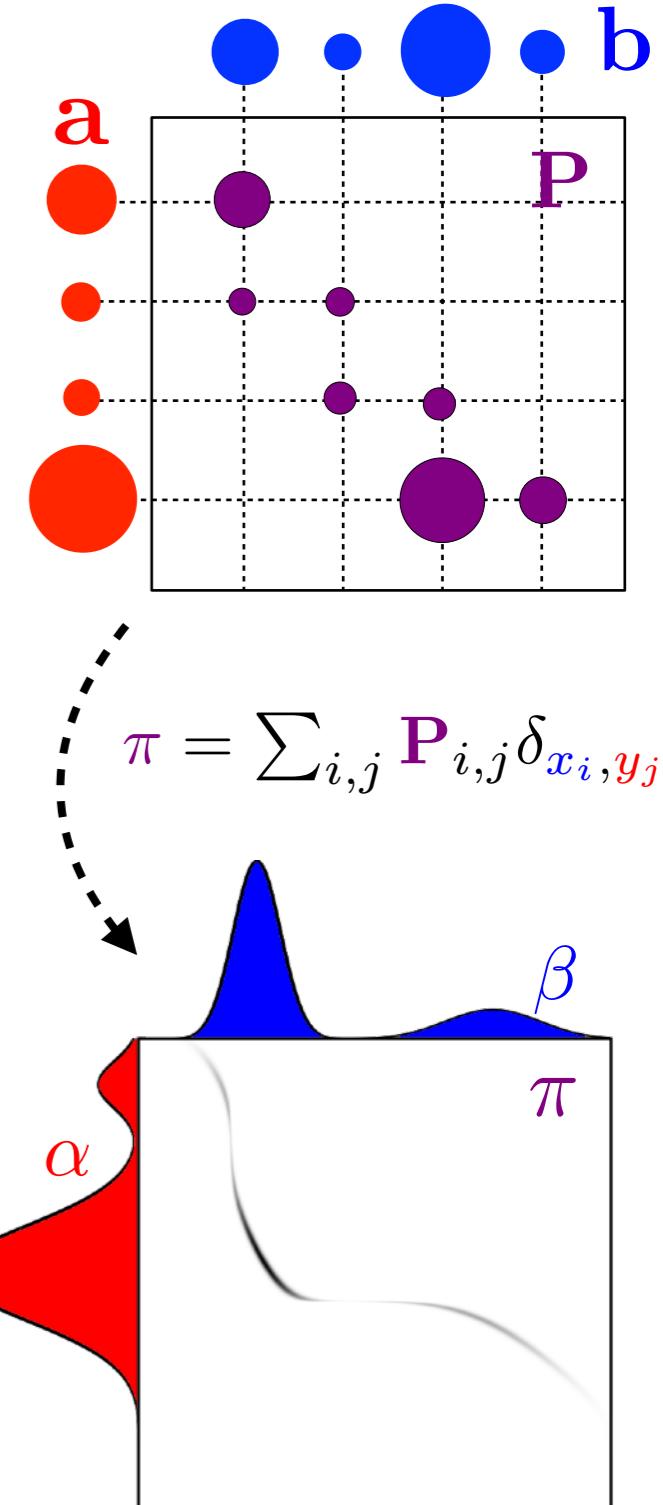
$$W_p(\alpha, \beta) \stackrel{\text{def.}}{=} \left(\min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} \right)^{\frac{1}{p}}$$

$$\min_{\pi \in \mathcal{M}_+^1(\mathcal{X}^2)} \left\{ \int_{\mathcal{X}^2} d(x, y)^p d\pi(x, y) ; \pi_1 = \alpha, \pi_2 = \beta \right\}$$

Convergence in law: $\alpha_n \rightarrow \beta$

$$\Leftrightarrow \forall f \in \mathcal{C}(\mathcal{X}), \int_{\mathcal{X}} f d\alpha_n \rightarrow \int_{\mathcal{X}} f d\beta$$

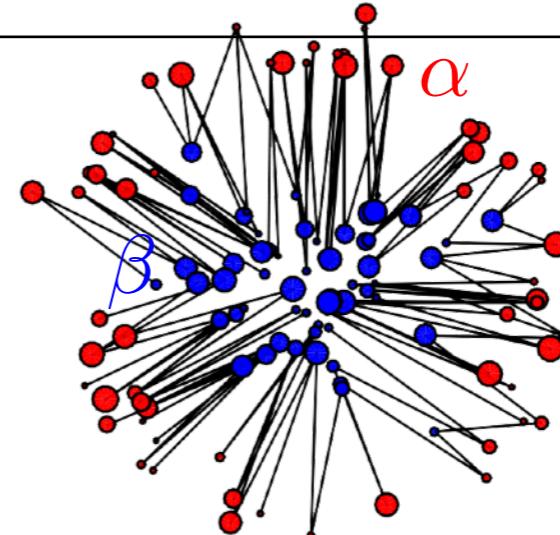
$$\|\delta_{x_n} - \delta_x\|_1 = 2 \quad \text{vs.} \quad W_p(\delta_{x_n}, \delta_x) = d(x_n, x)$$



Theorem: W_p is a distance and $\alpha_n \rightarrow \beta \Leftrightarrow W_p(\alpha_n, \beta) \rightarrow 0$

Algorithms

Linear programming: $O(n^3 \log(n)^2)$

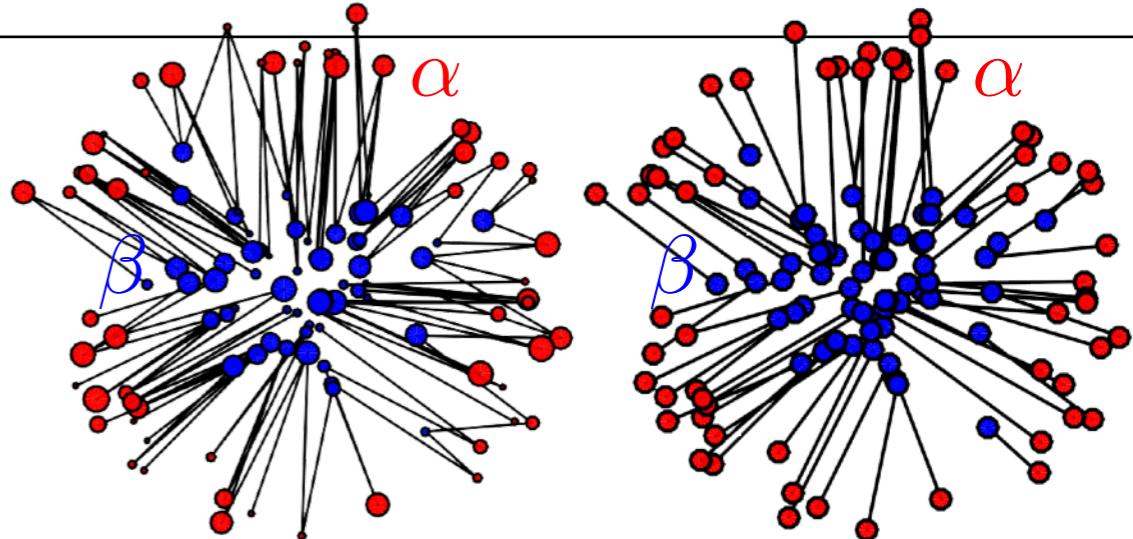


Algorithms

Linear programming: $O(n^3 \log(n)^2)$

Hungarian/Auction: $O(n^3)$

$$\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \beta = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$$



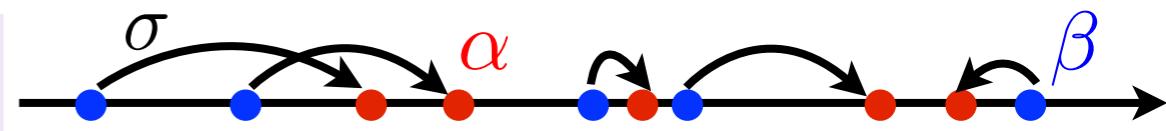
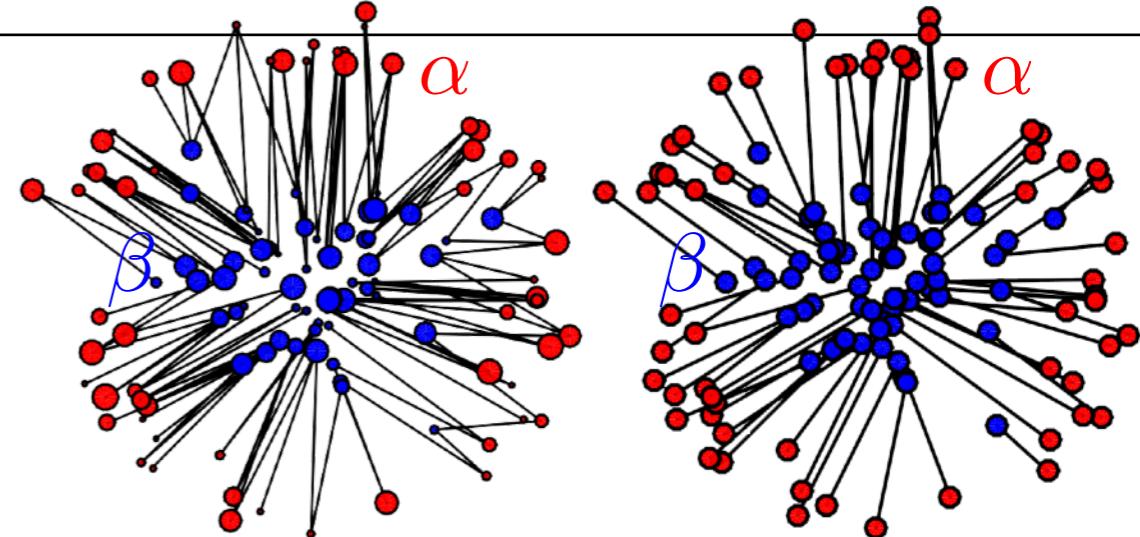
Algorithms

Linear programming: $O(n^3 \log(n)^2)$

Hungarian/Auction: $O(n^3)$

$$\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \beta = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$$

1-D case: sorting $O(n \log(n))$.



Algorithms

Linear programming: $O(n^3 \log(n)^2)$

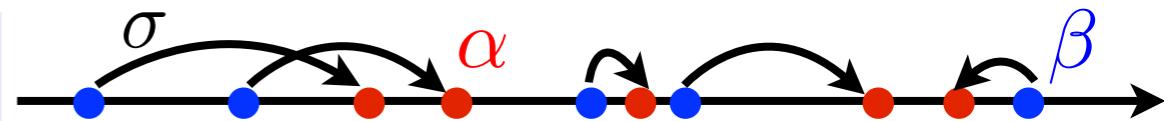
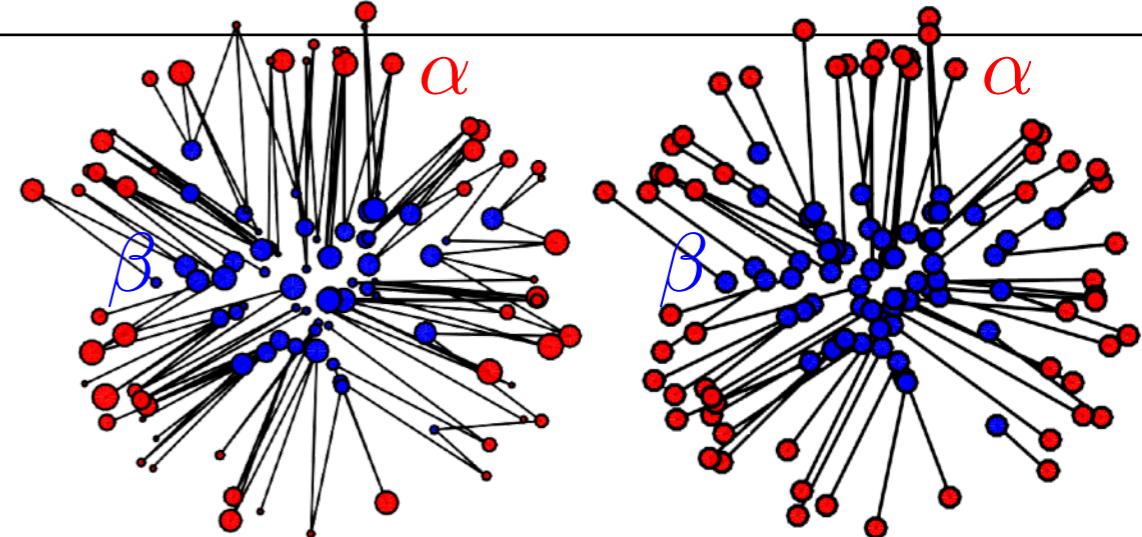
Hungarian/Auction: $O(n^3)$

$$\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \beta = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$$

1-D case: sorting $O(n \log(n))$.

$$\begin{array}{ll} p = 1 & \\ d = \|\cdot\| & W_1(\alpha, \beta) = \min_{\text{div}(u) = \alpha - \beta} \int \|u(x)\| dx \end{array}$$

\rightarrow min-cost flow, on graphs $O(n^2 \log(n))$.



Algorithms

Linear programming: $O(n^3 \log(n)^2)$

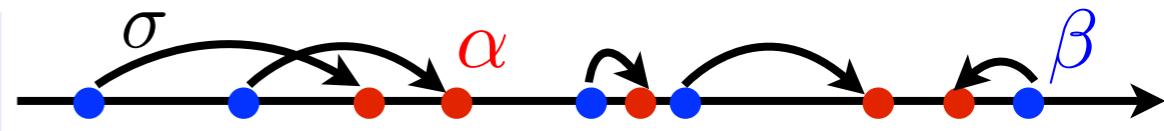
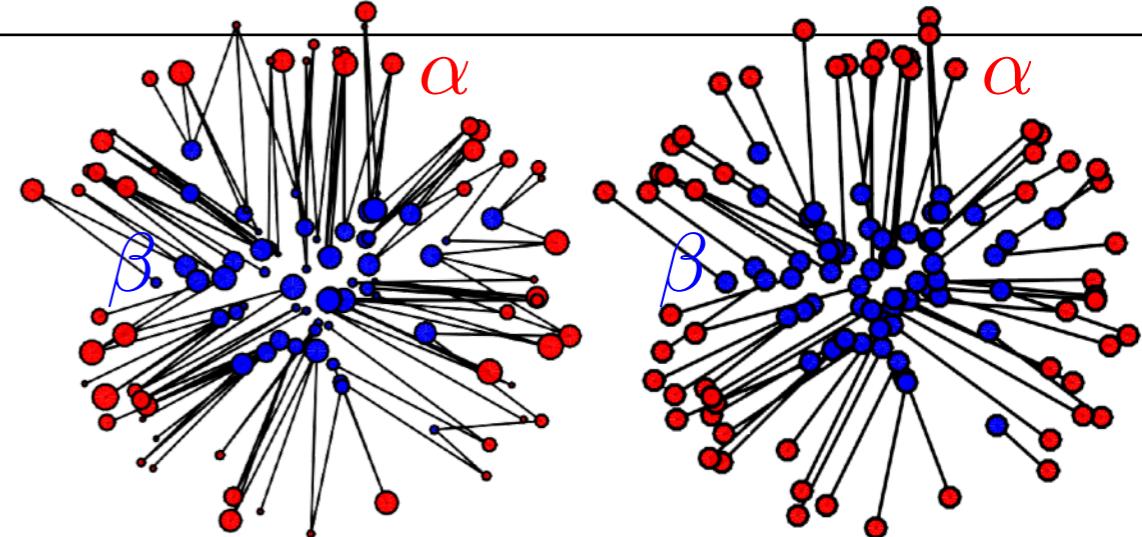
Hungarian/Auction: $O(n^3)$

$$\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \beta = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$$

1-D case: sorting $O(n \log(n))$.

$$\begin{aligned} p &= 1 \\ d &= \|\cdot\| \end{aligned} \quad W_1(\alpha, \beta) = \min_{\text{div}(u) = \alpha - \beta} \int \|u(x)\| dx$$

→ min-cost flow, on graphs $O(n^2 \log(n))$.



Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.

Algorithms

Linear programming: $O(n^3 \log(n)^2)$

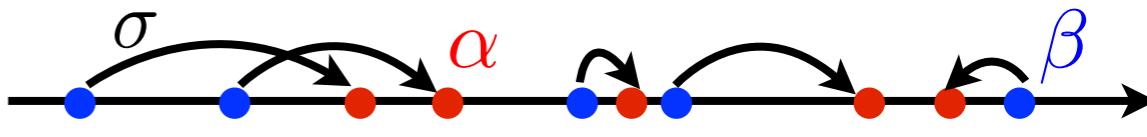
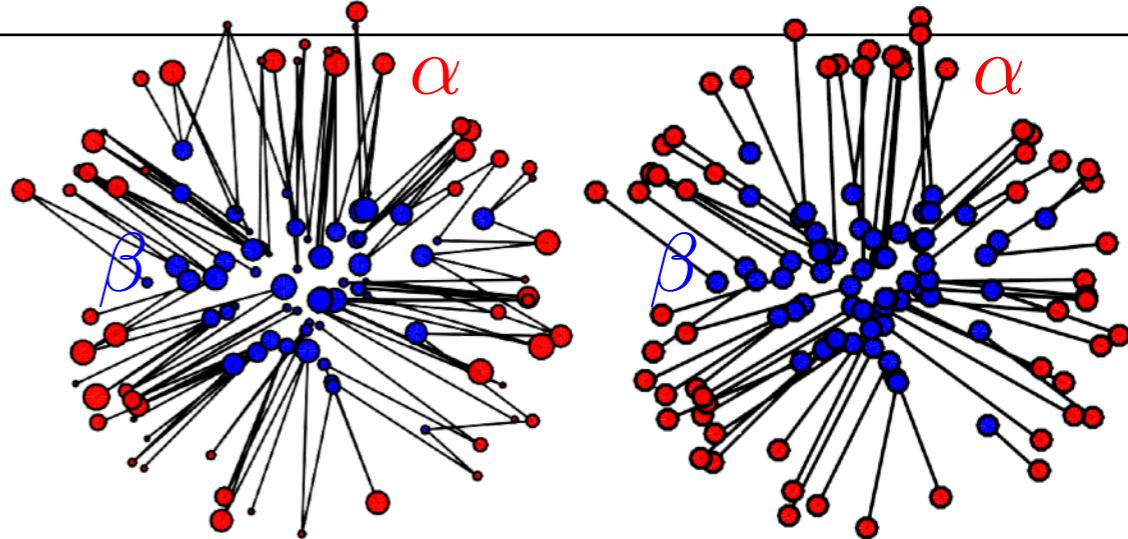
Hungarian/Auction: $O(n^3)$

$$\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \beta = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$$

1-D case: sorting $O(n \log(n))$.

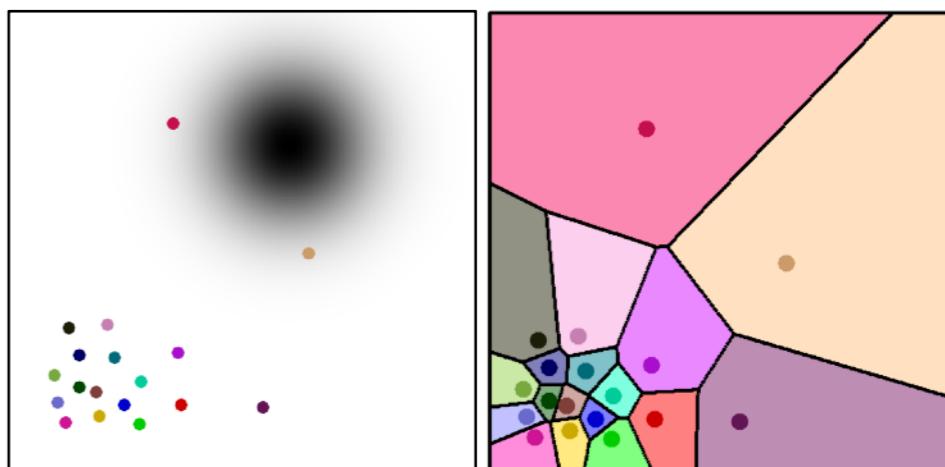
$$\begin{array}{ll} p = 1 \\ d = \|\cdot\| \end{array} \quad W_1(\alpha, \beta) = \min_{\text{div}(u) = \alpha - \beta} \int \|u(x)\| dx$$

\rightarrow min-cost flow, on graphs $O(n^2 \log(n))$.



Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.

Semi-discrete: Laguerre cells, $d = \|\cdot\|_2^2$.
[Merigot 2013]



Algorithms

Linear programming: $O(n^3 \log(n)^2)$

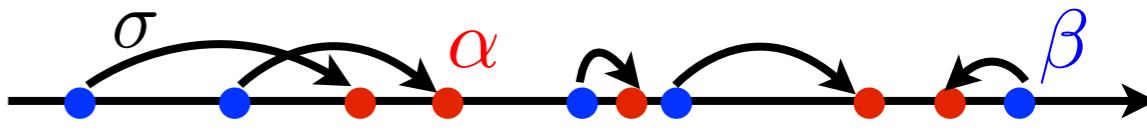
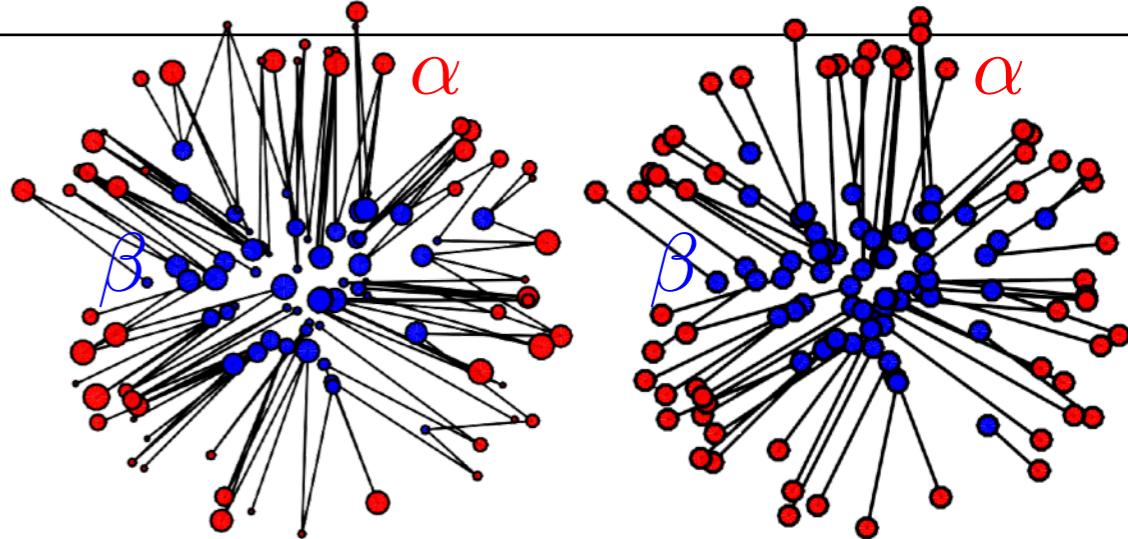
Hungarian/Auction: $O(n^3)$

$$\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \beta = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$$

1-D case: sorting $O(n \log(n))$.

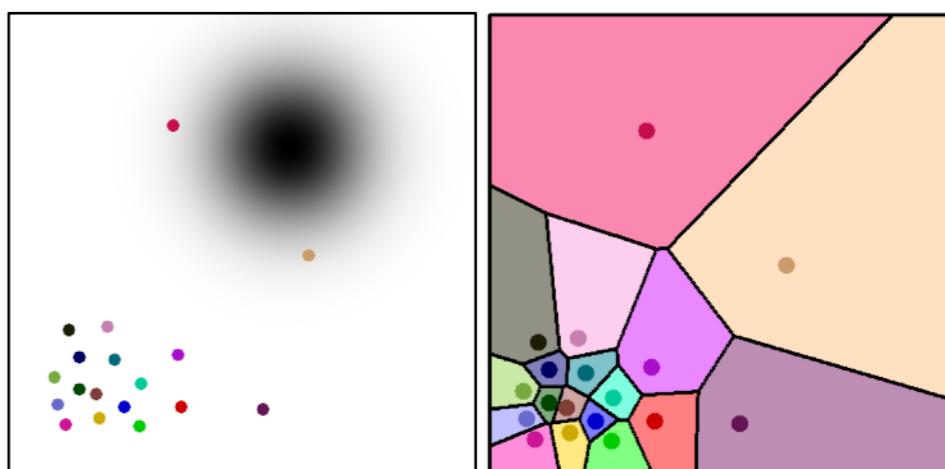
$$\begin{array}{ll} p = 1 \\ d = \|\cdot\| \end{array} \quad W_1(\alpha, \beta) = \min_{\text{div}(u) = \alpha - \beta} \int \|u(x)\| dx$$

\rightarrow min-cost flow, on graphs $O(n^2 \log(n))$.



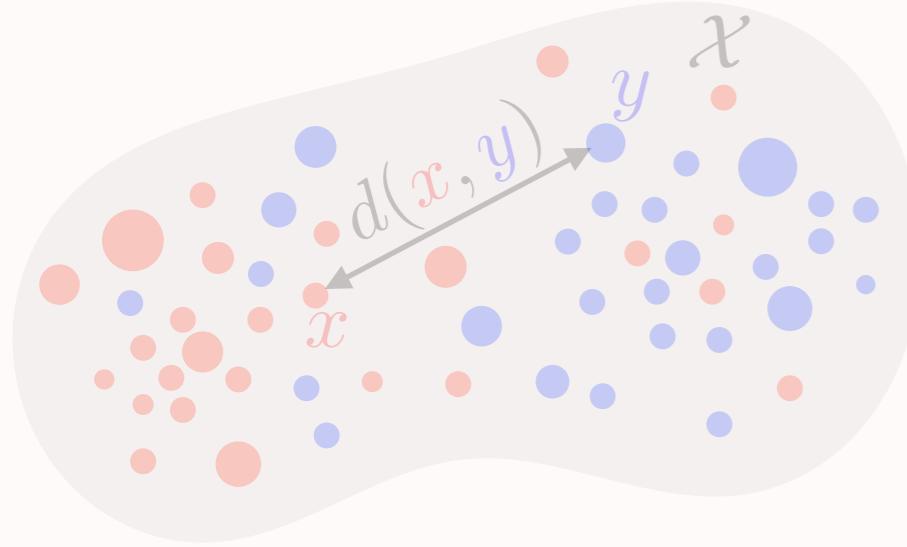
Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.

Semi-discrete: Laguerre cells, $d = \|\cdot\|_2^2$.
[Merigot 2013]

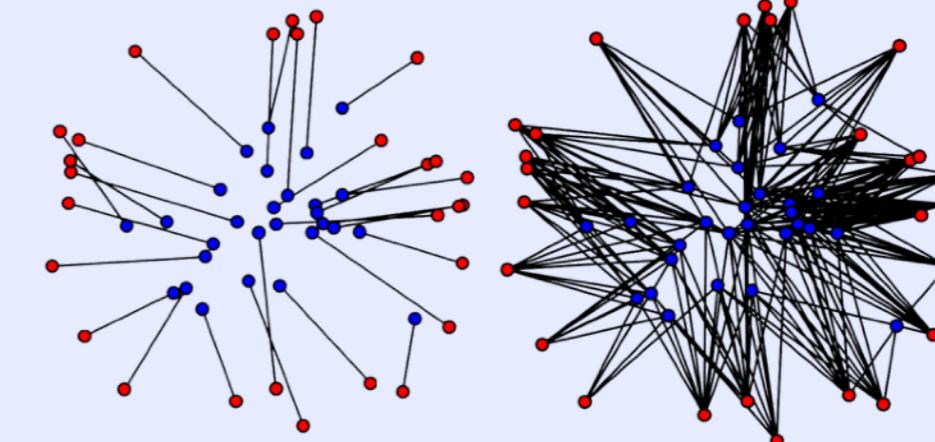


Need for fast approximate algorithms for generic $d(\mathbf{x}, \mathbf{y})$.

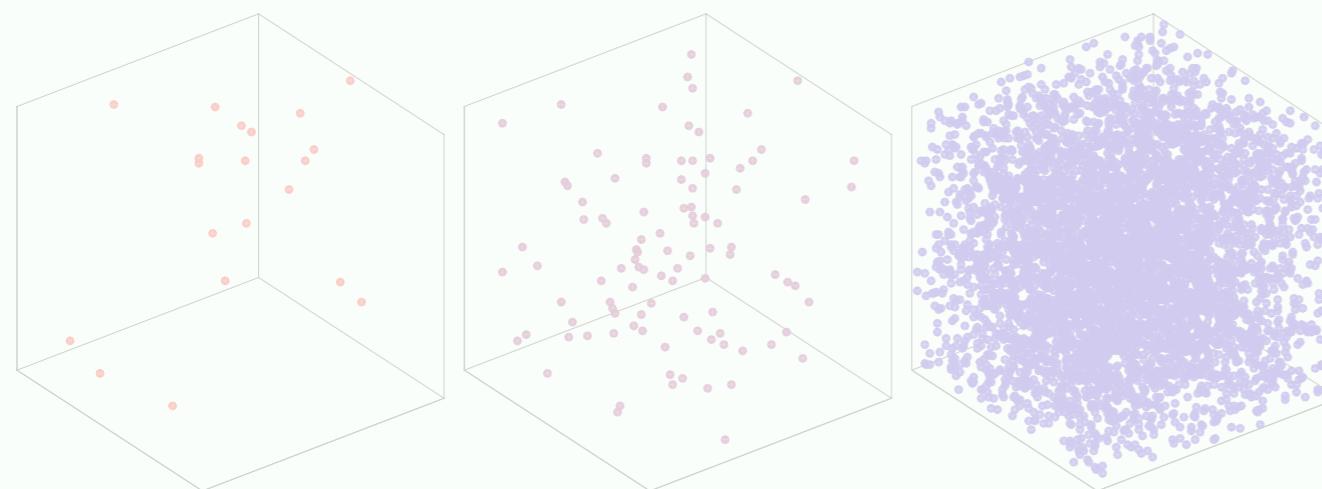
1. Optimal Transport



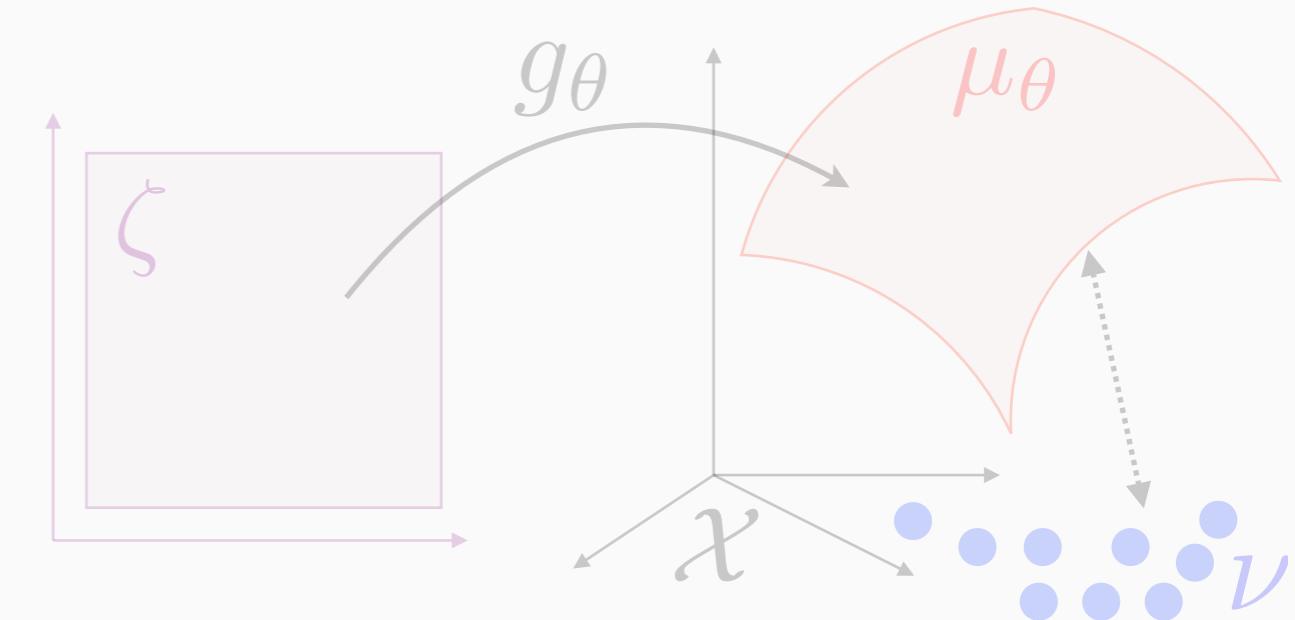
2. Entropic Regularization



3. Sinkhorn Divergences



4. Application to Generative Models

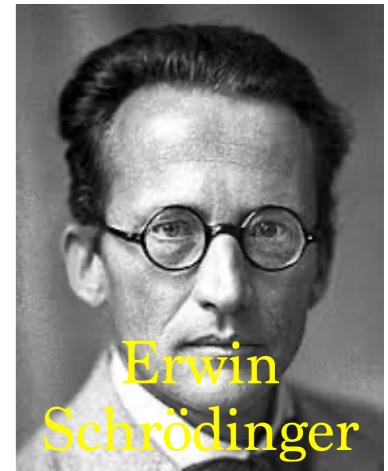


Entropic Regularization

Schrödinger's problem:

[1931]

$$\min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$



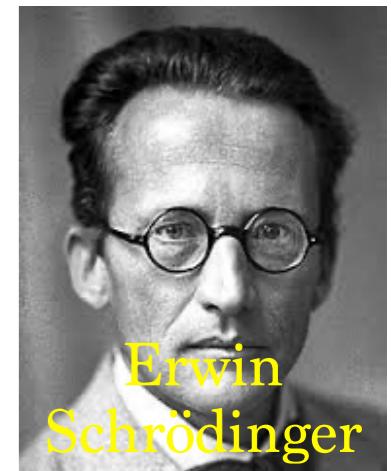
$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$

Entropic Regularization

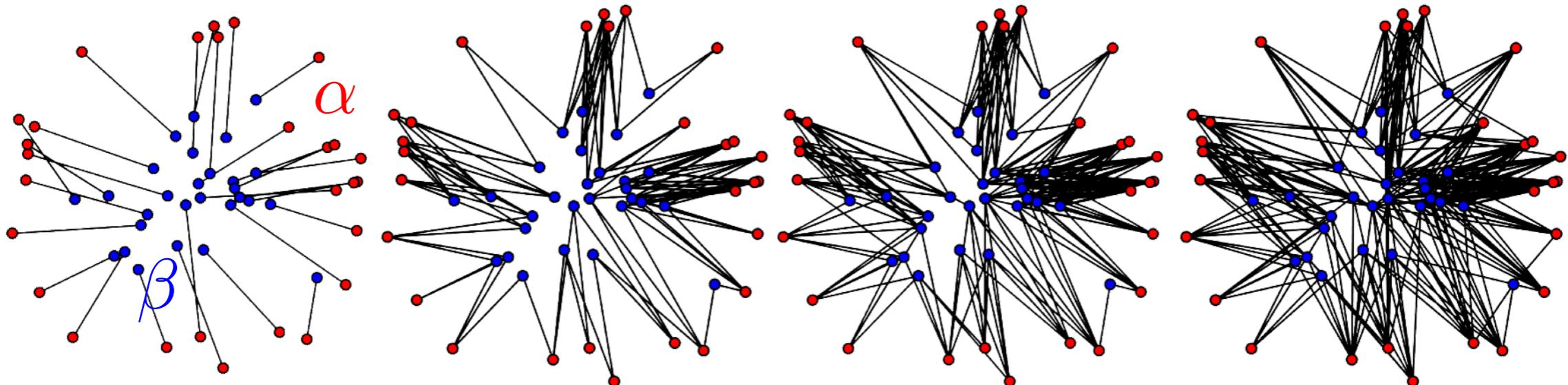
Schrödinger's problem:

[1931]

$$\min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$



$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$

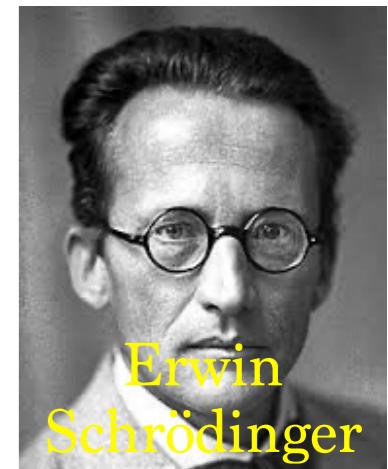


Entropic Regularization

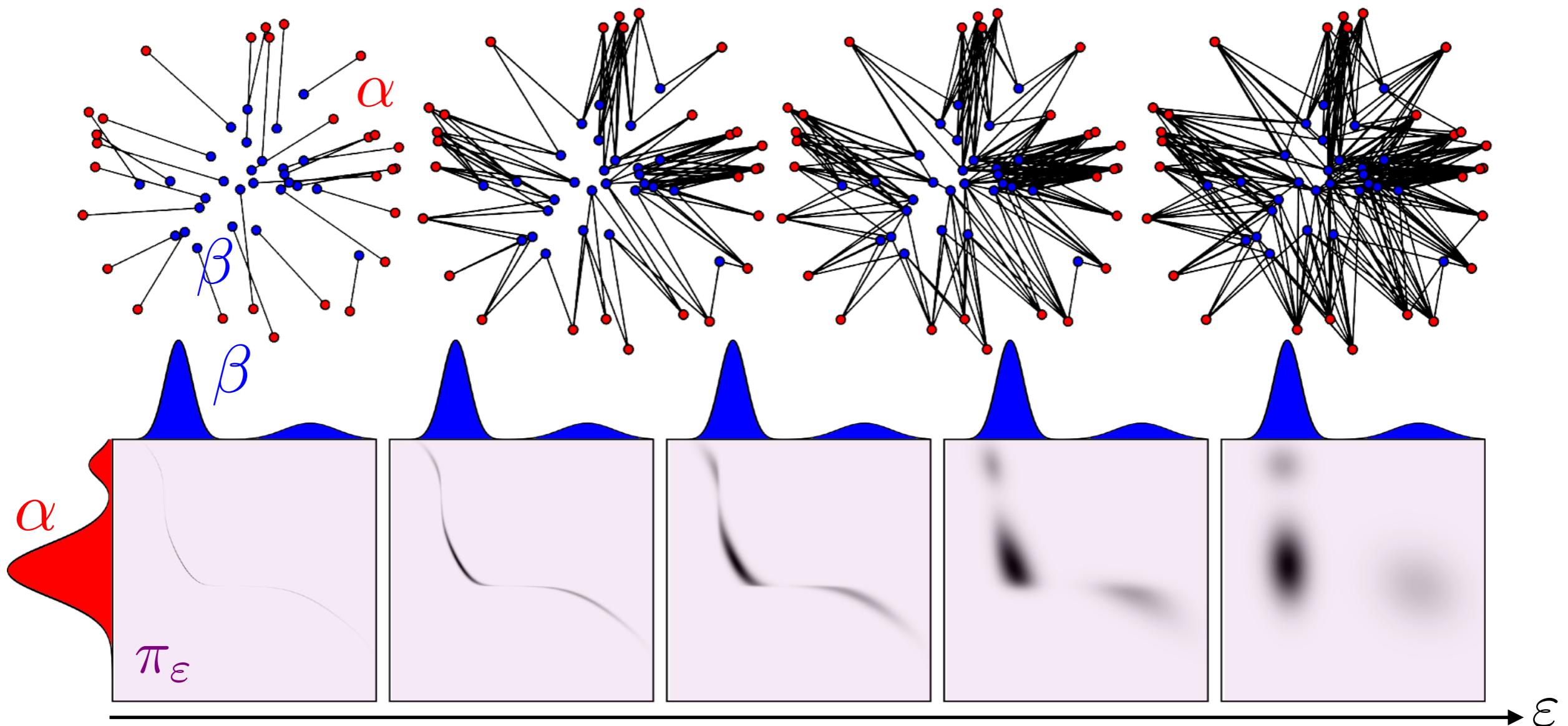
Schrödinger's problem:

[1931]

$$\min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$



$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$



Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) ; \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \right\}$$

Proposition: \mathbf{P} solution $\Leftrightarrow \begin{cases} \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j \end{cases} \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(\mathbf{x}_i, \mathbf{y}_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) ; \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \right\}$$

Proposition: \mathbf{P} solution $\Leftrightarrow \begin{cases} \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j \end{cases} \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(\mathbf{x}_i, \mathbf{y}_j)^p}{\varepsilon}}$

$$\mathbf{P} = \text{diag}(\mathbf{u}) \mathbf{K} \text{diag}(\mathbf{v}) \implies \mathbf{a} = \mathbf{P}\mathbf{1} = \text{diag}(\mathbf{u})(\mathbf{K}\mathbf{v}) = \mathbf{u} \odot (\mathbf{K}\mathbf{v})$$

Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) ; \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \right\}$$

Proposition: \mathbf{P} solution $\Leftrightarrow \begin{cases} \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j \end{cases} \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

$$\mathbf{P} = \text{diag}(\mathbf{u}) \mathbf{K} \text{diag}(\mathbf{v}) \implies \mathbf{a} = \mathbf{P}\mathbf{1} = \text{diag}(\mathbf{u})(\mathbf{K}\mathbf{v}) = \mathbf{u} \odot (\mathbf{K}\mathbf{v})$$

Row constraint: $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$

Col. constraint: $\mathbf{v} \odot (\mathbf{K}^\top \mathbf{u}) = \mathbf{b}$

Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(\mathbf{x}_i, \mathbf{y}_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) ; \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \right\}$$

Proposition: \mathbf{P} solution $\Leftrightarrow \begin{cases} \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j \end{cases} \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(\mathbf{x}_i, \mathbf{y}_j)^p}{\varepsilon}}$

$$\mathbf{P} = \text{diag}(\mathbf{u}) \mathbf{K} \text{diag}(\mathbf{v}) \implies \mathbf{a} = \mathbf{P}\mathbf{1} = \text{diag}(\mathbf{u})(\mathbf{K}\mathbf{v}) = \mathbf{u} \odot (\mathbf{K}\mathbf{v})$$

Row constraint: $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$

Col. constraint: $\mathbf{v} \odot (\mathbf{K}^\top \mathbf{u}) = \mathbf{b}$

Sinkhorn iterations:

$$\mathbf{u} \leftarrow \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$$

$$\mathbf{v} \leftarrow \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}$$

Theorem: [Sinkhorn 1964] (\mathbf{u}, \mathbf{v}) converges.

Sinkhorn's Algorithm

$$\min_{\mathbf{P}} \left\{ \sum_{i,j} d(\mathbf{x}_i, \mathbf{y}_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j}) ; \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \right\}$$

Proposition: \mathbf{P} solution $\Leftrightarrow \begin{cases} \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \text{ and} \\ \exists \mathbf{u}, \mathbf{v}, \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j \end{cases} \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(\mathbf{x}_i, \mathbf{y}_j)^p}{\varepsilon}}$

$$\mathbf{P} = \text{diag}(\mathbf{u}) \mathbf{K} \text{diag}(\mathbf{v}) \implies \mathbf{a} = \mathbf{P}\mathbf{1} = \text{diag}(\mathbf{u})(\mathbf{K}\mathbf{v}) = \mathbf{u} \odot (\mathbf{K}\mathbf{v})$$

Row constraint: $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$

Col. constraint: $\mathbf{v} \odot (\mathbf{K}^\top \mathbf{u}) = \mathbf{b}$

Sinkhorn iterations:

$$\mathbf{u} \leftarrow \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$$

$$\mathbf{v} \leftarrow \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}$$

Theorem: [Sinkhorn 1964] (\mathbf{u}, \mathbf{v}) converges.

Matrix/vector multiplications: $\rightarrow O(n^2/\varepsilon^2)$ complexity.

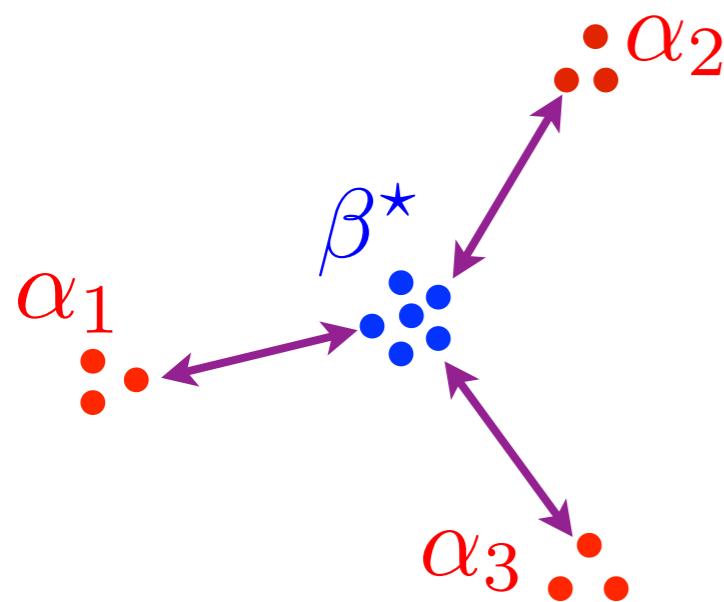
\rightarrow Parallelizable on GPUs.

\rightarrow Convolution on regular grids, separable kernels.

Wasserstein Barycenters

Barycenters of measures $(\alpha_s)_s$: $\sum_s \lambda_s = 1$

$$\beta^* \in \operatorname{argmin}_{\beta} \sum_s \lambda_s W_p^p(\alpha_s, \beta)$$

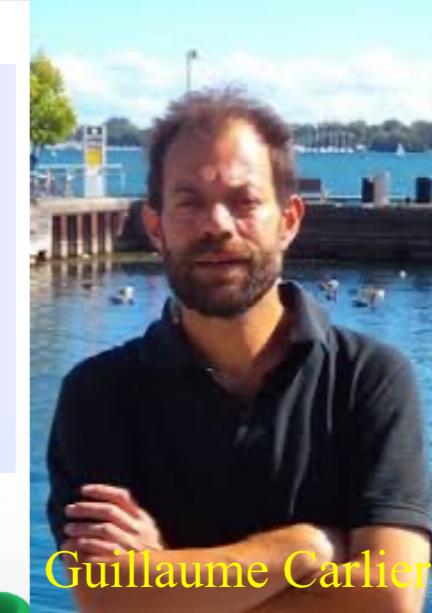
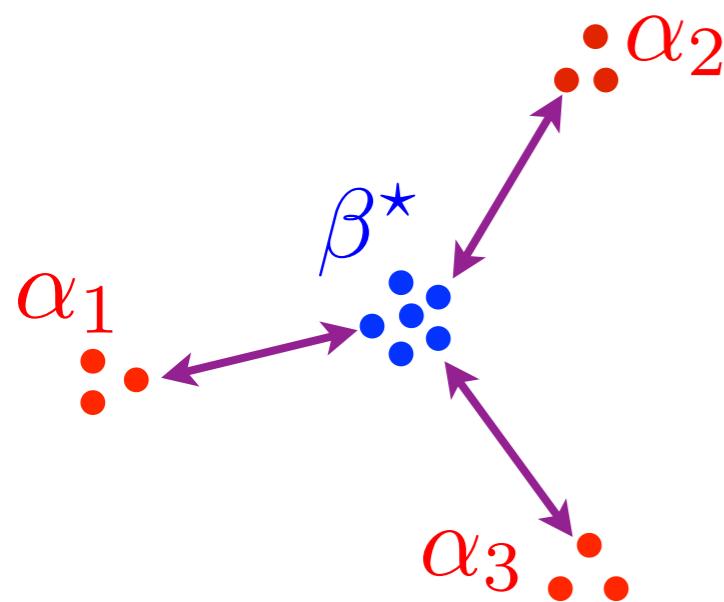


[Solomon et al, SIGGRAPH 2015]

Wasserstein Barycenters

Barycenters of measures $(\alpha_s)_s$: $\sum_s \lambda_s = 1$

$$\beta^* \in \operatorname{argmin}_{\beta} \sum_s \lambda_s W_p^p(\alpha_s, \beta)$$

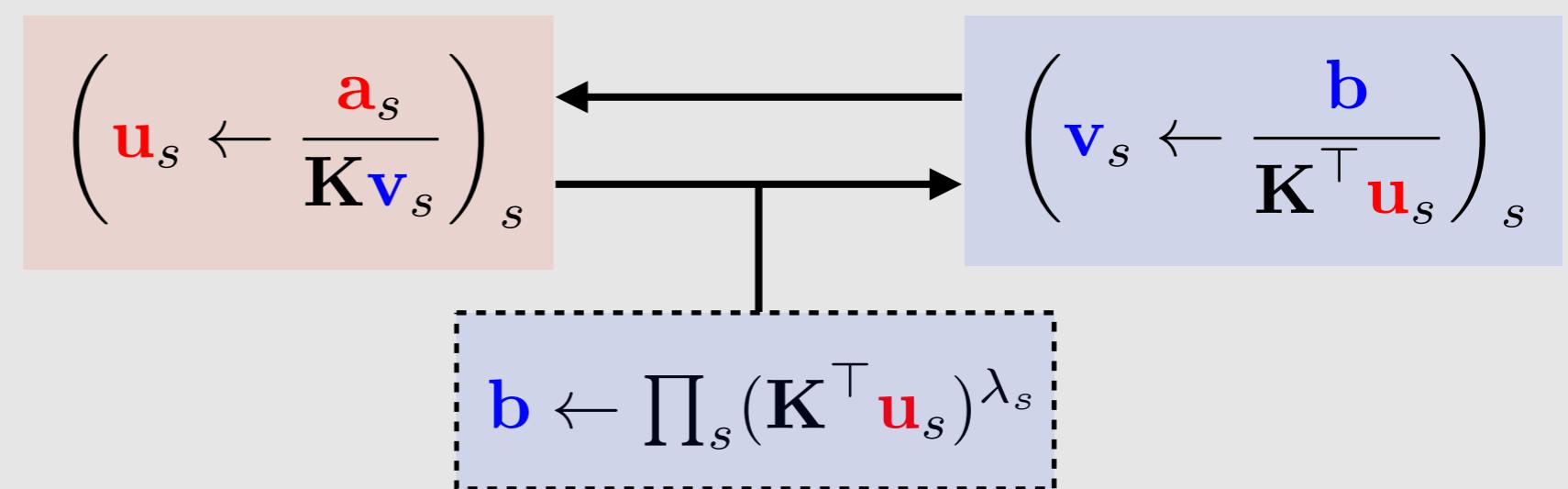


Guillaume Carlier

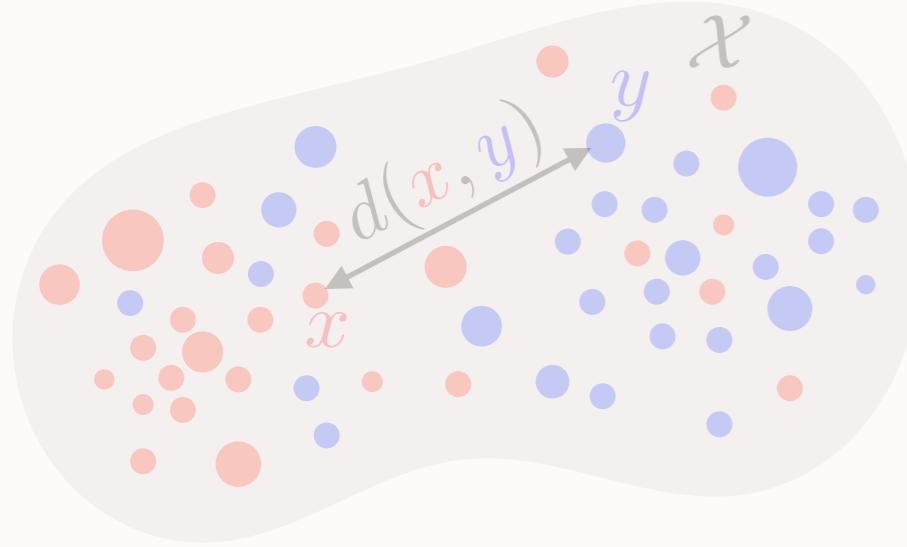


Martial Aguech

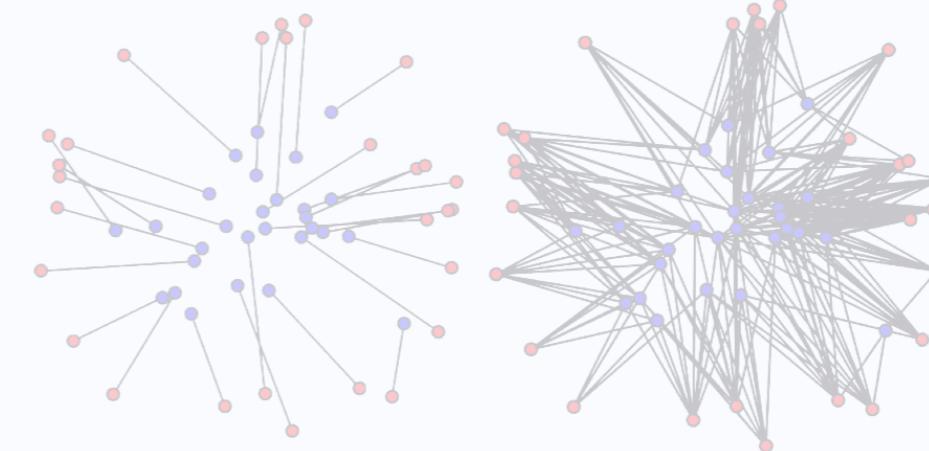
Sinkhorn's algorithm:



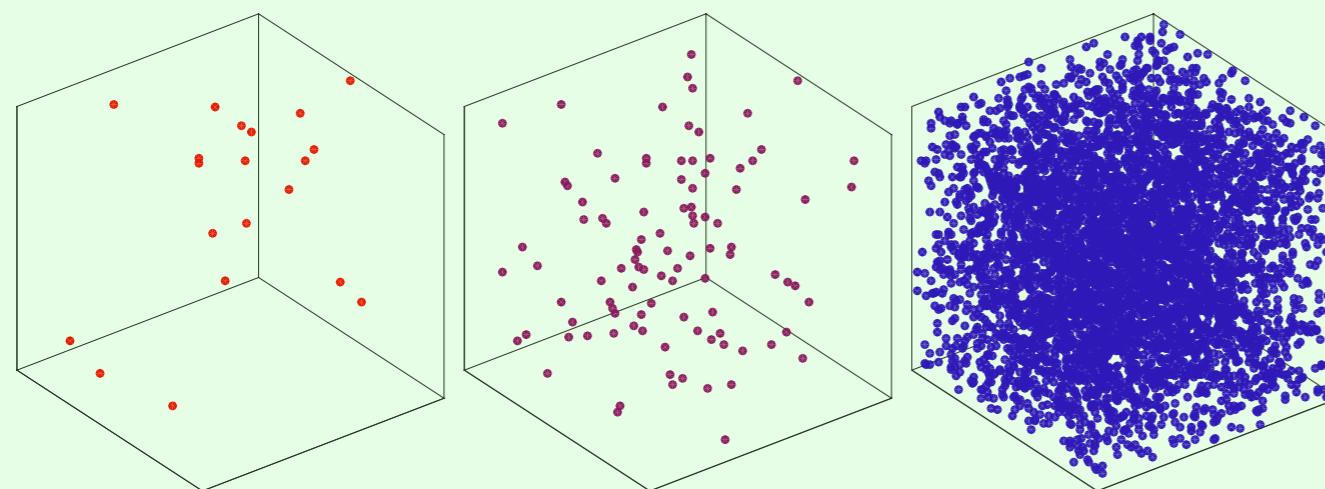
1. Optimal Transport



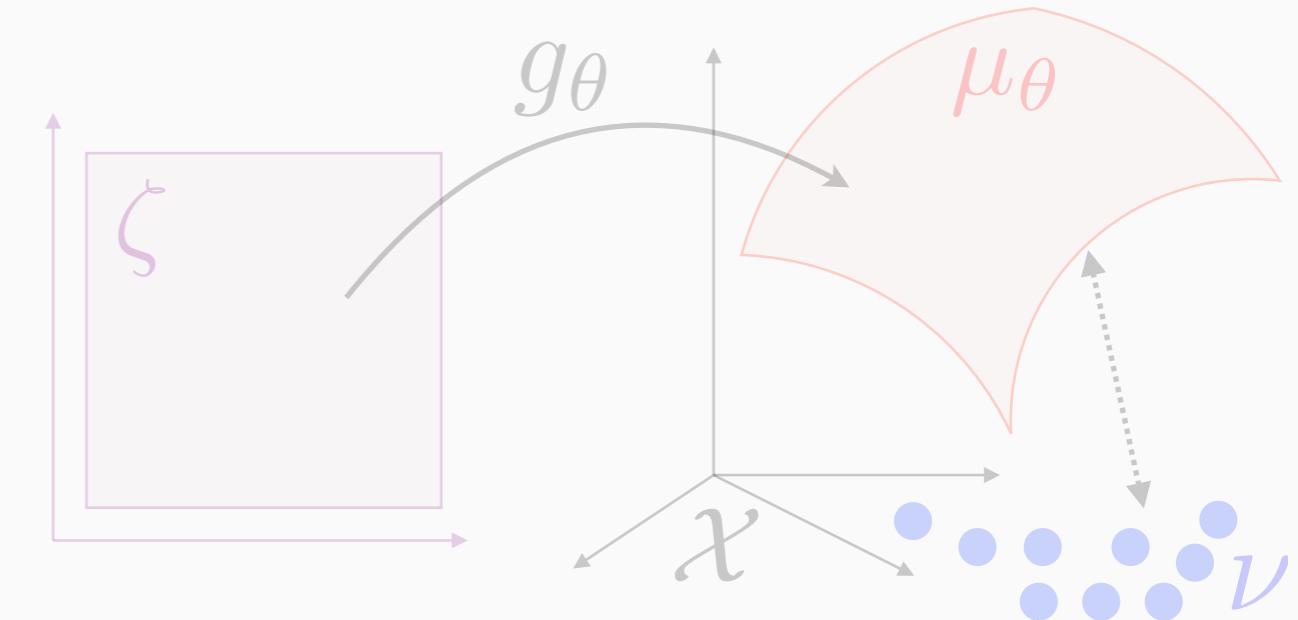
2. Entropic Regularization



3. Sinkhorn Divergences



4. Application to Generative Models

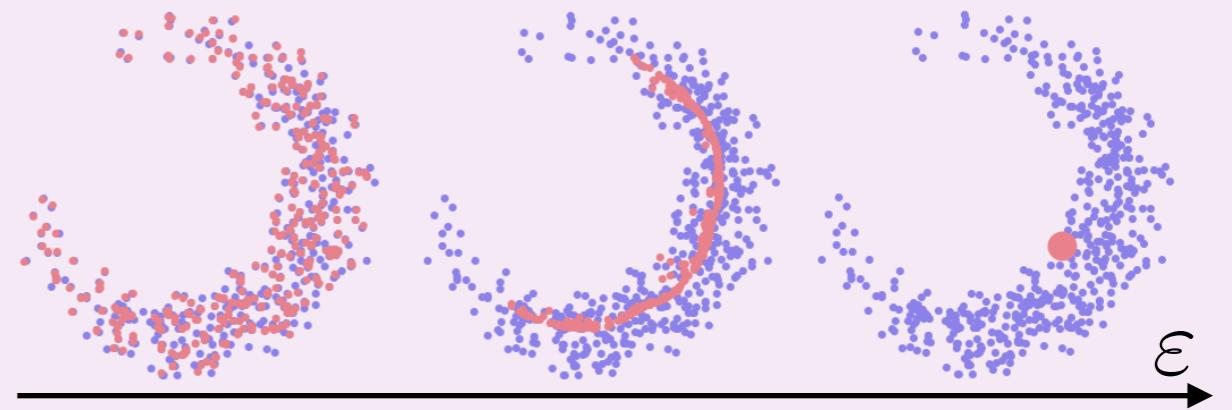


Sinkhorn Divergences

$$W_{\varepsilon,p}^p(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$

Problem: $W_\varepsilon(\alpha, \alpha) \neq 0$

$$\min_{\alpha} W_{\varepsilon,p}^p(\alpha, \beta)$$

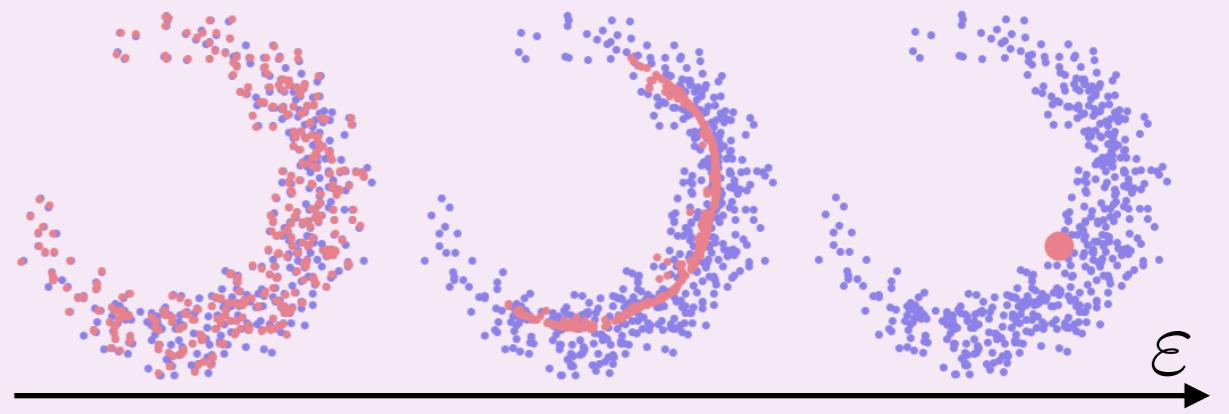


Sinkhorn Divergences

$$W_{\varepsilon,p}^p(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$

Problem: $W_\varepsilon(\alpha, \alpha) \neq 0$

$$\min_{\alpha} W_{\varepsilon,p}^p(\alpha, \beta)$$



$$\overline{W}_{p,\varepsilon}^p(\alpha, \beta) \stackrel{\text{def.}}{=} W_{p,\varepsilon}^p(\alpha, \beta) - \frac{1}{2} W_{p,\varepsilon}^p(\alpha, \alpha) - \frac{1}{2} W_{p,\varepsilon}^p(\beta, \beta)$$

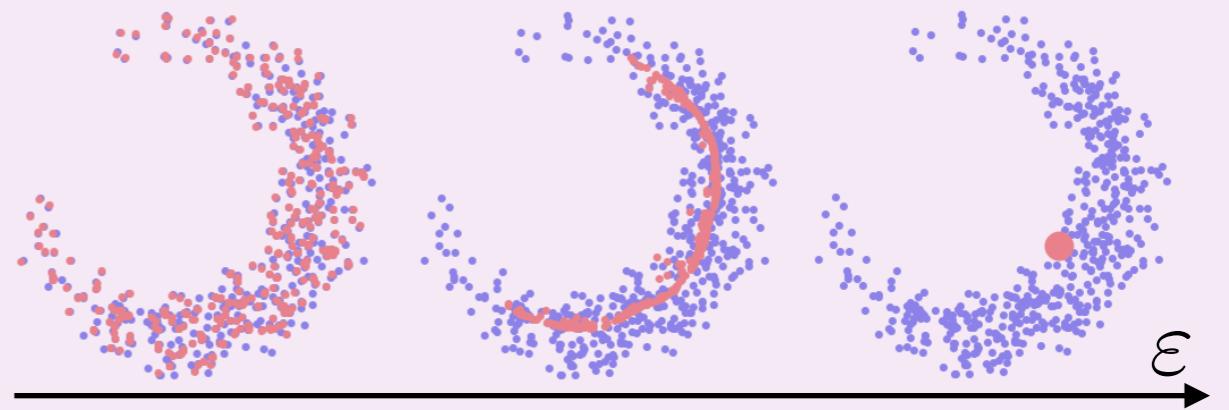
[Ramdas, García Trillos, Cuturi, 2017]

Sinkhorn Divergences

$$W_{\varepsilon,p}^p(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})$$

Problem: $W_\varepsilon(\alpha, \alpha) \neq 0$

$$\min_{\alpha} W_{\varepsilon,p}^p(\alpha, \beta)$$



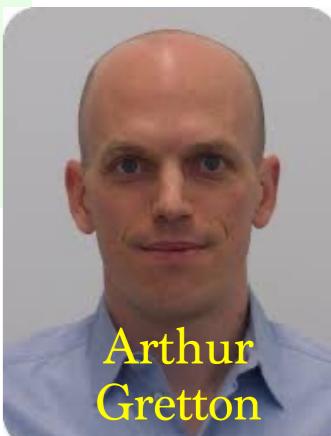
$$\overline{W}_{p,\varepsilon}^p(\alpha, \beta) \stackrel{\text{def.}}{=} W_{p,\varepsilon}^p(\alpha, \beta) - \frac{1}{2} W_{p,\varepsilon}^p(\alpha, \alpha) - \frac{1}{2} W_{p,\varepsilon}^p(\beta, \beta)$$

[Ramdas, García Trillos, Cuturi, 2017]

$$\text{Theorem: } W_p^p(\alpha, \beta) \xleftarrow[\substack{\text{[Léonard 2012]} \\ \text{[Carlier et al 2017]}}]{\varepsilon \rightarrow 0} \overline{W}_{\varepsilon,p}^p(\alpha, \beta) \xrightarrow{\substack{\text{[Ramdas, García Trillos,} \\ \text{Cuturi, 2017]}}} \|\alpha - \beta\|_{-d^p}^2$$

Kernel norms (MMD): $\|\xi\|_{-d^p}^2 \stackrel{\text{def.}}{=} - \int_{\mathcal{X}^2} d(x, y)^p d\xi(x) d\xi(y)$

Proposition: $\|\cdot\|_{-\|\cdot\|^p}$ is a norm for $0 < p < 2$.



Sinkhorn Divergences Positivity

$$\overline{W}_{p,\varepsilon}^p(\alpha, \beta) \stackrel{\text{def.}}{=} W_{p,\varepsilon}^p(\alpha, \beta) - \frac{1}{2} W_{p,\varepsilon}^p(\alpha, \alpha) - \frac{1}{2} W_{p,\varepsilon}^p(\beta, \beta)$$

↓ concave ↓ concave

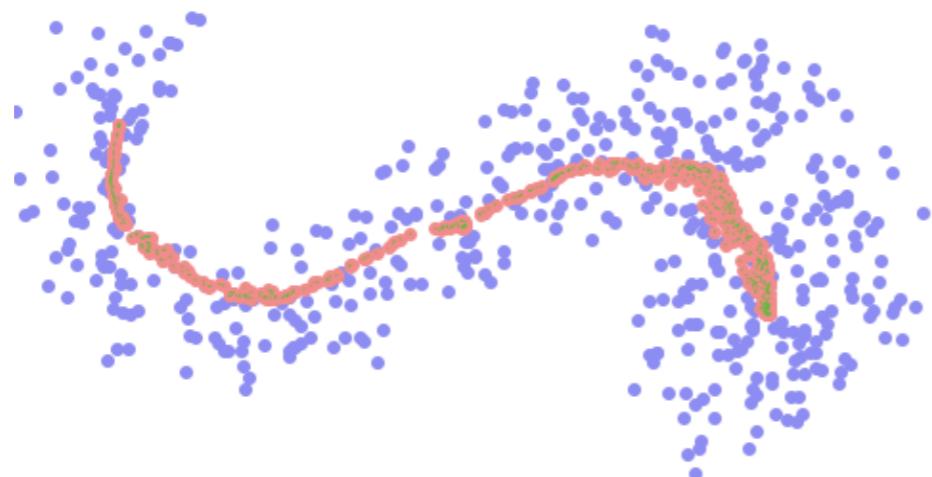
Theorem: [Feydy, Séjourné, P, Vialard, Trouvé, Amari 2018]

If $e^{-\frac{d^p}{\varepsilon}}$ is positive:

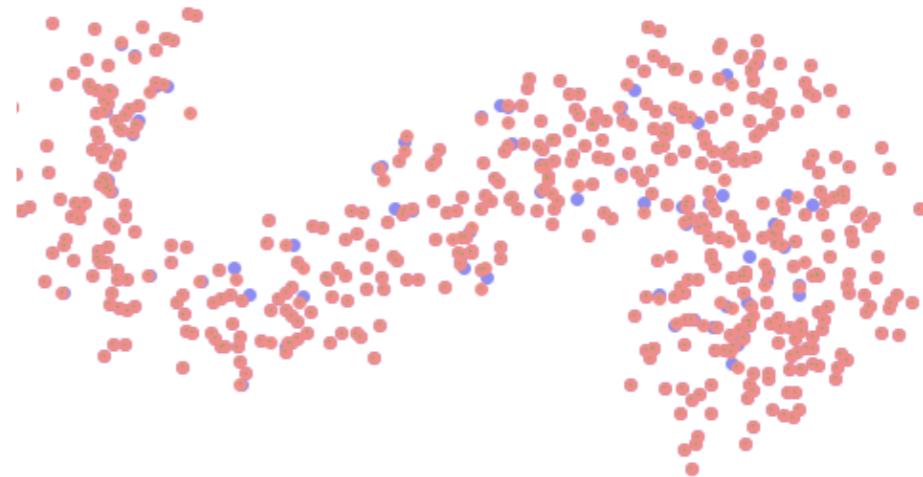
$\overline{W}_{\varepsilon,p} \geqslant 0$ and $\overline{W}_{\varepsilon,p}^p(\cdot, \beta)$ is convex.

$\overline{W}_{\varepsilon,p}(\alpha_n, \beta) \rightarrow 0 \iff \alpha_n \xrightarrow{\text{weak*}} \beta$

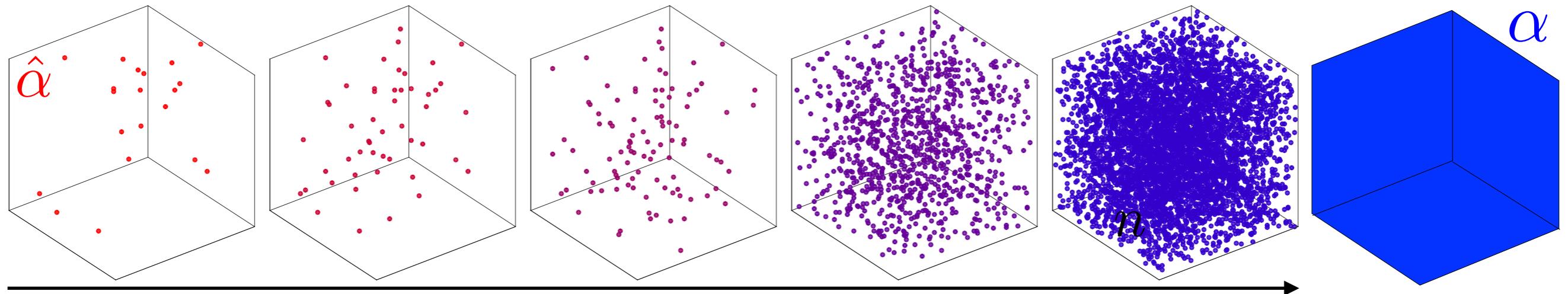
$$\min_{\alpha} W_{\varepsilon,p}^p(\alpha, \beta)$$



$$\min_{\alpha} \overline{W}_{\varepsilon,p}^p(\alpha, \beta)$$

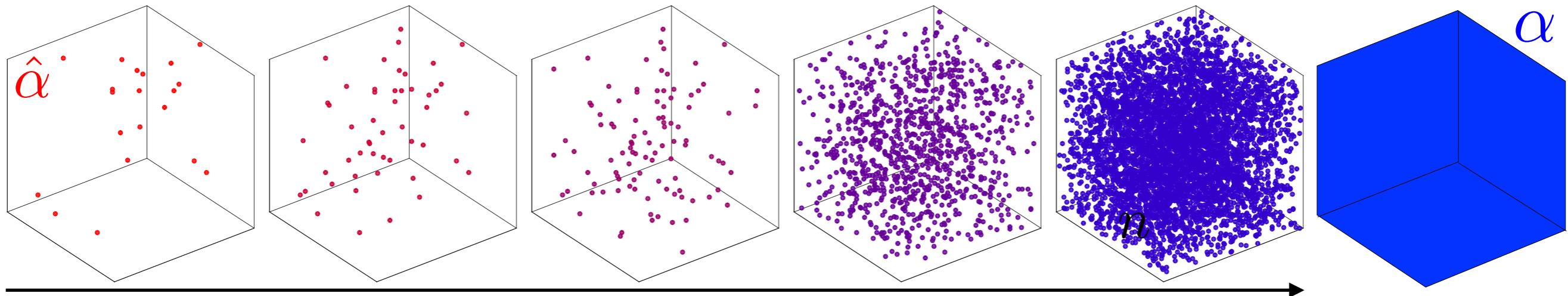


Sample Complexity



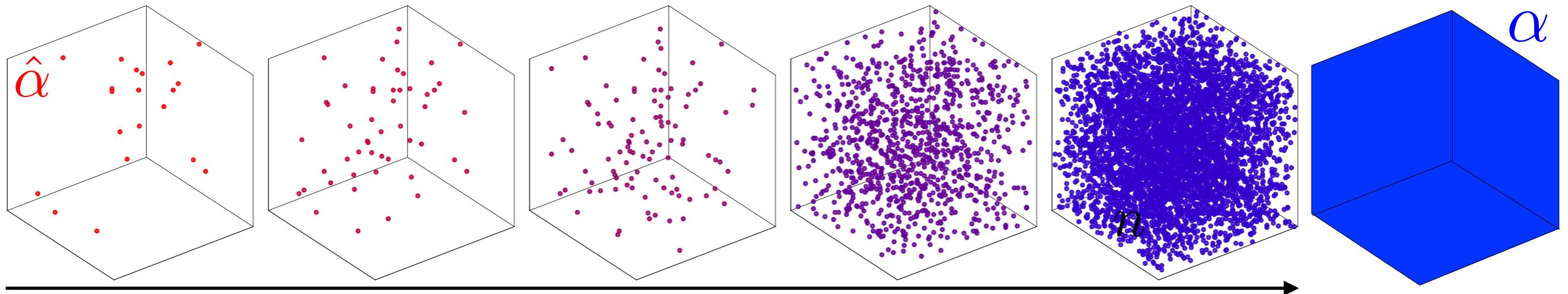
Theorem: $\mathbb{E}(|W_p(\hat{\alpha}, \hat{\beta}) - W_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$

Sample Complexity



Theorem: $\mathbb{E}(|W_p(\hat{\alpha}, \hat{\beta}) - W_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$ ↗ if $\alpha \neq \beta$
 $\mathbb{E}(|W_2(\hat{\alpha}, \hat{\beta}) - W_2(\alpha, \beta)|) = O(n^{-\frac{2}{d}})$ ↗
[Chizat, Roussillon, Léger, Vialard, P. 2020]

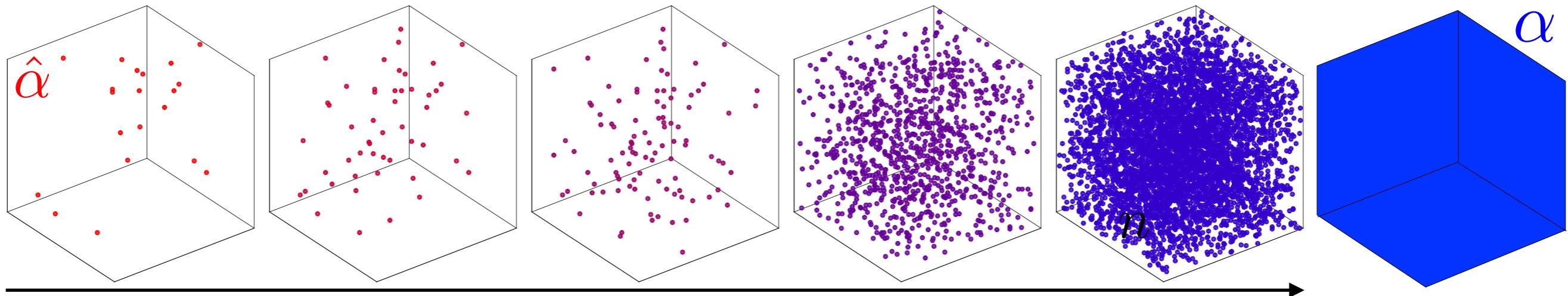
Sample Complexity



Theorem: $\mathbb{E}(|W_p(\hat{\alpha}, \hat{\beta}) - W_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$ ↗ if $\alpha \neq \beta$
 $\mathbb{E}(|W_2(\hat{\alpha}, \hat{\beta}) - W_2(\alpha, \beta)|) = O(n^{-\frac{2}{d}})$ ↗
[Chizat, Roussillon, Léger, Vialard, P. 2020]

$$\mathbb{E}(|\|\hat{\alpha} - \hat{\beta}\|_k - \|\alpha - \beta\|_k|) = O(n^{-\frac{1}{2}})$$

Sample Complexity



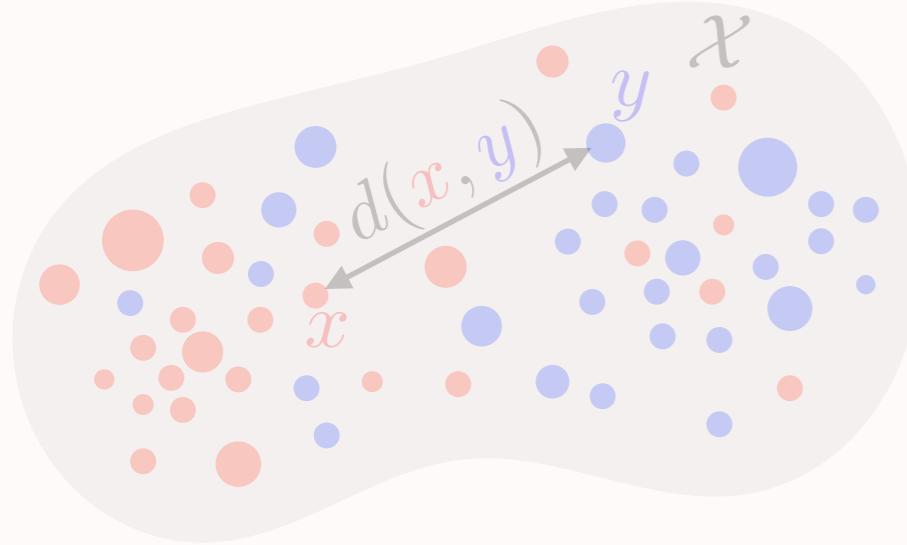
Theorem: $\mathbb{E}(|W_p(\hat{\alpha}, \hat{\beta}) - W_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$ ↗ if $\alpha \neq \beta$
 $\mathbb{E}(|W_2(\hat{\alpha}, \hat{\beta}) - W_2(\alpha, \beta)|) = O(n^{-\frac{2}{d}})$ ↗
[Chizat, Roussillon, Léger, Vialard, P. 2020]

$$\mathbb{E}(|\|\hat{\alpha} - \hat{\beta}\|_k - \|\alpha - \beta\|_k|) = O(n^{-\frac{1}{2}})$$

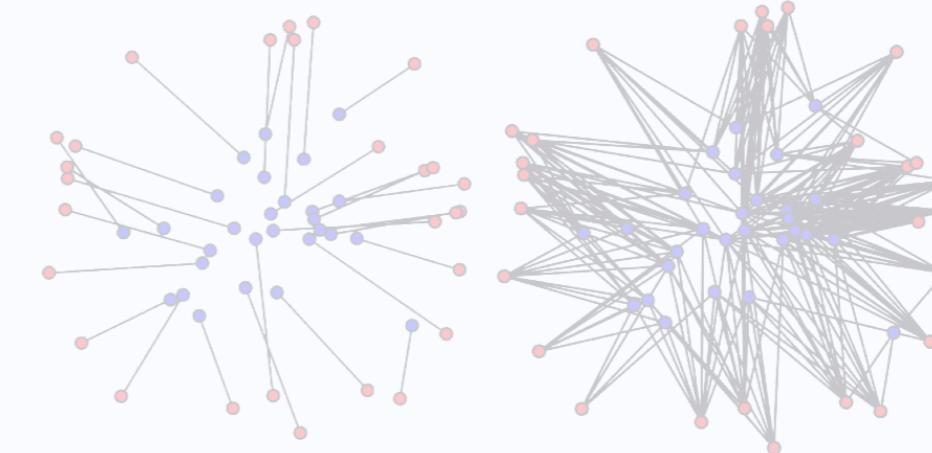
$$\mathbb{E}(|\overline{W}_p^\varepsilon(\hat{\alpha}, \hat{\beta}) - \overline{W}_p^\varepsilon(\alpha, \beta)|) = O(\varepsilon^{-\frac{d}{2}} n^{-\frac{1}{2}})$$

[Genevay, Bach, P, Cuturi, 2019]

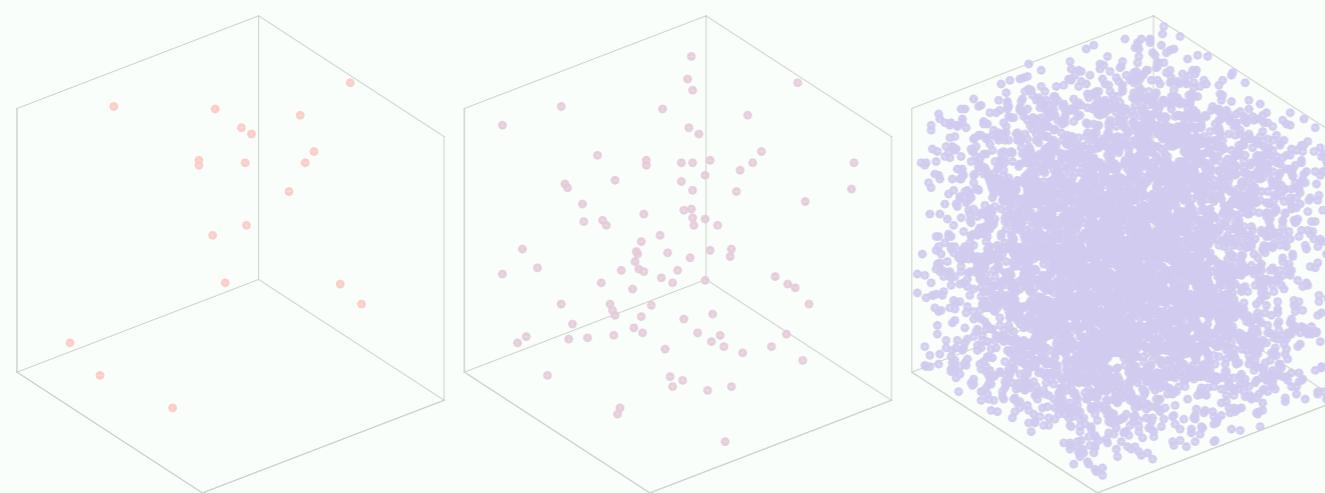
1. Optimal Transport



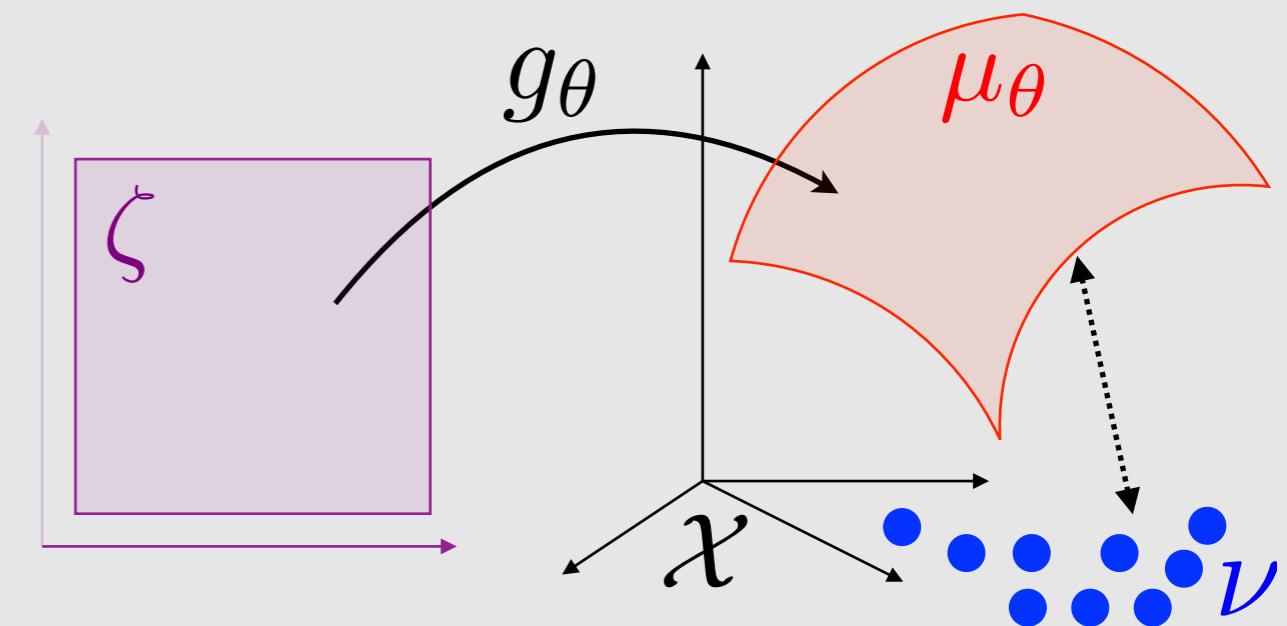
2. Entropic Regularization



3. Sinkhorn Divergences



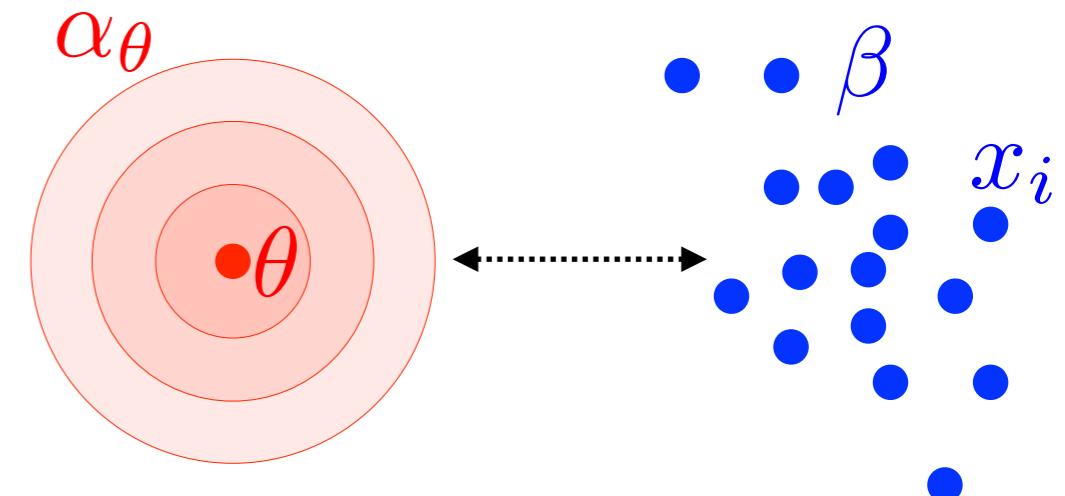
4. Application to Generative Models



Density Fitting and Generative Models

Observations: $\beta \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$

Parametric model: $\theta \mapsto \alpha_\theta$



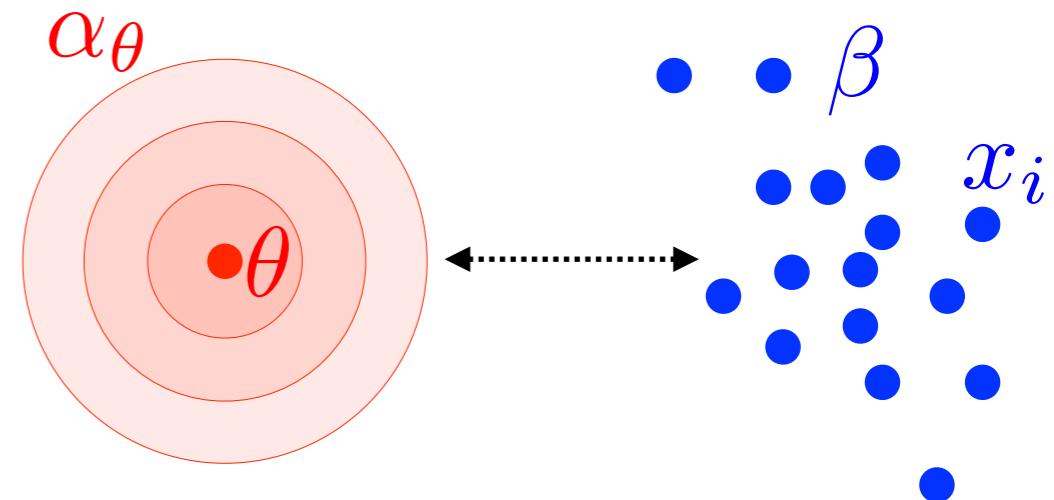
Density Fitting and Generative Models

Observations: $\beta \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$

Parametric model: $\theta \mapsto \alpha_\theta$

Density fitting: $d\alpha_\theta(x) = \rho_\theta(x)dx$

$$\min_{\theta} - \sum_i \log(\rho_\theta(x_i)) \xrightarrow{n \rightarrow +\infty} \text{KL}(\beta | \alpha_\theta)$$



Maximum likelihood (MLE)

Density Fitting and Generative Models

Observations: $\beta \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$

Parametric model: $\theta \mapsto \alpha_\theta$

Density fitting: $d\alpha_\theta(x) = \rho_\theta(x)dx$

$$\min_{\theta} - \sum_i \log(\rho_\theta(x_i)) \xrightarrow{n \rightarrow +\infty} \text{KL}(\beta | \alpha_\theta)$$

Maximum likelihood (MLE)

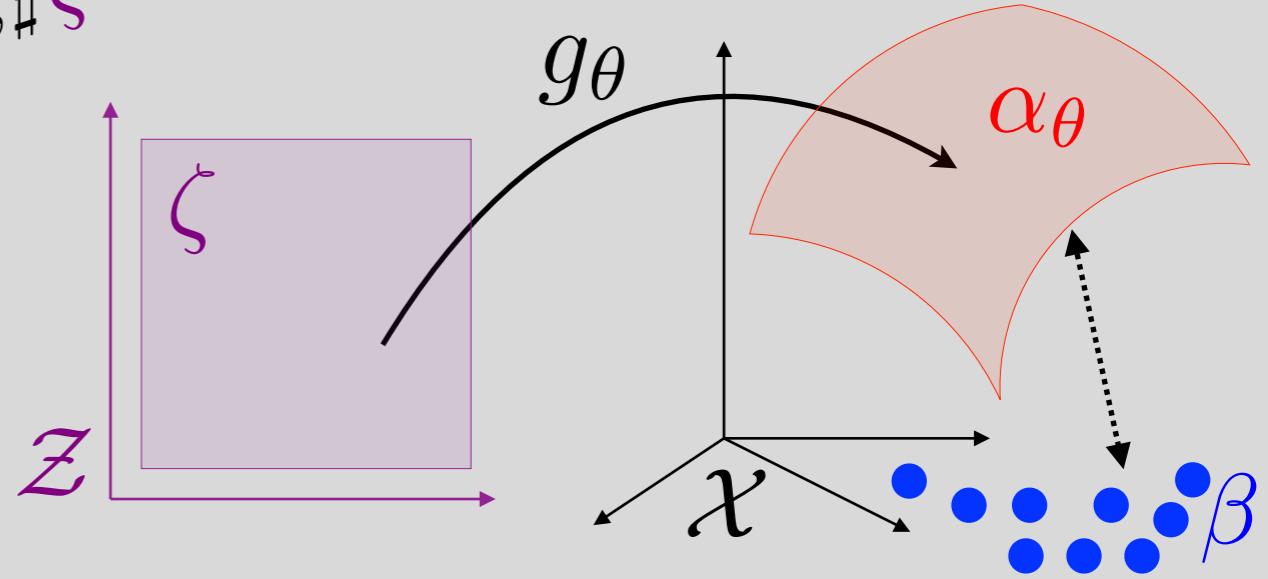
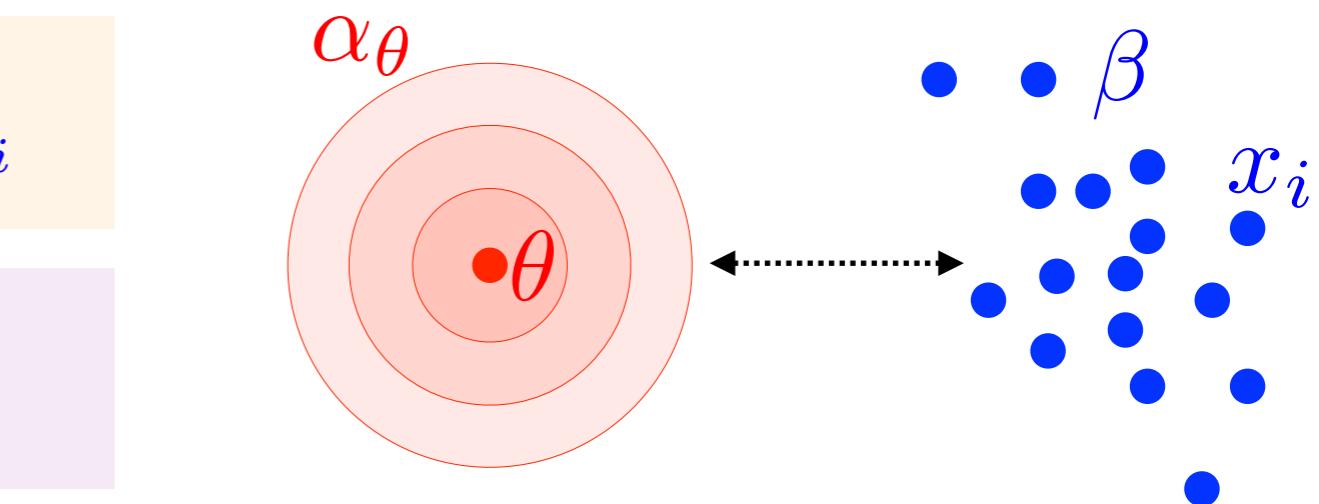
Generative model fit: $\alpha_\theta = g_{\theta, \sharp} \zeta$

$$\text{KL}(\beta | \alpha_\theta) = +\infty$$

→ MLE undefined.

→ Need a weaker metric.

$$\min_{\theta} \overline{W}_{\varepsilon, p}^p(\alpha_\theta, \beta)$$



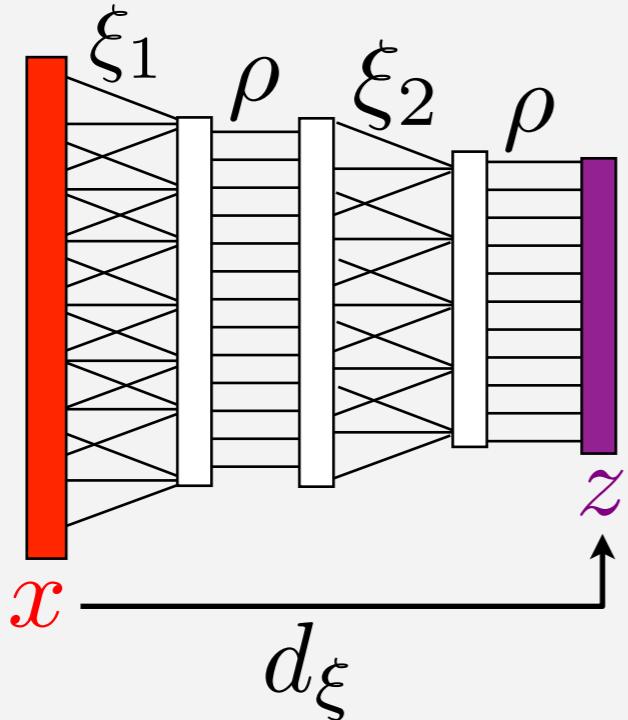
Deep Discriminative vs Generative Models

Deep networks:

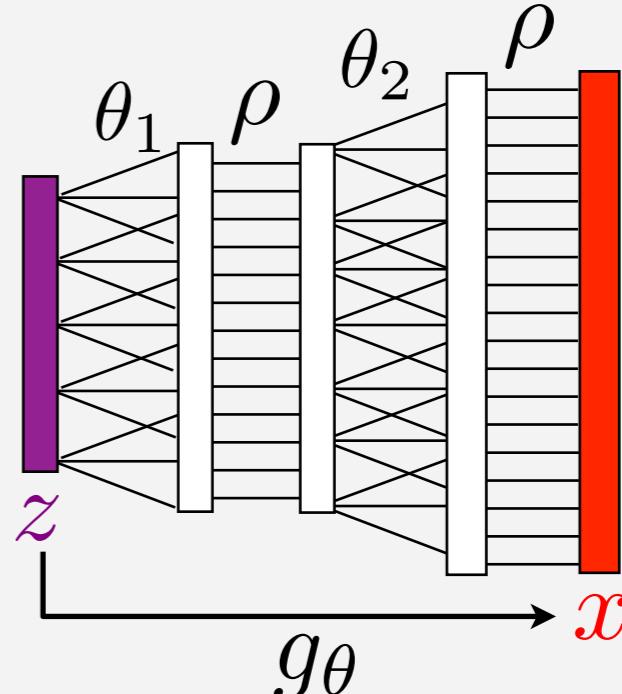
$$d_\xi(\textcolor{red}{x}) = \rho(\xi_K(\dots \rho(\xi_2(\rho(\xi_1(\textcolor{red}{x}) \dots)$$

$$g_\theta(\textcolor{violet}{z}) = \rho(\theta_K(\dots \rho(\theta_2(\rho(\theta_1(\textcolor{violet}{z}) \dots)$$

Discriminative



Generative

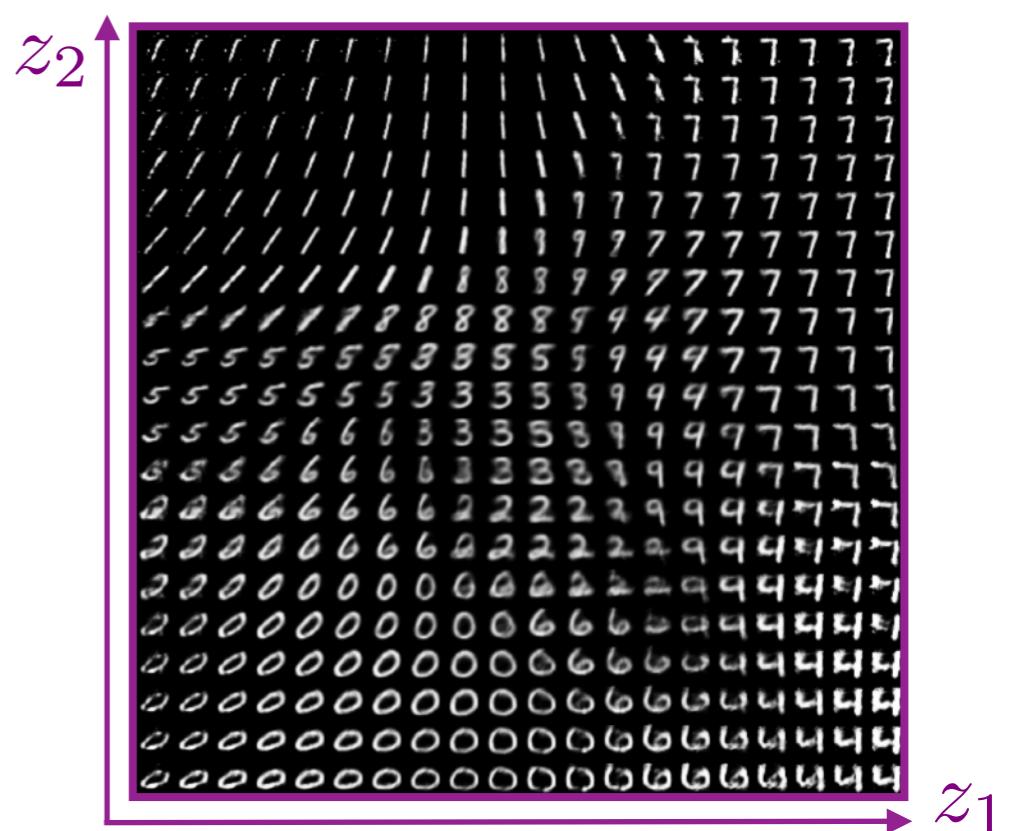
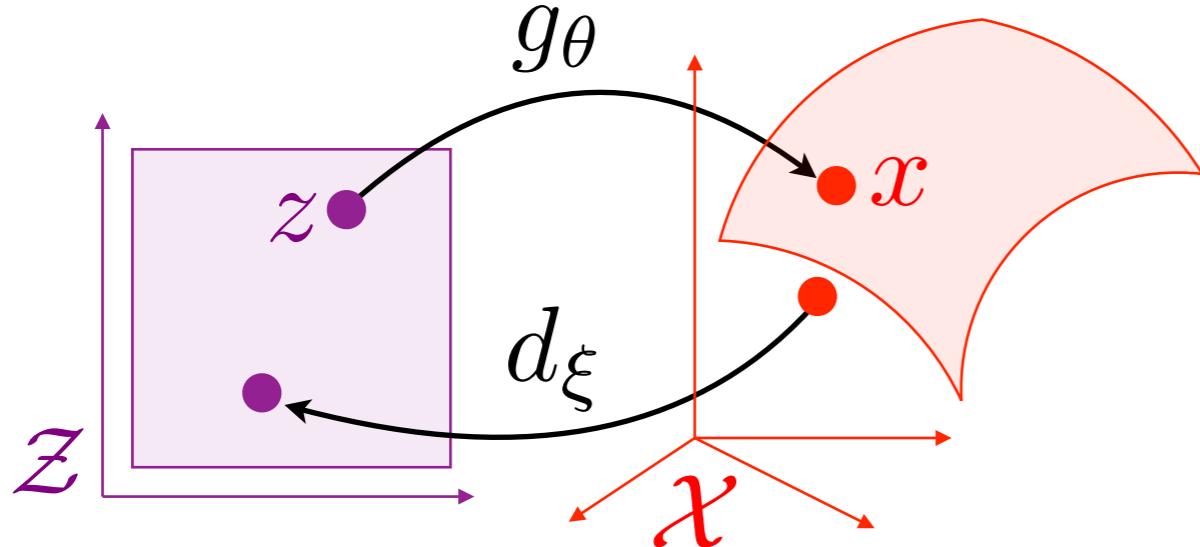
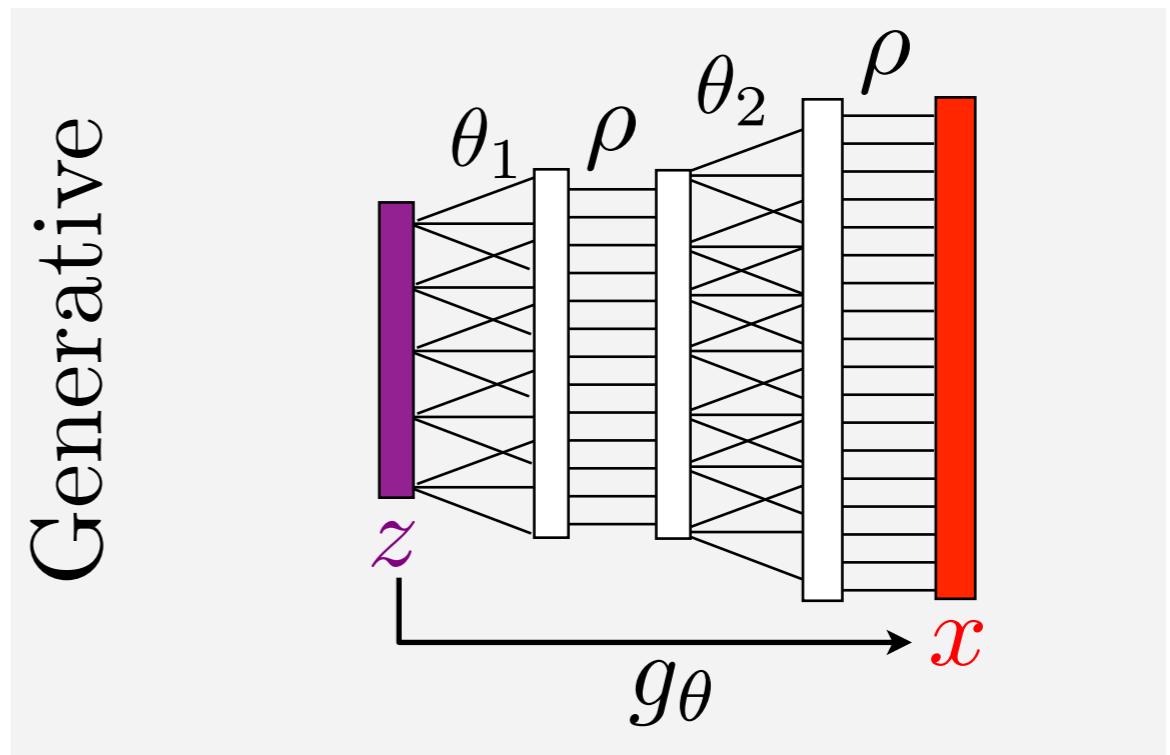
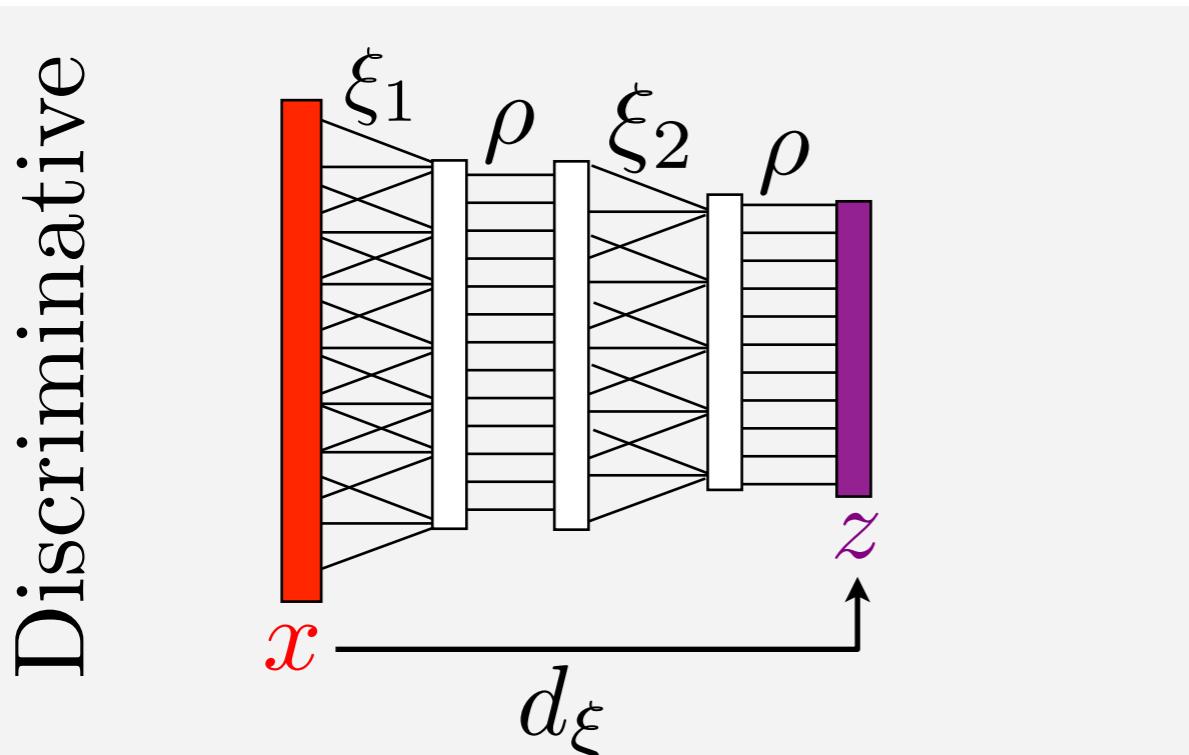


Deep Discriminative vs Generative Models

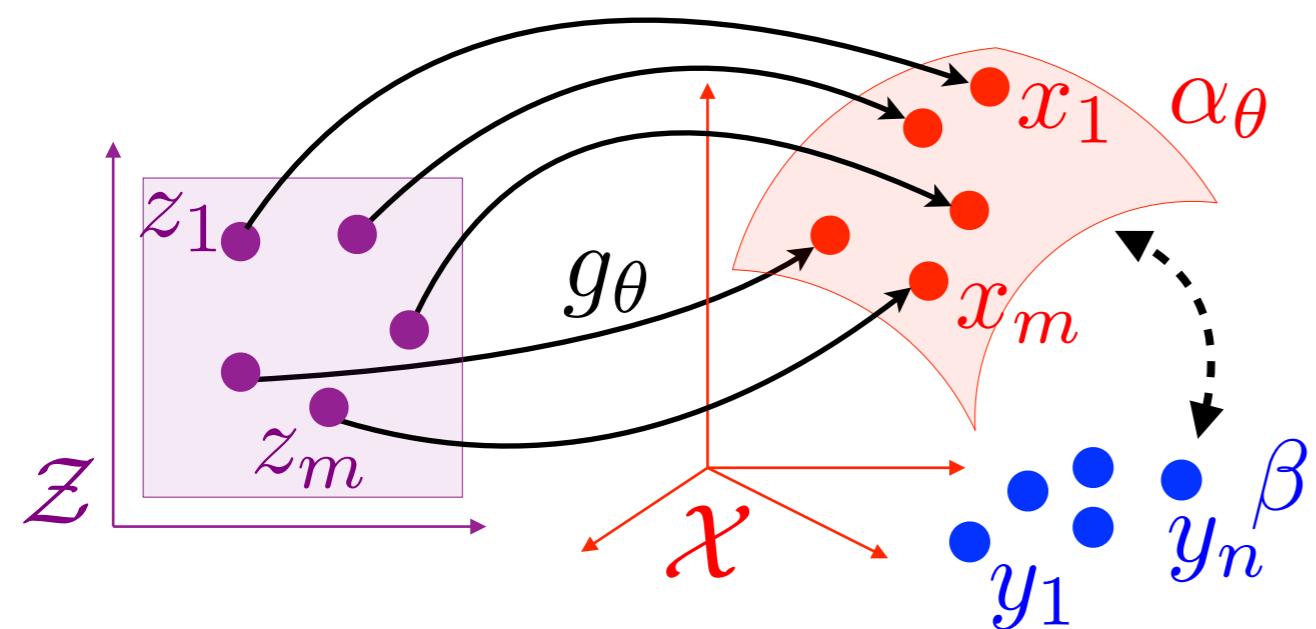
Deep networks: <

$$d_\xi(x) = \rho(\xi_K(\dots \rho(\xi_2(\rho(\xi_1(x) \dots)$$

$$g_{\theta}(z) = \rho(\theta_K(\dots \rho(\theta_2(\rho(\theta_1(z)\dots)$$



Training Architecture



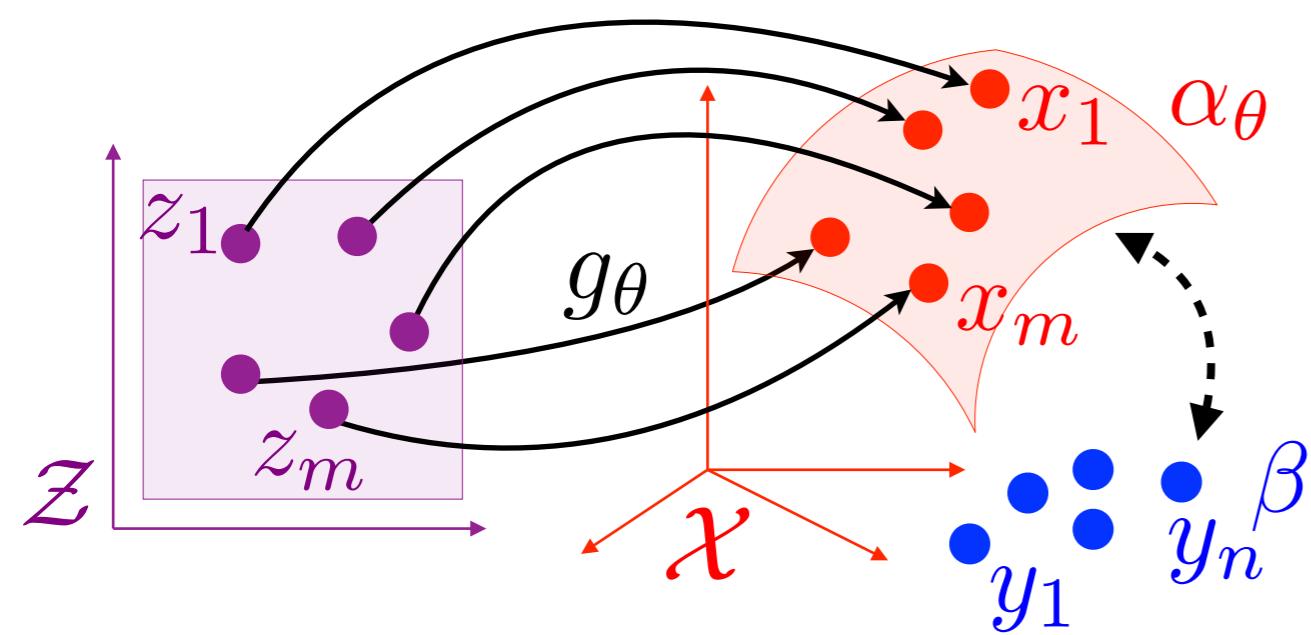
$$\min_{\theta} \mathcal{E}(\theta) \stackrel{\text{def.}}{=} \overline{\mathbf{W}}_{\varepsilon,p}^p(\alpha_\theta, \beta)$$

Stochastic gradient descent

$$\theta \leftarrow \theta - \tau \nabla \hat{\mathcal{E}}(\theta)$$

$$\hat{\mathcal{E}}(\theta) \stackrel{\text{def.}}{=} \overline{\mathbf{W}}_{\varepsilon,p}^p\left(\frac{1}{m} \sum_i \delta_{g_\theta(z_i)}, \beta\right)$$

Training Architecture

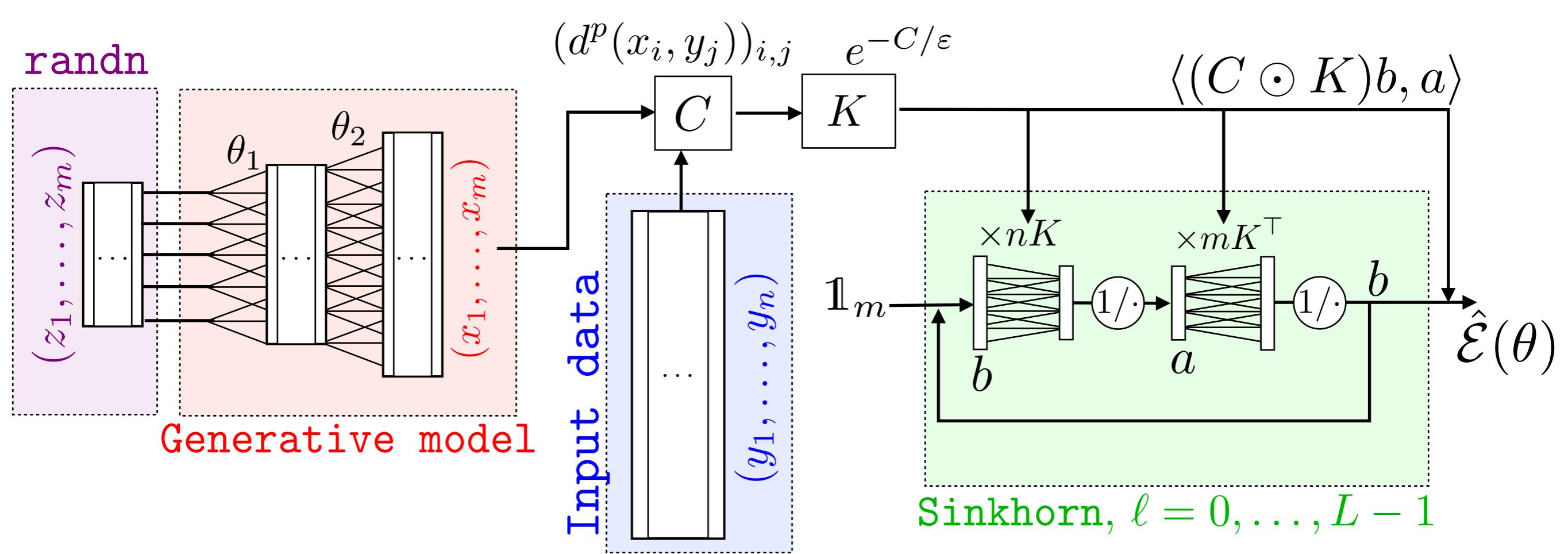


$$\min_{\theta} E(\theta) \stackrel{\text{def.}}{=} \overline{W}_{\epsilon,p}^p(\alpha_\theta, \beta)$$

Stochastic gradient descent

$$\theta \leftarrow \theta - \tau \nabla \hat{E}(\theta)$$

$$\hat{E}(\theta) \stackrel{\text{def.}}{=} \overline{W}_{\epsilon,p}^p\left(\frac{1}{m} \sum_i \delta_{g_\theta(z_i)}, \beta\right)$$



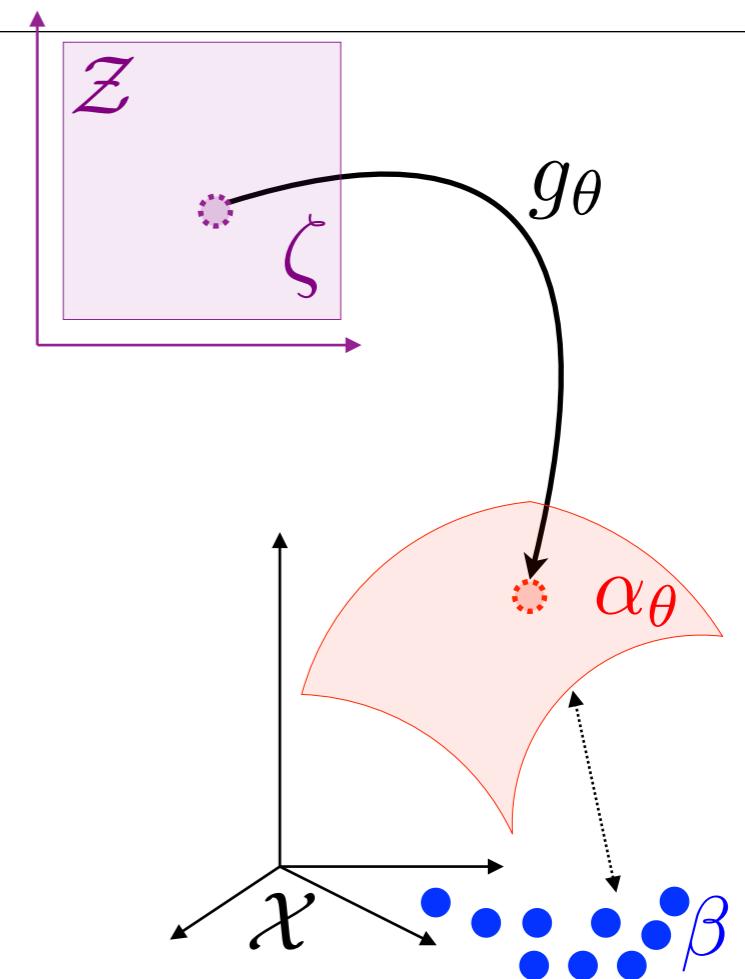
Examples of Images Generation

Inputs β

3	4	2	1	9	5	6	2	1
8	9	1	2	5	0	0	6	6
6	7	0	1	6	3	6	3	7
3	7	7	9	4	6	6	1	8
2	9	3	4	3	9	8	7	2
1	5	9	8	3	6	5	7	2
9	3	1	9	1	5	8	0	8
5	6	2	6	8	5	8	8	9
3	7	7	0	9	4	8	5	4

Generated α_θ

9	4	7	3	3	9	6	8
5	5	1	0	8	1	2	0
5	4	0	8	0	0	5	9
8	2	6	0	7	2	4	7
3	9	0	6	1	9	1	8
4	2	6	7	9	3	6	7
8	0	0	2	4	8	5	7
2	6	0	5	3	4	0	3

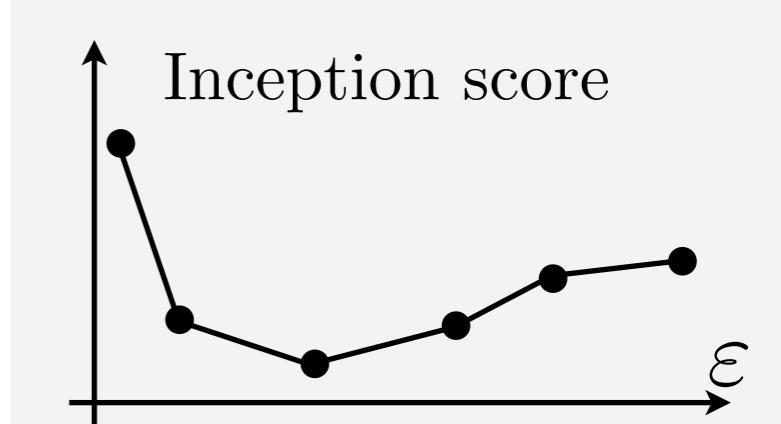
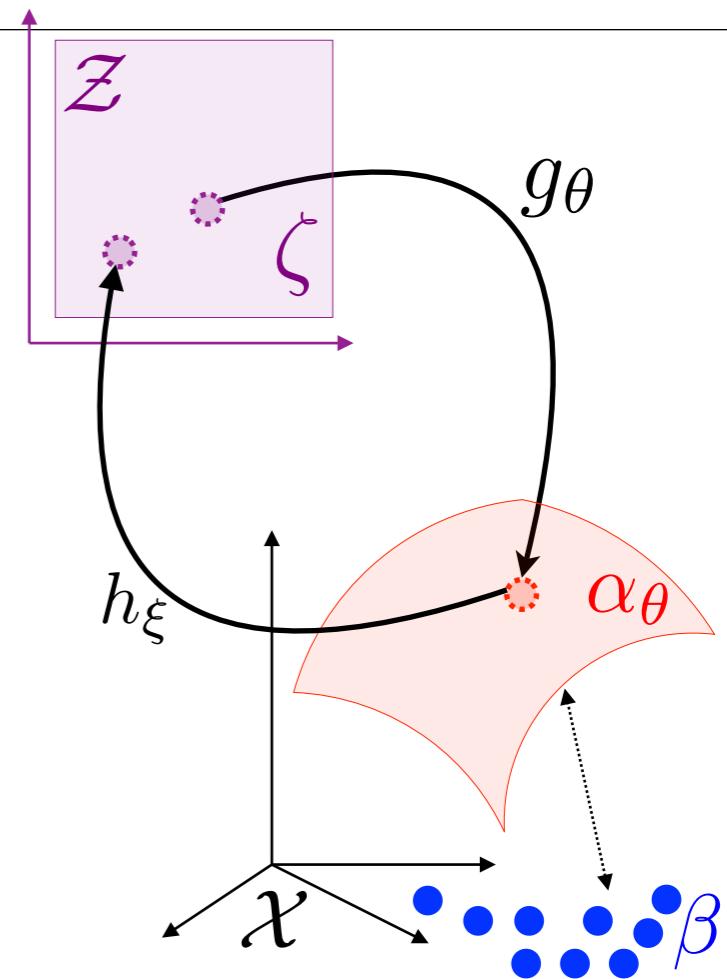
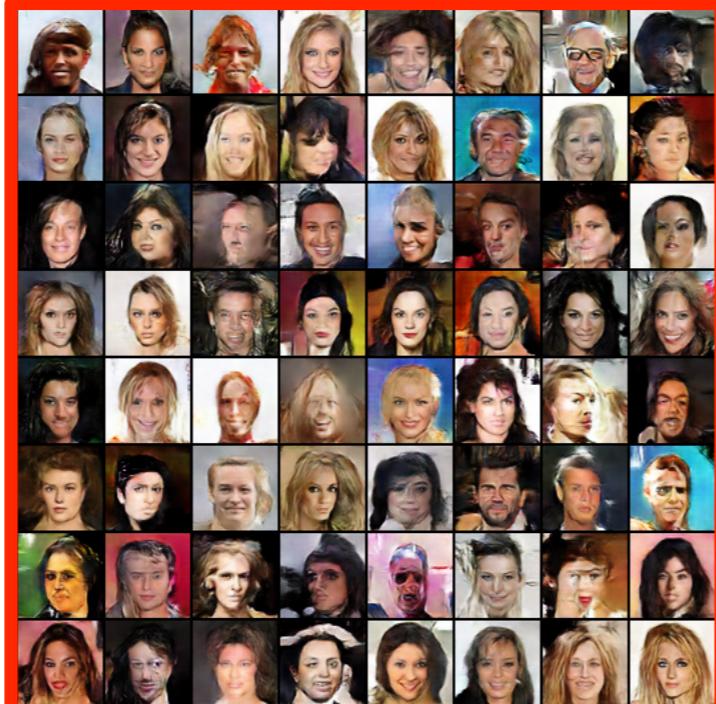


Examples of Images Generation

Inputs β



Generated α_θ

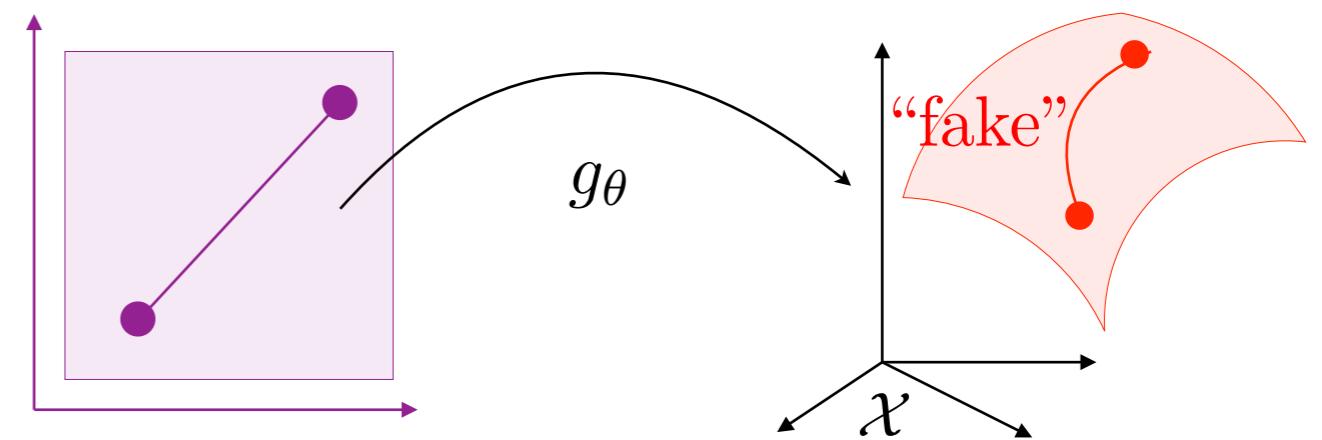


- Need to learn the metric $d(x, y) = \|h_\xi(x) - h_\xi(y)\|$ (GANs)
- Influence of ϵ ?
- Performance evaluation of generative models is an open problem.



Progressive Growing of GANs for Improved Quality, Stability, and Variation

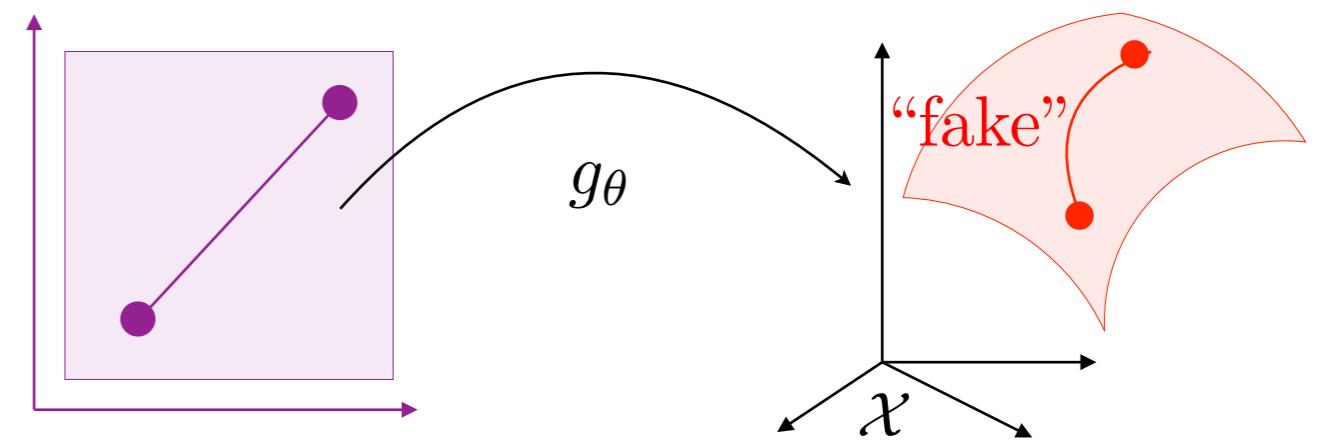
Tero Karras, Timo Aila, Samuli Laine,
Jaakko Lehtinen, ICLR 2018



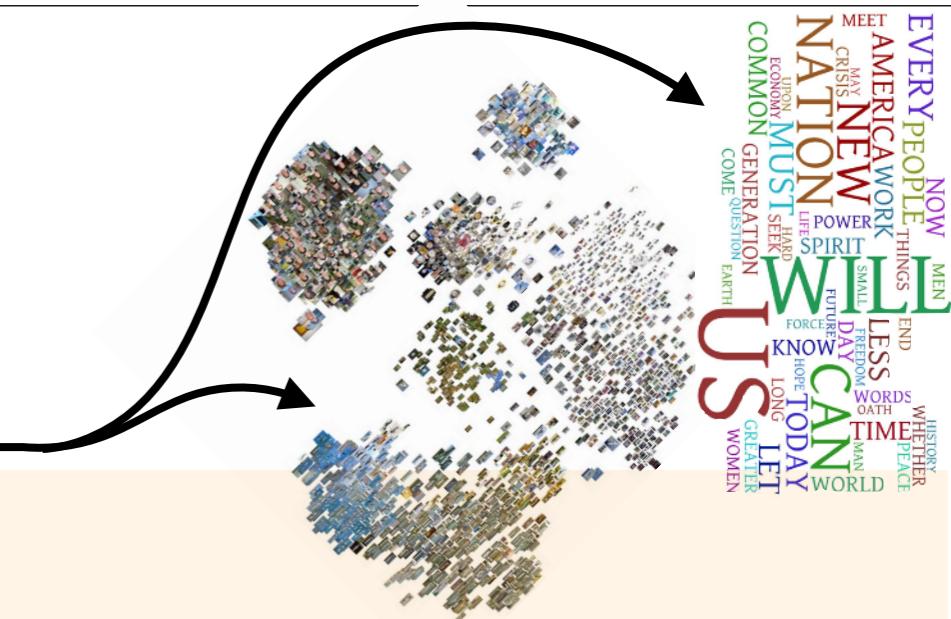


Progressive Growing of GANs for Improved Quality, Stability, and Variation

Tero Karras, Timo Aila, Samuli Laine,
Jaakko Lehtinen, ICLR 2018



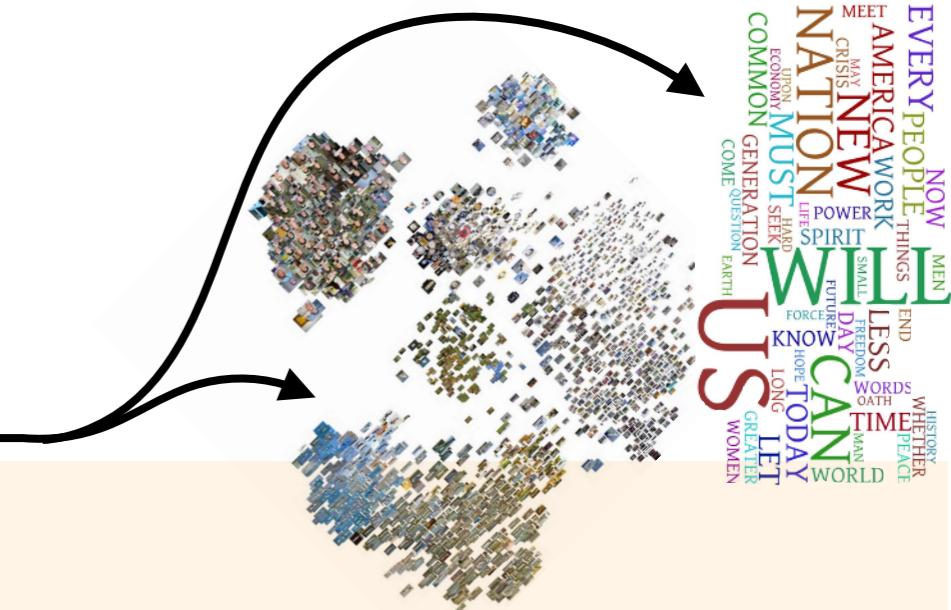
Conclusion



Toward high-dimensional OT:

- Scalable geometrical loss functions in high dimension?
- Performance quality measures for unsupervised learning?

Conclusion

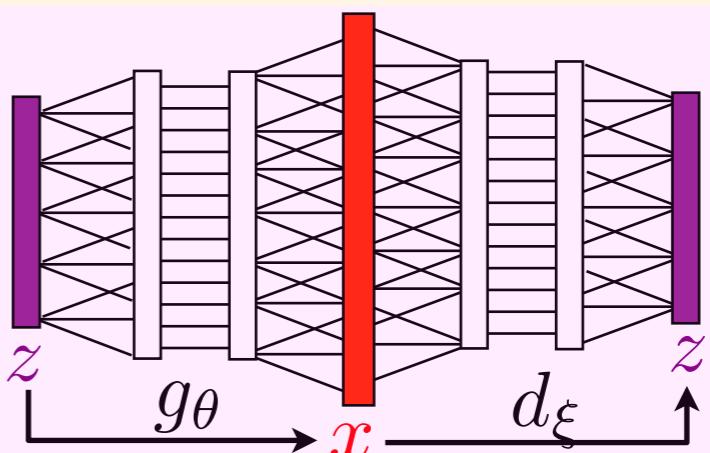


Toward high-dimensional OT:

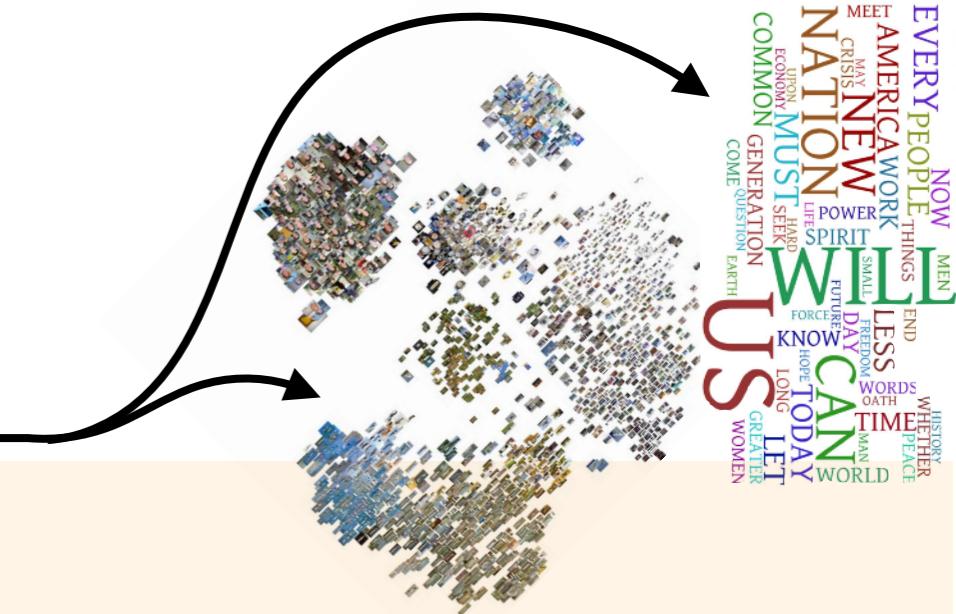
- Scalable geometrical loss functions in high dimension?
- Performance quality measures for unsupervised learning?

Metric learning for OT:

- Adversarial training to leverage multi scale priors?



Conclusion

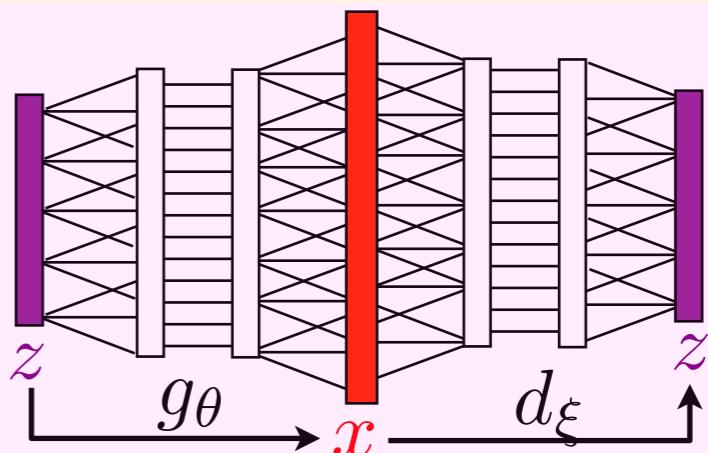


Toward high-dimensional OT:

- Scalable geometrical loss functions in high dimension?
- Performance quality measures for unsupervised learning?

Metric learning for OT:

- Adversarial training to leverage multi scale priors?



Beyond comparing measures:

- Learning for surfaces, graphs, metric spaces?
- Using Gromov-Wasserstein geometry?

