Computational Optimal Transport

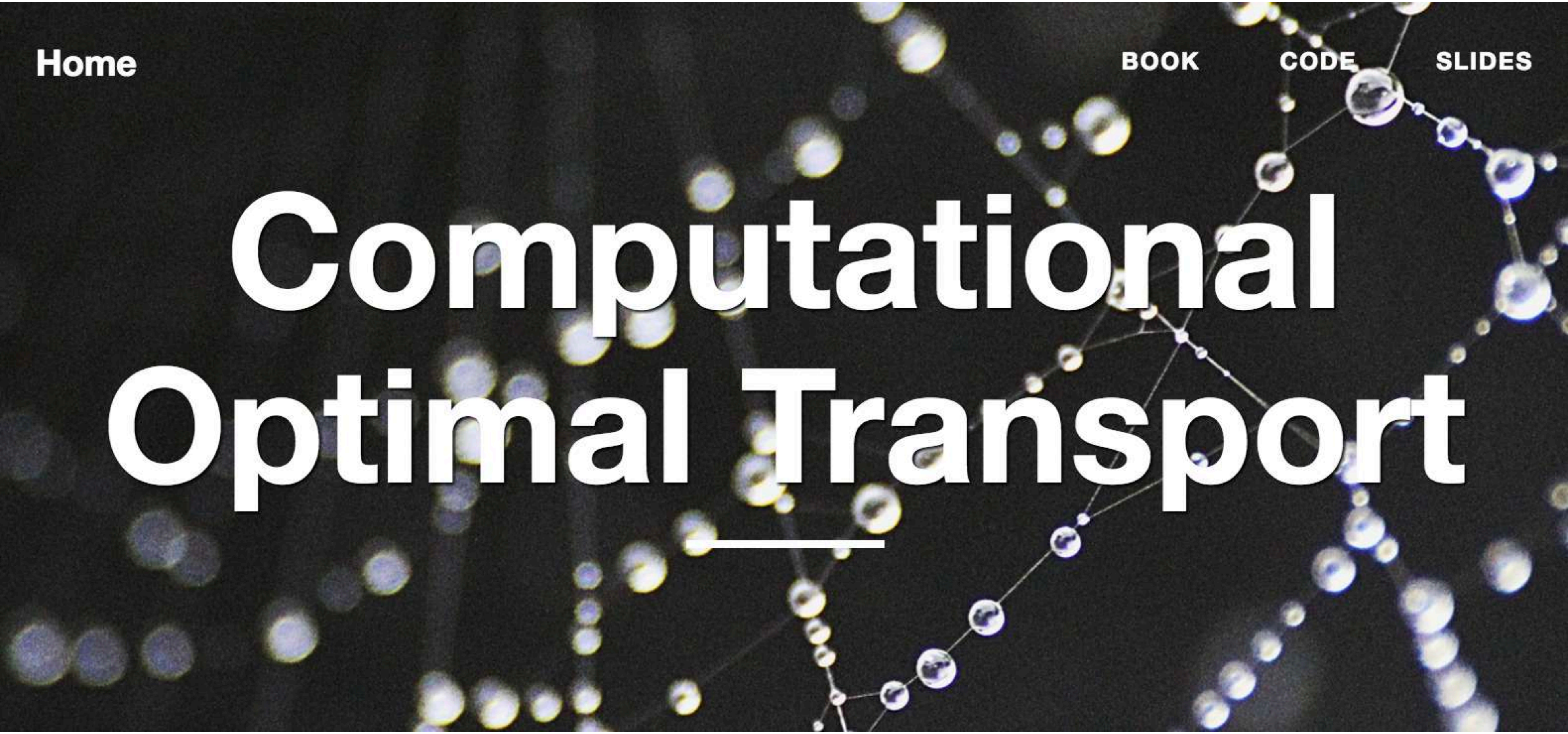
Gabriel Peyré

cnrs dépasser les frontières

ENS ÉCOLE NORMALE SUPÉRIEURE

Joint work with:

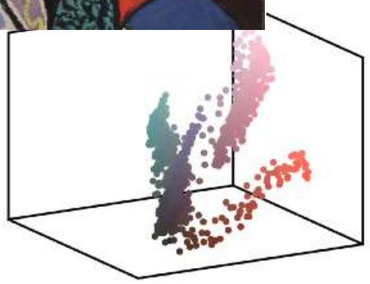
https://optimaltransport.github.io

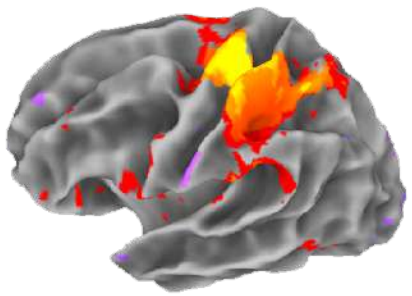


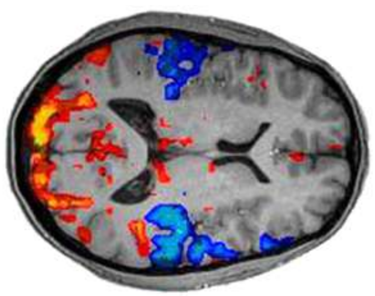
Probability Distributions in Data Sciences

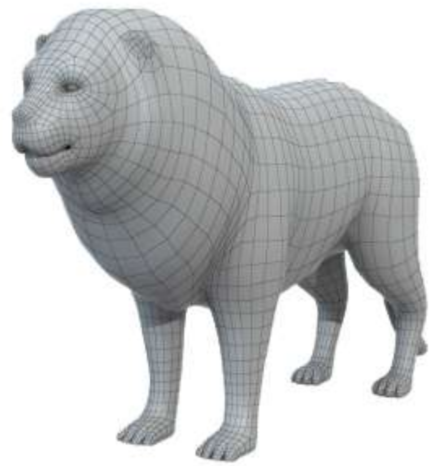
Probability distributions and histograms

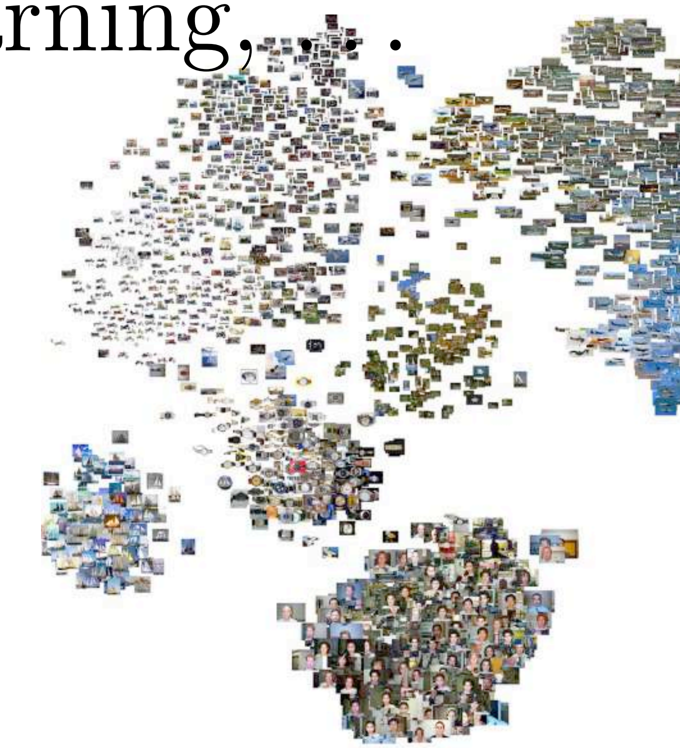
$\rightarrow$ images, vision, graphics and machine learning, . . .







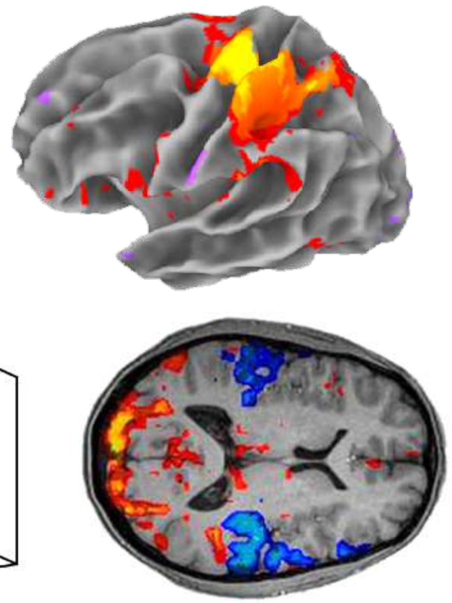
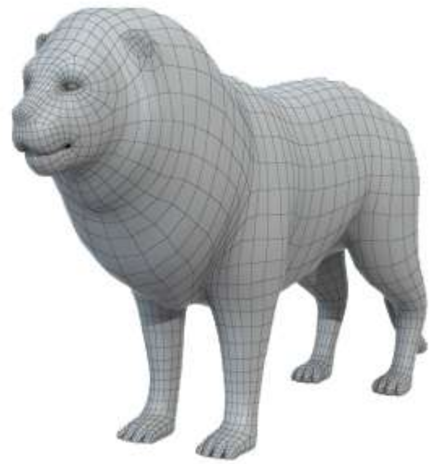


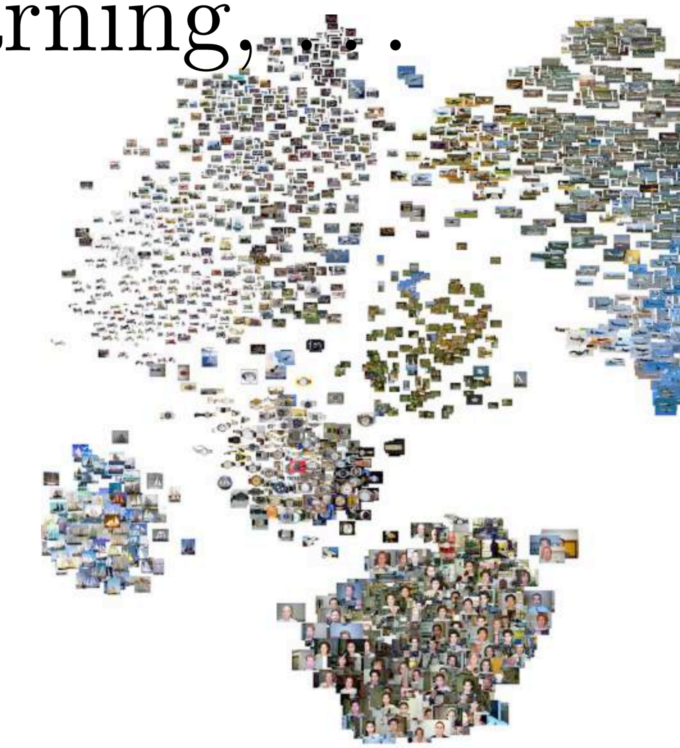


Probability Distributions in Data Sciences

Probability distributions and histograms

→ images, vision, graphics and machine learning, . . .

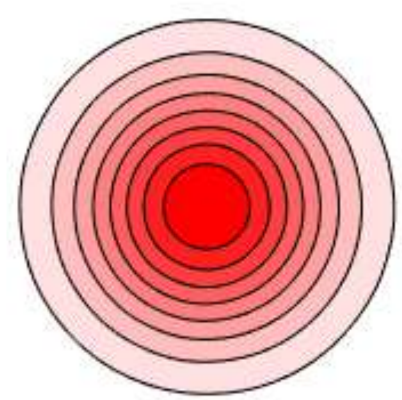



Unsupervised learning

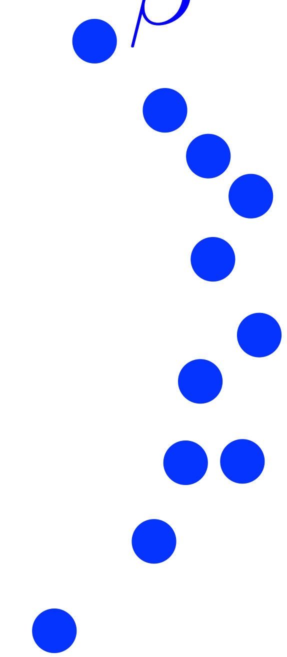
Observations: $\beta \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

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

α_θ



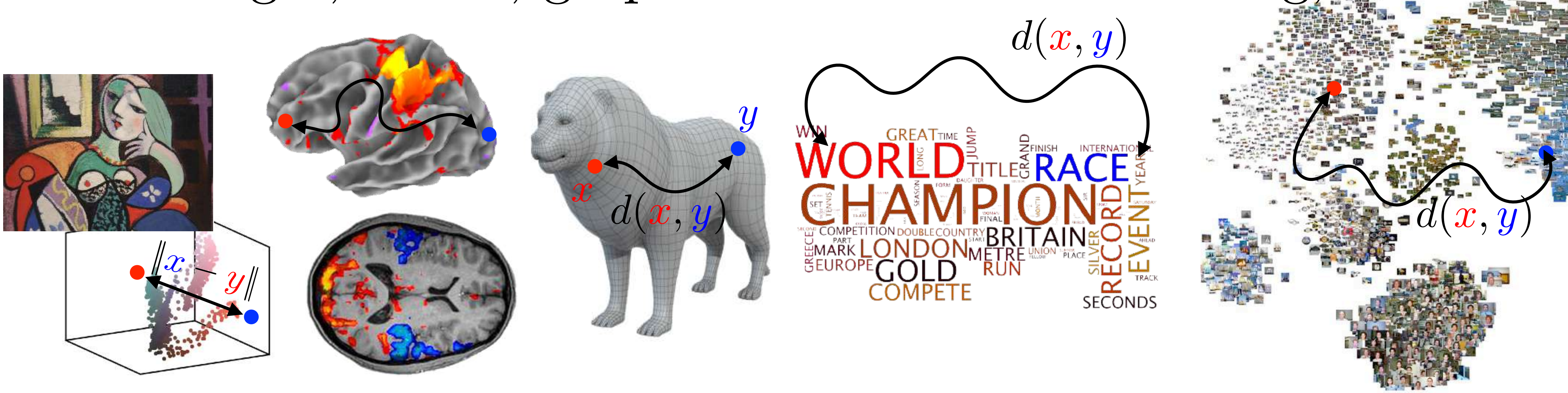
β



Probability Distributions in Data Sciences

Probability distributions and histograms

→ images, vision, graphics and machine learning, . . .



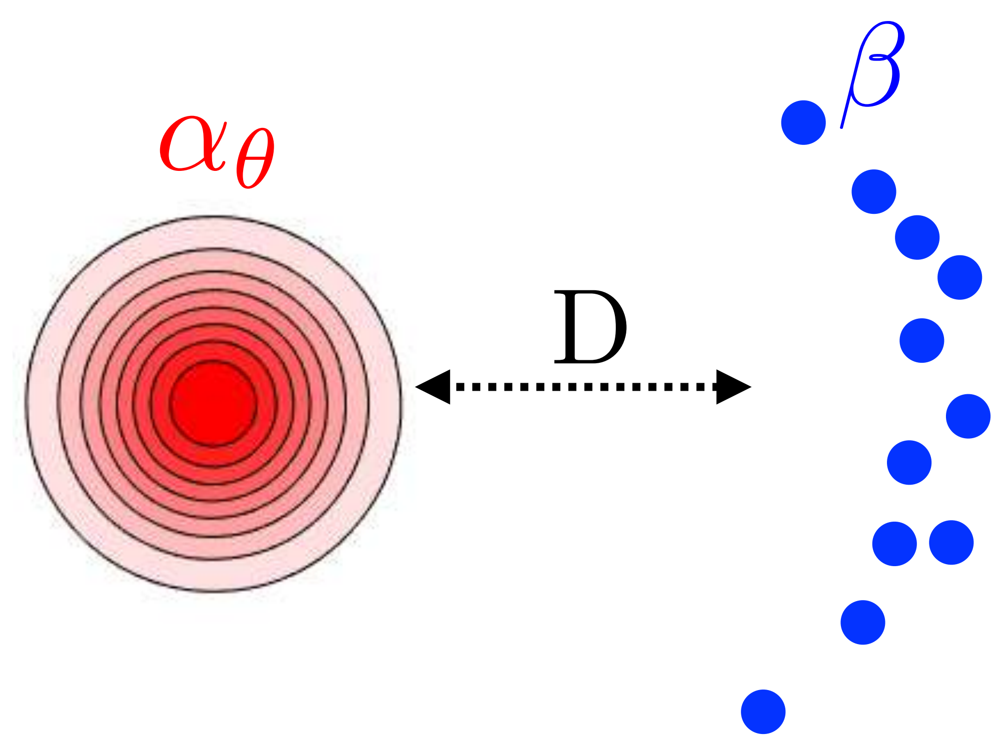
Unsupervised learning

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

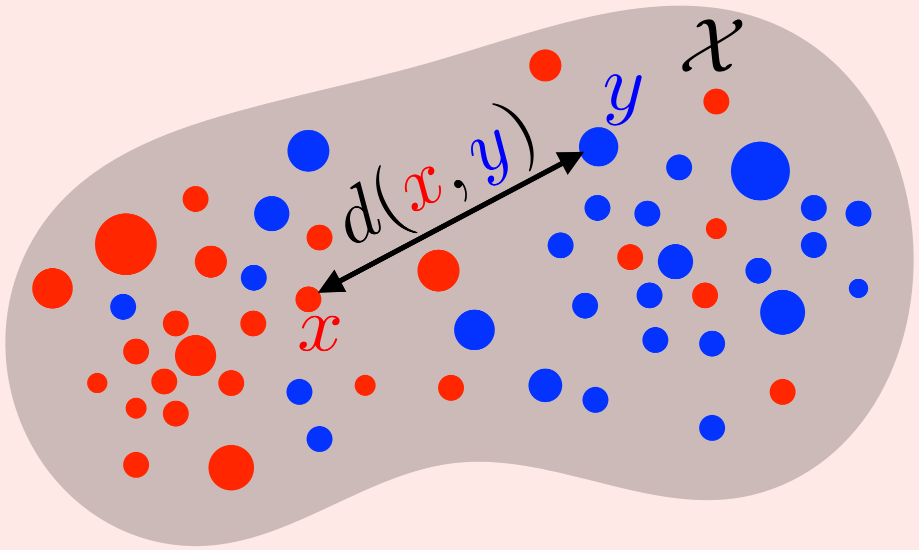
Parametric model: $\theta \mapsto \alpha_\theta$

Density fitting: $\min_{\theta} \mathrm{D}(\alpha_\theta, \beta)$

→ takes into account a metric d.



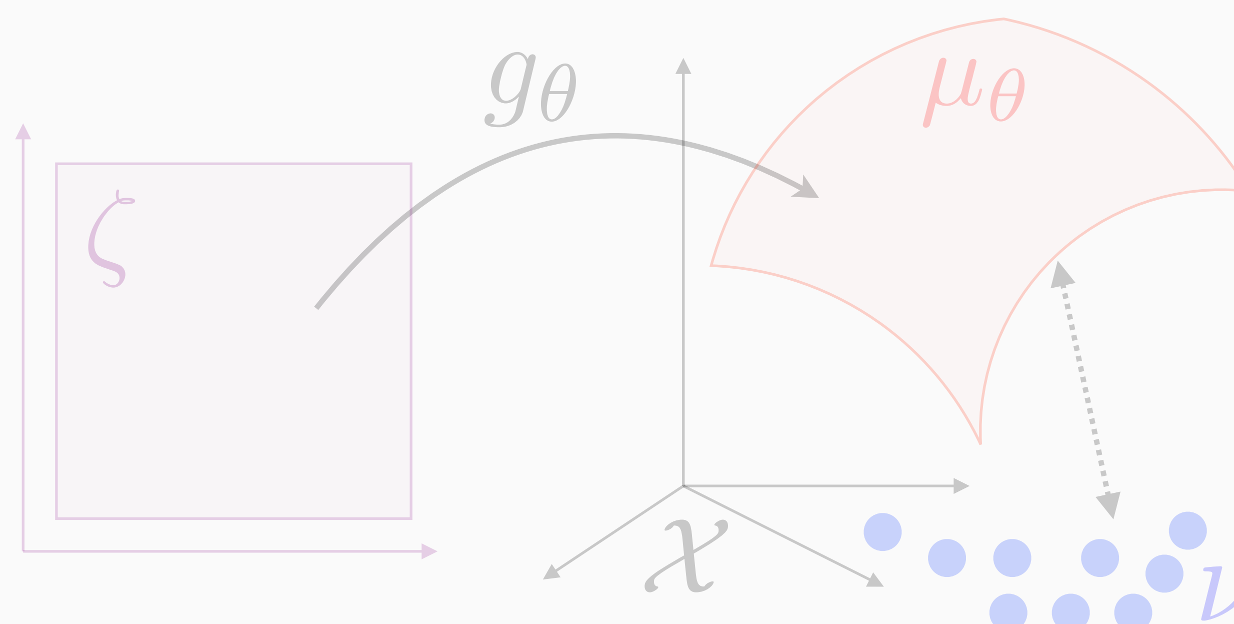
1. Optimal Transport



2. Entropic Regularization

3. Sinkhorn Divergences

4. Application to Generative Models

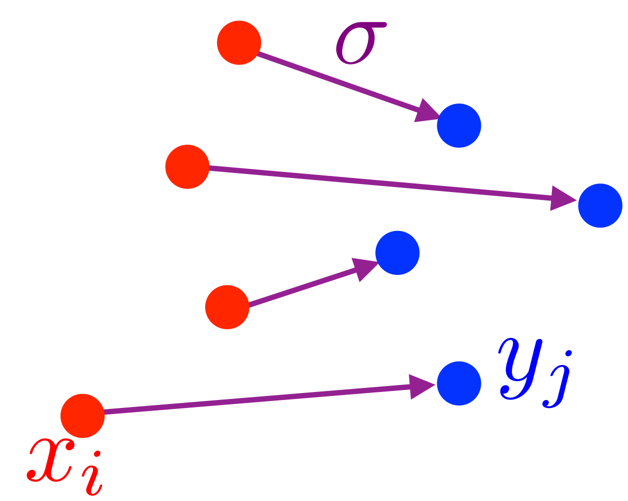


Monge's Problem

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

Permutation:

$$\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$$

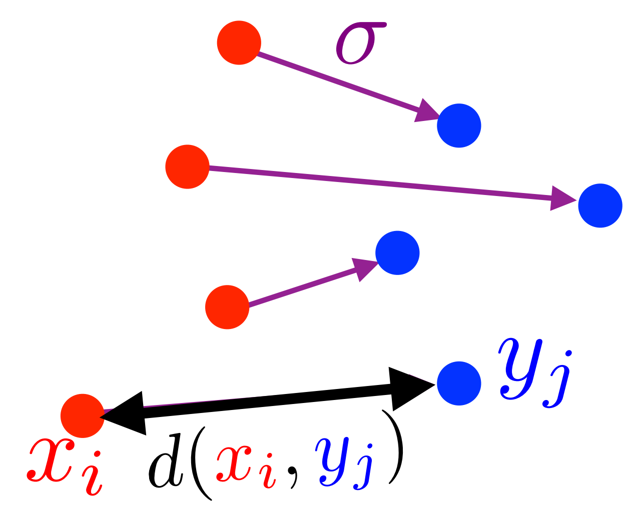


Monge's Problem

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

Permutation:

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



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

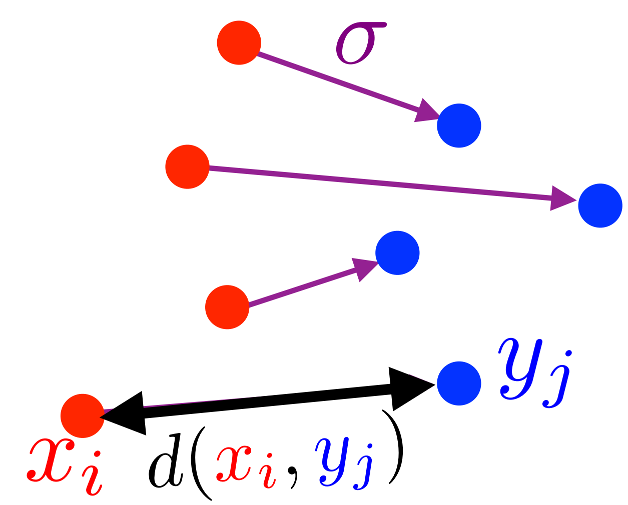
[Monge 1784]

Monge's Problem

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

Permutation:

$$\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$$



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

$\to$ Seems intractable: $n!$ possibilities.

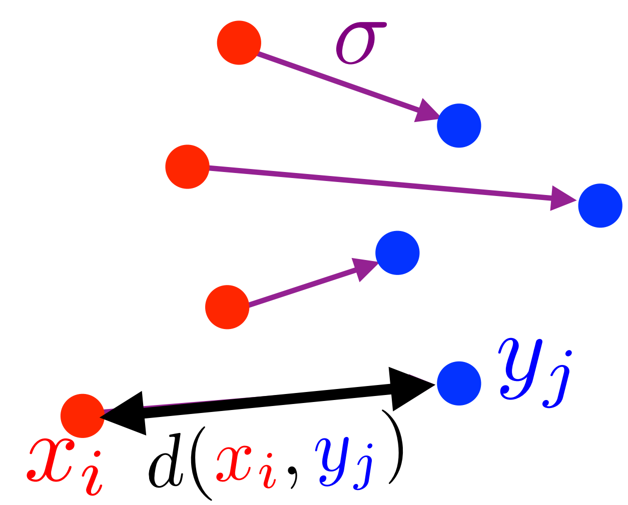
[Monge 1784]

Monge's Problem

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

Permutation:

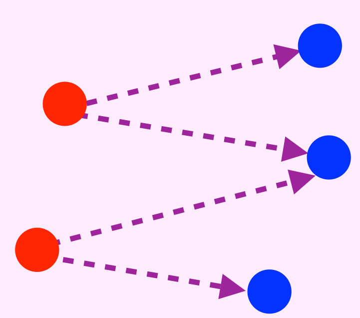
$$\sigma : \{1, \ldots, n\} \rightarrow \{1, \ldots, n\}$$



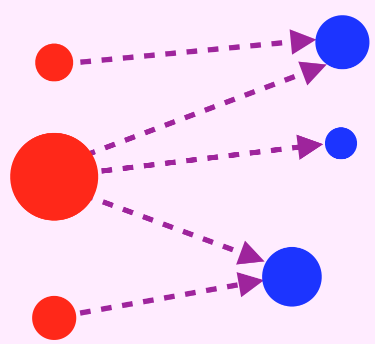
Monge optimal matching: $$\min_{\sigma} \sum_{i=1}^{n} d(x_i, y_{\sigma(i)})$$

$\rightarrow$ Seems intractable: $n!$ possibilities.

[Monge 1784]



Different # points?



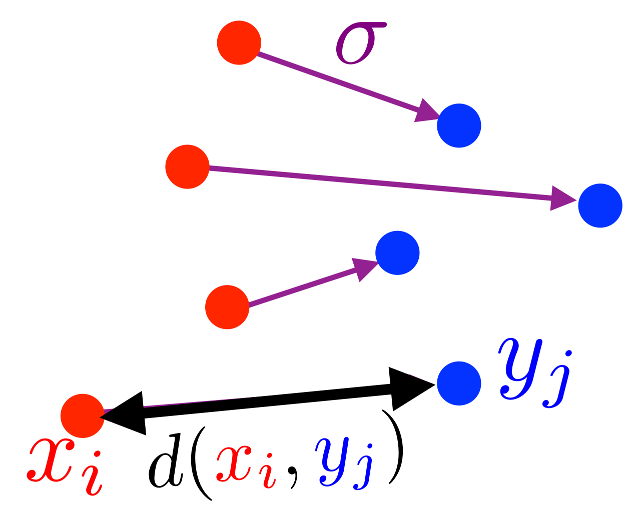
Weights?

Monge's Problem

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

Permutation:

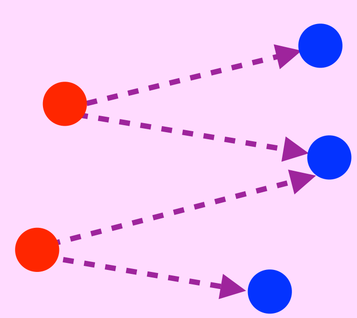
$$\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$$



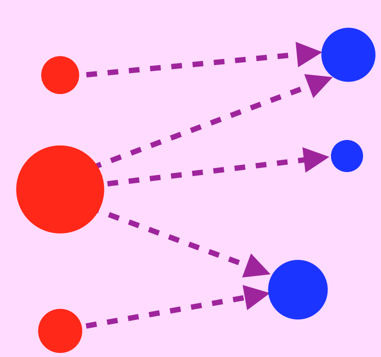
Monge optimal matching: $\displaystyle \min_{\sigma} \sum_{i=1}^{n} d(x_i, y_{\sigma(i)})$

$\to$ Seems intractable: $n!$ possibilities.

[Monge 1784]



Different # points?



Weights?

$\Longrightarrow$

"Relax"

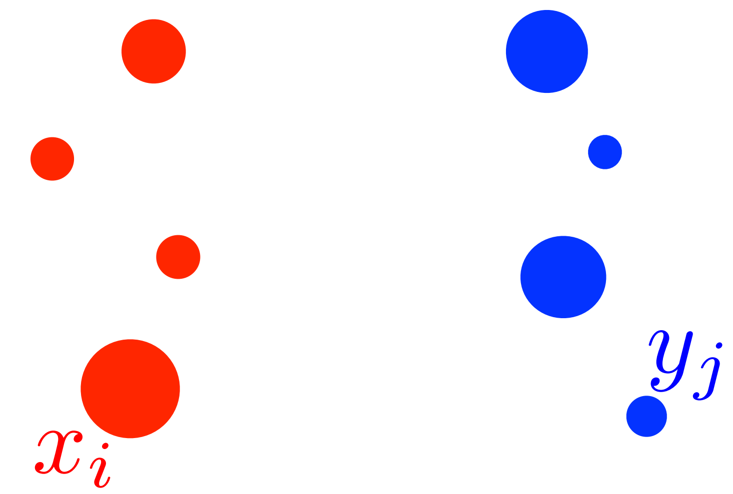
points $\to$ mass

permutation $\to$ 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 $(x_i)_i$, $(y_j)_j$

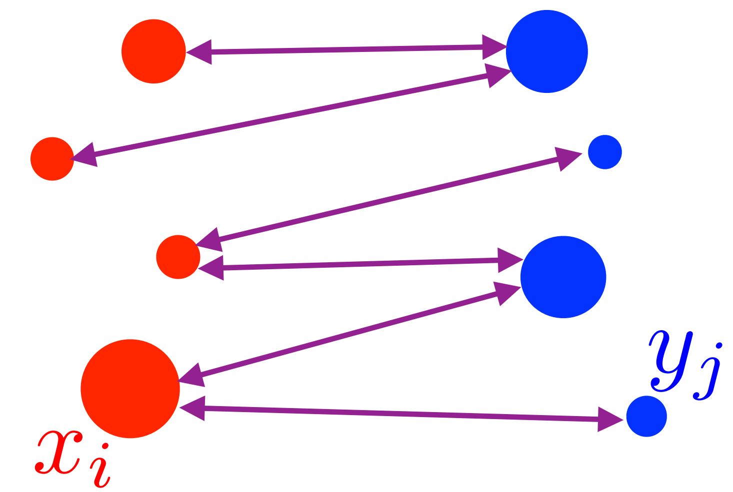
Weights $\mathbf{a}_i \geqslant 0$, $\mathbf{b}_j \geqslant 0$.

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

Kantorovitch's Formulation

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

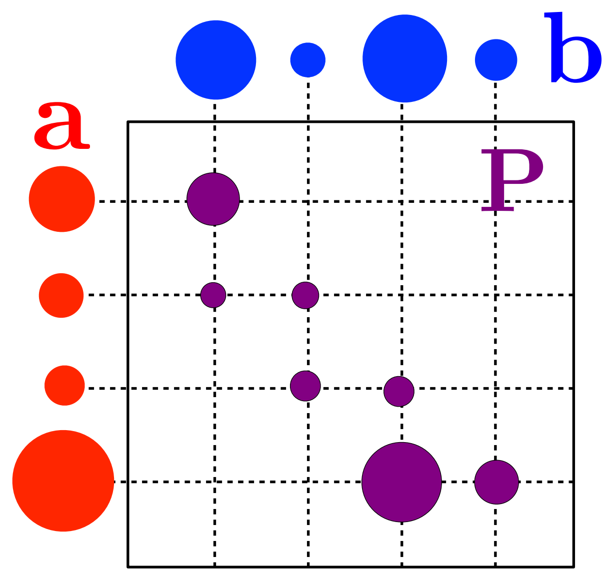
$\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$



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

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

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



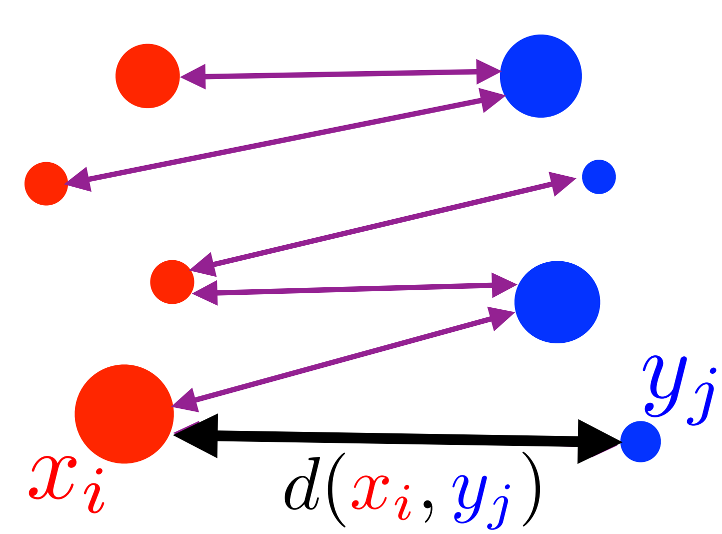
Couplings: $\sum_j \mathbf{P}_{i,j} = \mathbf{a}_i$ $\sum_i \mathbf{P}_{i,j} = \mathbf{b}_j$

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

Kantorovitch's Formulation

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

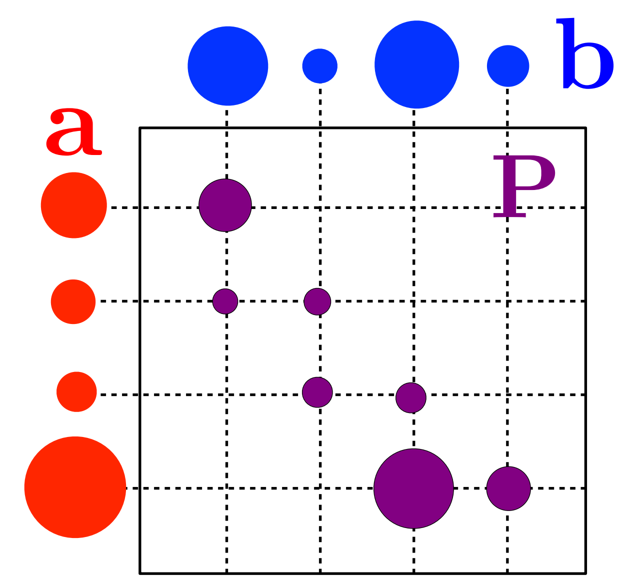
$\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$



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

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

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



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

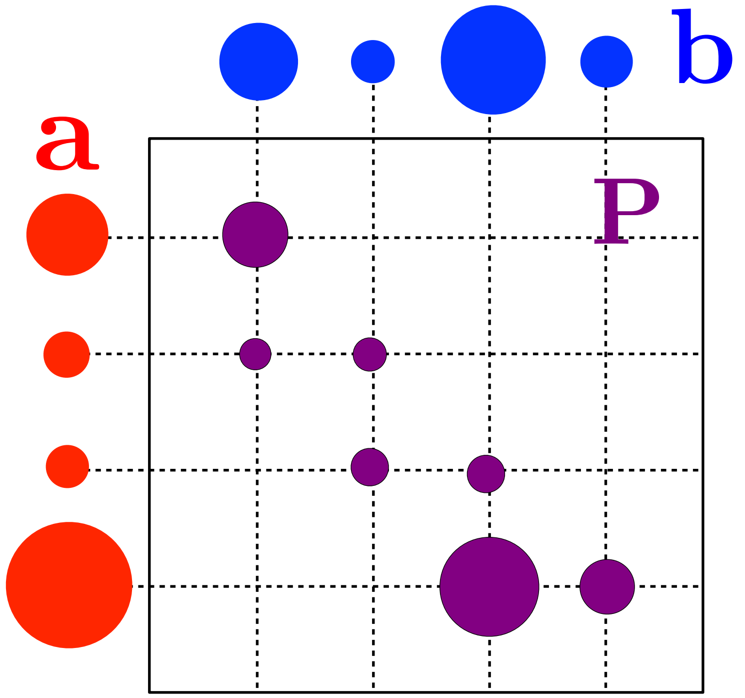
[Kantorovich 1942]

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

Optimal Transport Distances

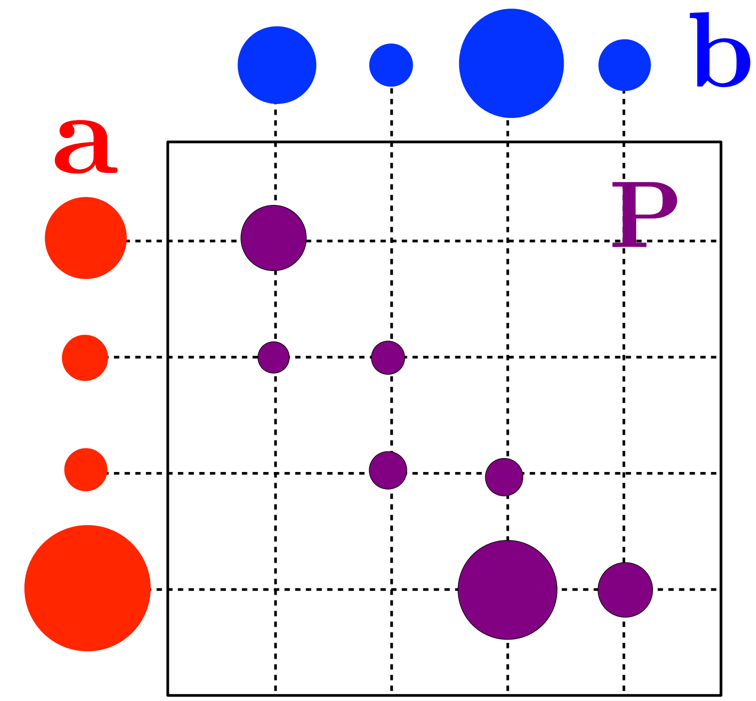
$$\mathrm{W}_p(\alpha, \beta) \overset{\text{def.}}{=}$$

$$\left(\min_{\mathbf{P}\mathbb{1}=\mathbf{a}, \mathbf{P}^\top\mathbb{1}=\mathbf{b}} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{I,j} \right)^{\frac{1}{p}}$$



Optimal Transport Distances

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

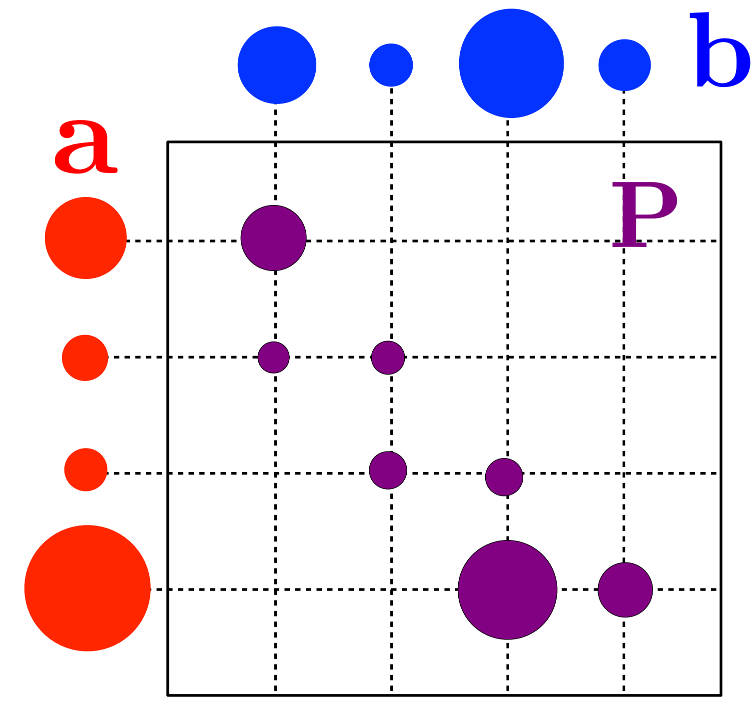


Convergence in law: $\alpha_n \rightharpoonup \beta \Leftrightarrow \forall f \in \mathcal{C}(\mathcal{X}), \int_\mathcal{X} f \mathrm{d}\alpha_n \to \int_\mathcal{X} f \mathrm{d}\beta$

$\|\delta_{x_n} - \delta_x\|_1 = 2$ **vs.** $\mathrm{W}_p(\delta_{x_n}, \delta_x) = d(x_n, x)$

Optimal Transport Distances

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



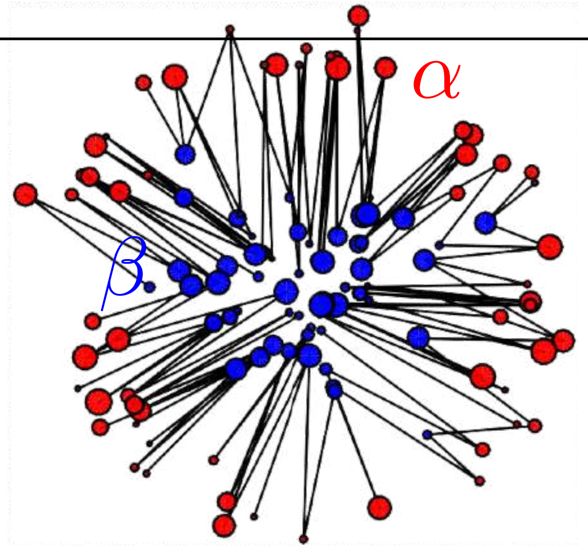
Convergence in law: $\alpha_n \rightharpoonup \beta \Leftrightarrow$ $\forall f \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} f \mathrm{d}\alpha_n \to \int_{\mathcal{X}} f \mathrm{d}\beta$

$\|\delta_{x_n} - \delta_x\|_1 = 2$ **vs.** $\mathrm{W}_p(\delta_{x_n}, \delta_x) = d(x_n, x)$

Theorem: W_p is a distance and $\alpha_n \rightharpoonup \beta \Leftrightarrow \mathrm{W}_p(\alpha_n, \beta) \to 0$

Algorithms

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

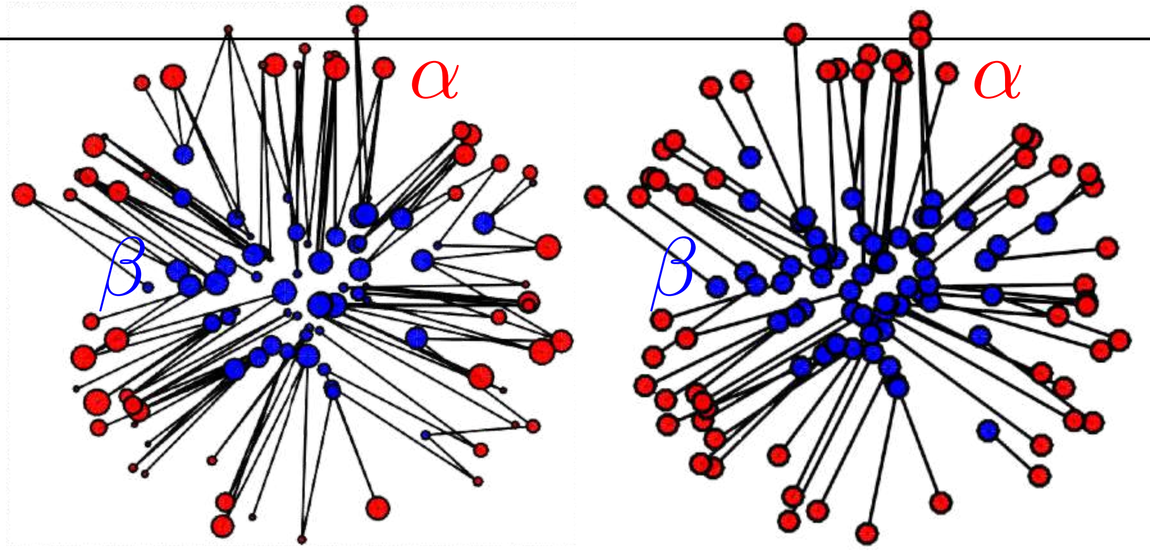


Algorithms

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

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

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

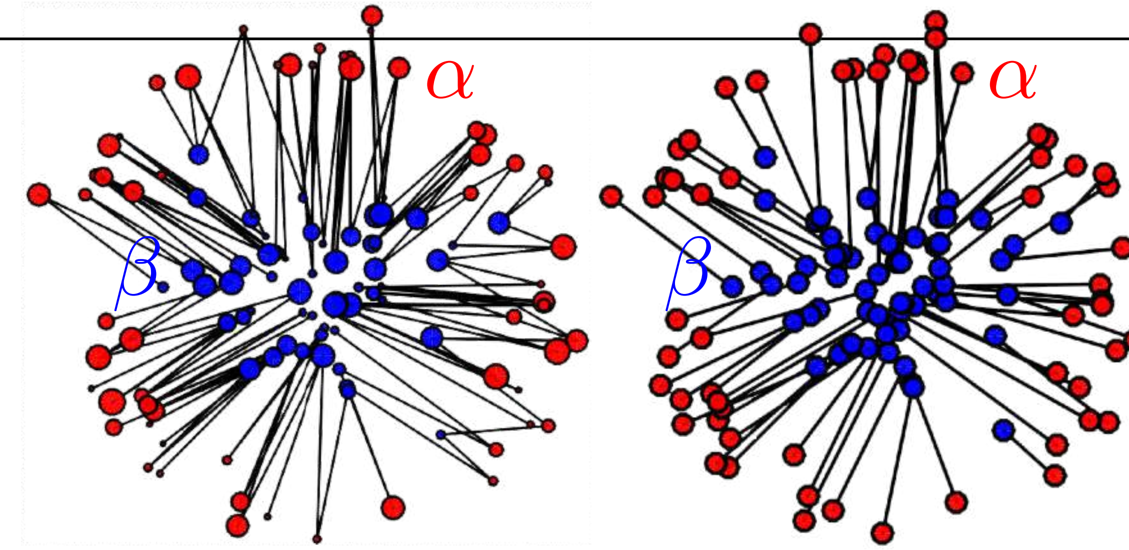


Algorithms

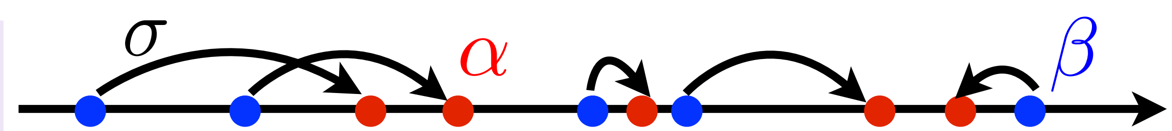
Linear programming: $O(n^3 \log(n)^2)$

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

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



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

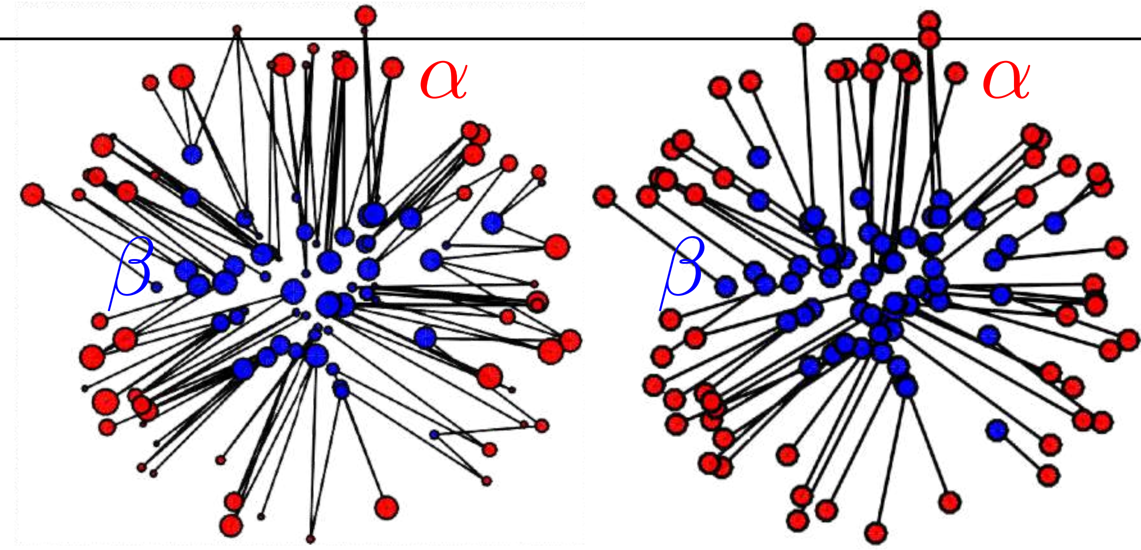


Algorithms

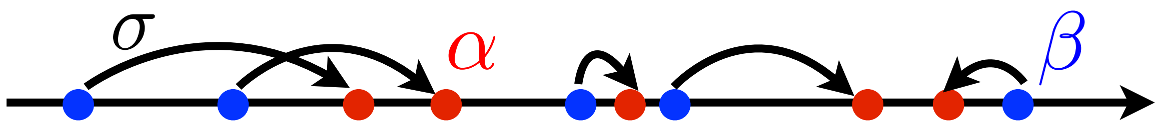
Linear programming: $O(n^3 \log(n)^2)$

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

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



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



$p = 1$
$d = \| \cdot \|$

$$W_1(\alpha, \beta) = \min_{\mathrm{div}(u) = \alpha - \beta} \int \|u(x)\| \mathrm{d}x$$

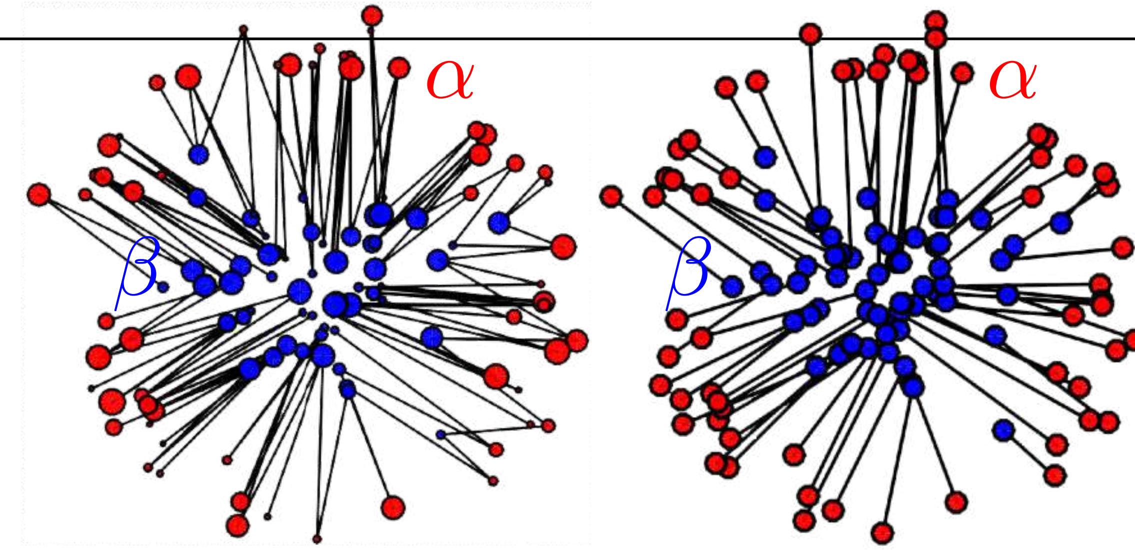
$\rightarrow$ min-cost flow, on graphs $O(n^2 \log(n))$.

Algorithms

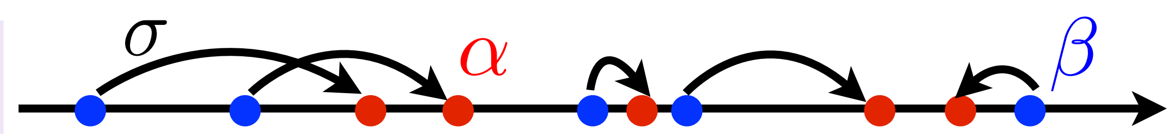
Linear programming: $O(n^3 \log(n)^2)$

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

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



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



$p = 1$
$d = \|\cdot\|$

$$W_1(\alpha,\beta) = \min_{\text{div}(u)=\alpha-\beta} \int \|u(x)\| \text{d}x$$

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

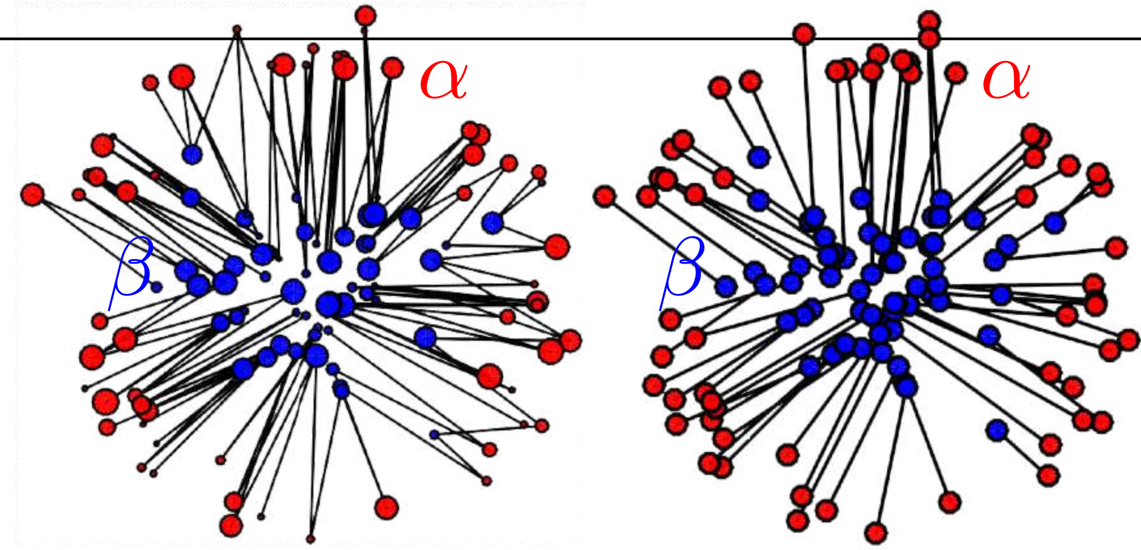
Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.

Algorithms

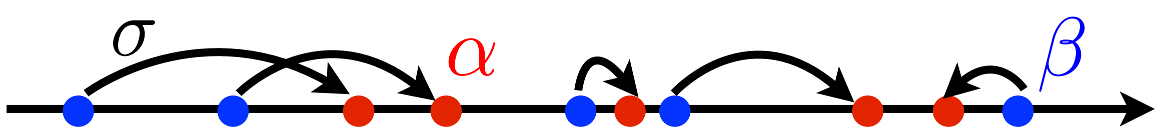
Linear programming: $O(n^3 \log(n)^2)$

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

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



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



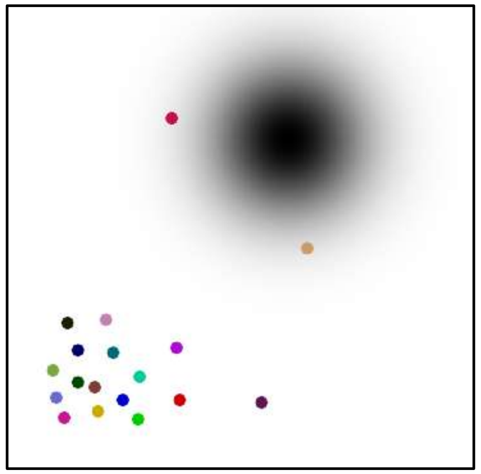
$p = 1$
$d = \|\cdot\|$

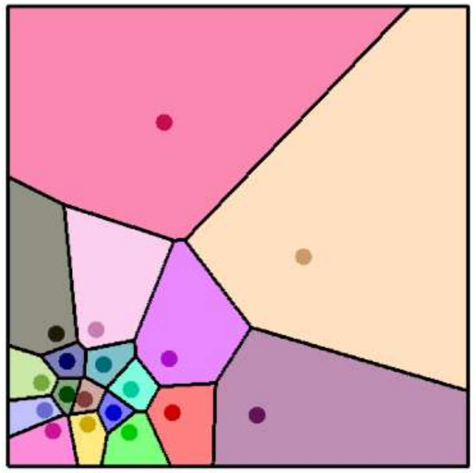
$$W_1(\alpha, \beta) = \min_{\text{div}(u) = \alpha - \beta} \int \|u(x)\| \mathrm{d}x$$

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

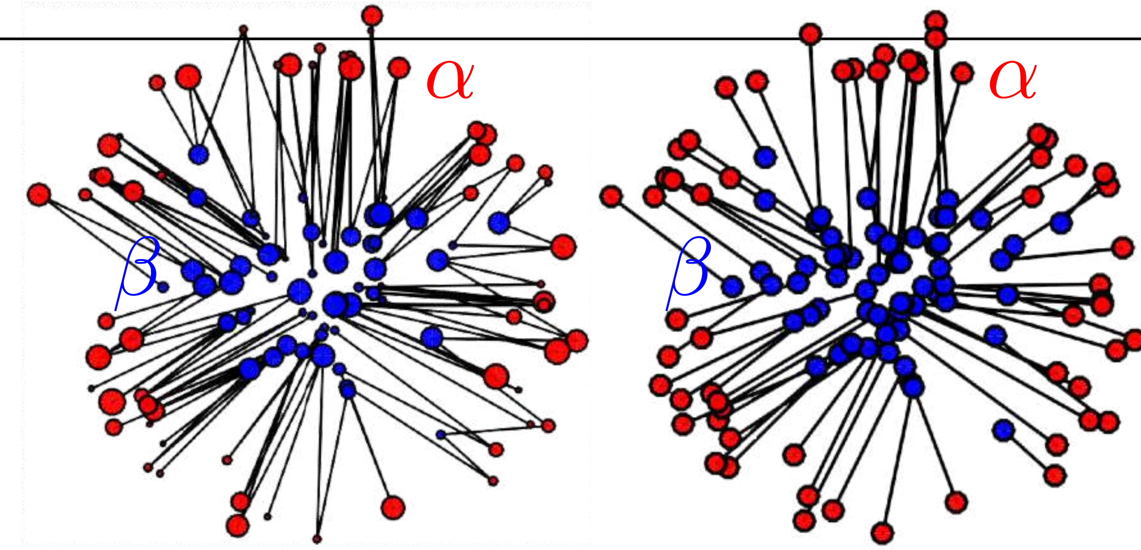
Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.

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





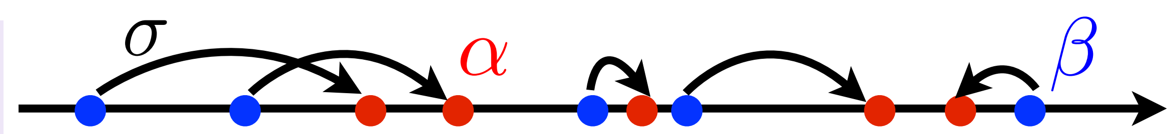
Algorithms



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

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

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

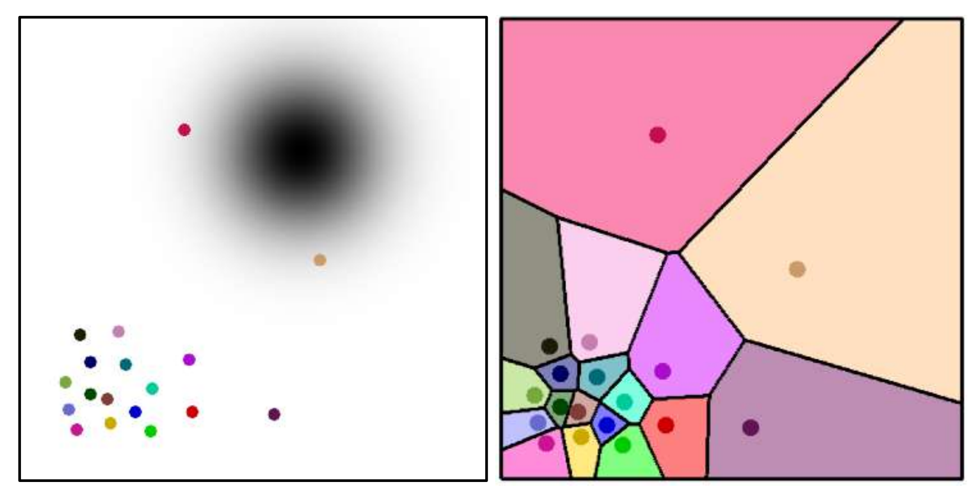


1-D case: sorting $O(n \log(n))$.

$p = 1$
$d = \| \cdot \|$

$$W_1(\alpha, \beta) = \min_{\mathrm{div}(u) = \alpha - \beta} \int \|u(x)\| \mathrm{d}x$$

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

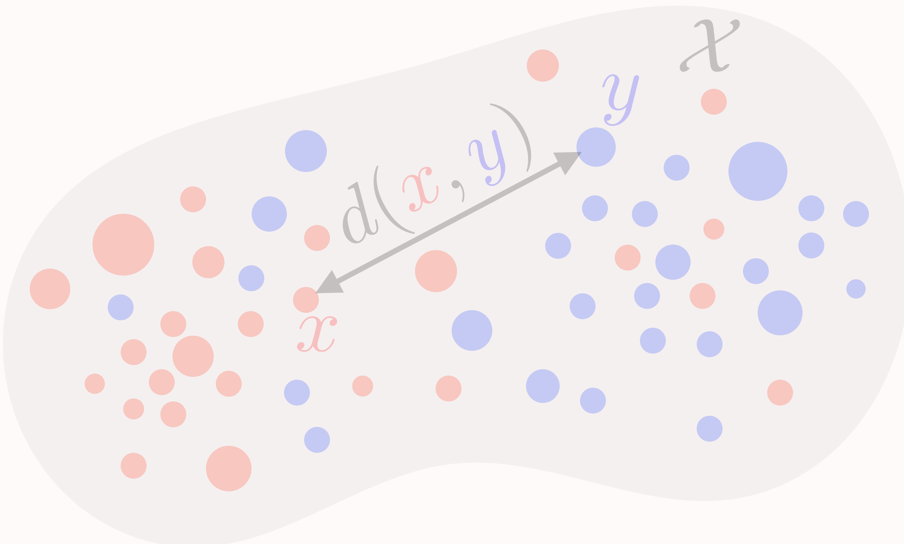


Monge-Ampère/Benamou-Brenier, $d = \| \cdot \|_2^2$.

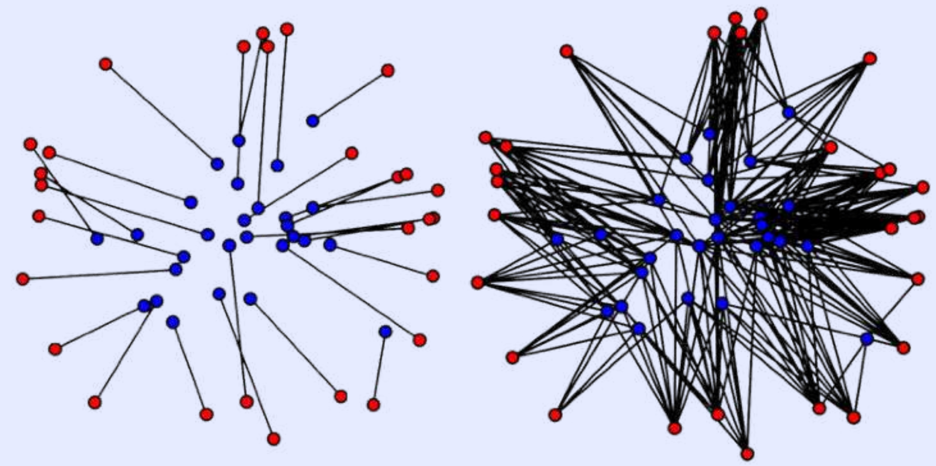
Semi-discrete: Laguerre cells, $d = \| \cdot \|_2^2$.
[Merigot 2013]

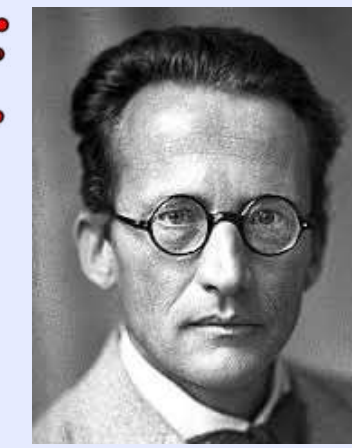
Need for fast approximate algorithms for generic $d(x, y)$.

1. Optimal Transport



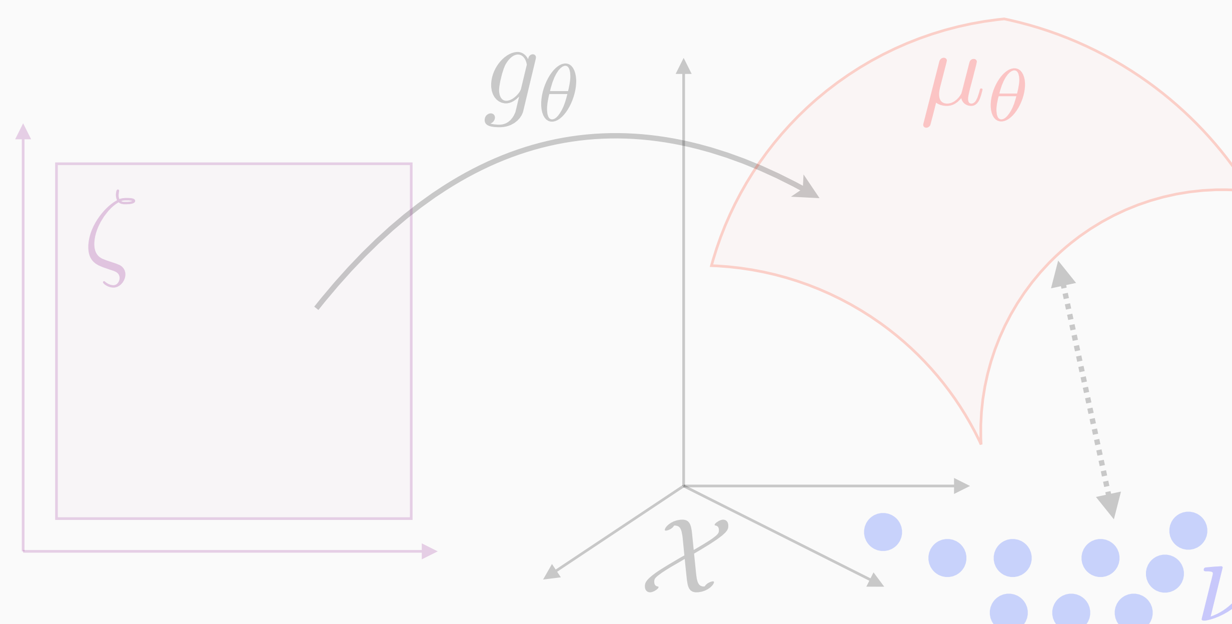
2. Entropic Regularization





3. Sinkhorn Divergences

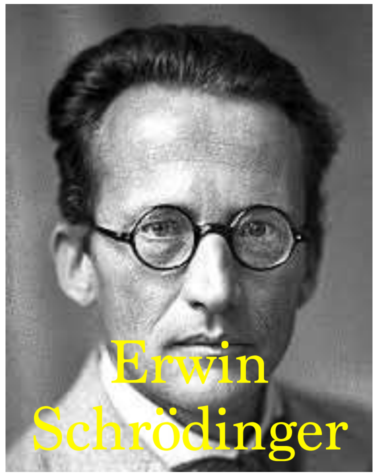
4. Application to Generative Models



Entropic Regularization

Schrödinger's problem: [1931]

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

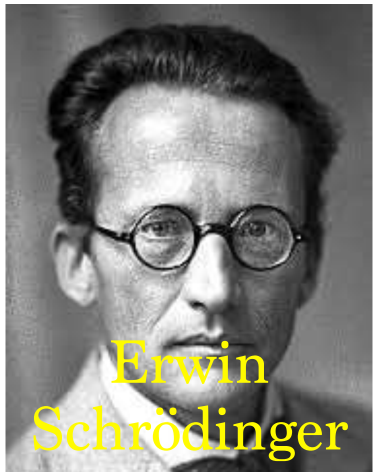


$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$

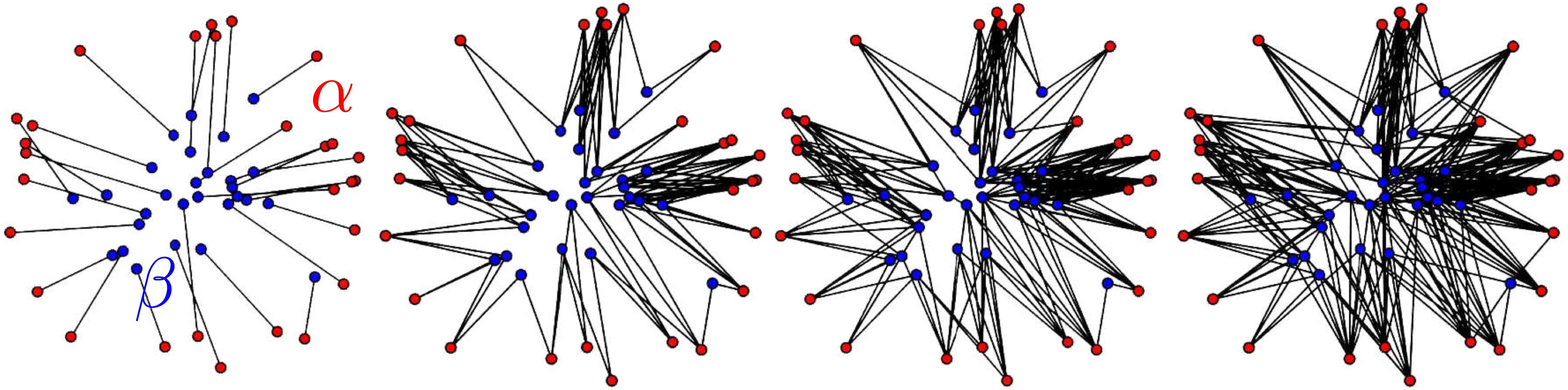
Entropic Regularization

Schrödinger's problem: [1931]

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



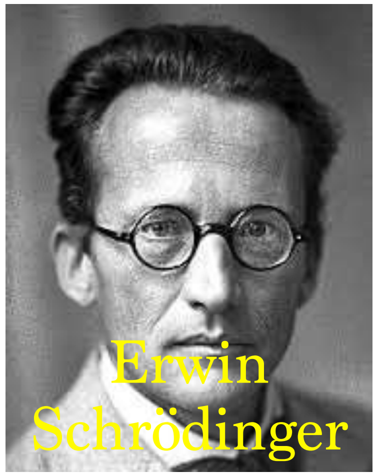
$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$



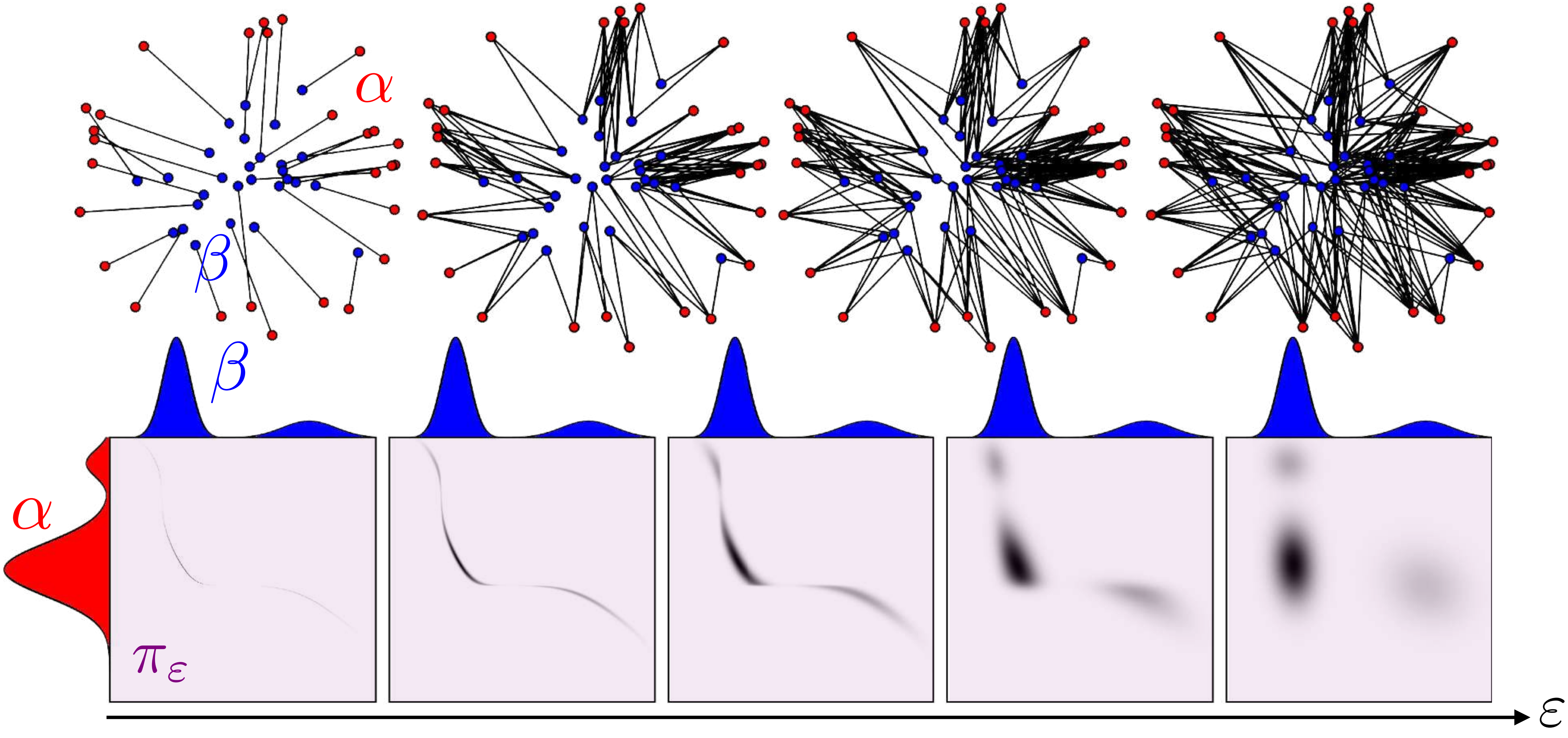
Entropic Regularization

Schrödinger's problem: [1931]

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



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



Sinkhorn's Algorithm

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

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j \qquad \mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

Sinkhorn's Algorithm

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

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j \qquad \mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

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

Sinkhorn's Algorithm

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

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j$ $\qquad \mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

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

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

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

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j$ $\quad \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

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

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: $\mathbf{u} \leftarrow \dfrac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$ $\quad \mathbf{v} \leftarrow \dfrac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}$

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

Sinkhorn's Algorithm

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

Proposition: $\mathbf{P}_{i,j} = \mathbf{u}_i \, \mathbf{K}_{i,j} \, \mathbf{v}_j$ $\qquad \mathbf{K}_{i,j} \stackrel{\text{def.}}{=} e^{-\frac{d(x_i, y_j)^p}{\varepsilon}}$

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

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: $\mathbf{u} \leftarrow \dfrac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$ $\qquad \mathbf{v} \leftarrow \dfrac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}$

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

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

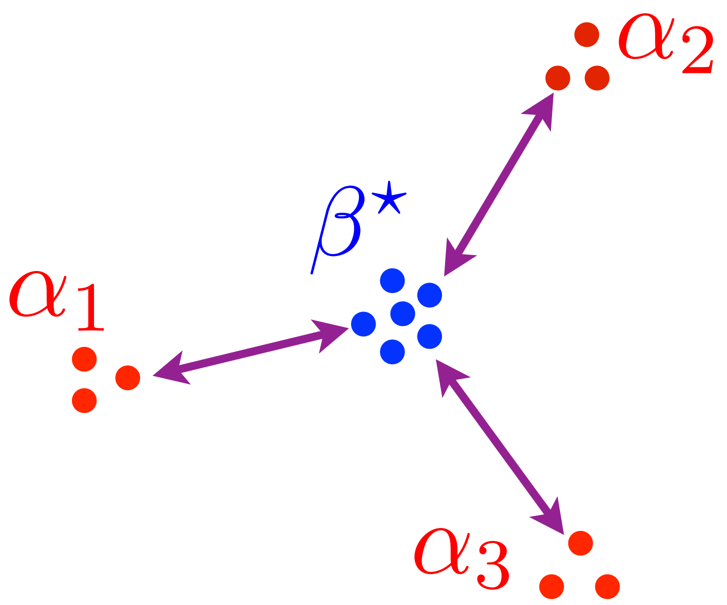
$\rightarrow$ Parallelizable on GPUs.

$\rightarrow$ Convolution on regular grids, separable kernels.

Wasserstein Barycenters

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

$$\beta^\star \in \underset{\beta}{\operatorname{argmin}} \sum_s \lambda_s \mathrm{W}_p^p(\alpha_s, \beta)$$

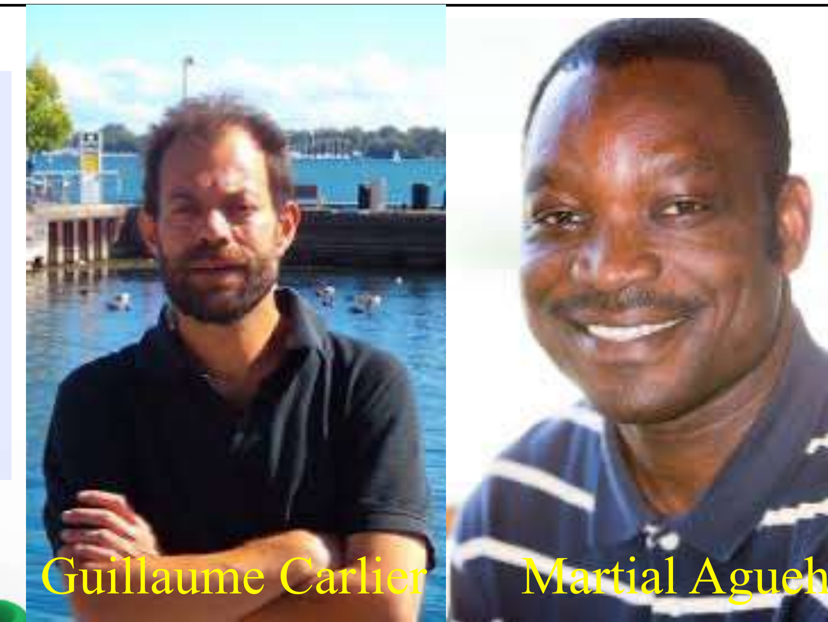


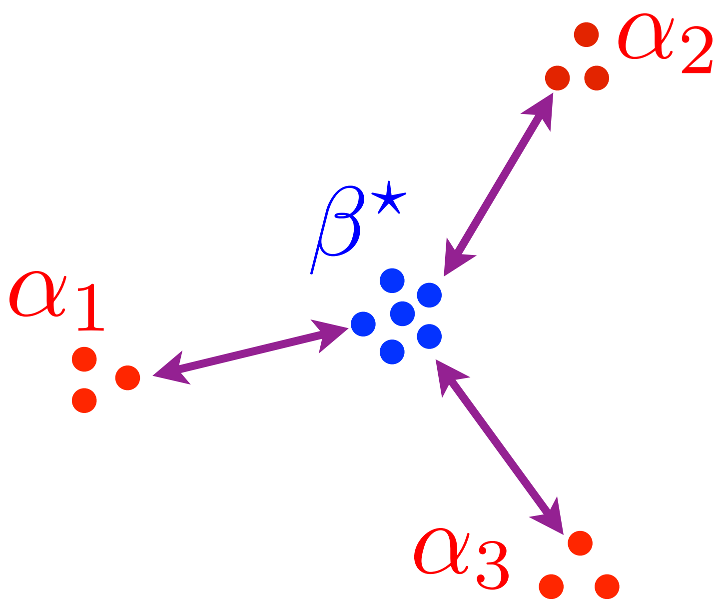
[Solomon et al, SIGGRAPH 2015]

Wasserstein Barycenters

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

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





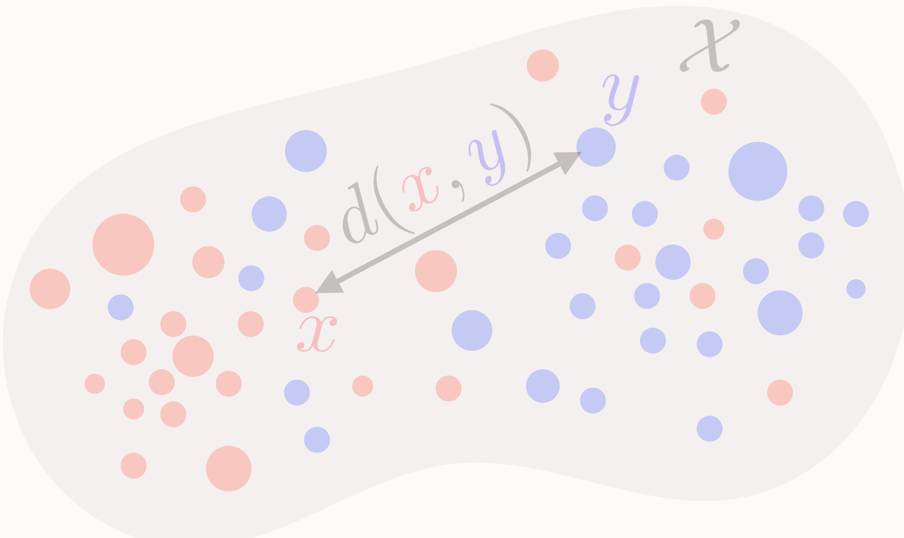
[Solomon et al, SIGGRAPH 2015]

Sinkhorn's algorithm:

$$\left(\mathbf{u}_s \leftarrow \frac{\mathbf{a}_s}{\mathbf{K}\mathbf{v}_s}\right)_s \quad \longleftrightarrow \quad \left(\mathbf{v}_s \leftarrow \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}_s}\right)_s$$

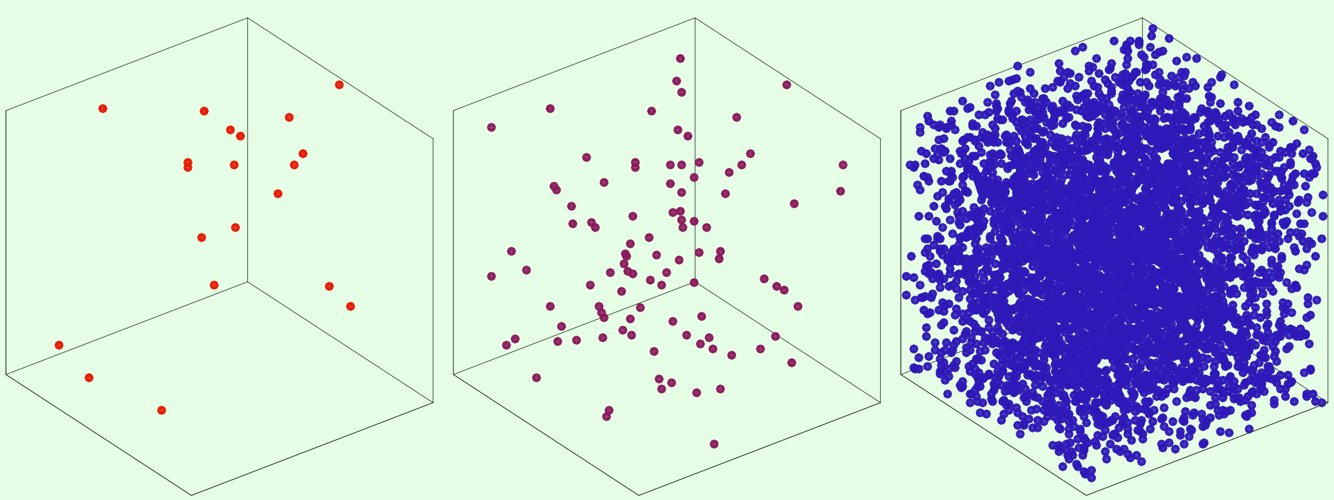
$$\mathbf{b} \leftarrow \prod_s (\mathbf{K}^\top \mathbf{u}_s)^{\lambda_s}$$

1. Optimal Transport

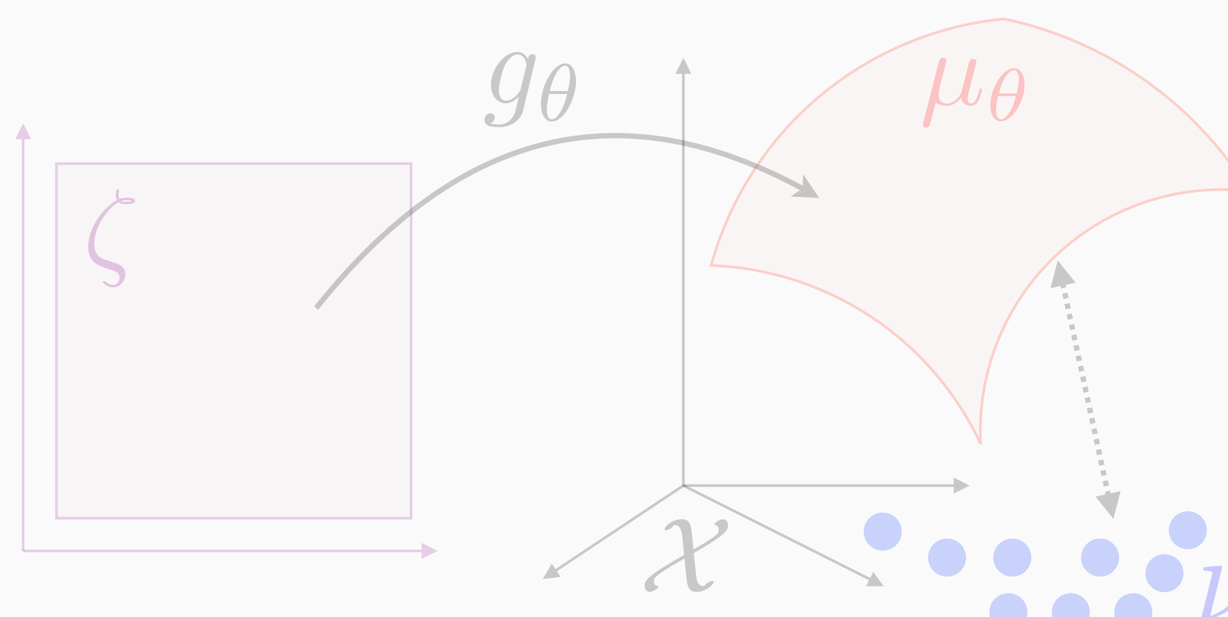


2. Entropic Regularization

3. Sinkhorn Divergences



4. Application to Generative Models

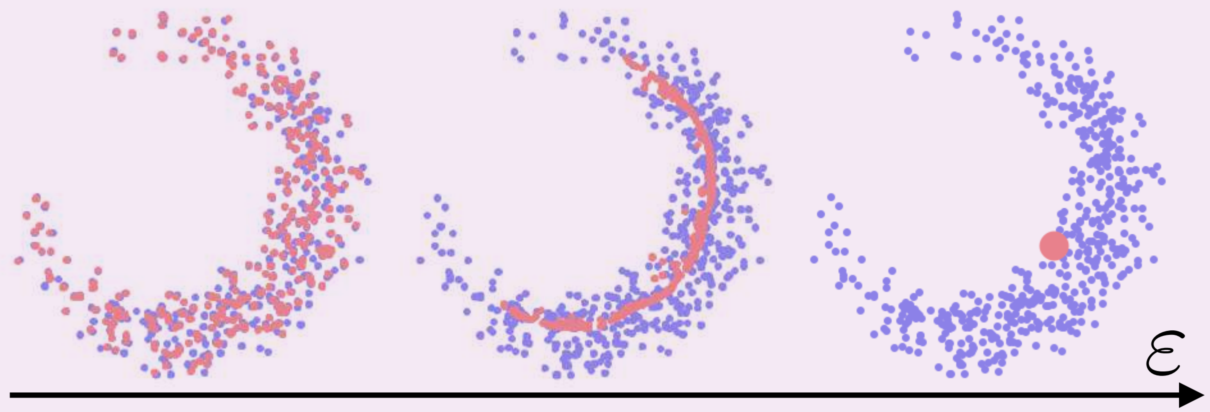


Sinkhorn Divergences

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

Problem: $\mathrm{W}_\varepsilon(\alpha,\alpha) \neq 0$

$$\min_{\alpha} \mathrm{W}^p_{\varepsilon,p}(\alpha,\beta)$$

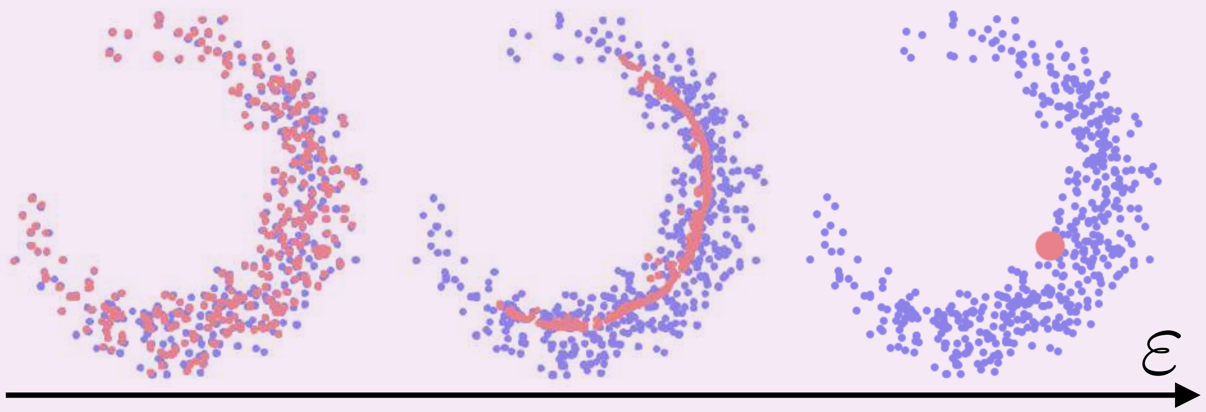


Sinkhorn Divergences

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

Problem: $\mathrm{W}_\varepsilon(\alpha,\alpha) \neq 0$

$$\min_{\alpha} \mathrm{W}^p_{\varepsilon,p}(\alpha,\beta)$$



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

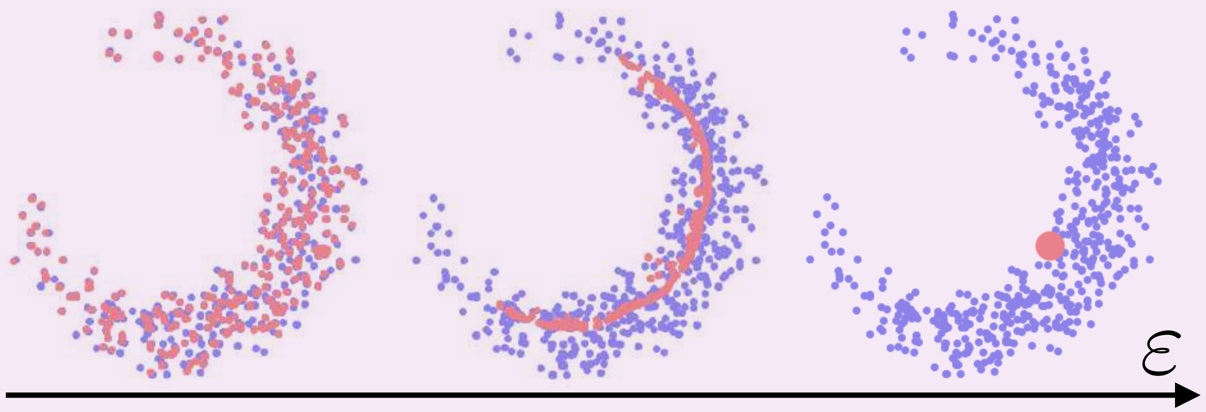
[Ramdas, García Trillos, Cuturi, 2017]

Sinkhorn Divergences

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

Problem: $\mathrm{W}_\varepsilon(\alpha,\alpha) \neq 0$

$$\min_{\alpha} \mathrm{W}^p_{\varepsilon,p}(\alpha,\beta)$$



$$\overline{\mathrm{W}}^p_{p,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} \mathrm{W}^p_{p,\varepsilon}(\alpha,\beta) - \tfrac{1}{2}\mathrm{W}^p_{p,\varepsilon}(\alpha,\alpha) - \tfrac{1}{2}\mathrm{W}^p_{p,\varepsilon}(\beta,\beta)$$

[Ramdas, García Trillos, Cuturi, 2017]

Theorem: $\mathrm{W}^p_p(\alpha,\beta) \xleftarrow{\varepsilon \to 0} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha,\beta) \xrightarrow{\varepsilon \to +\infty} \|\alpha-\beta\|^2_{-d^p}$

[Léonard 2012] [Carlier et al 2017]

[Ramdas, García Trillos, Cuturi, 2017]

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

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

Sinkhorn Divergences

$$\overline{\mathrm{W}}^p_{p,\varepsilon}(\alpha, \beta) \overset{\text{def.}}{=} \mathrm{W}^p_{p,\varepsilon}(\alpha, \beta) - \frac{1}{2}\mathrm{W}^p_{p,\varepsilon}(\alpha, \alpha) - \frac{1}{2}\mathrm{W}^p_{p,\varepsilon}(\beta, \beta)$$

concave   concave

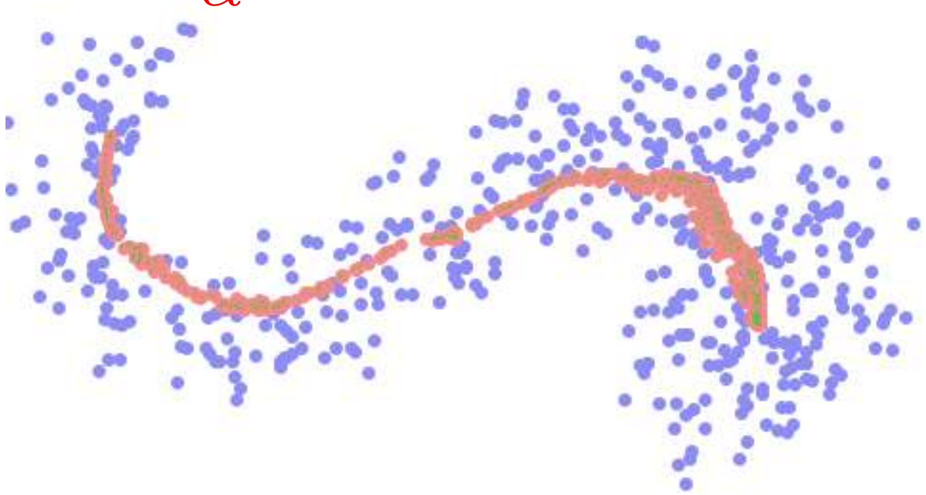
Theorem: [Feydy, Séjourné, P, Vialard, Trouvé, Amari 2018]

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

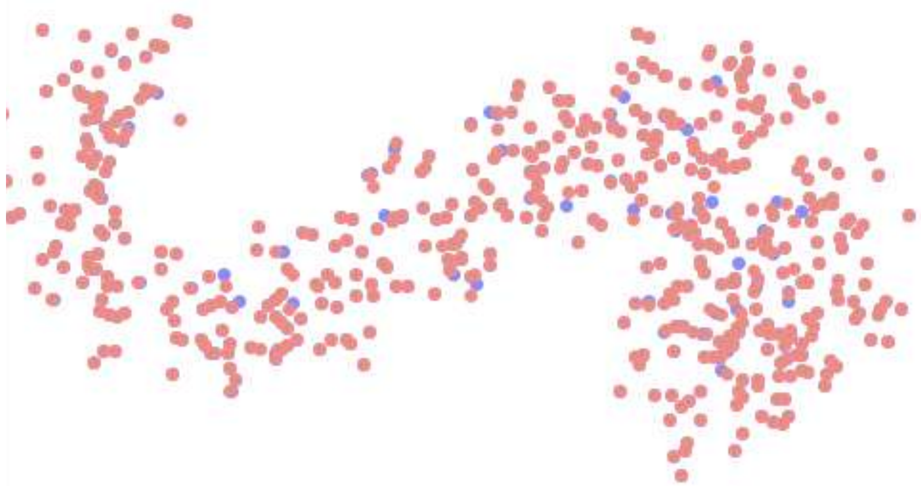
$\overline{\mathrm{W}}_{\varepsilon,p} \geqslant 0$ and $\overline{\mathrm{W}}^p_{\varepsilon,p}(\cdot, \beta)$ is convex.

$$\overline{\mathrm{W}}_{\varepsilon,p}(\alpha_n, \beta) \to 0 \iff \alpha_n \overset{\text{weak}*}{\rightharpoonup} \beta$$

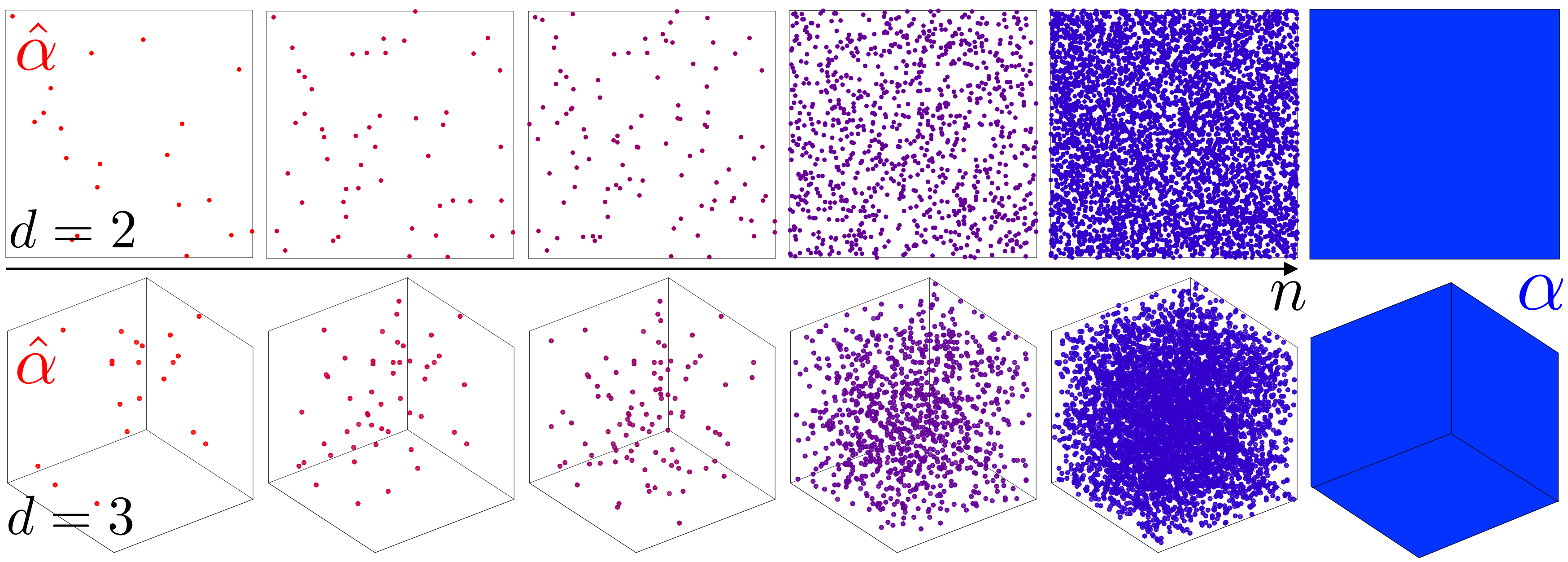
$\min_{\alpha} \mathrm{W}^p_{\varepsilon,p}(\alpha, \beta)$

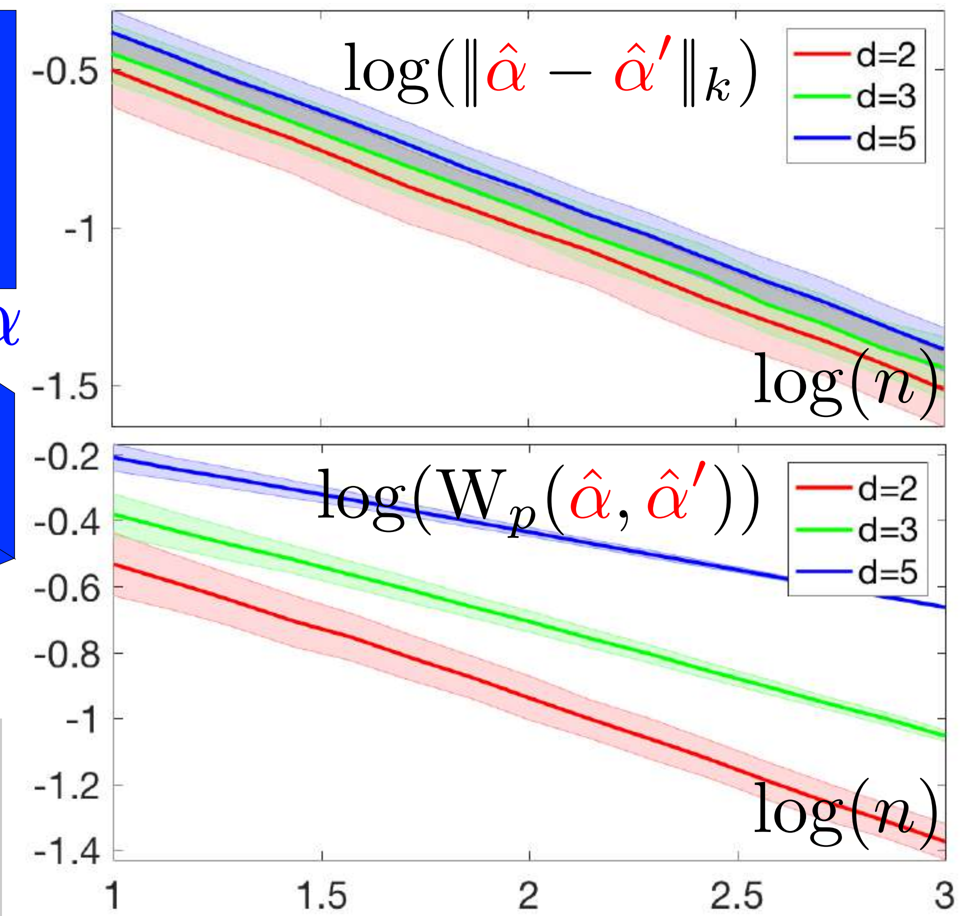


$\min_{\alpha} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha, \beta)$



Sample Complexity





Theorem:

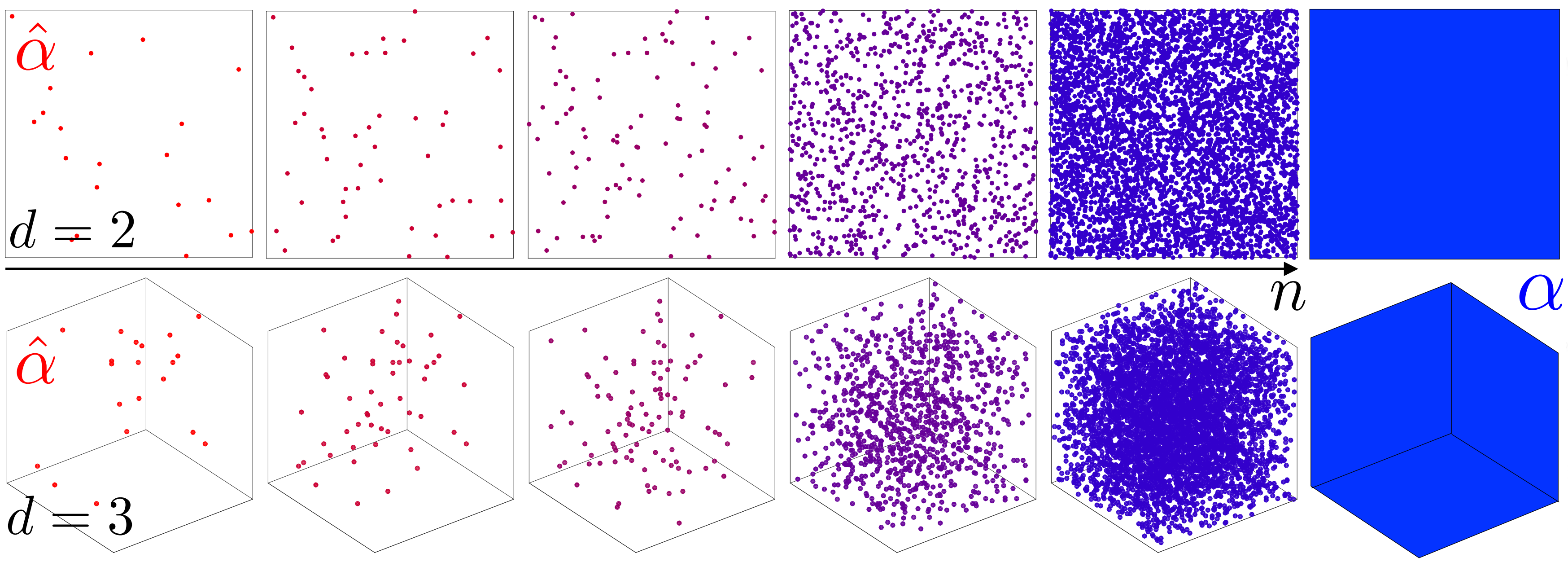
$$\mathbb{E}(|W_p(\hat\alpha, \hat\beta) - W_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$$

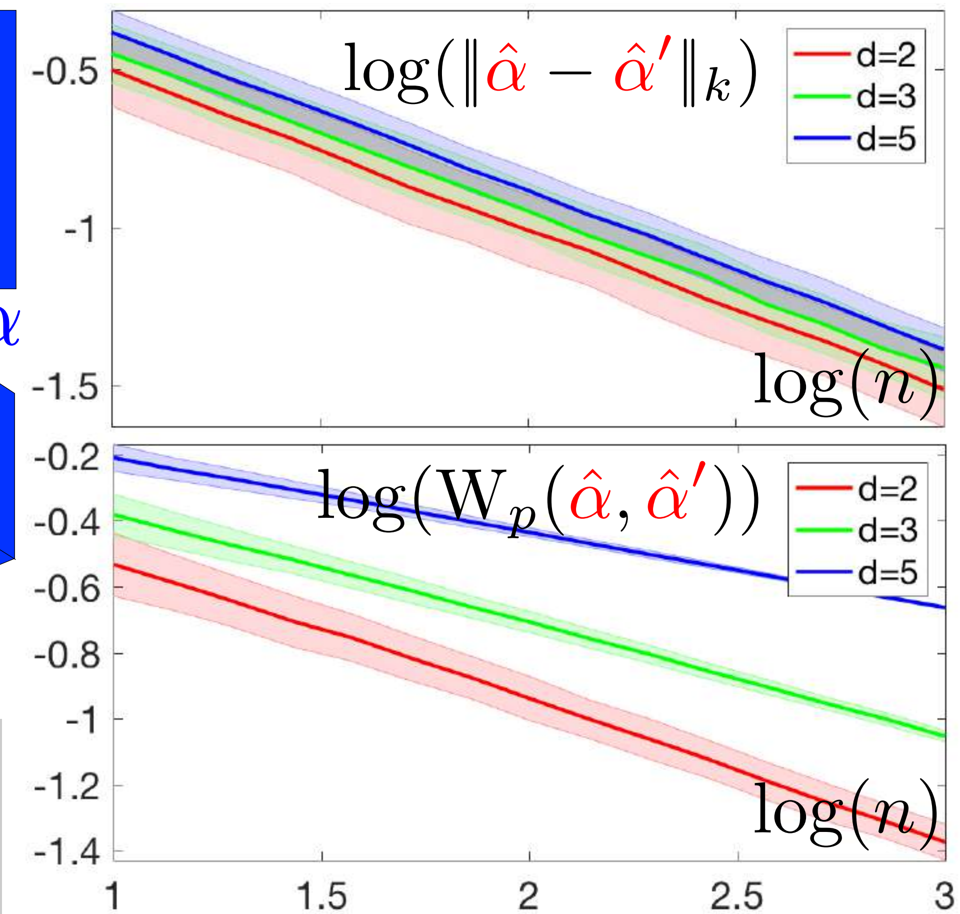
$$\mathbb{E}(|\|\hat\alpha - \hat\beta\|_k - \|\alpha - \beta\|_k|) = O(n^{-\frac{1}{2}})$$

Optimal transport: suffers from curse of dimensionality.

$\rightarrow$ Adapt to support dimensionality [Weed, Bach 2017]

Sample Complexity

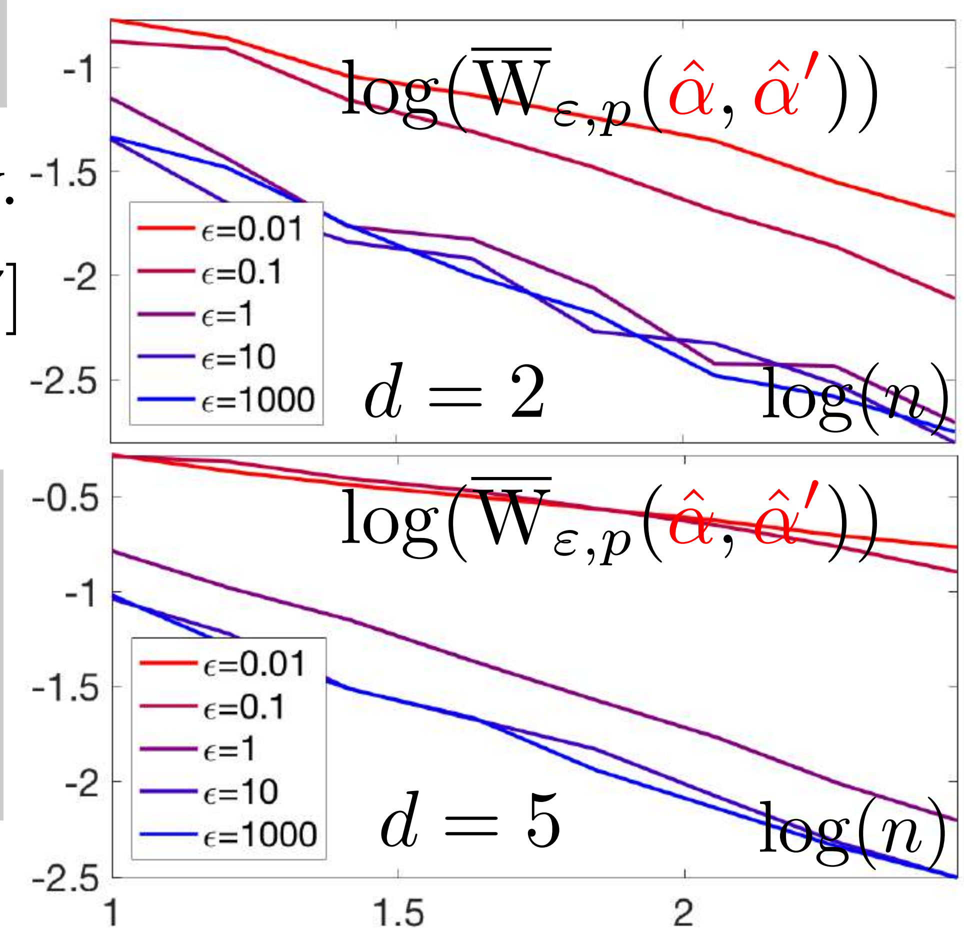




Theorem:
$$\mathbb{E}(|W_p(\hat{\alpha}, \hat{\beta}) - W_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$$
$$\mathbb{E}(|\|\hat{\alpha} - \hat{\beta}\|_k - \|\alpha - \beta\|_k|) = O(n^{-\frac{1}{2}})$$

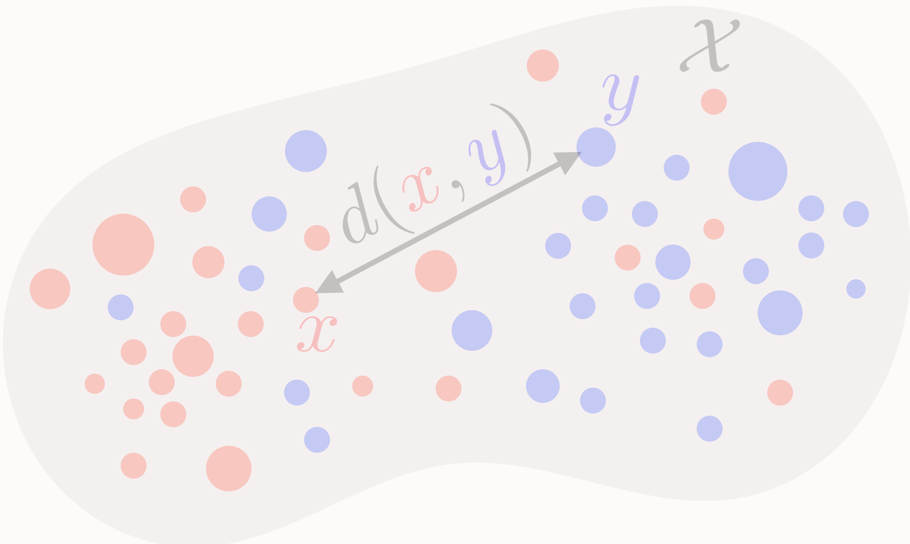
Optimal transport: suffers from curse of dimensionality.

→ Adapt to support dimensionality [Weed, Bach 2017]



Theorem: [Genevay, Bach, P, Cuturi]
$$\mathbb{E}(|\overline{W}_{\varepsilon,p}(\hat{\alpha}, \hat{\beta}) - \overline{W}_{\varepsilon,p}(\alpha, \beta)|) = O(\varepsilon^{-\frac{d}{2}} n^{-\frac{1}{2}})$$

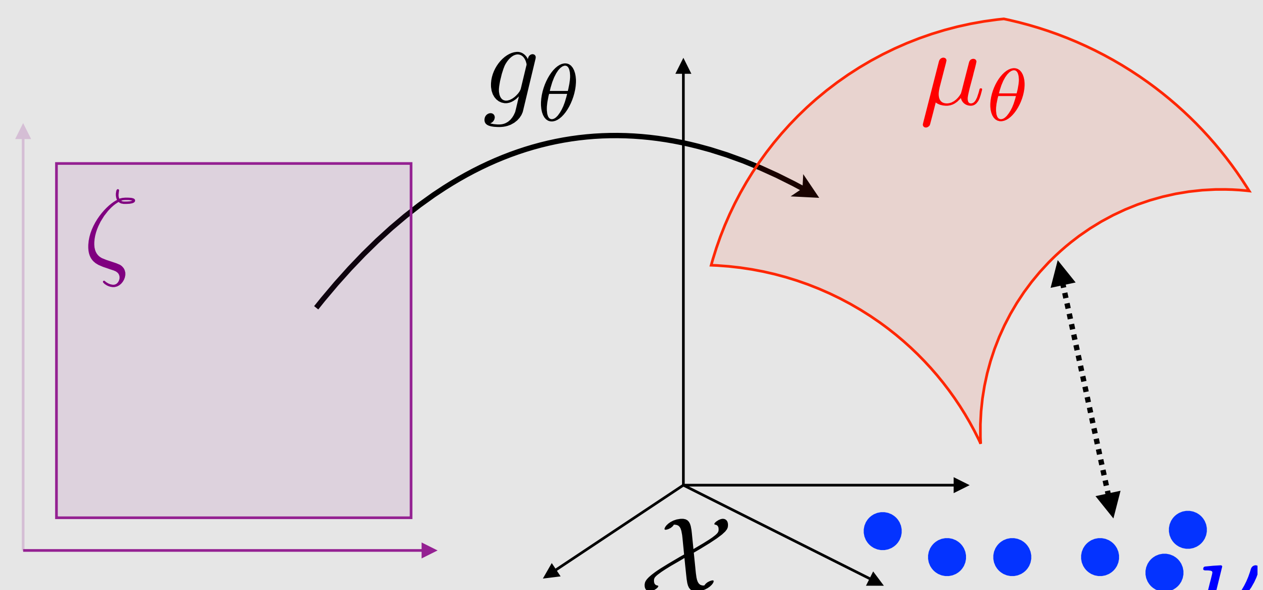
1. Optimal Transport



2. Entropic Regularization

3. Sinkhorn Divergences

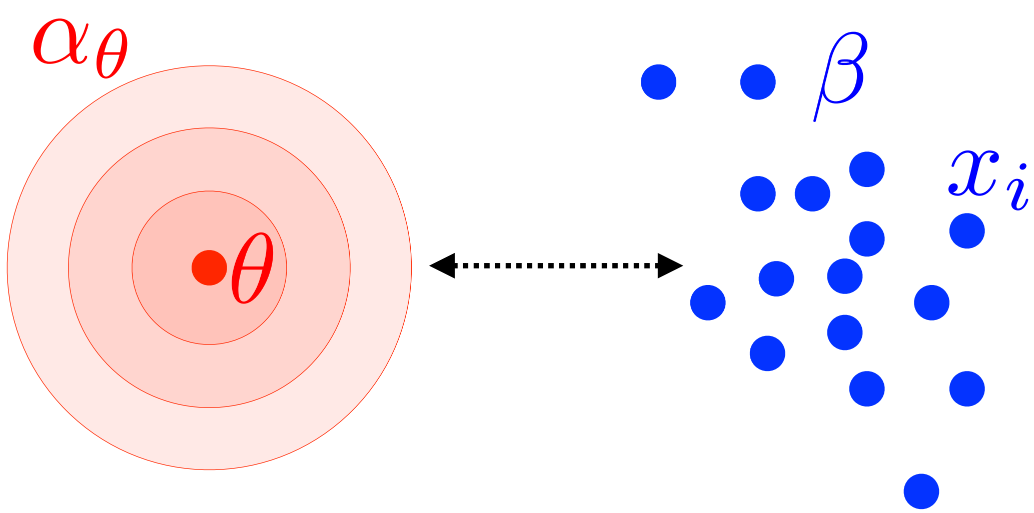
4. Application to Generative Models



Density Fitting and Generative Models

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

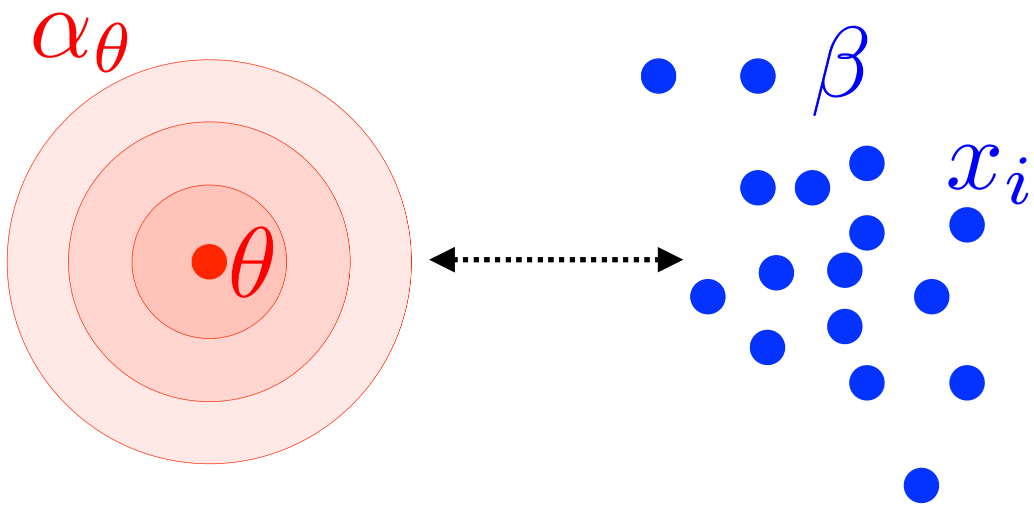
Parametric model: $\theta \mapsto \alpha_\theta$



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: $\mathrm{d}\alpha_\theta(x) = \rho_\theta(x)\mathrm{d}x$

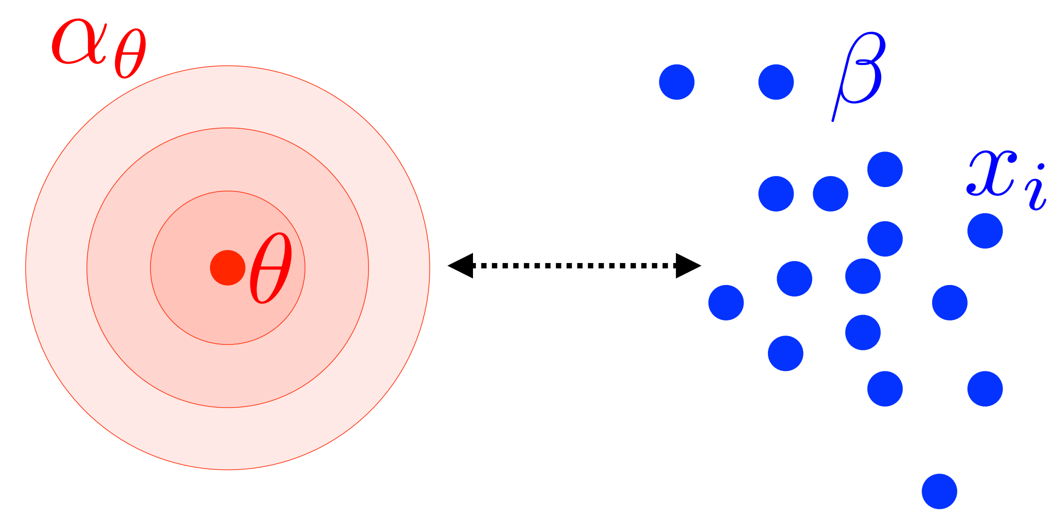
$$\min_\theta \widehat{\mathrm{KL}}(\alpha_\theta|\beta) \stackrel{\text{def.}}{=} -\sum_i \log(\rho_\theta(x_i))$$

Maximum likelihood (MLE)

Density Fitting and Generative Models

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

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



Density fitting: $\mathrm{d}\alpha_\theta(x) = \rho_\theta(x)\mathrm{d}x$

$$\min_\theta \widehat{\mathrm{KL}}(\alpha_\theta|\beta) \overset{\text{def.}}{=} -\sum_i \log(\rho_\theta(x_i))$$

Maximum likelihood (MLE)

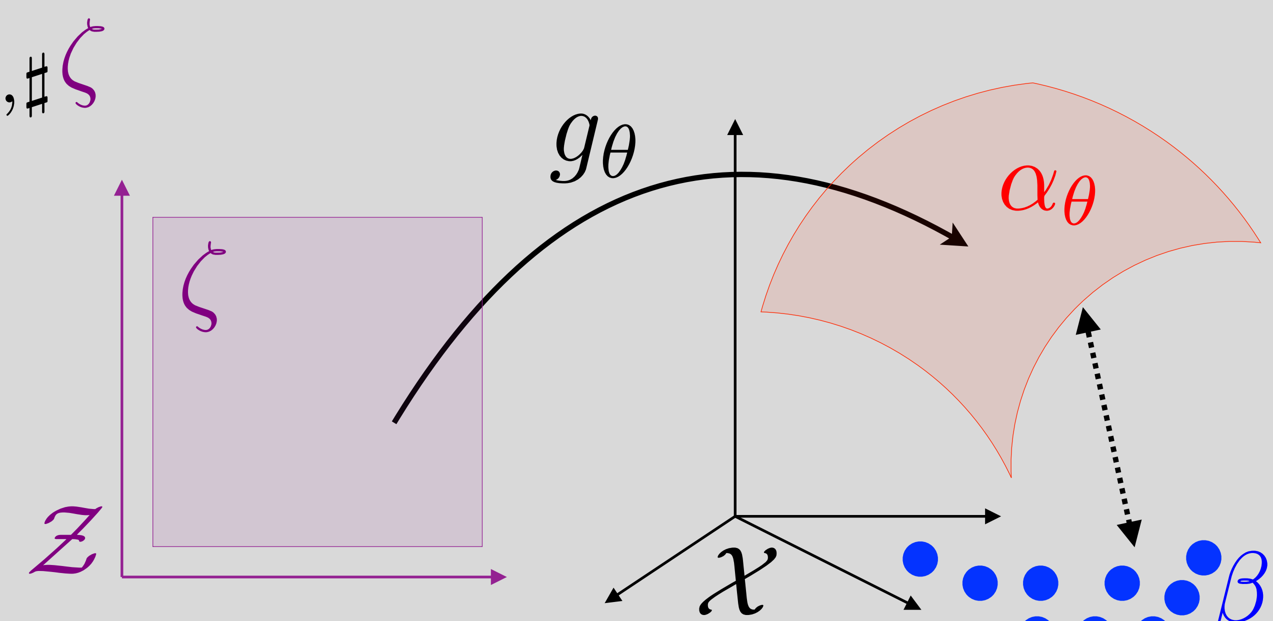
Generative model fit: $\alpha_\theta = g_{\theta,\sharp}\zeta$

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

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.

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

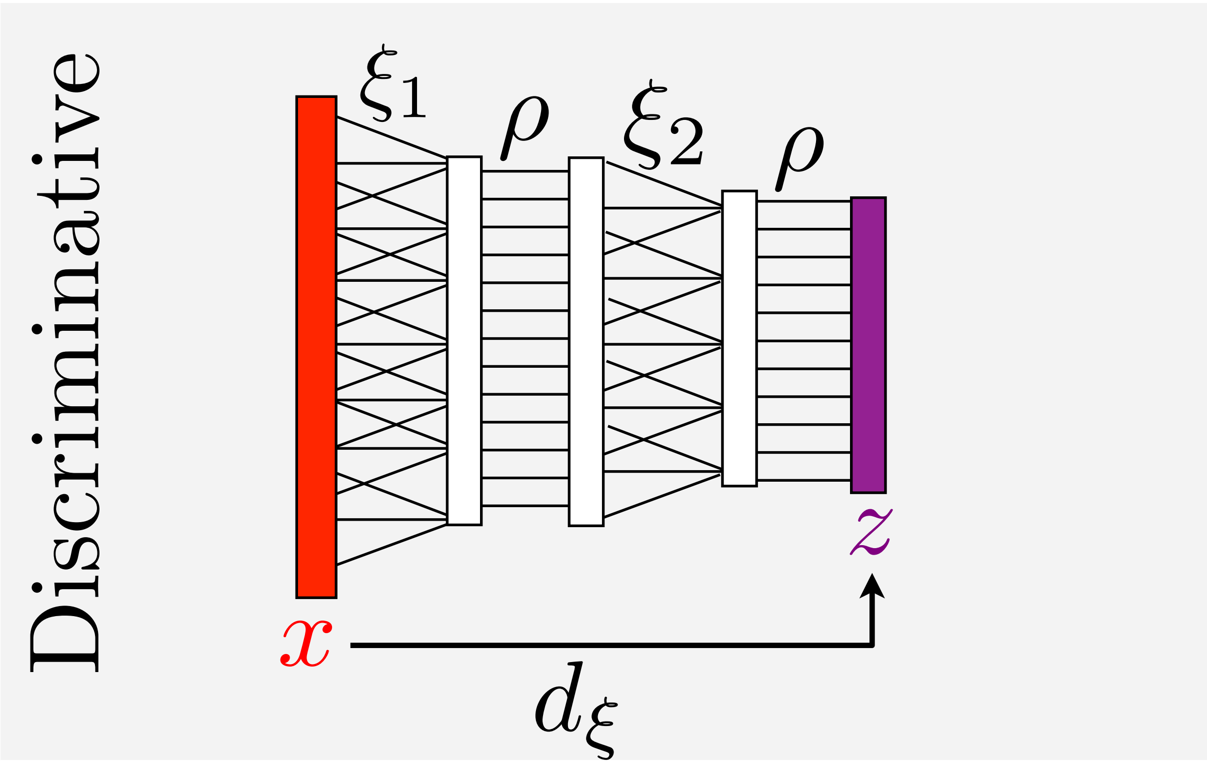


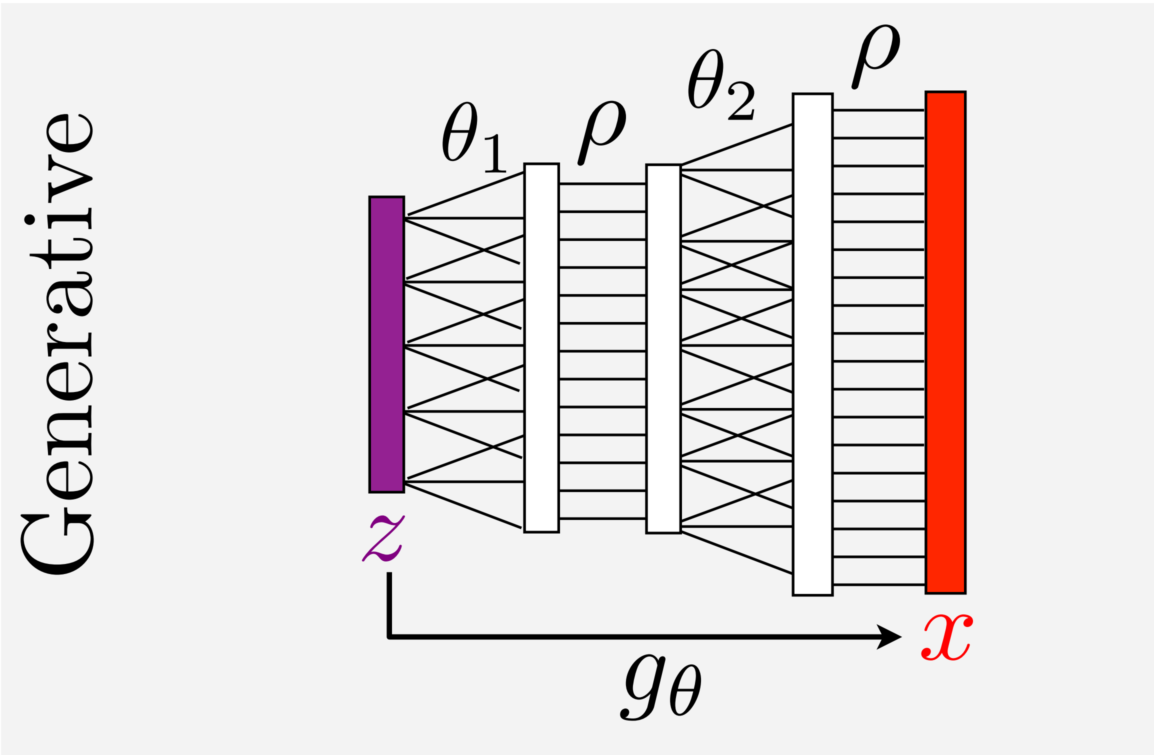
Deep Discriminative vs Generative Models

Deep networks:

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

$$g_\theta(z) = \rho(\theta_K(\ldots\rho(\theta_2(\rho(\theta_1(z))\ldots)$$



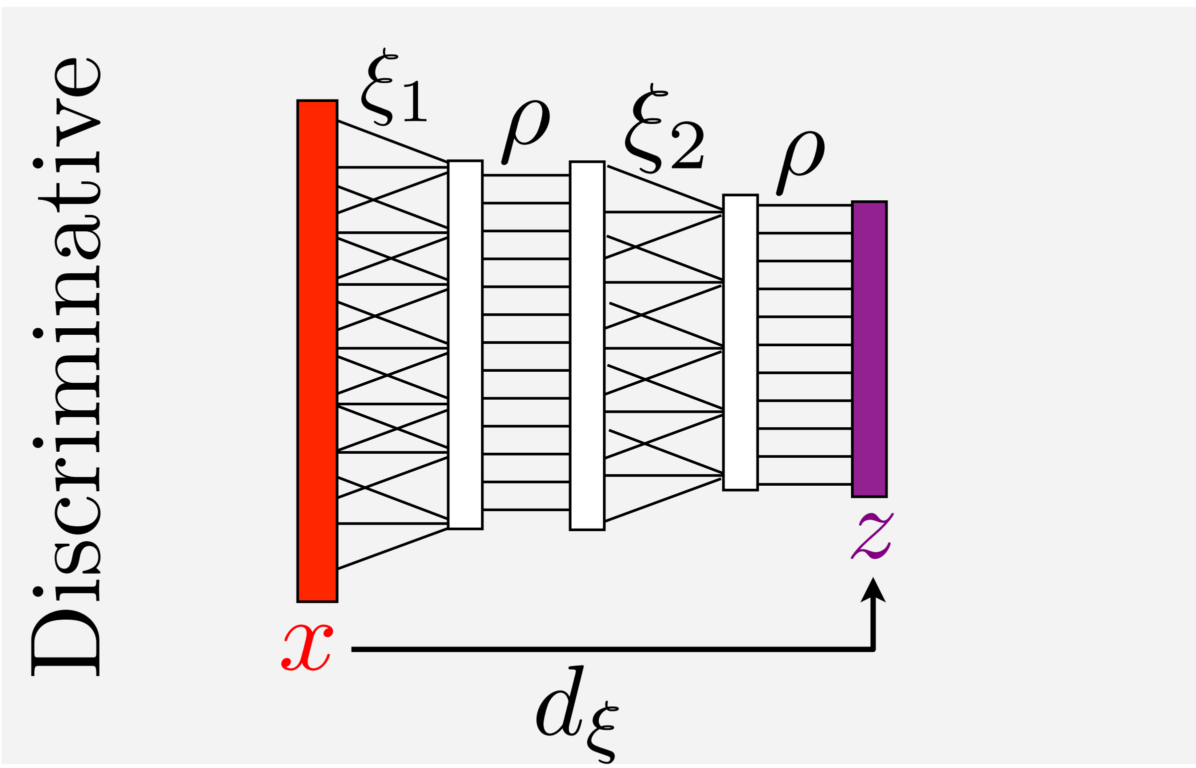


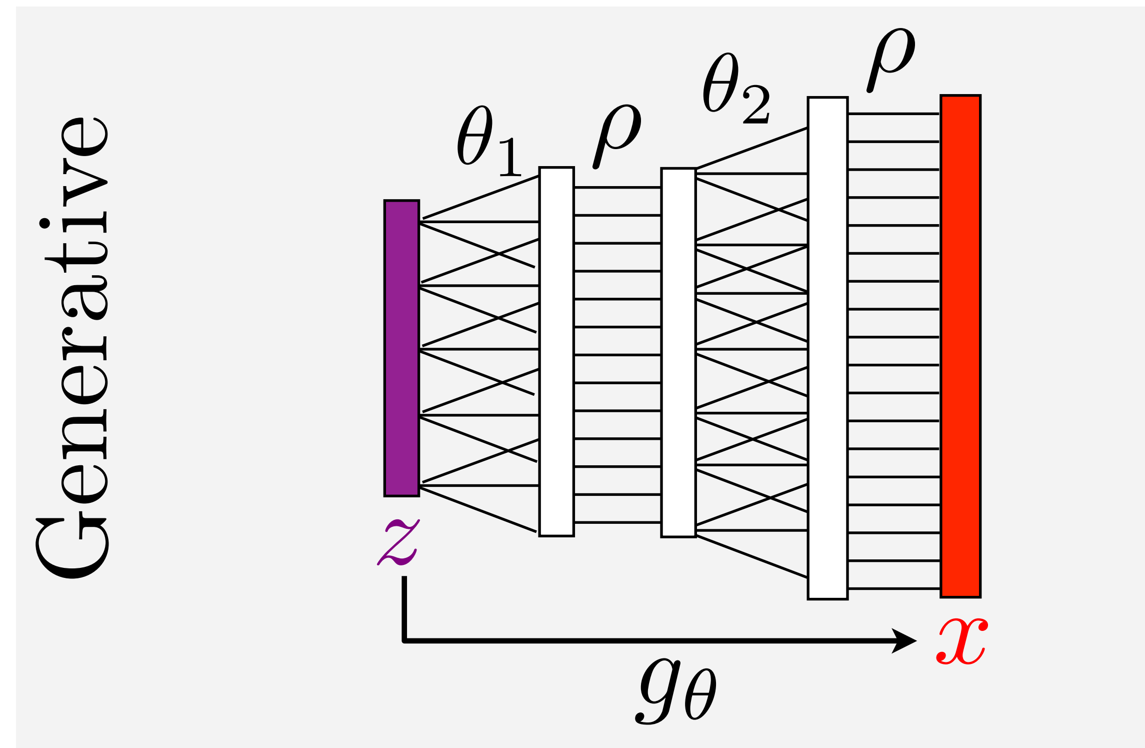
Deep Discriminative vs Generative Models

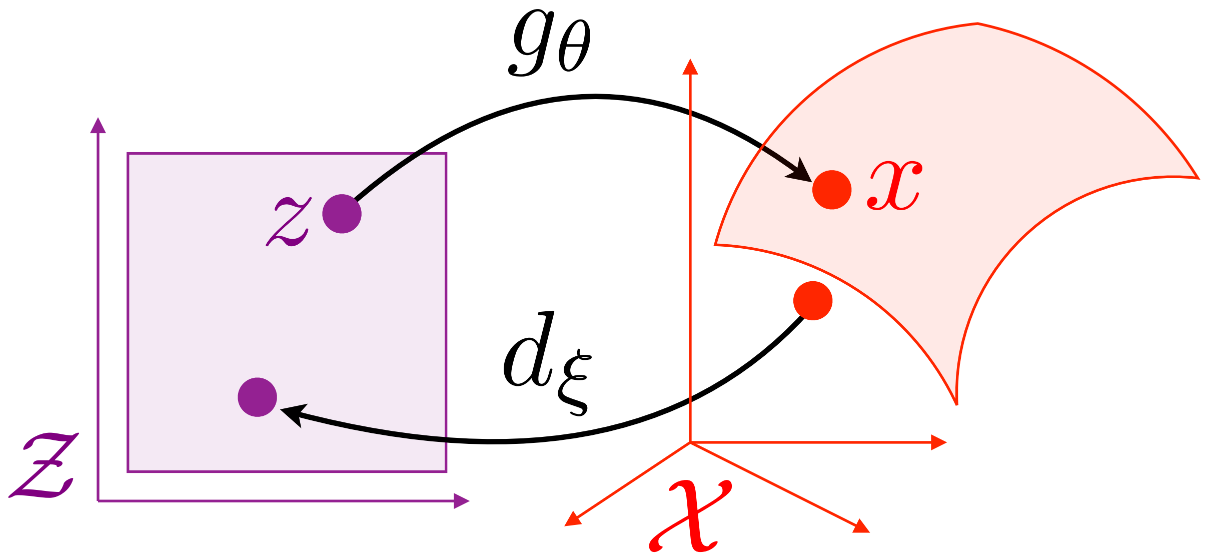
Deep networks:

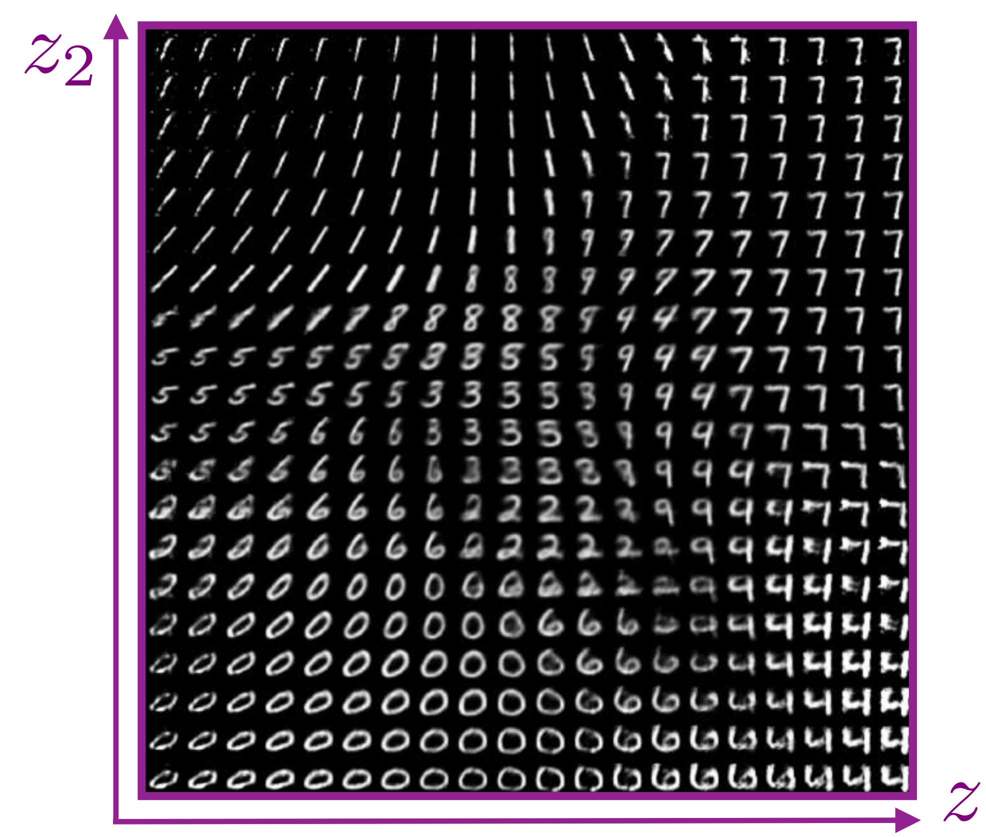
$$d_{\xi}(x) = \rho(\xi_K(\dots\rho(\xi_2(\rho(\xi_1(x))\dots)$$

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

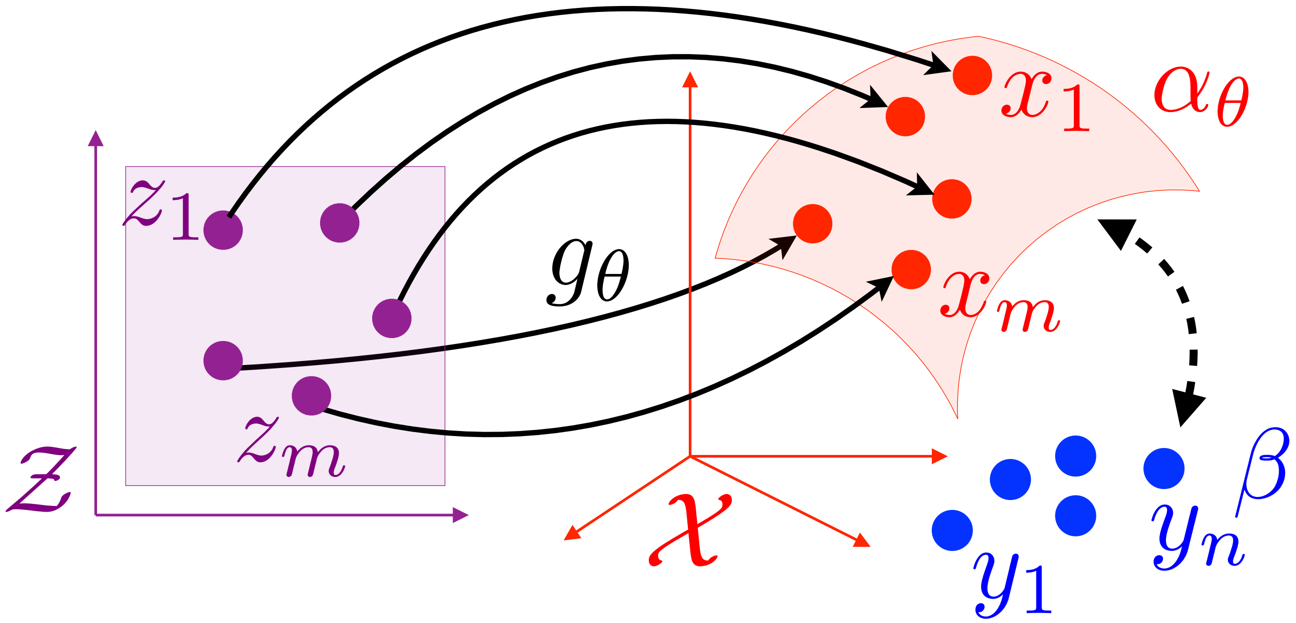








Training Architecture



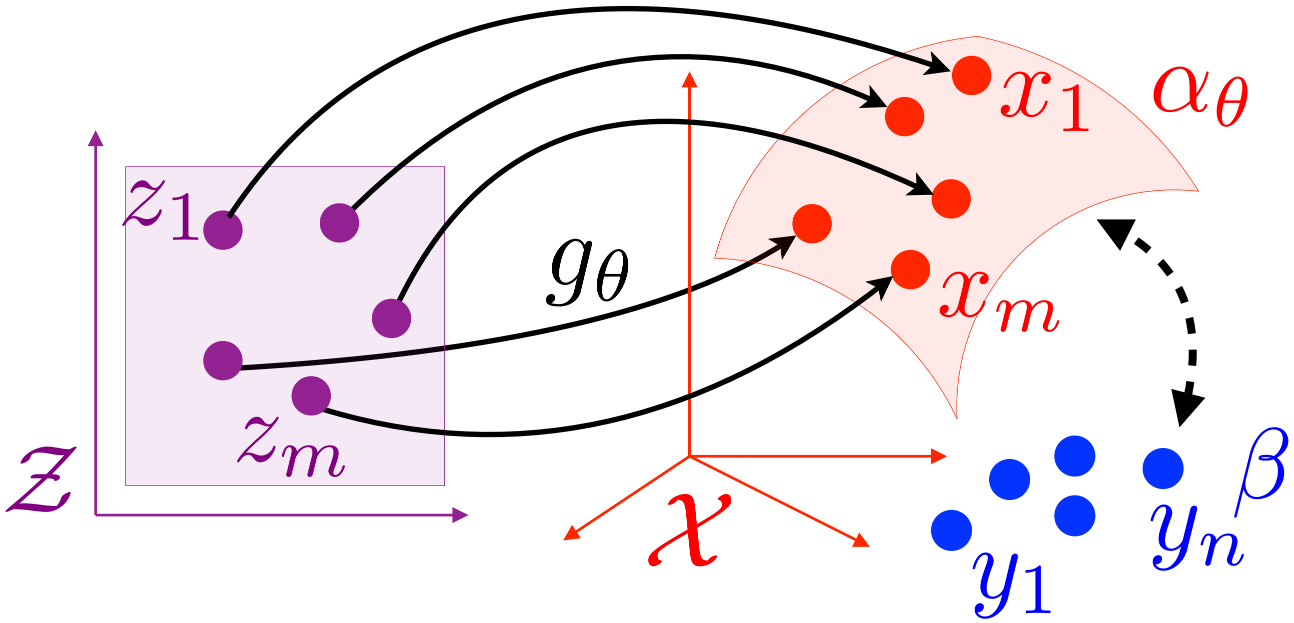
$$\min_\theta \mathcal{E}(\theta) \overset{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

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

$$\hat{\mathcal{E}}(\theta) \overset{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\tfrac{1}{m} \textstyle\sum_i \delta_{g_\theta(z_i)}, \beta)$$

Training Architecture

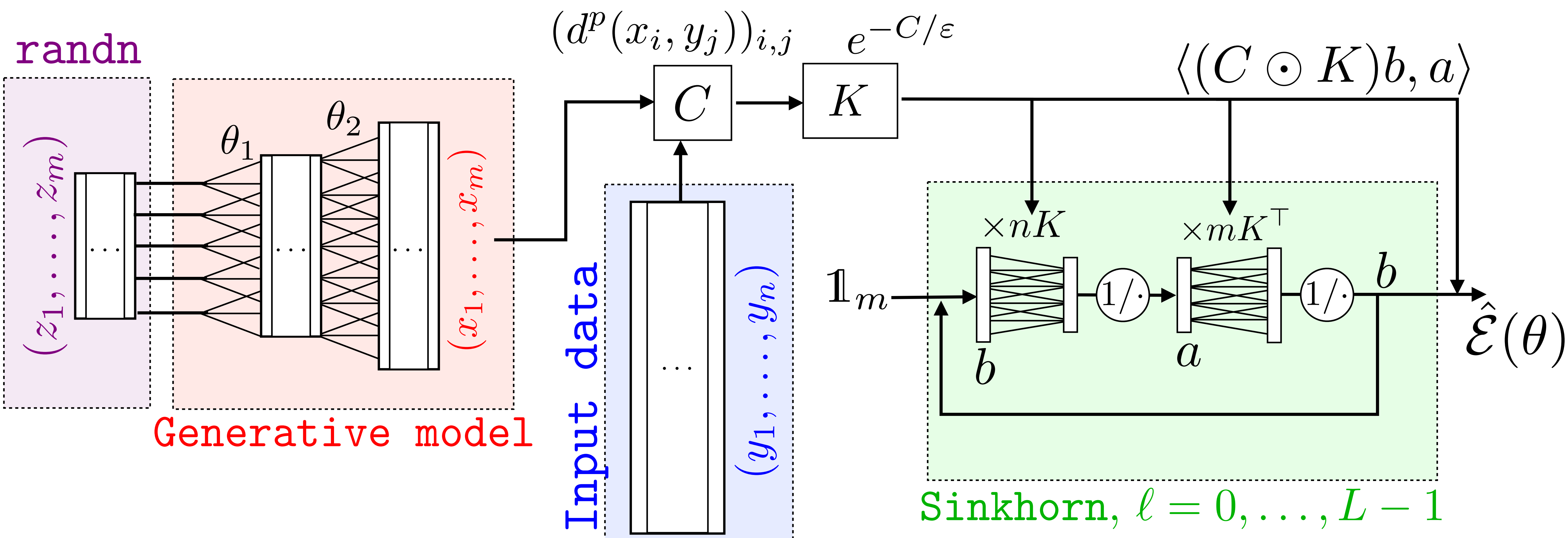


$$\min_\theta \mathcal{E}(\theta) \stackrel{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

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

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



Automatic Differentiation

Setup: $f : \mathbb{R}^n \to \mathbb{R}$ computable in K operations.

Hypothesis: elementary operations $(a \times b, \log(a), \sqrt{a} \ldots)$ and their derivatives cost $O(1)$.

Question: What is the complexity of computing $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$?

Automatic Differentiation

Setup: $f : \mathbb{R}^n \to \mathbb{R}$ computable in K operations.

Hypothesis: elementary operations $(a \times b, \log(a), \sqrt{a} \dots)$ and their derivatives cost $O(1)$.

Question: What is the complexity of computing $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$?

Finite differences: $$\nabla f(x) \approx \frac{1}{\varepsilon}(f(x + \varepsilon\delta_1) - f(x), \dots f(x + \varepsilon\delta_n) - f(x))$$

$K(n+1)$ operations, intractable for large n.

Automatic Differentiation

Setup: $f : \mathbb{R}^n \to \mathbb{R}$ computable in K operations.

Hypothesis: elementary operations $(a \times b, \log(a), \sqrt{a} \dots)$ and their derivatives cost $O(1)$.

Question: What is the complexity of computing $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$?

Finite differences: $$\nabla f(x) \approx \frac{1}{\varepsilon}(f(x+\varepsilon\delta_1) - f(x), \dots f(x+\varepsilon\delta_n) - f(x))$$
$K(n+1)$ operations, intractable for large n.

Theorem: there is an algorithm to compute ∇f in $O(K)$ operations. [Seppo Linnainmaa, 1970]

This algorithm is reverse mode automatic differentiation

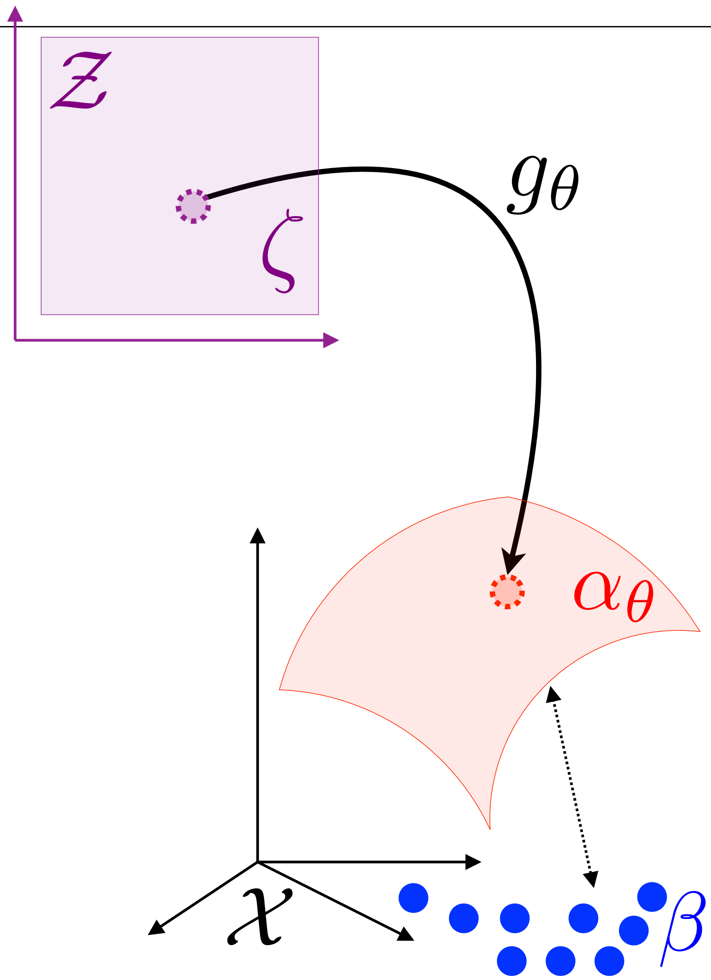
→ it is not numerical calculus (exact computations).

→ it is not formal calculus (algorithms matter).

Examples of Images Generation

Inputs β

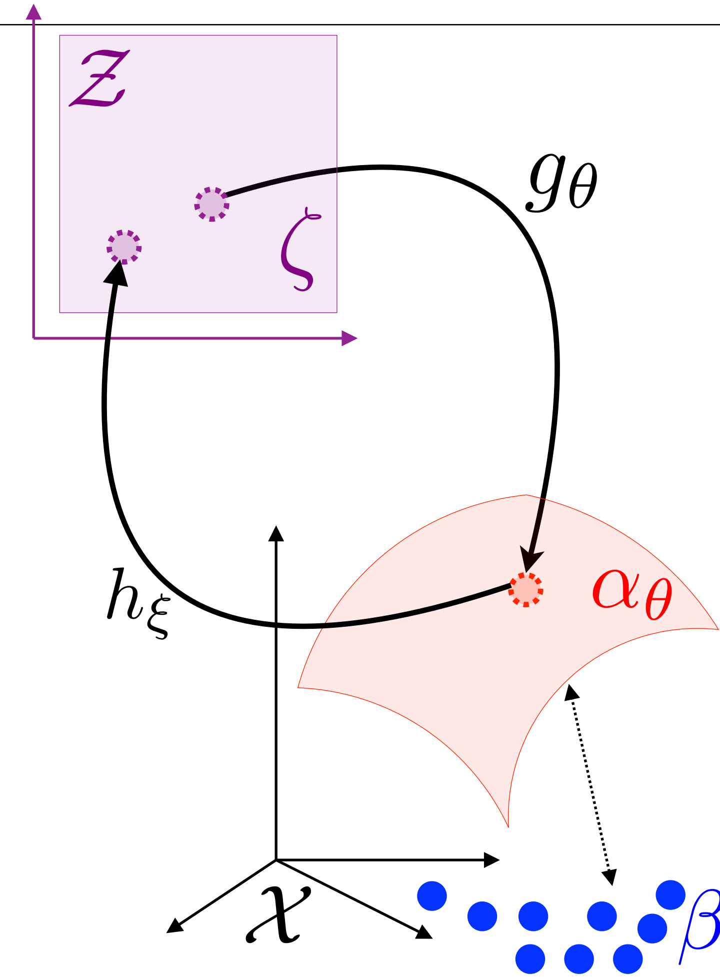
Generated α_θ

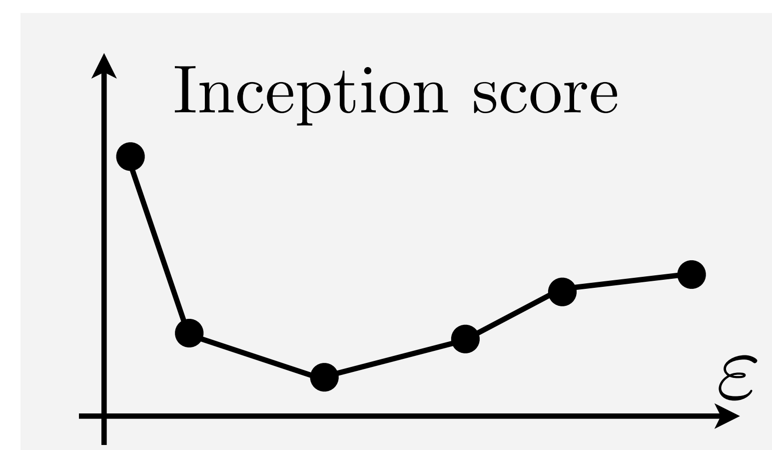


Examples of Images Generation

Inputs β

Generated α_θ



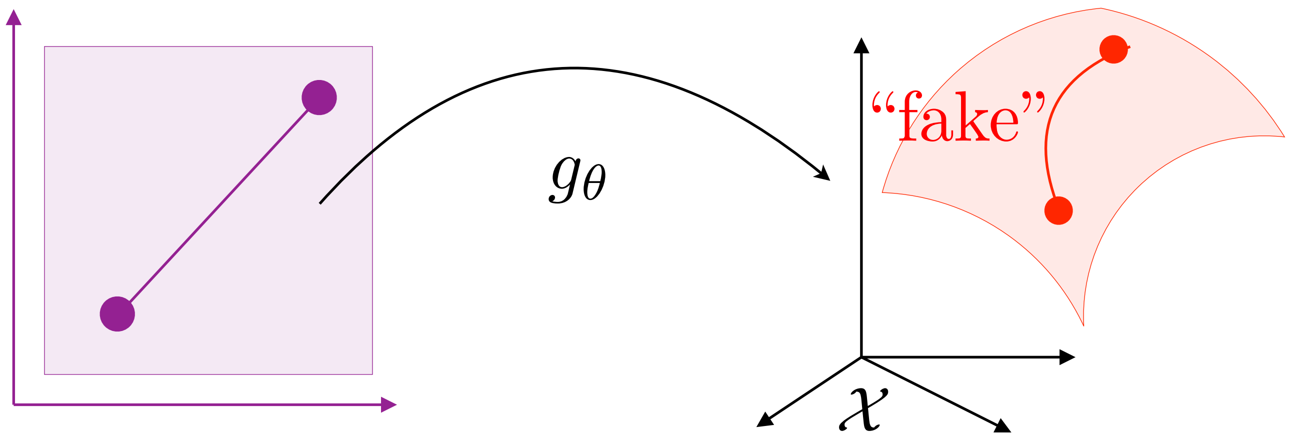


$\rightarrow$ Need to learn the metric $d(x, y) = \|h_\xi(x) - h_\xi(y)\|$ (GANs)

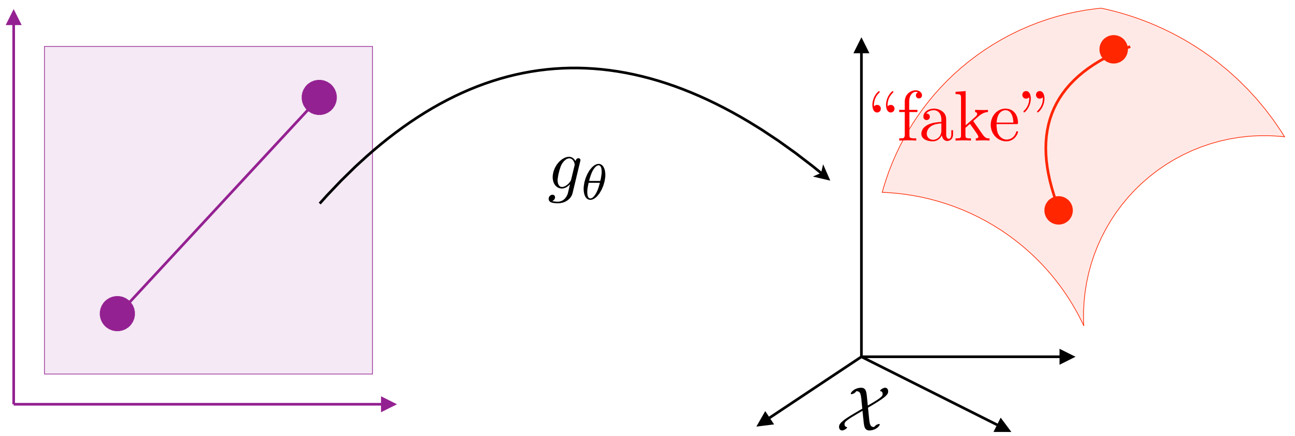
$\rightarrow$ Influence of ε?

$\rightarrow$ Performance evaluation of generative models is an open problem.

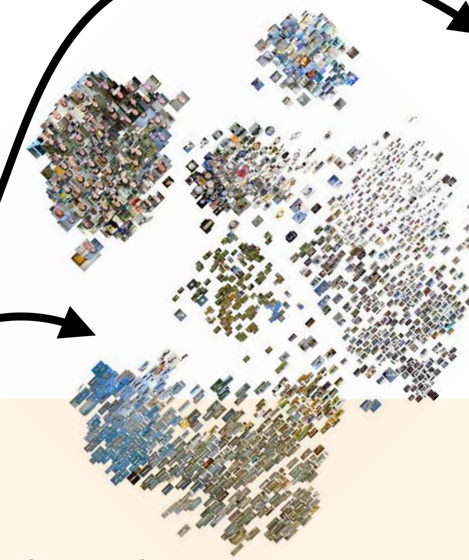
Progressive Growing of GANs for Improved Quality, Stability, and Variation
Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, ICLR 2018



Progressive Growing of GANs for Improved Quality, Stability, and Variation
Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, ICLR 2018



Open Problems

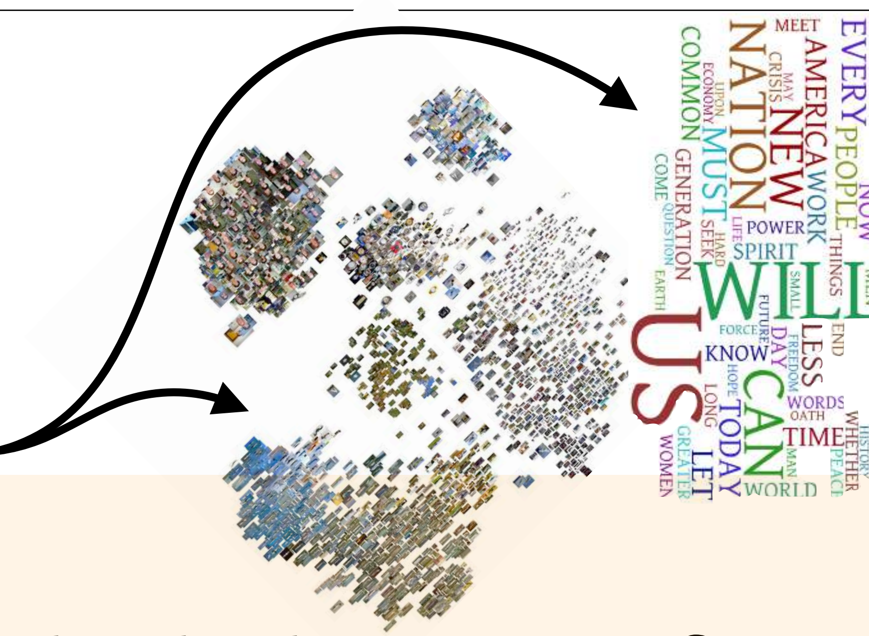


Toward high-dimensional OT:

→ Scalable geometrical loss functions in high dimension?

→ Performance quality measures for unsupervised learning?

Open Problems



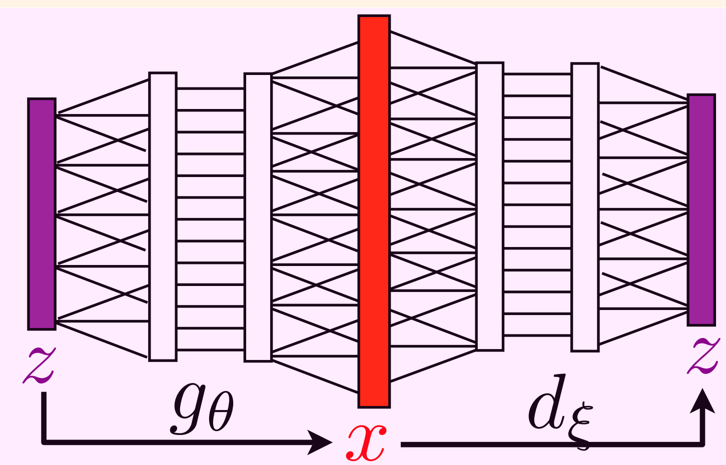
Toward high-dimensional OT:

→ Scalable geometrical loss functions in high dimension?

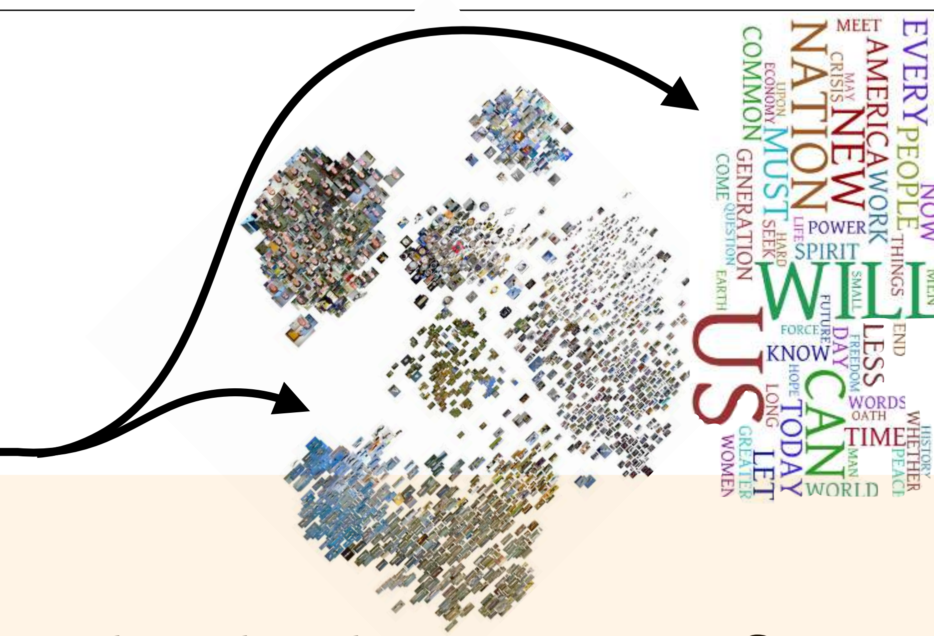
→ Performance quality measures for unsupervised learning?

Metric learning for OT:

→ Adversarial training to leverage multi scale priors?



Open Problems



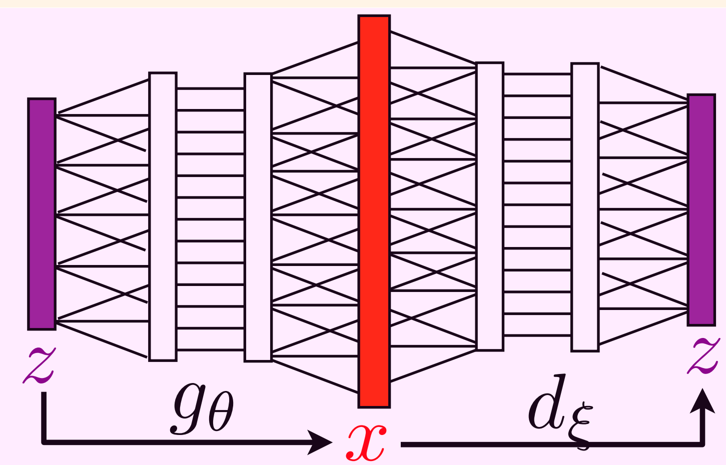
Toward high-dimensional OT:

→ 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?