Optimal Transport for Imaging and Learning

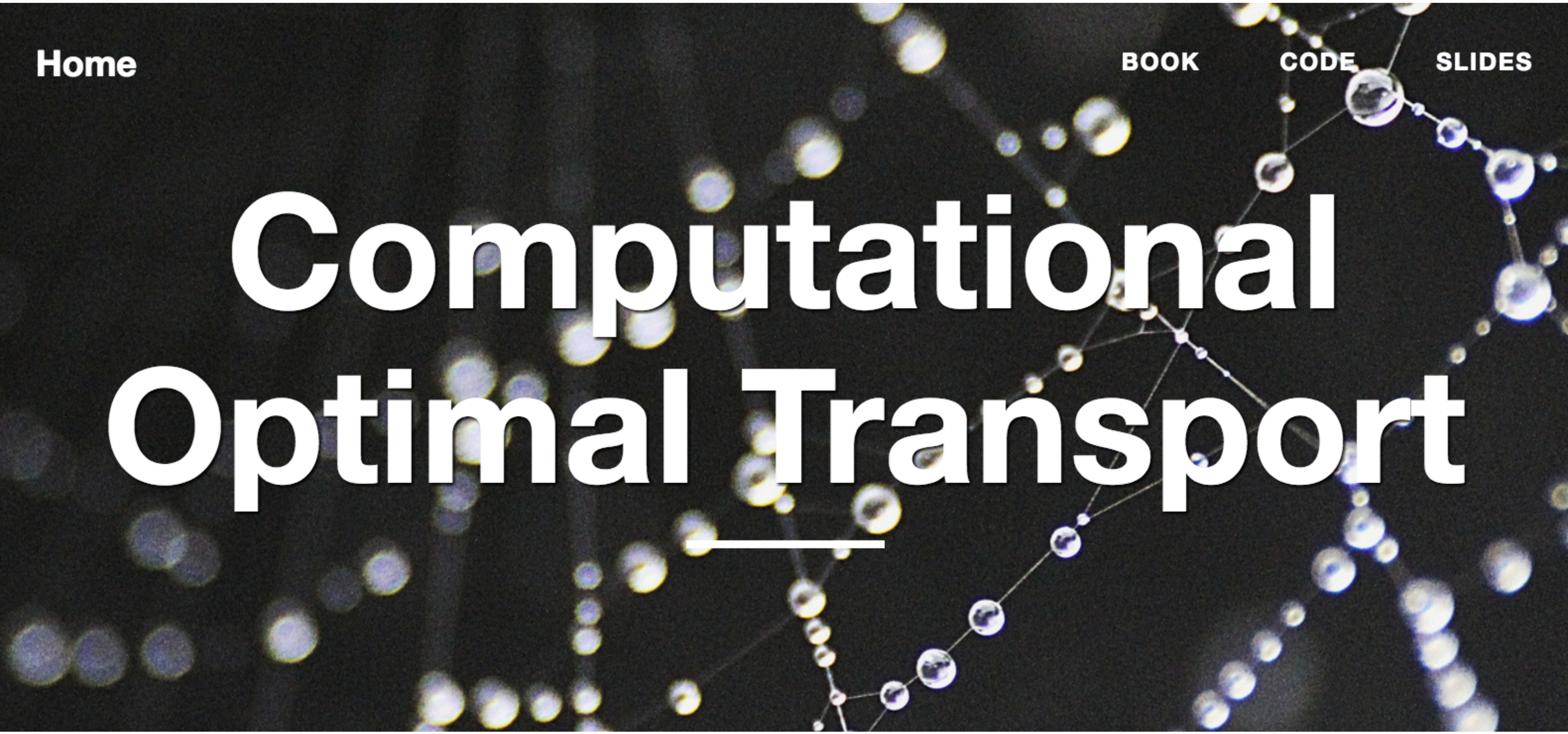
Gabriel Peyré

cnrs dépasser les frontières

ENS
ÉCOLE NORMALE SUPÉRIEURE

Joint works with:

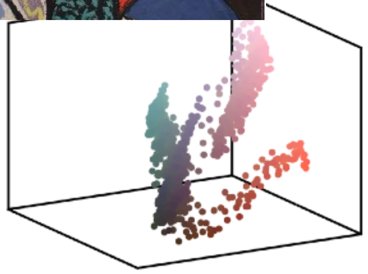
https://optimaltransport.github.io

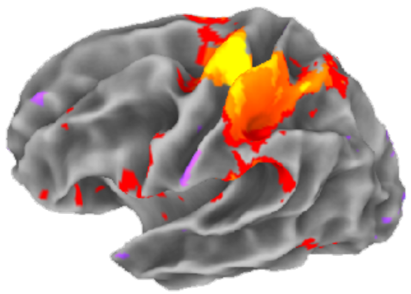


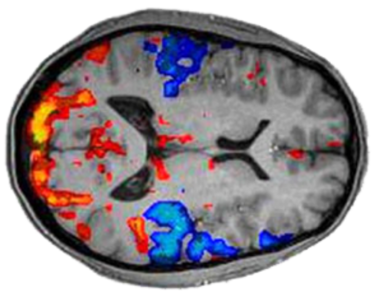
Probability Distributions in Data Sciences

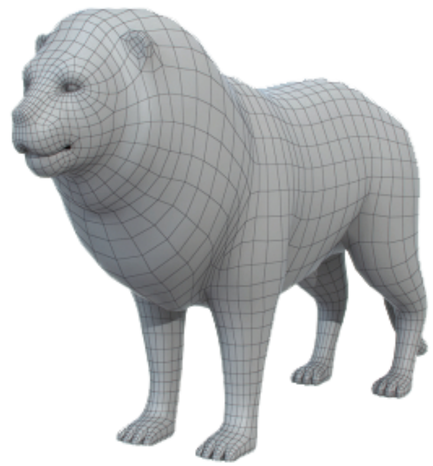
Probability distributions and histograms

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





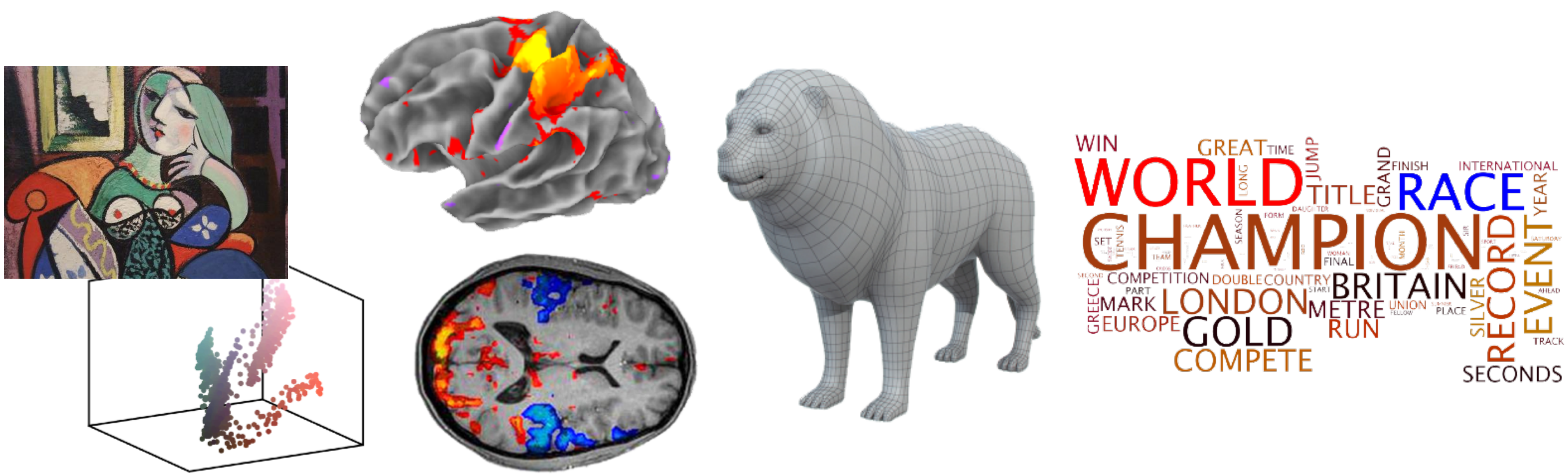




Probability Distributions in Data Sciences

Probability distributions and histograms

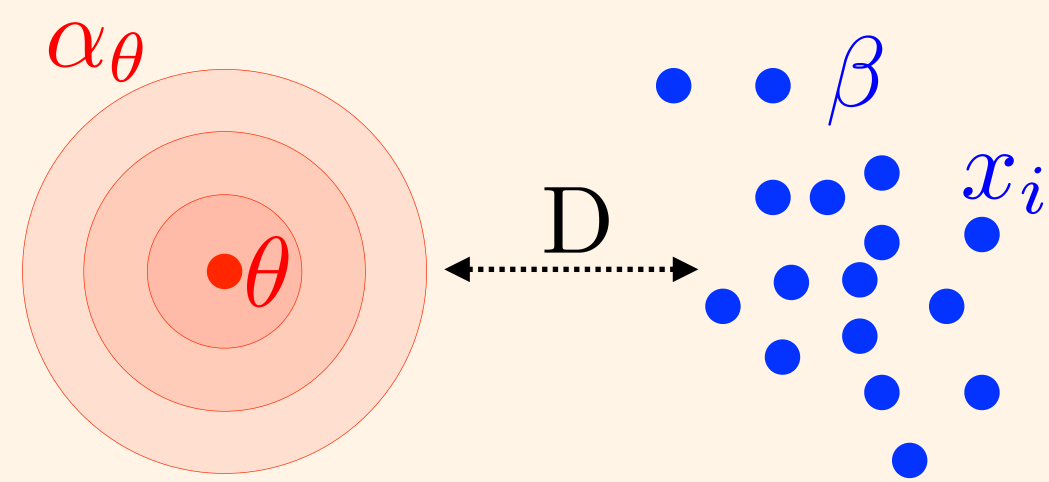
→ images, vision, graphics and machine learning, . . .



Unsupervised learning

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

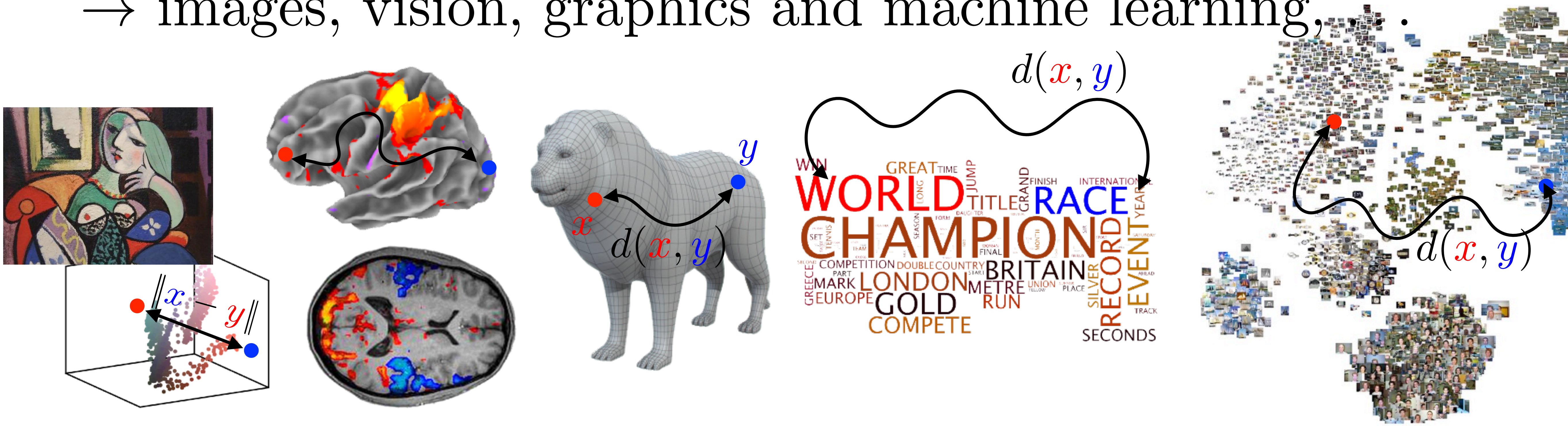
Parametric model: $\theta \mapsto \alpha_\theta$



Probability Distributions in Data Sciences

Probability distributions and histograms

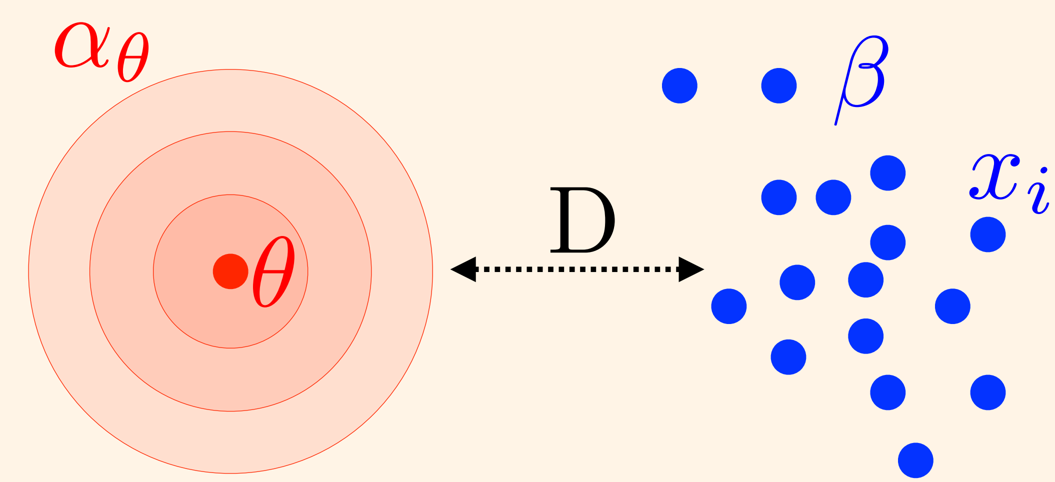
$\rightarrow$ images, vision, graphics and machine learning, . . .



Unsupervised learning

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

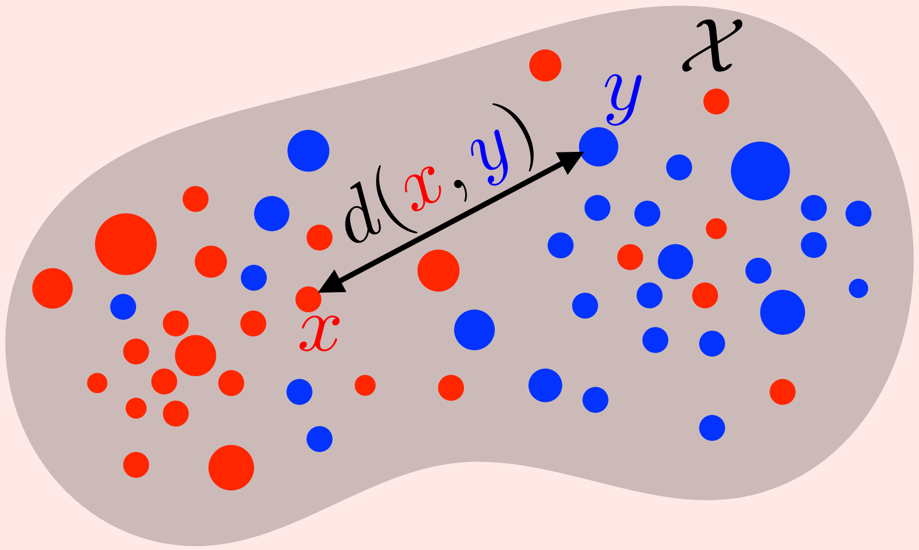
Parametric model: $\theta \mapsto \alpha_\theta$



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

$\rightarrow$ takes into account a metric d.

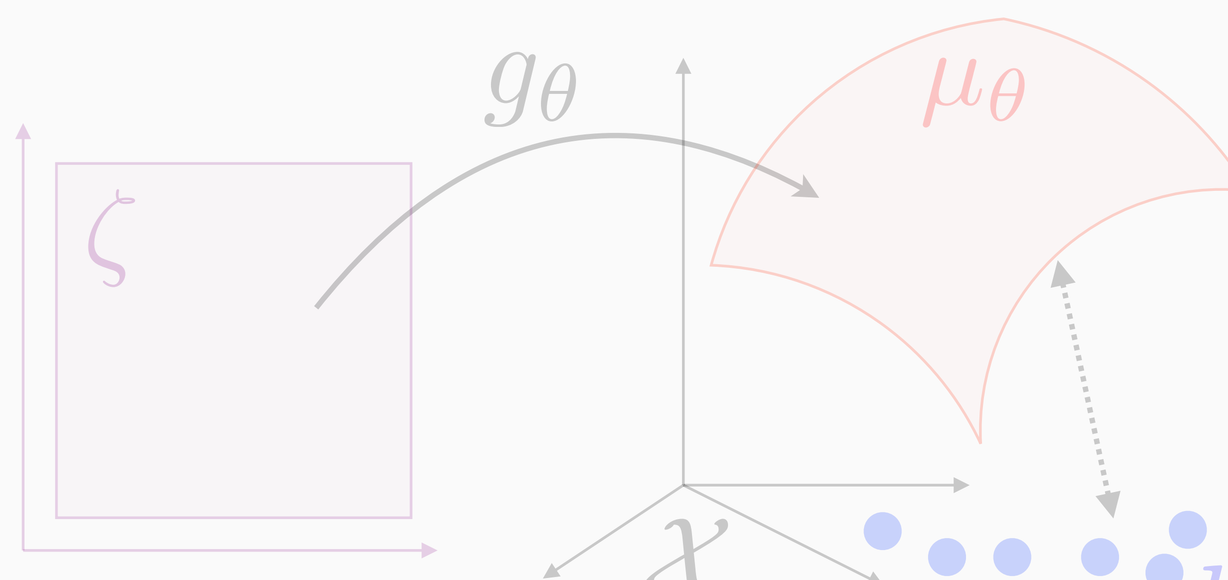
1. Optimal Transport



2. Entropic Regularization

3. Sinkhorn Divergences

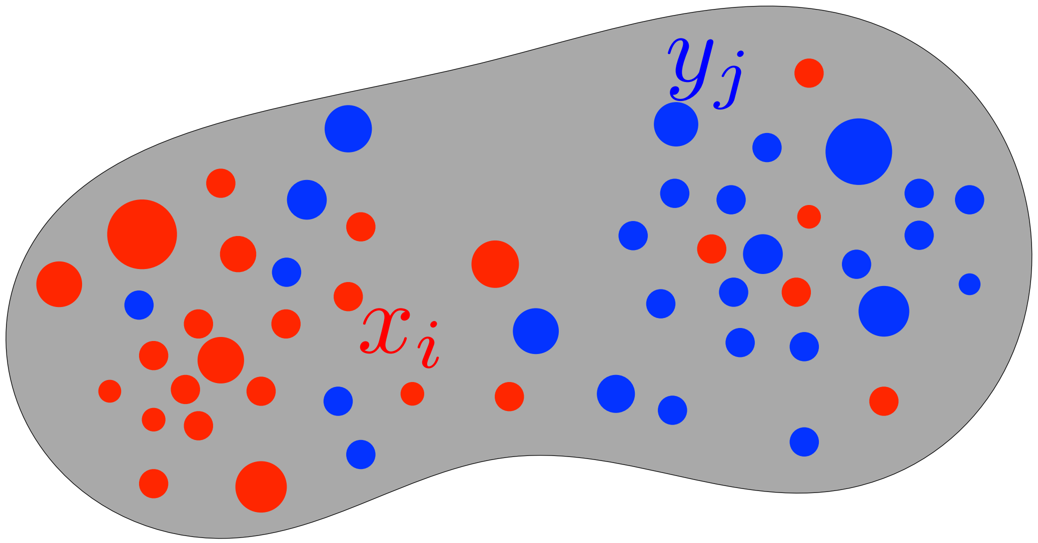
4. Application to Generative Models



Kantorovitch's Formulation

Input distributions

$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$



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

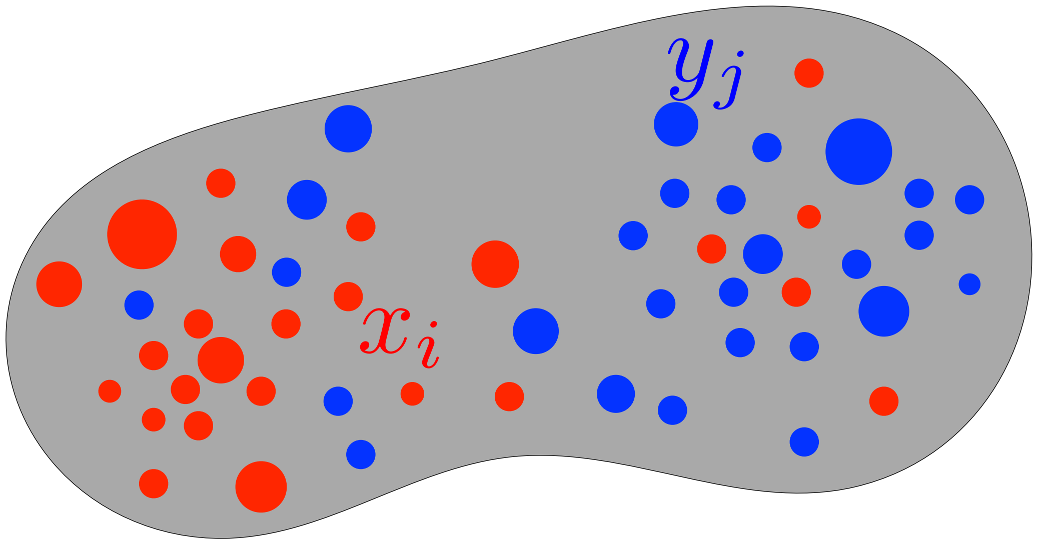
Weights $\mathbf{a}_i \geqslant 0$, $\mathbf{b}_j \geqslant 0$.

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

Kantorovitch's Formulation

Input distributions

$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i} \quad \beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$



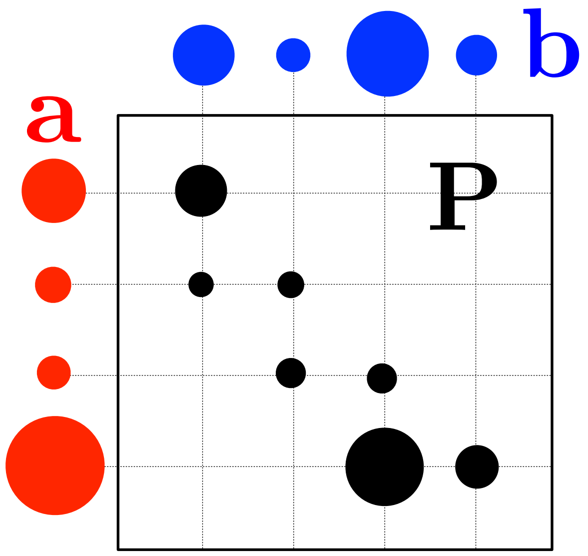
Points $(x_i)_i$, $(y_j)_j$

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:

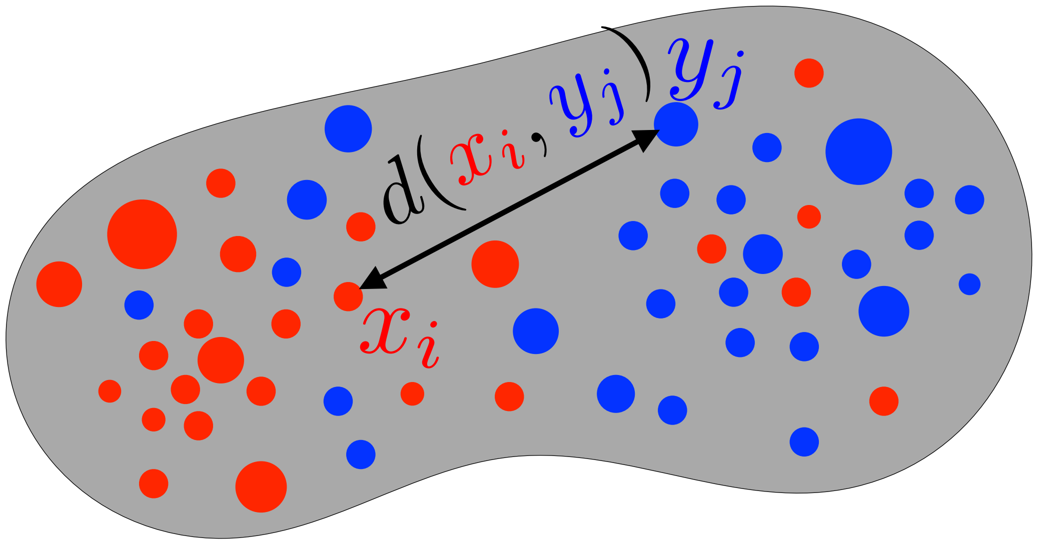
$$\mathbf{U}(\mathbf{a},\mathbf{b}) \stackrel{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n\times m} \; ; \; \mathbf{P}\mathbb{1}_n = \mathbf{a}, \mathbf{P}^\top \mathbb{1}_m = \mathbf{b} \right\}$$



Kantorovitch's Formulation

Input distributions

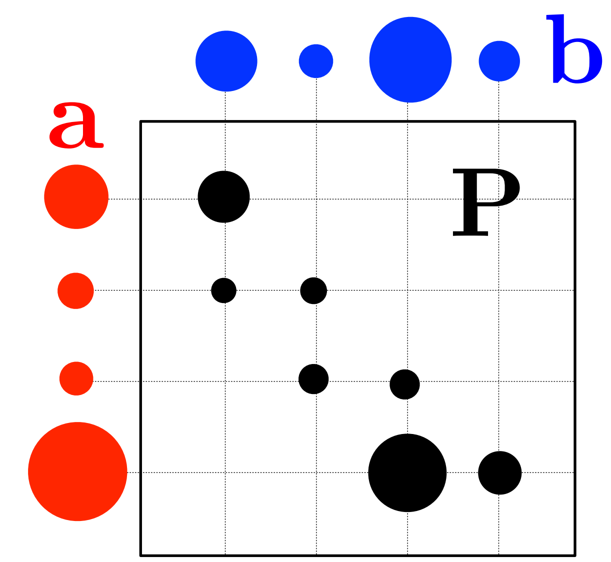
$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$



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

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

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

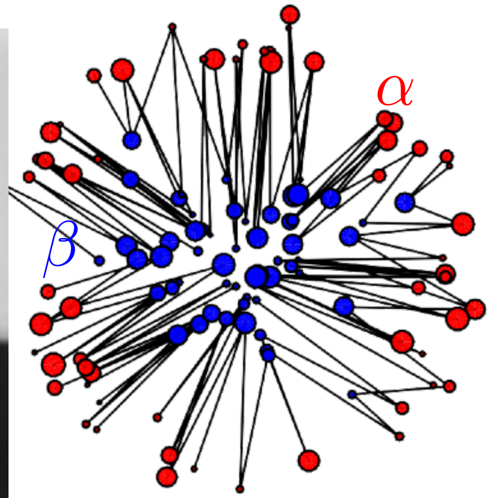


Couplings:

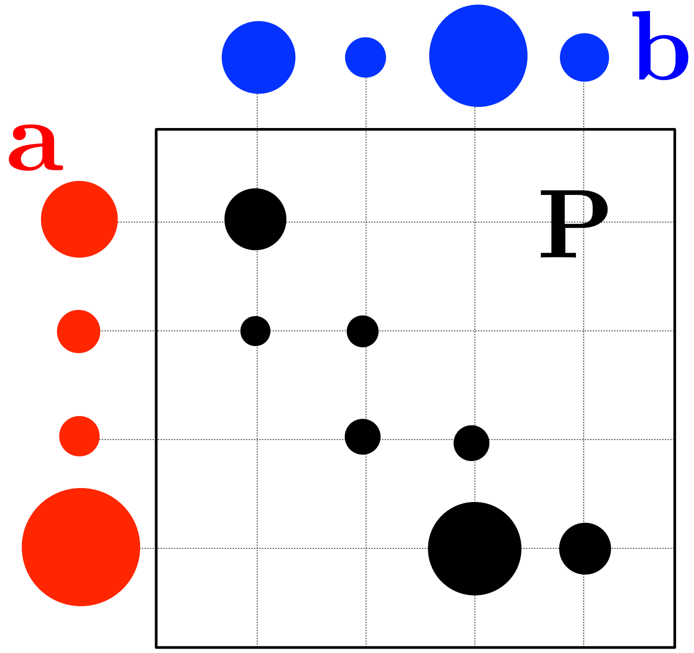
$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \; ; \; \mathbf{P}\mathbb{1}_n = \mathbf{a}, \mathbf{P}^\top \mathbb{1}_m = \mathbf{b} \right\}$$

[Kantorovich 1942]

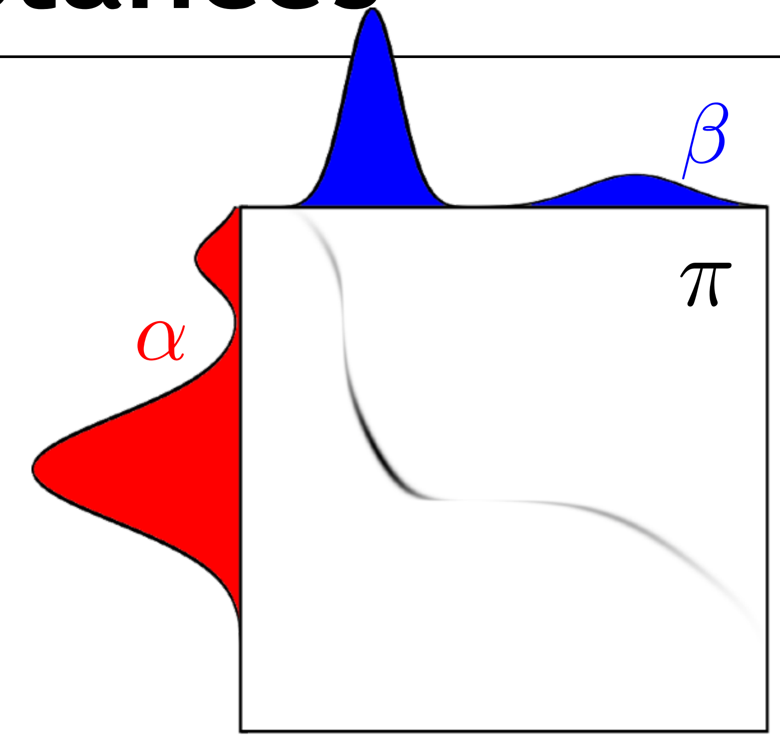
$$\min \left\{ \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} \; ; \; \mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b}) \right\}$$



Optimal Transport Distances

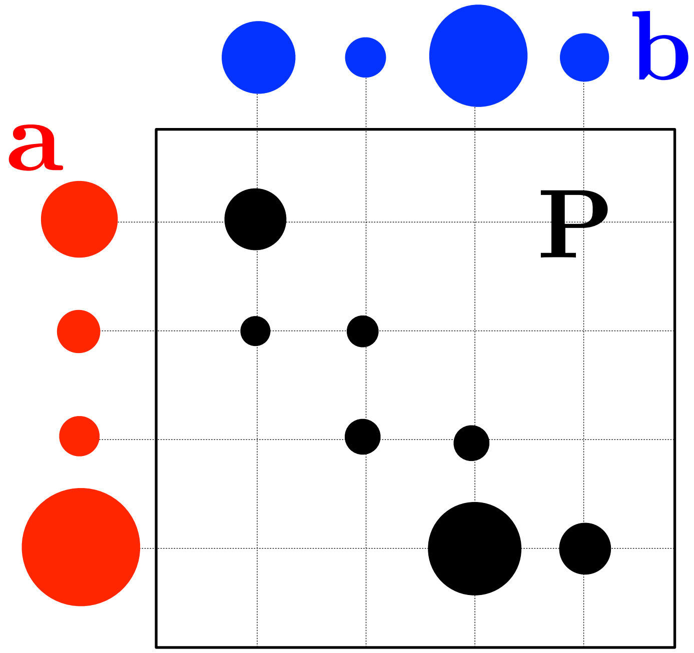


$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$

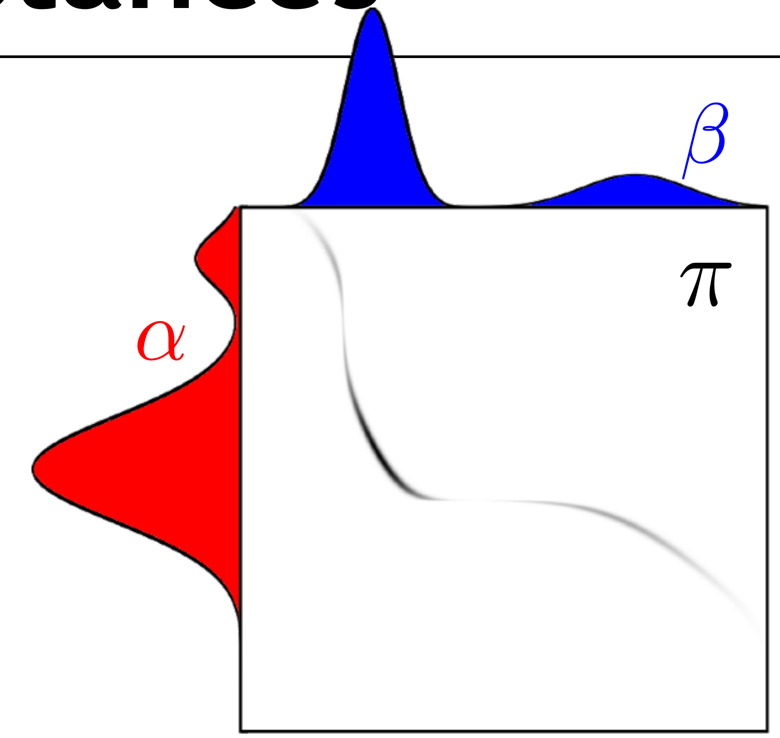


$$\mathrm{W}_p(\alpha, \beta)^p \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{M}_+^1(\mathcal{X}^2)} \left\{ \int_{\mathcal{X}^2} d(x, y)^p \mathrm{d}\pi(x, y) \; ; \; \pi_1 = \alpha, \pi_2 = \beta \right\}$$

Optimal Transport Distances



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



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

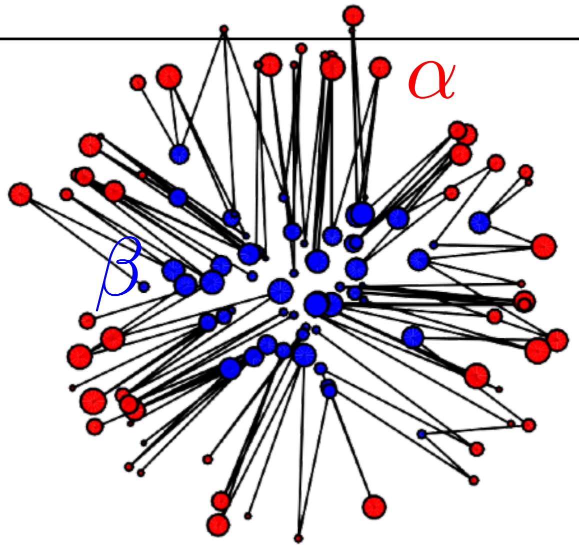
Theorem: W_p is a distance and $\alpha_n \rightharpoonup \beta \Leftrightarrow \mathrm{W}_p(\alpha_n, \beta) \to 0$

Weak* (aka in law) convergence: $\alpha_n \rightharpoonup \beta \Leftrightarrow \begin{matrix} \forall f \in \mathcal{C}(\mathcal{X}), \\ \int_{\mathcal{X}} f \mathrm{d}\alpha_n \to \int_{\mathcal{X}} f \mathrm{d}\beta \end{matrix}$

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

Algorithms

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

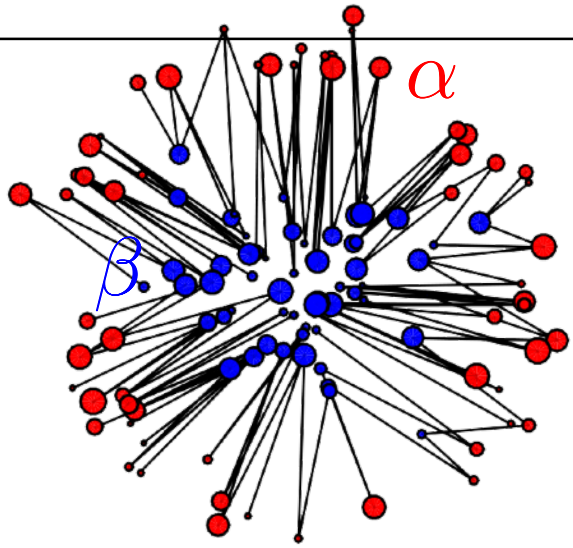


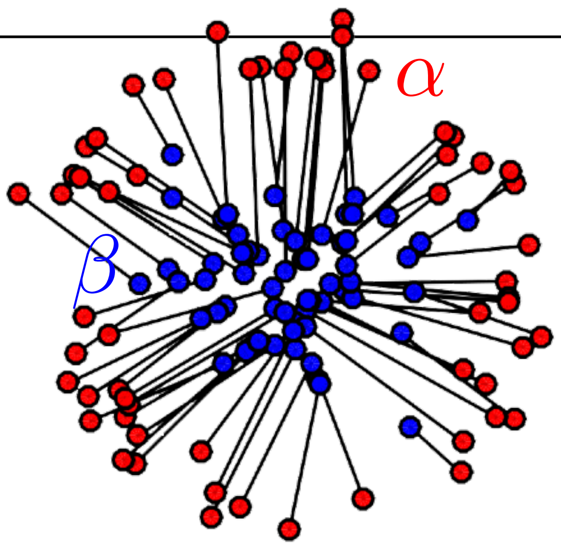
Algorithms

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

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

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



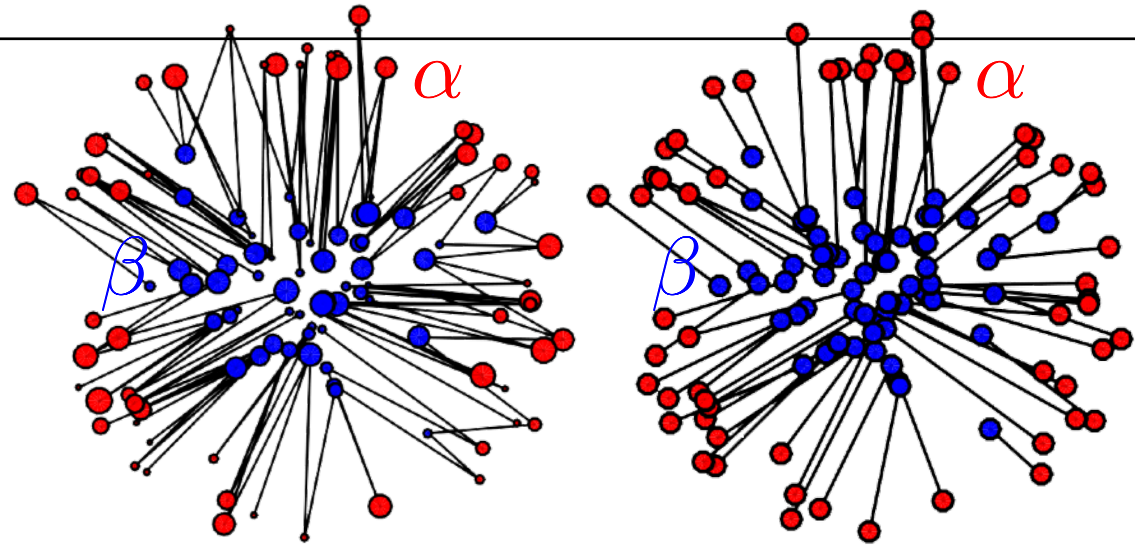


Algorithms

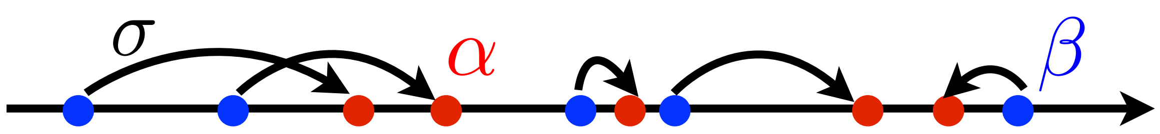
Linear programming: $O(n^3 \log(n)^2)$

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 $O(n \log(n))$.

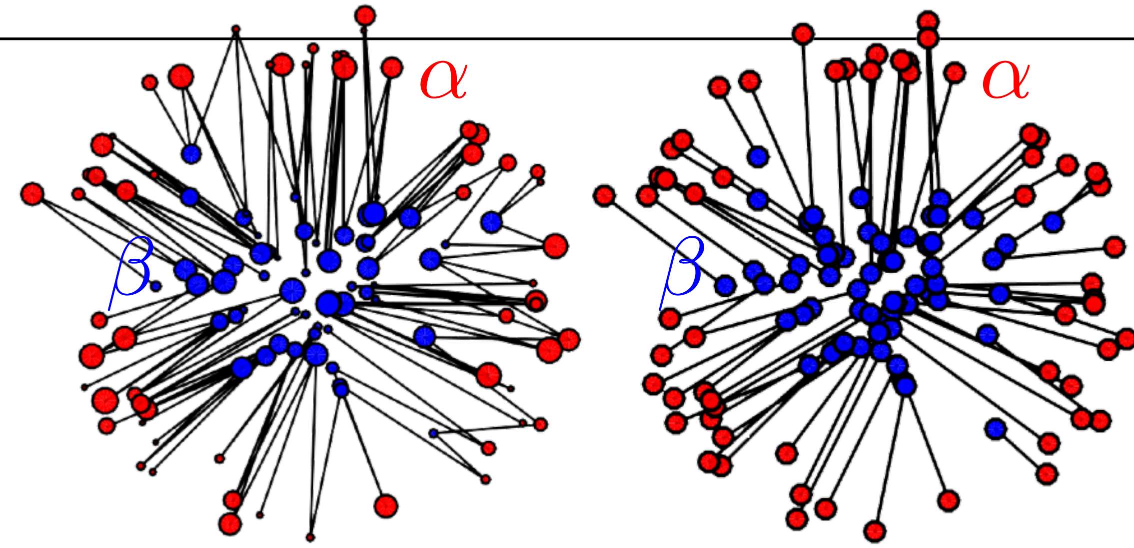


Algorithms

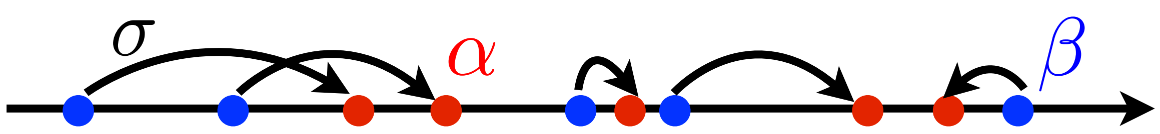
Linear programming: $O(n^3 \log(n)^2)$

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

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



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



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

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

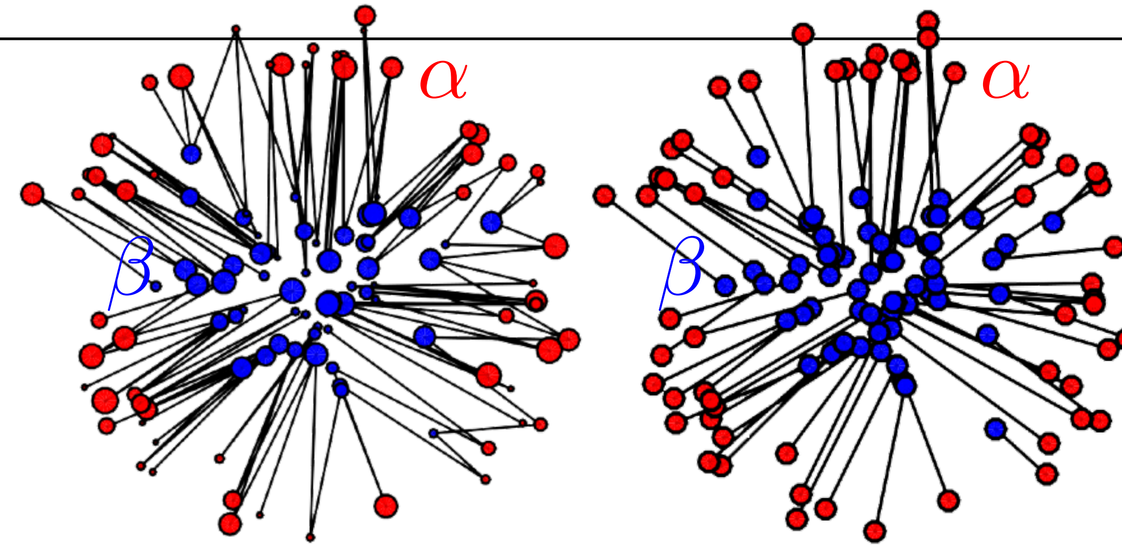
$\rightarrow$ min-cost flow, on graphs $O(n^2 \log(n))$.

Algorithms

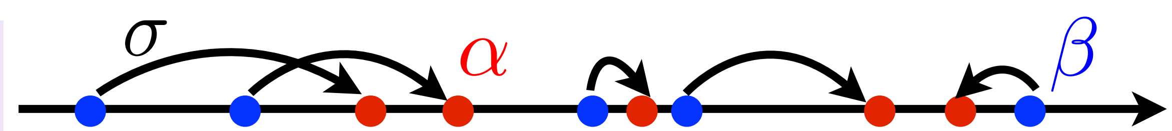
Linear programming: $O(n^3 \log(n)^2)$

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 $O(n \log(n))$.



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

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

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

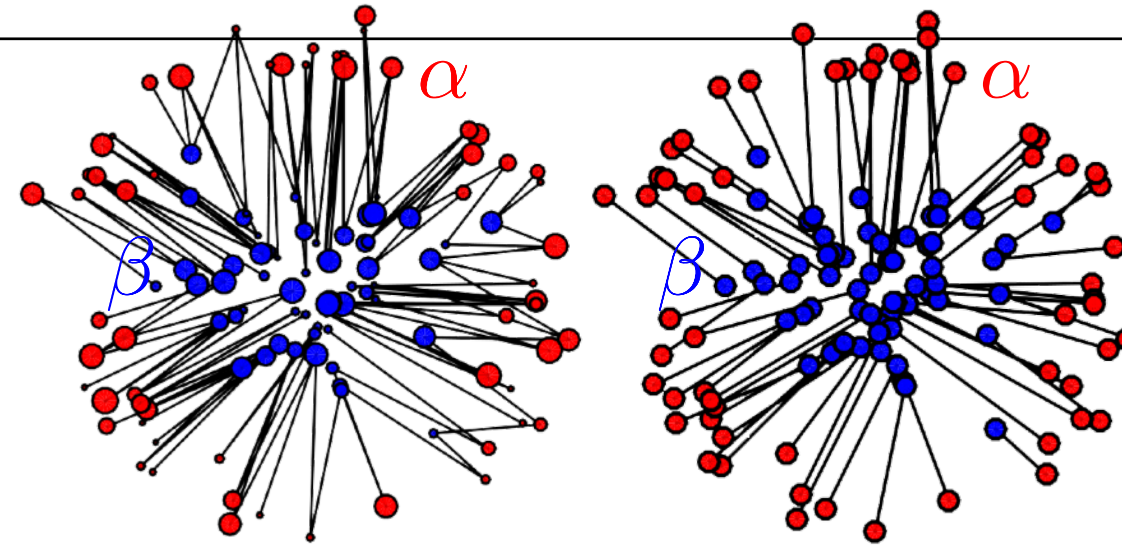
Monge-Ampère/Benamou-Brenier, $d = \| \cdot \|_2^2$.

Algorithms

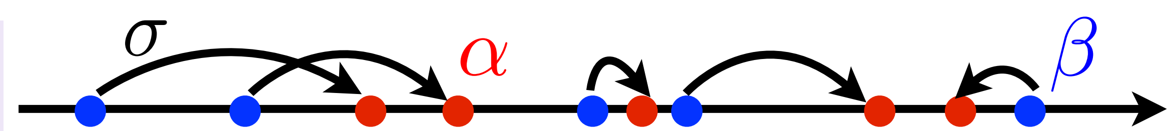
Linear programming: $O(n^3 \log(n)^2)$

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 $O(n\log(n))$.



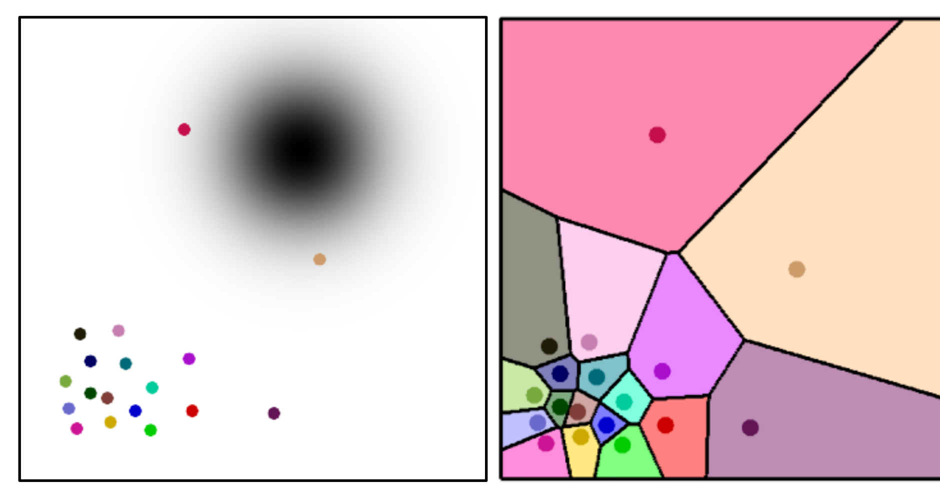
$p = 1$
$d = \|\cdot\|$

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

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

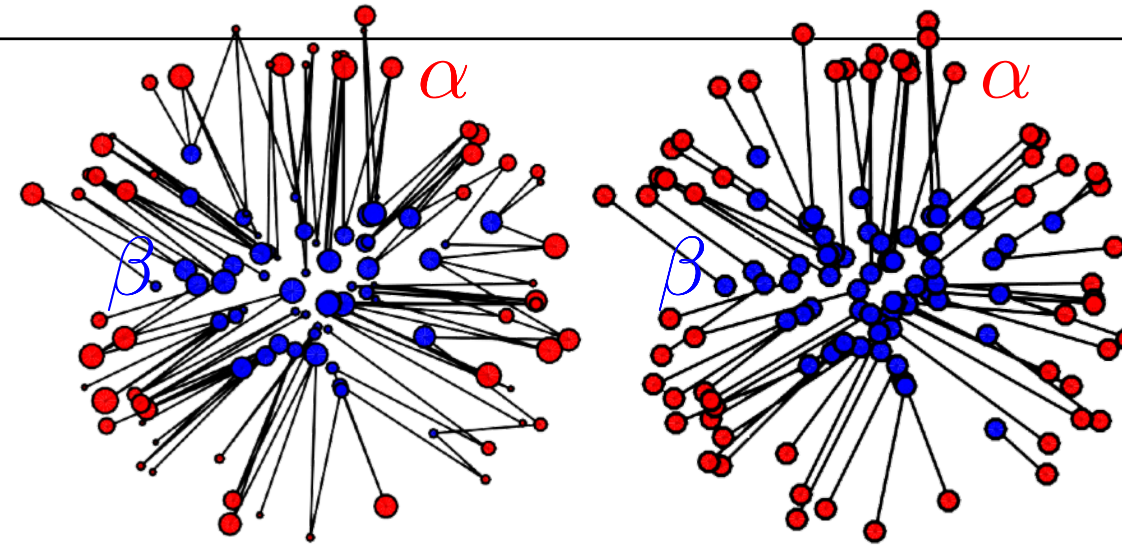
Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.

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



Algorithms

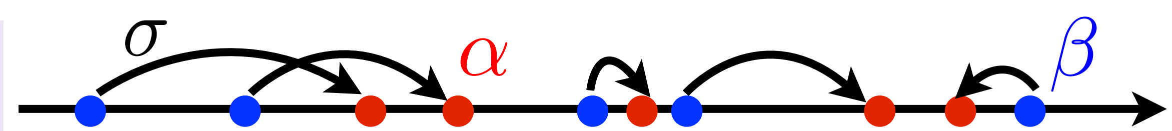
Linear programming: $O(n^3 \log(n)^2)$



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 $O(n\log(n))$.



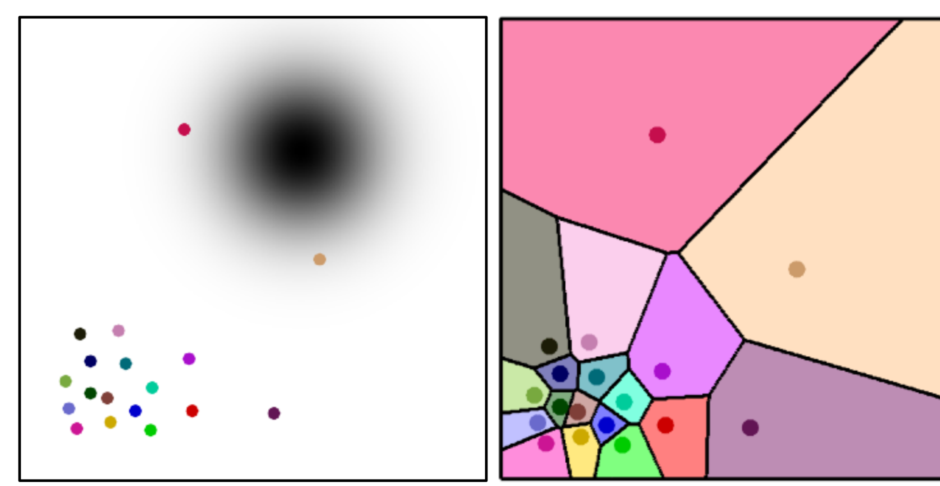
$p = 1$
$d = \|\cdot\|$

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

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

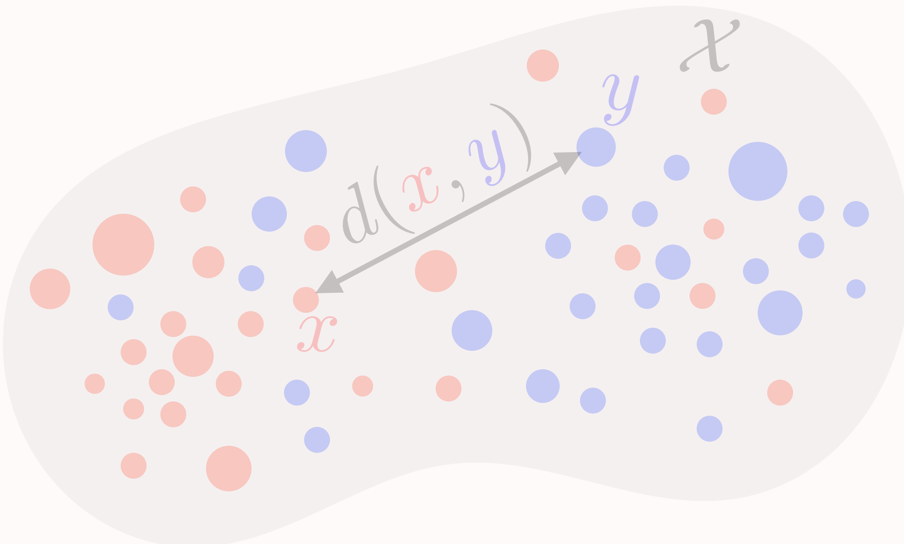
Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.

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

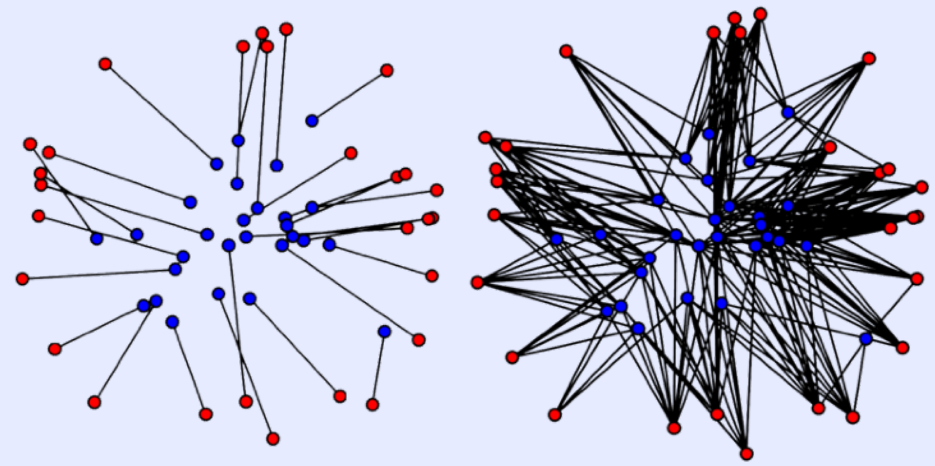


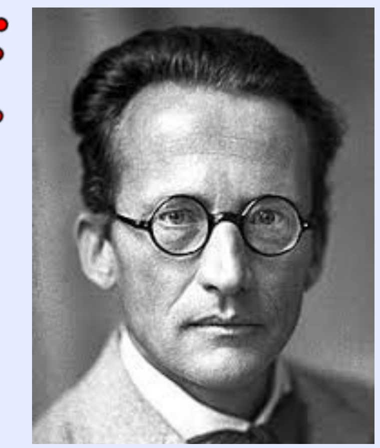
Need for fast approximate algorithms for generic c.

1. Optimal Transport



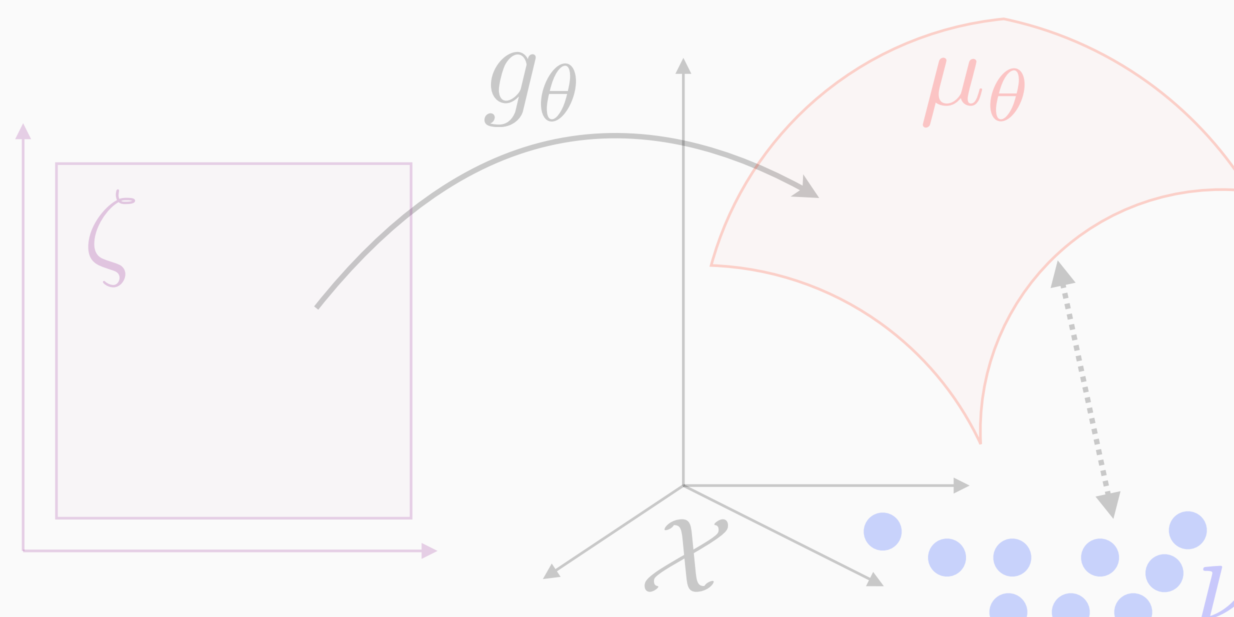
2. Entropic Regularization





3. Sinkhorn Divergences

4. Application to Generative Models

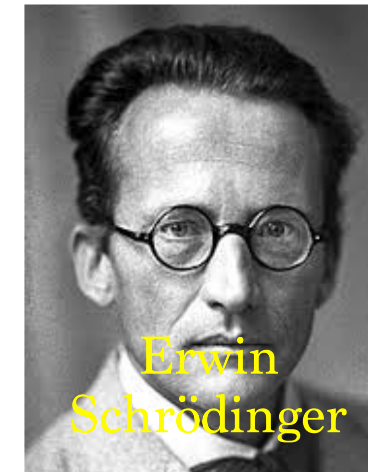


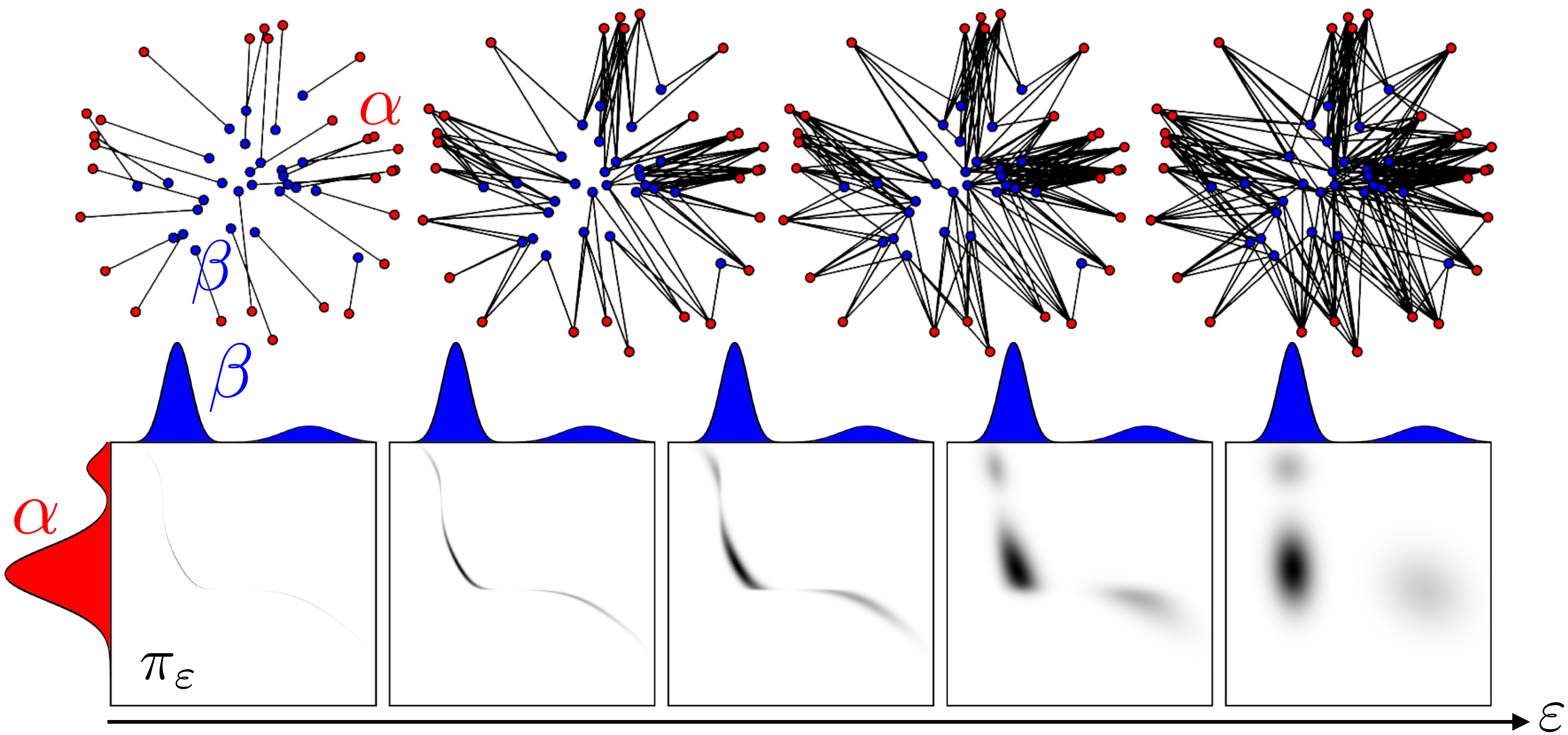
Entropic Regularization

Relative-entropy: $\mathrm{KL}(\pi|\alpha \otimes \beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}^2} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta}(x,y)\right) \mathrm{d}\pi(x,y)$

Schrödinger's problem: [1931]

$$\mathrm{W}^p_{\varepsilon,p}(\alpha,\beta) \stackrel{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} d^p(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\alpha \otimes \beta)$$





Sinkhorn's Algorithm

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

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

Sinkhorn's Algorithm

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

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

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

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

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

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

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

Sinkhorn's Algorithm

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} d(x_i,y_j)^p \mathbf{P}_{i,j} + \varepsilon \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{a}_i \mathbf{b}_j}\right)$$

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

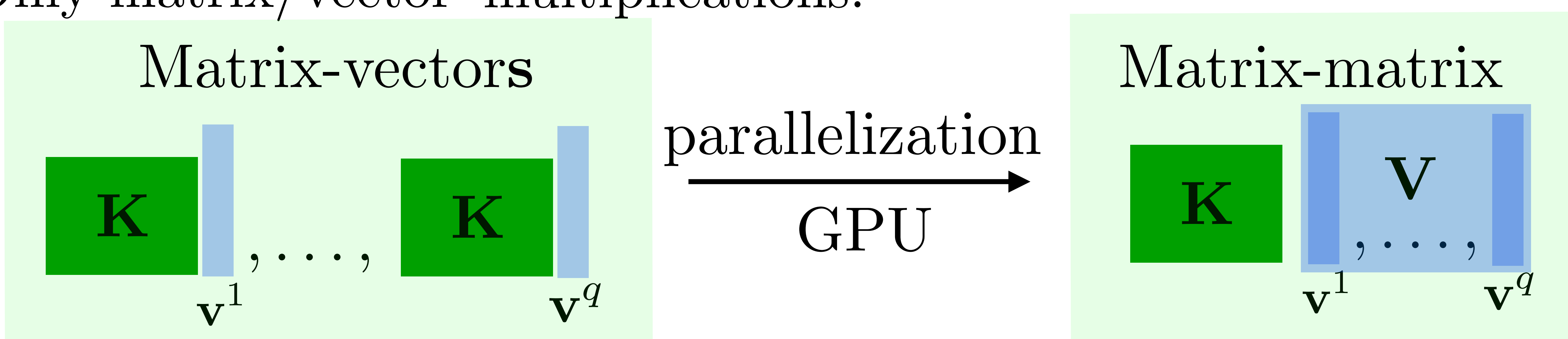
Row constraint: $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}$

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

Sinkhorn iterations: $\mathbf{u} \leftarrow \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$ $\quad \mathbf{v} \leftarrow \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}$

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

Only matrix/vector multiplications.



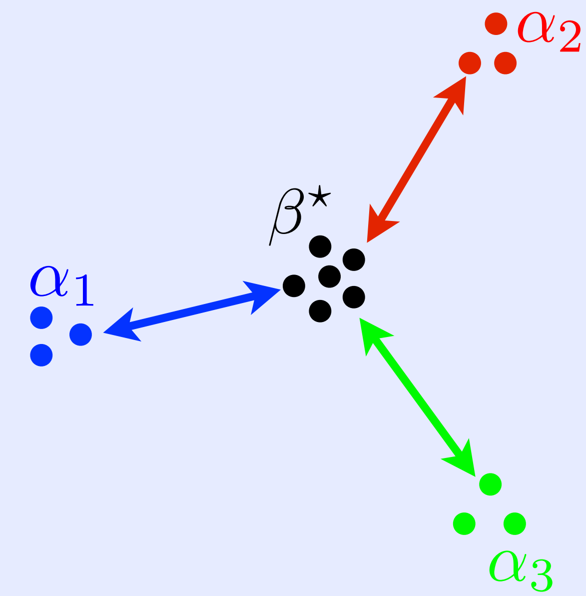
→ Convolution on regular grids, separable kernels.

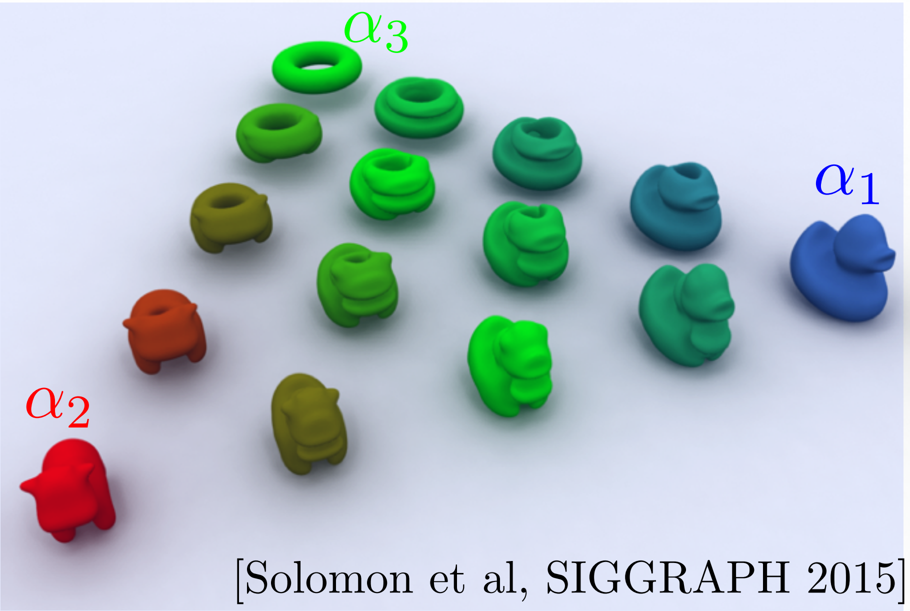
Generalizations

OT barycenters:

$$\min_{\beta} \sum_k \lambda_k \mathrm{W}_p^p(\alpha_k, \beta)$$

[Agueh, Carlier 2010]



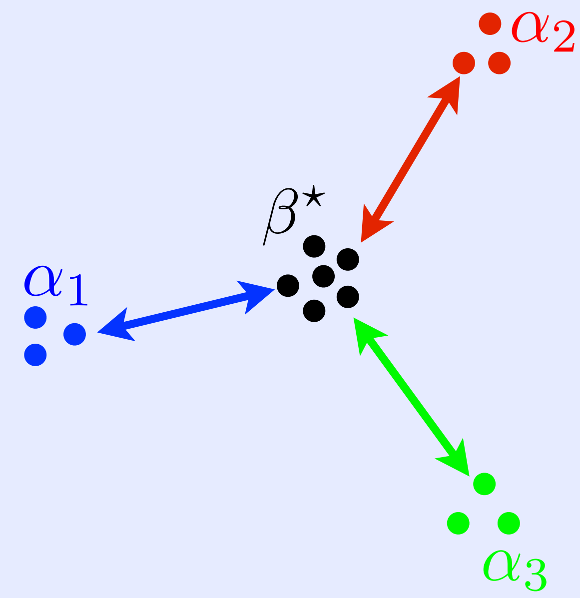


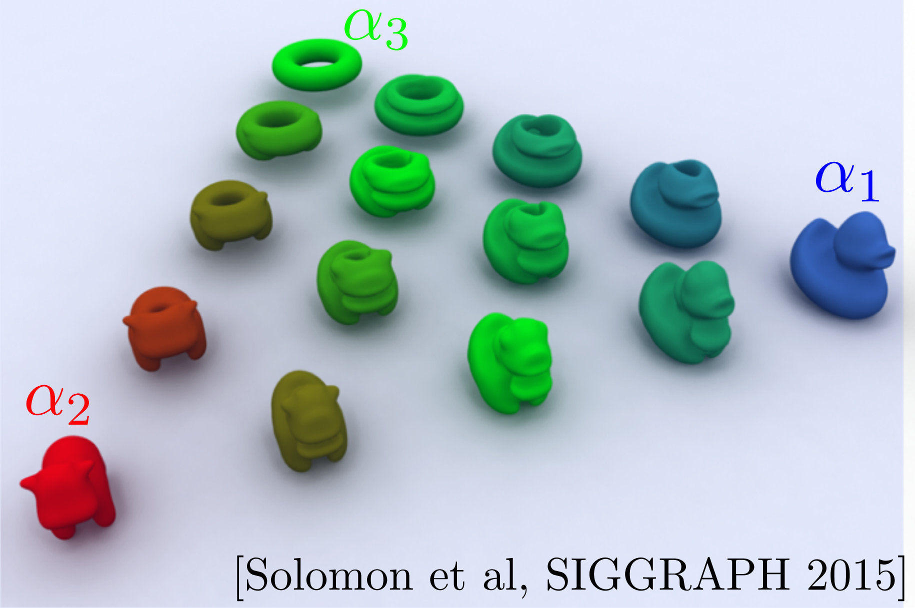
Generalizations

OT barycenters:

$$\min_{\beta} \sum_k \lambda_k W_p^p(\alpha_k, \beta)$$

[Agueh, Carlier 2010]



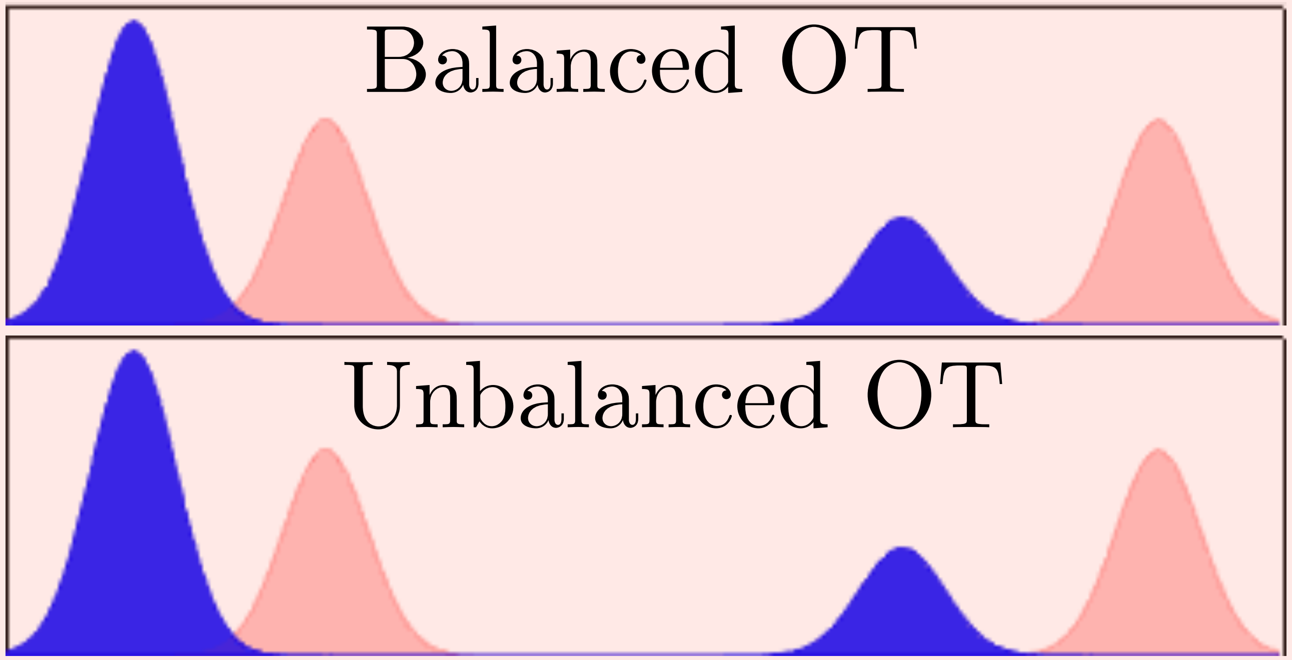


[Solomon et al, SIGGRAPH 2015]

Unbalanced transport:

$$\min_{\pi} \int c \mathrm{d}\pi +$$
$$\rho \mathrm{KL}(\pi_1 | \alpha) + \rho \mathrm{KL}(\pi_2 | \beta)$$

[Liereo, Mielke, Savaré 2015]
[Chizat, Schmitzer, Peyré, Vialard 2015]

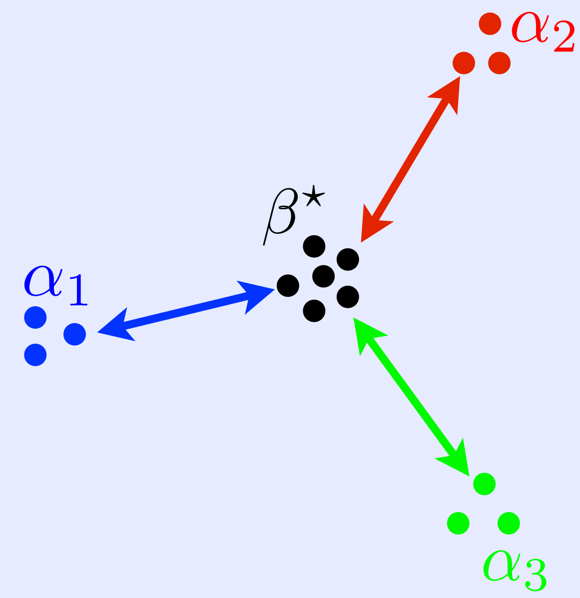


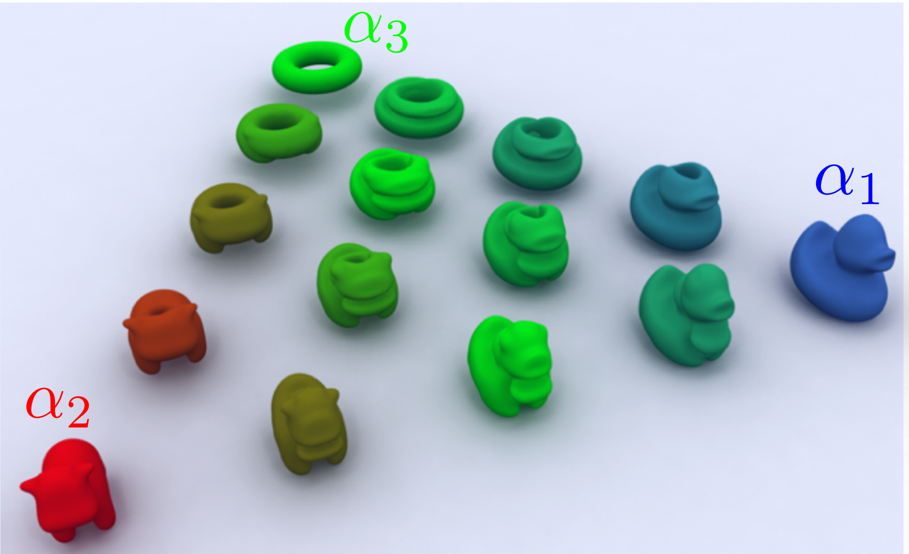
Generalizations

OT barycenters:

$$\min_{\beta} \sum_k \lambda_k \mathrm{W}_p^p(\alpha_k, \beta)$$

[Agueh, Carlier 2010]



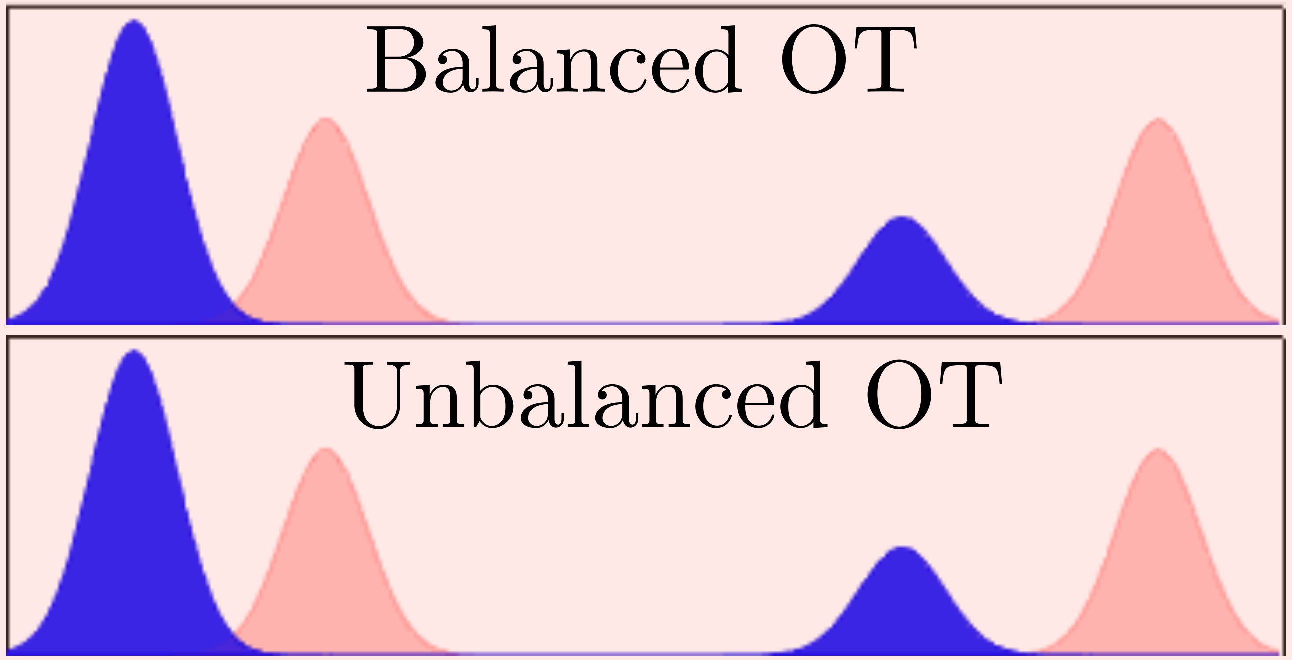


[Solomon et al, SIGGRAPH 2015]

Unbalanced transport:

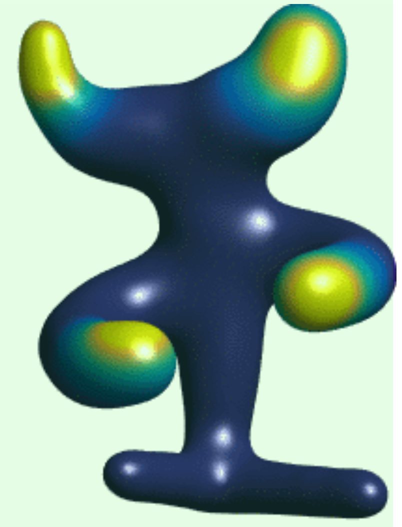
$$\min_{\pi} \int c \mathrm{d}\pi + \rho \mathrm{KL}(\pi_1 | \alpha) + \rho \mathrm{KL}(\pi_2 | \beta)$$

[Liereo, Mielke, Savaré 2015]
[Chizat, Schmitzer, Peyré, Vialard 2015]

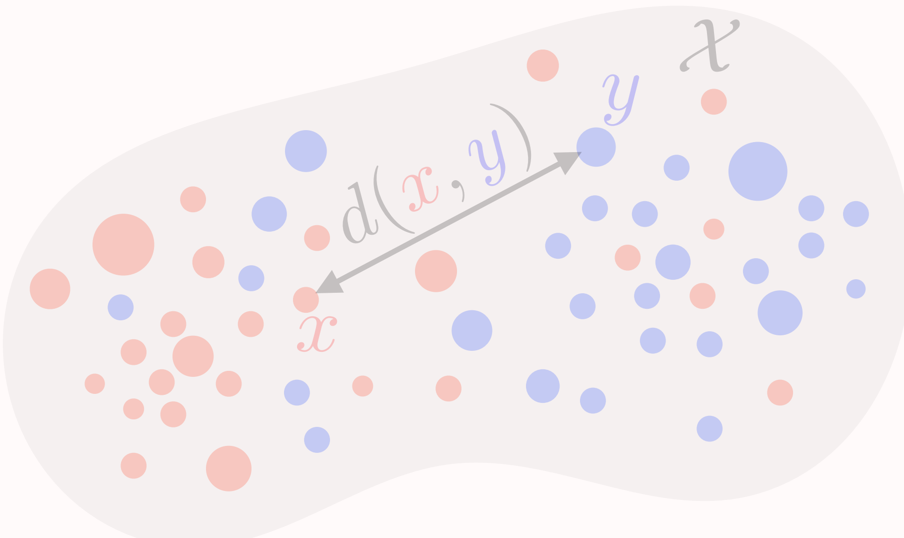


Gradient flows:

$$\alpha_{t+\tau} = \min_{\alpha} \mathrm{W}_p^p(\alpha_t, \alpha) + \tau f(\alpha)$$

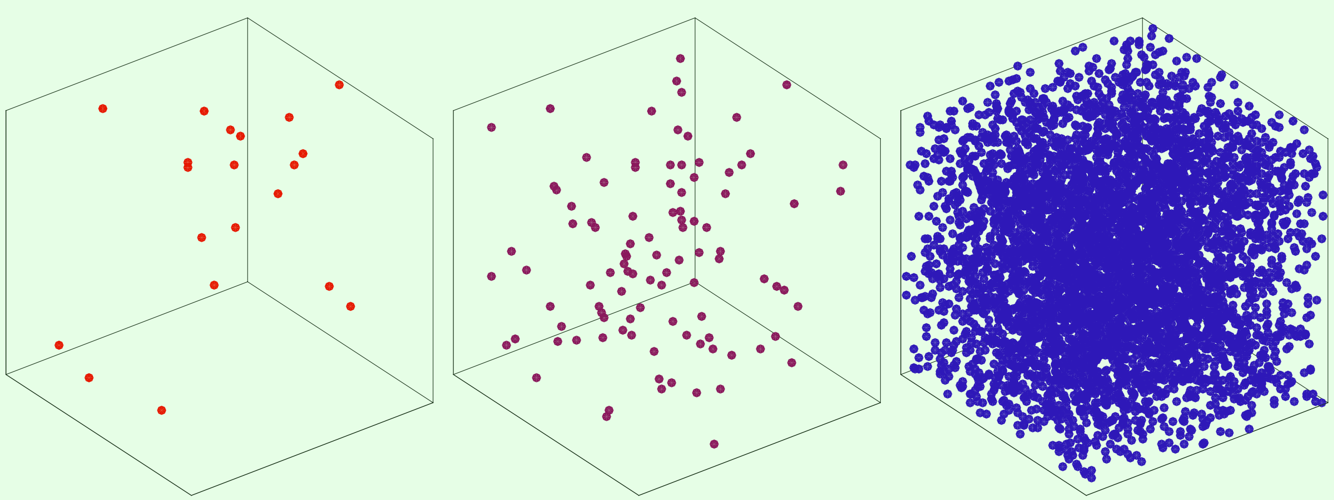


1. Optimal Transport

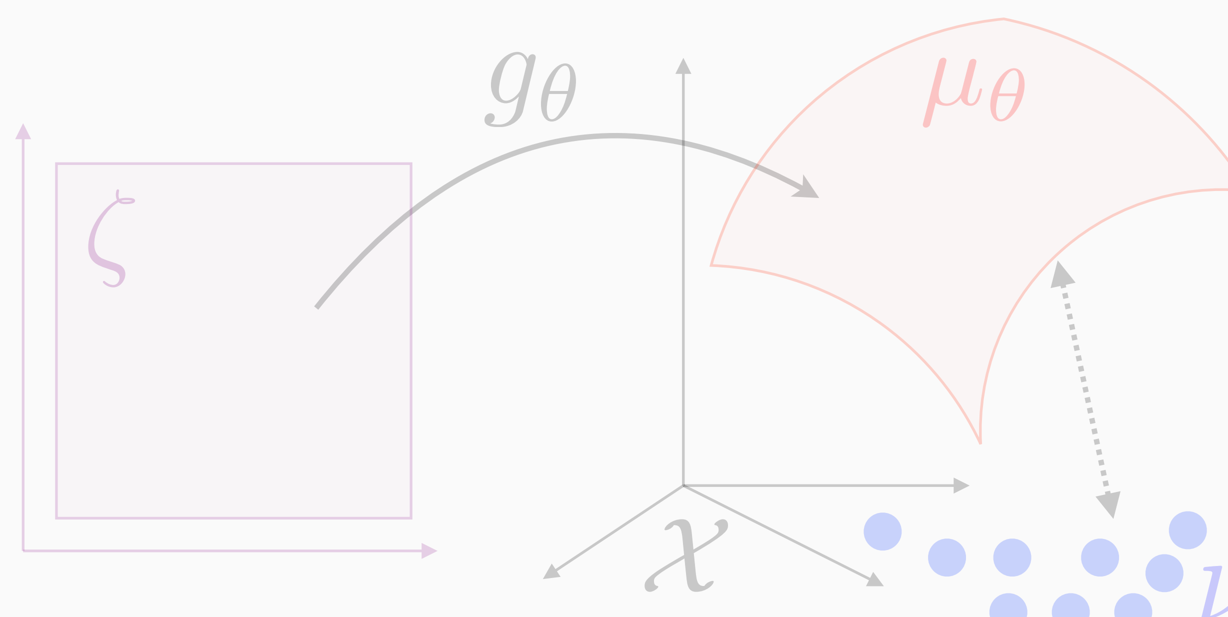


2. Entropic Regularization

3. Sinkhorn Divergences



4. Application to Generative Models

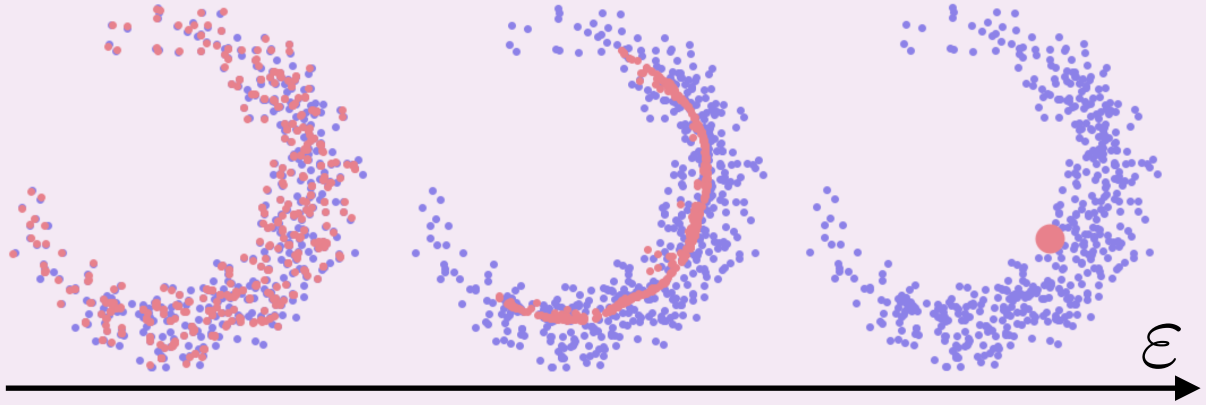


Sinkhorn Divergences

$$\mathrm{W}_{\varepsilon,p}^p(\alpha,\beta) \stackrel{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} d^p(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\xi)$$

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

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

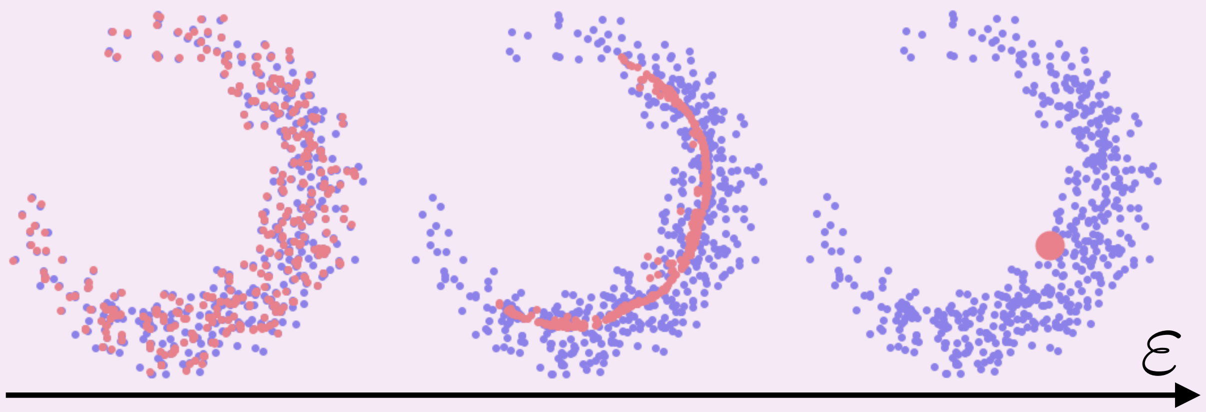


Sinkhorn Divergences

$$W^p_{\varepsilon,p}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} d^p(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\xi)$$

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

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



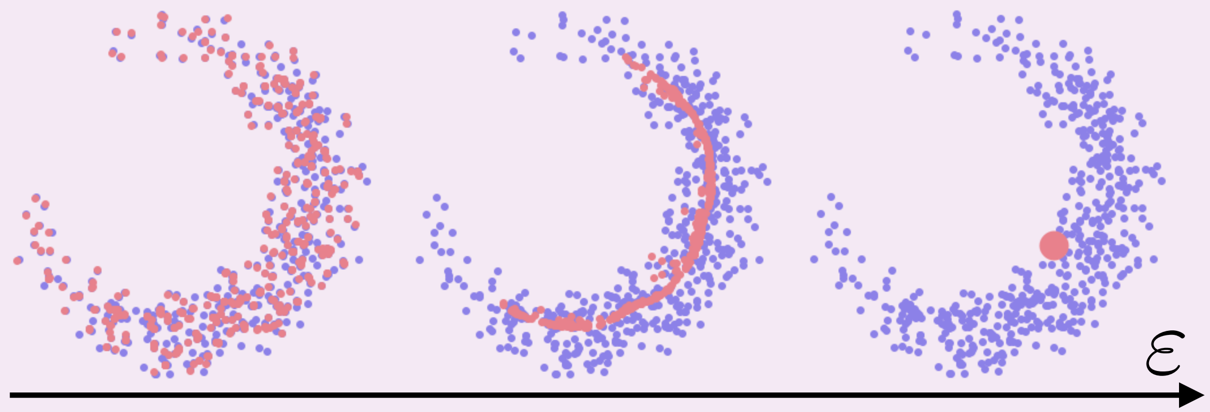
$$\overline{W}^p_{p,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} W^p_{p,\varepsilon}(\alpha,\beta) - \frac{1}{2}W^p_{p,\varepsilon}(\alpha,\alpha) - \frac{1}{2}W^p_{p,\varepsilon}(\beta,\beta)$$

Sinkhorn Divergences

$$\mathrm{W}^p_{\varepsilon,p}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} d^p(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\xi)$$

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

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



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

Theorem: [Genevay, P, Cuturi, 2017] $\overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha,\beta) \xrightarrow{\varepsilon\to 0} \mathrm{W}^p_p(\alpha,\beta)$, $\quad \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha,\beta) \xrightarrow{\varepsilon\to+\infty} \|\alpha-\beta\|^2_{-d^p}$

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

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

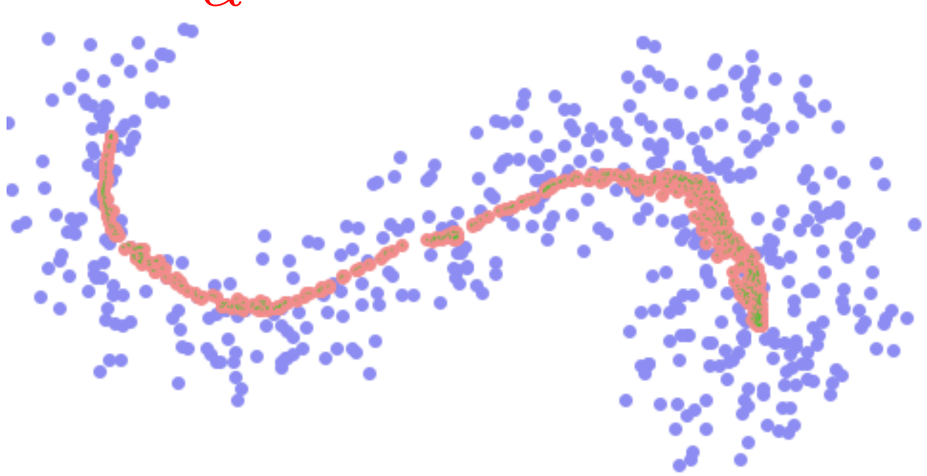
concave concave

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

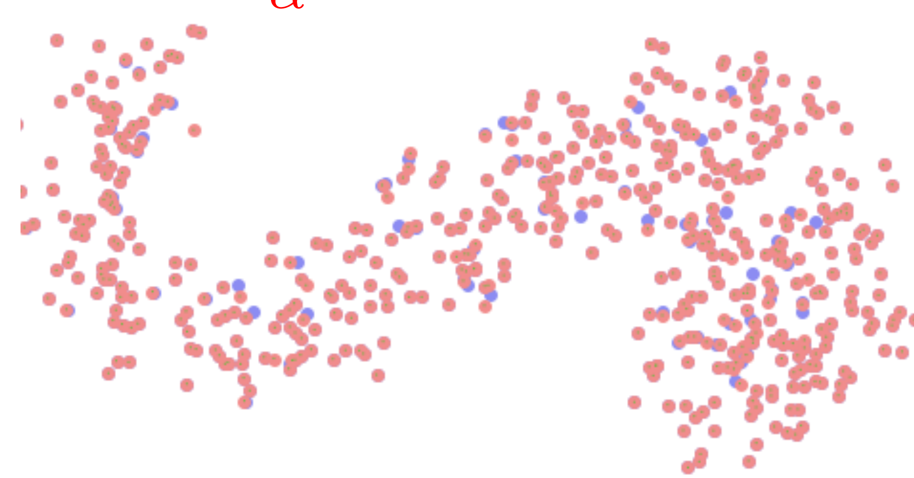
$$\overline{\mathrm{W}}_{\varepsilon,p} \geqslant 0 \text{ and } \overline{\mathrm{W}}^p_{\varepsilon,p}(\cdot,\beta) \text{ is convex.}$$

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

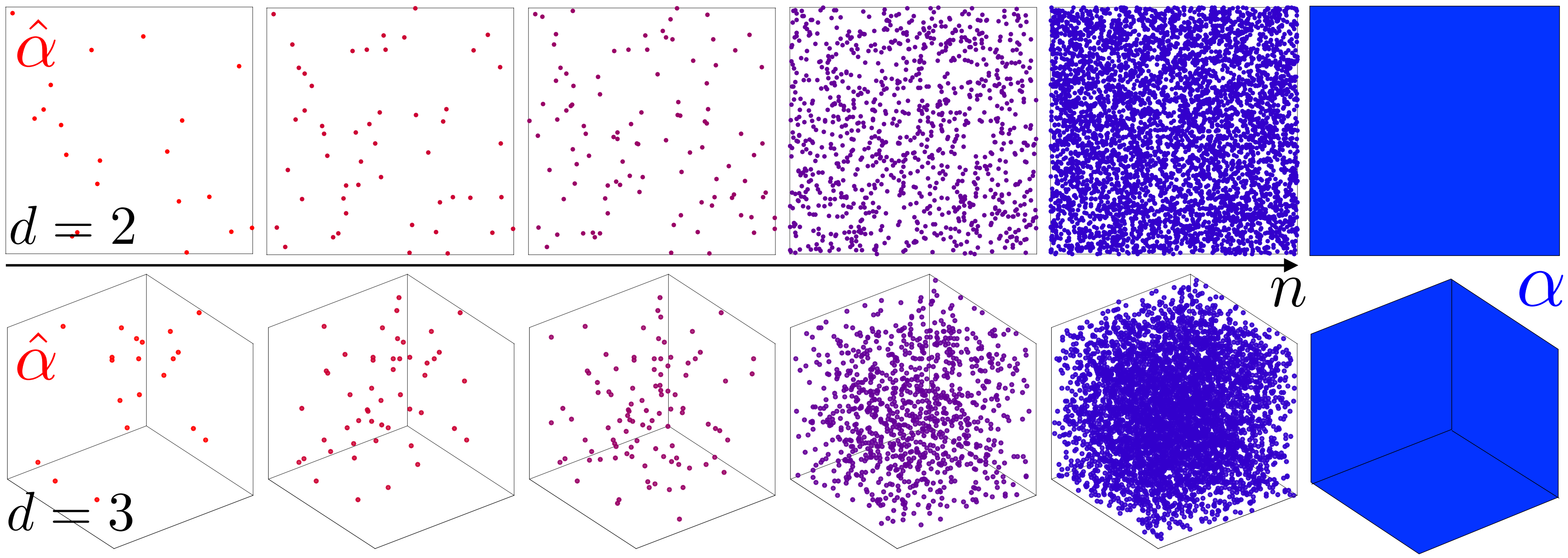
$$\min_\alpha \mathrm{W}^p_{\varepsilon,p}(\alpha,\beta)$$

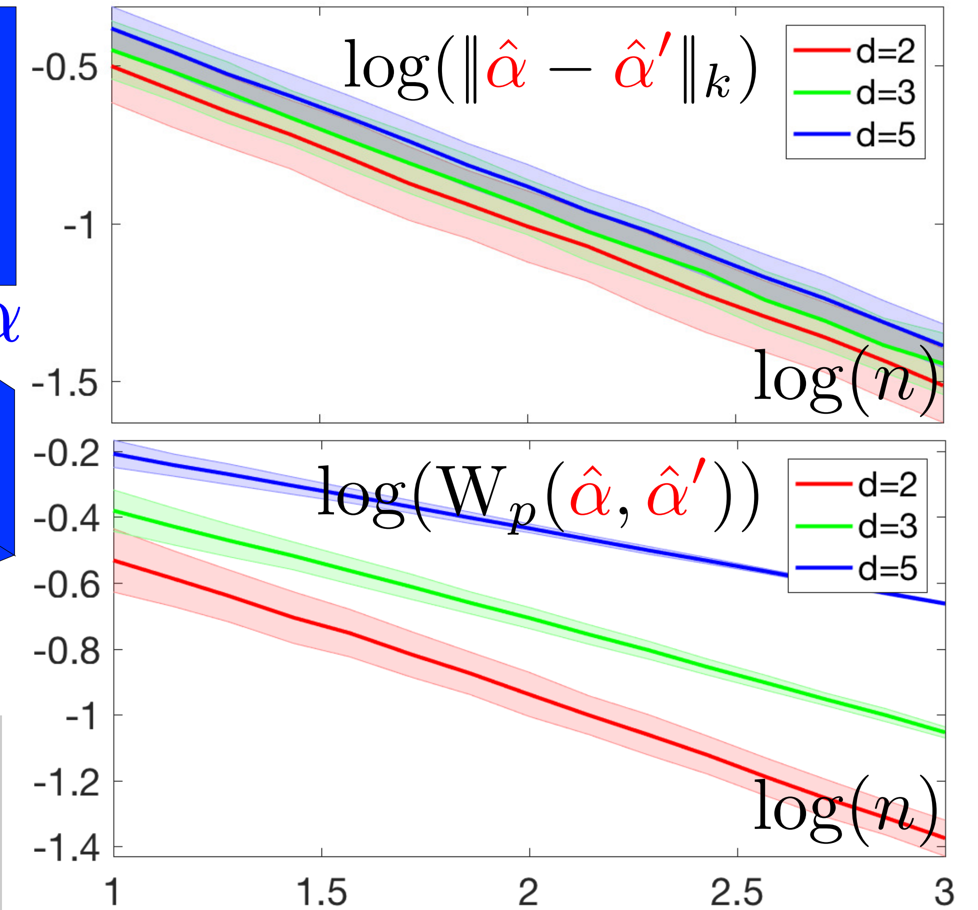


$$\min_\alpha \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha,\beta)$$



Sample Complexity





Theorem:

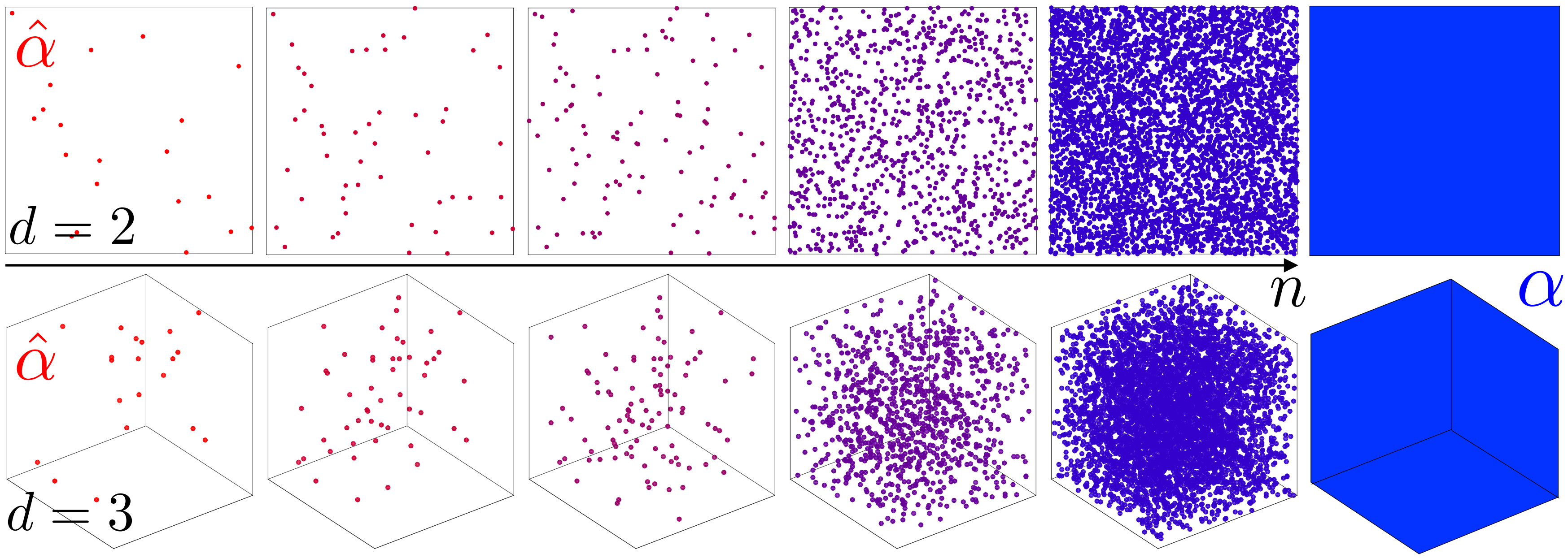
$$\mathbb{E}(|W_p(\hat\alpha, \hat\beta) - W_p(\alpha, \beta)|) = O(n^{-\frac{1}{d}})$$

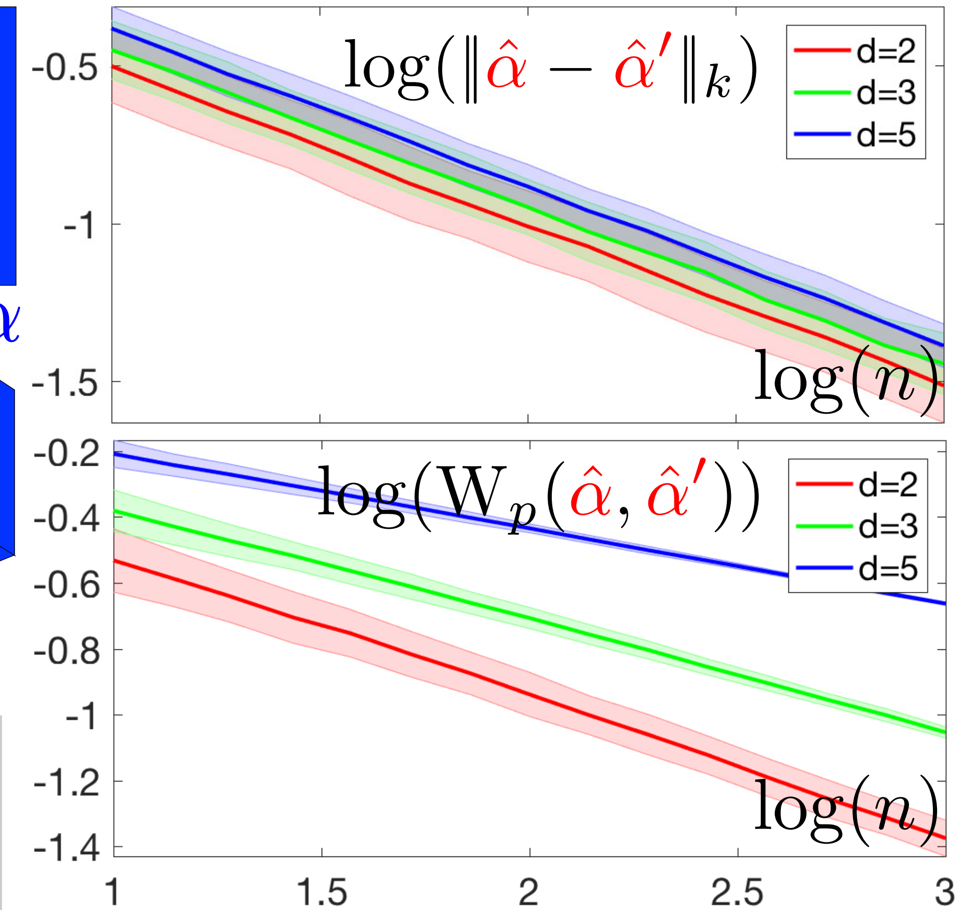
$$\mathbb{E}(|\|\hat\alpha - \hat\beta\|_k - \|\alpha - \beta\|_k|) = O(n^{-\frac{1}{2}})$$

Optimal transport: suffers from curse of dimensionality.

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

Sample Complexity





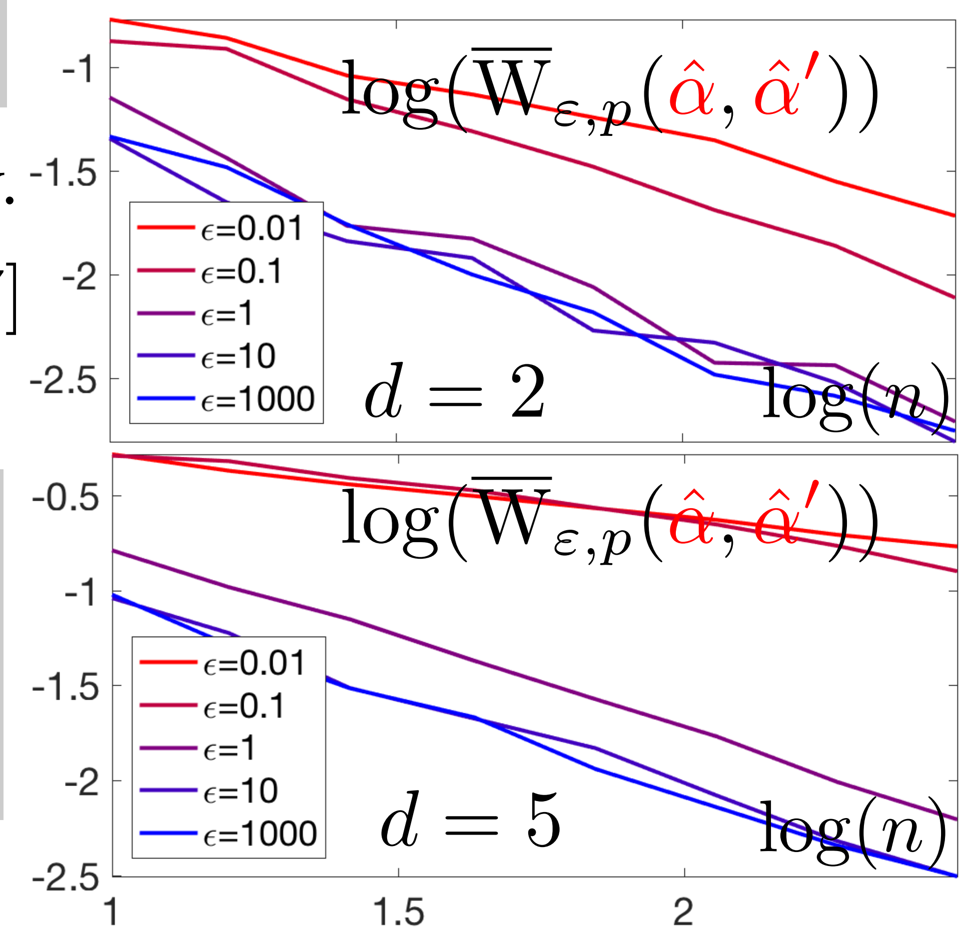
Theorem:

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

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

Optimal transport: suffers from curse of dimensionality.

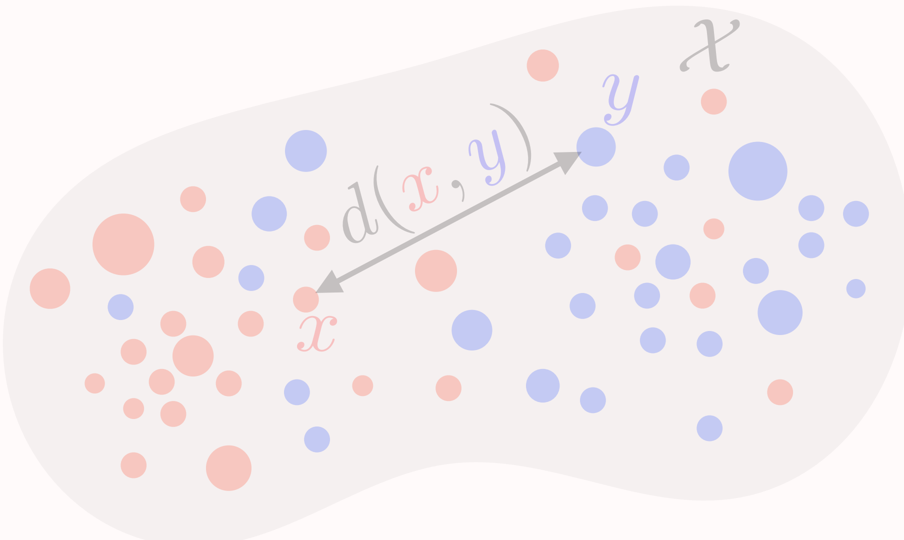
→ Adapt to support dimensionality [Weed, Bach 2017]



Theorem: [Genevay, Bach, P, Cuturi]

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

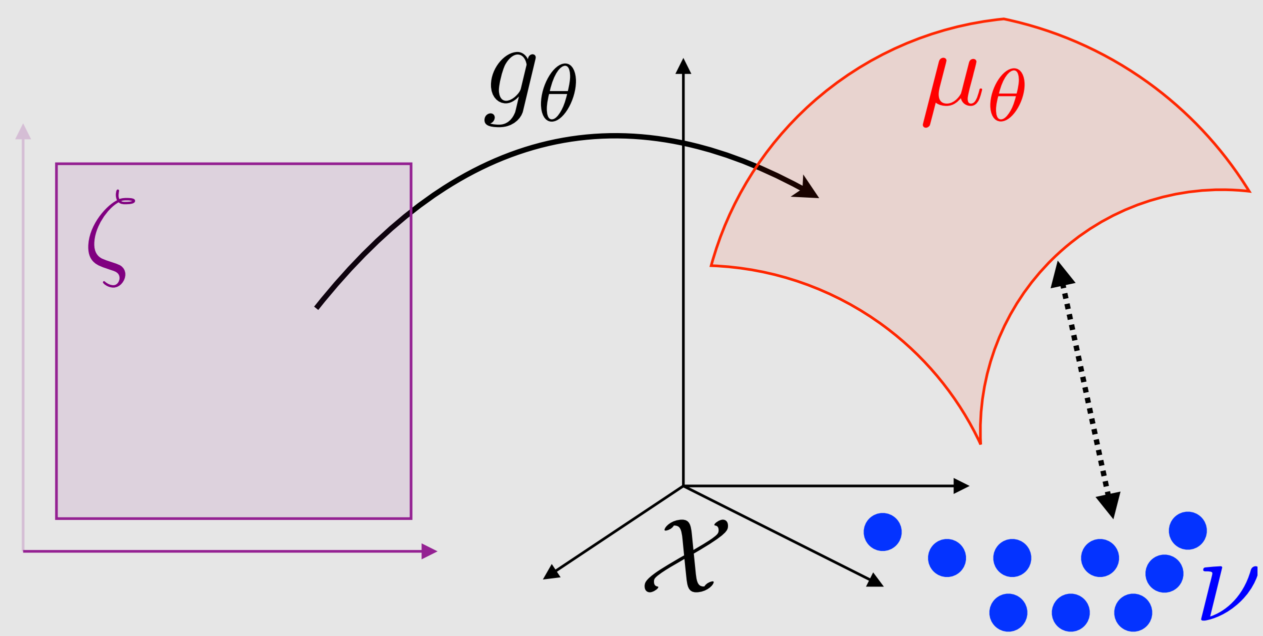
1. Optimal Transport



2. Entropic Regularization

3. Sinkhorn Divergences

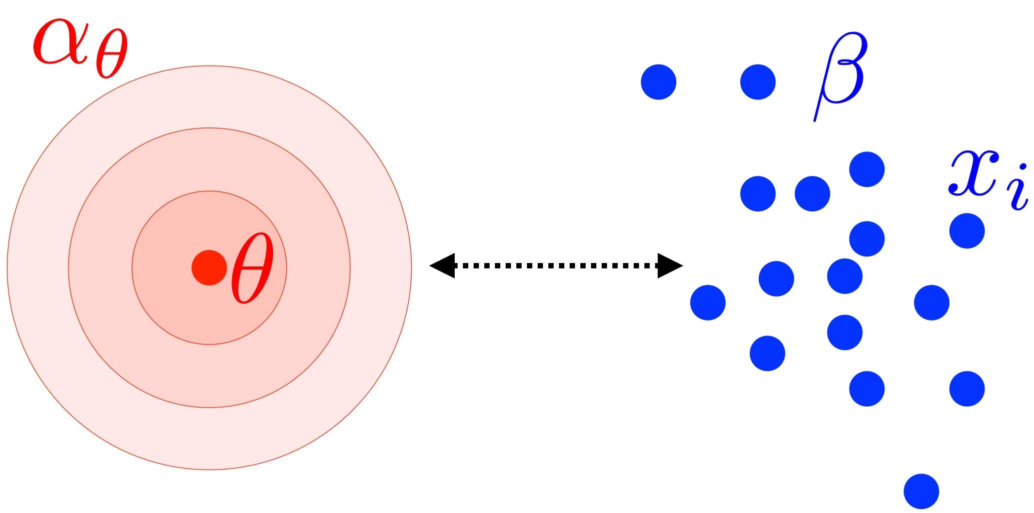
4. Application to Generative Models



Density Fitting and Generative Models

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

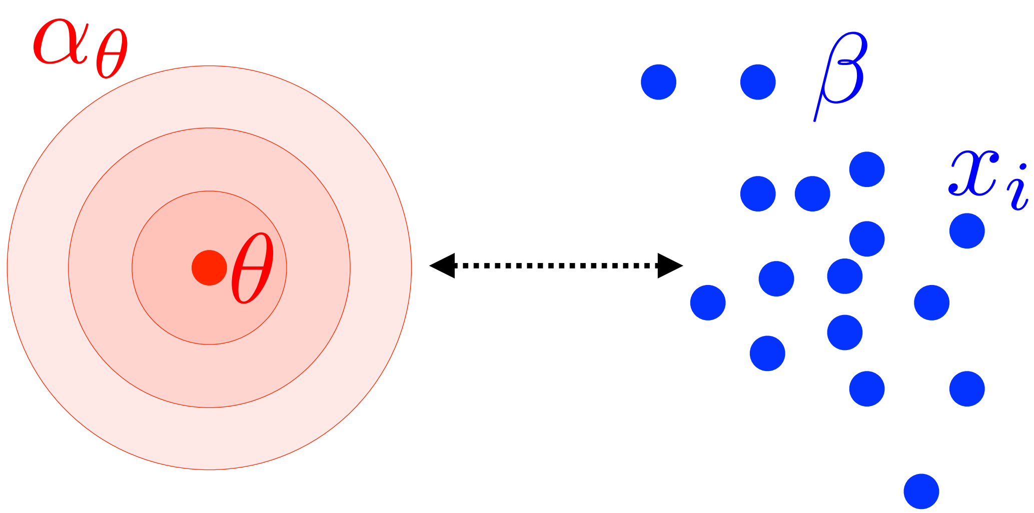
Parametric model: $\theta \mapsto \alpha_\theta$



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

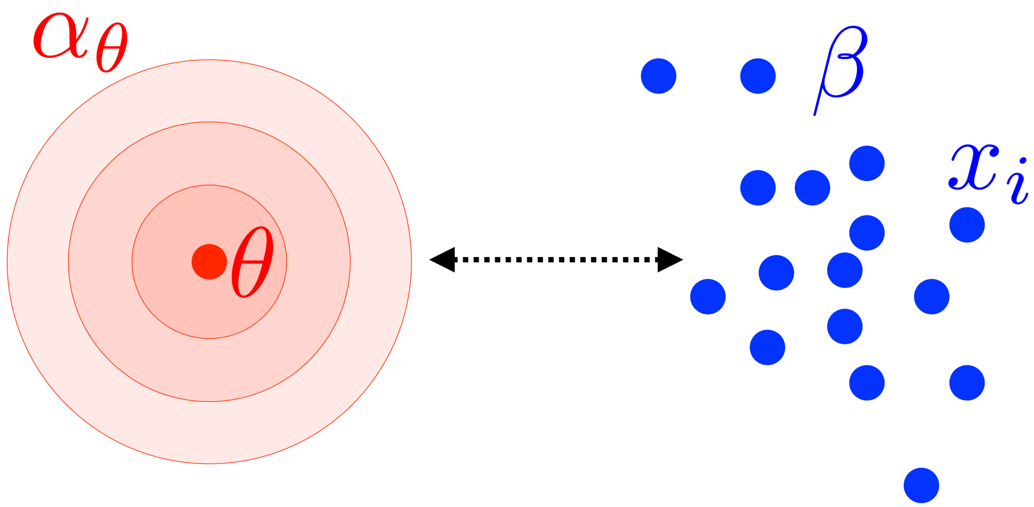
$$\min_\theta \widehat{\mathrm{KL}}(\beta|\alpha_\theta) \stackrel{\text{def.}}{=} -\sum_i \log(\rho_\theta(x_i))$$

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

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

Maximum likelihood (MLE)

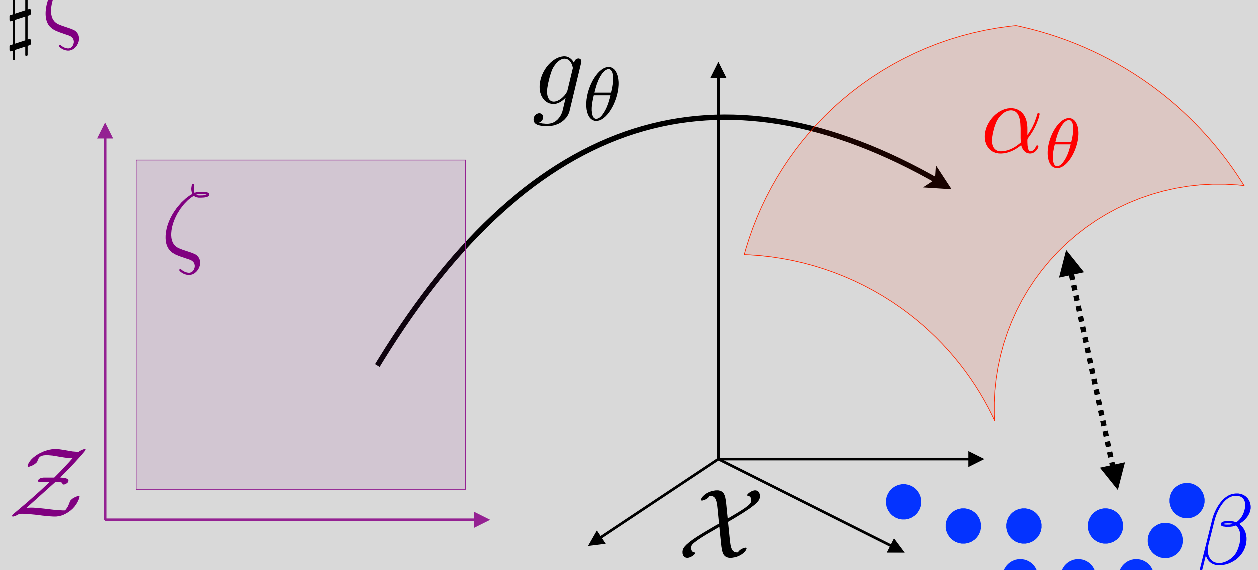
Generative model fit: $\alpha_\theta = g_{\theta,\sharp}\zeta$

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

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.

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

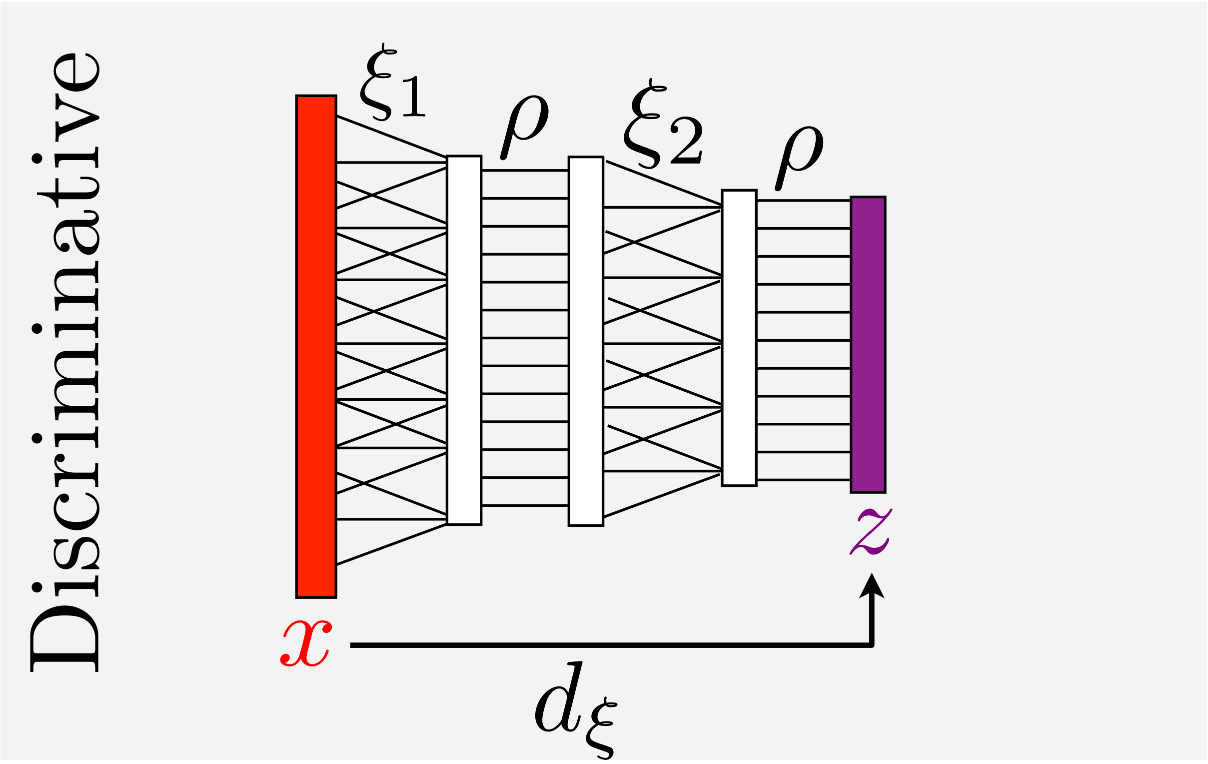


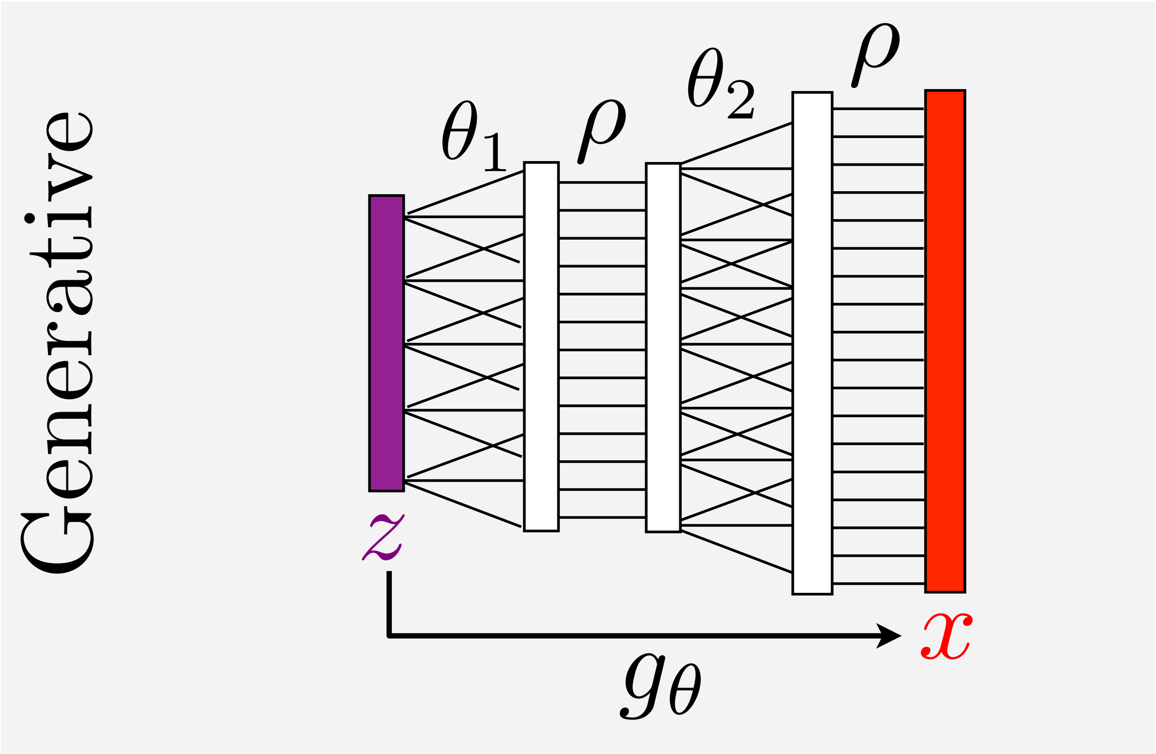
Deep Discriminative vs Generative Models

Deep networks:

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

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



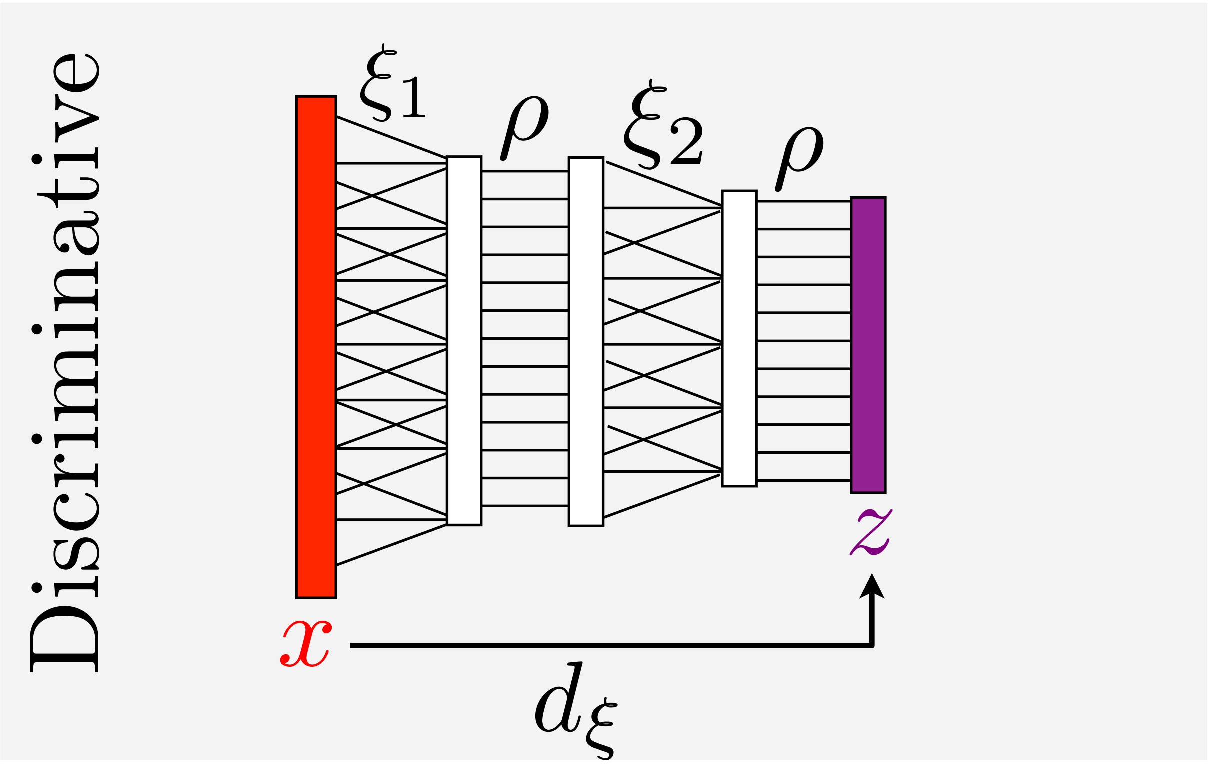


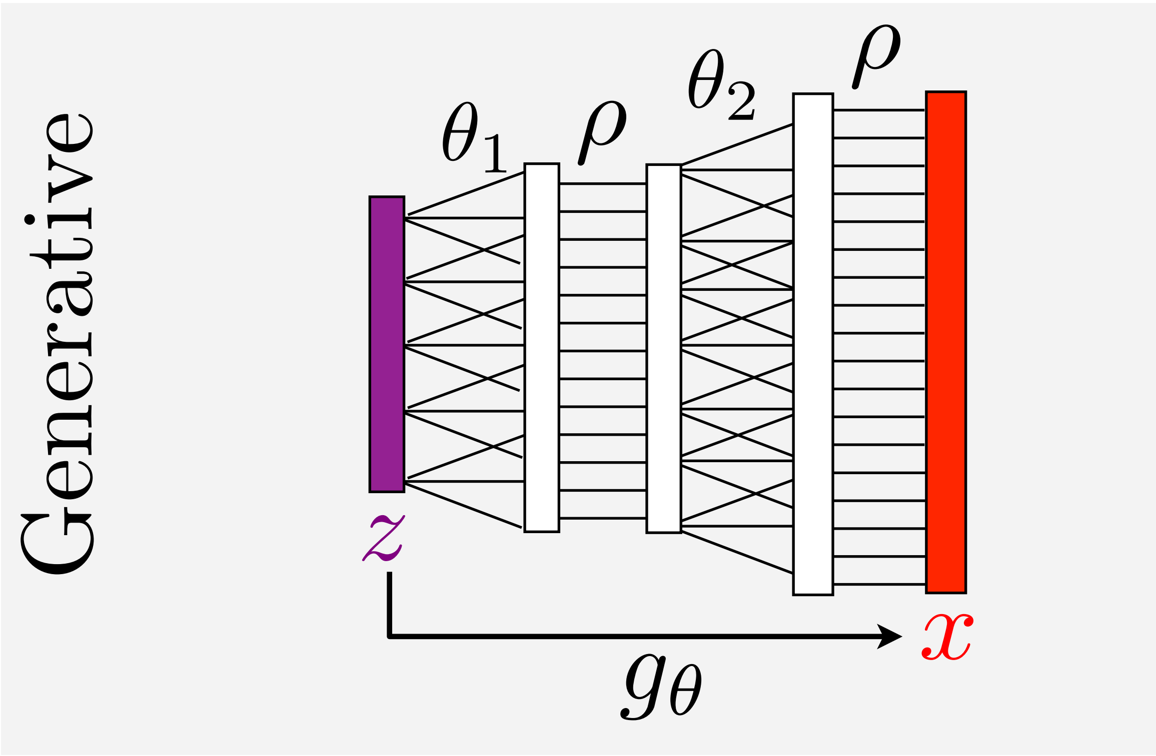
Deep Discriminative vs Generative Models

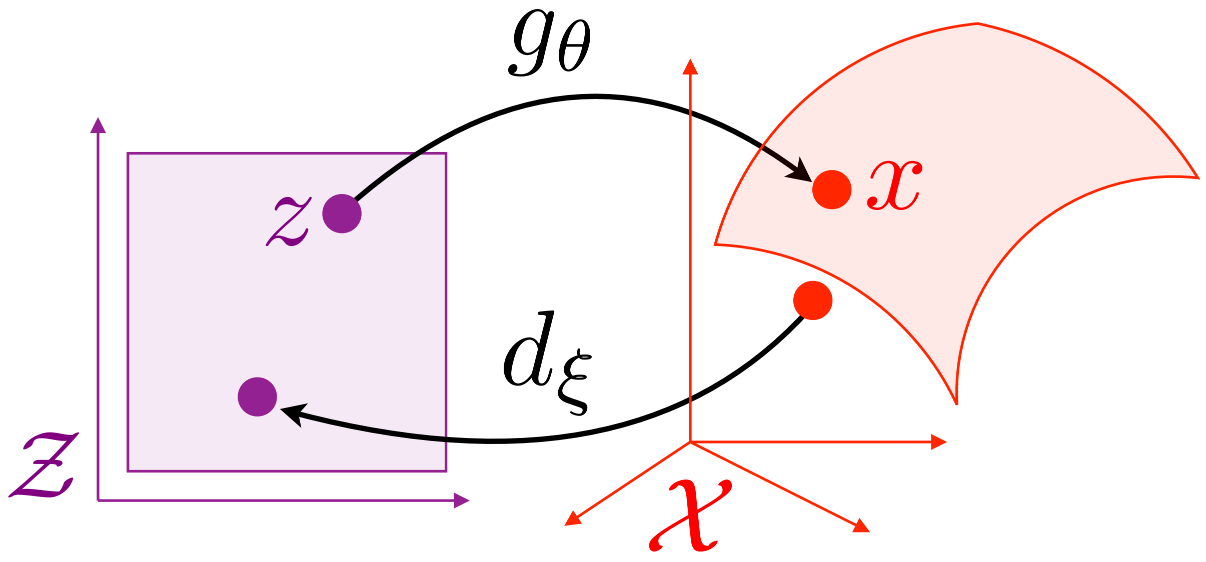
Deep networks:

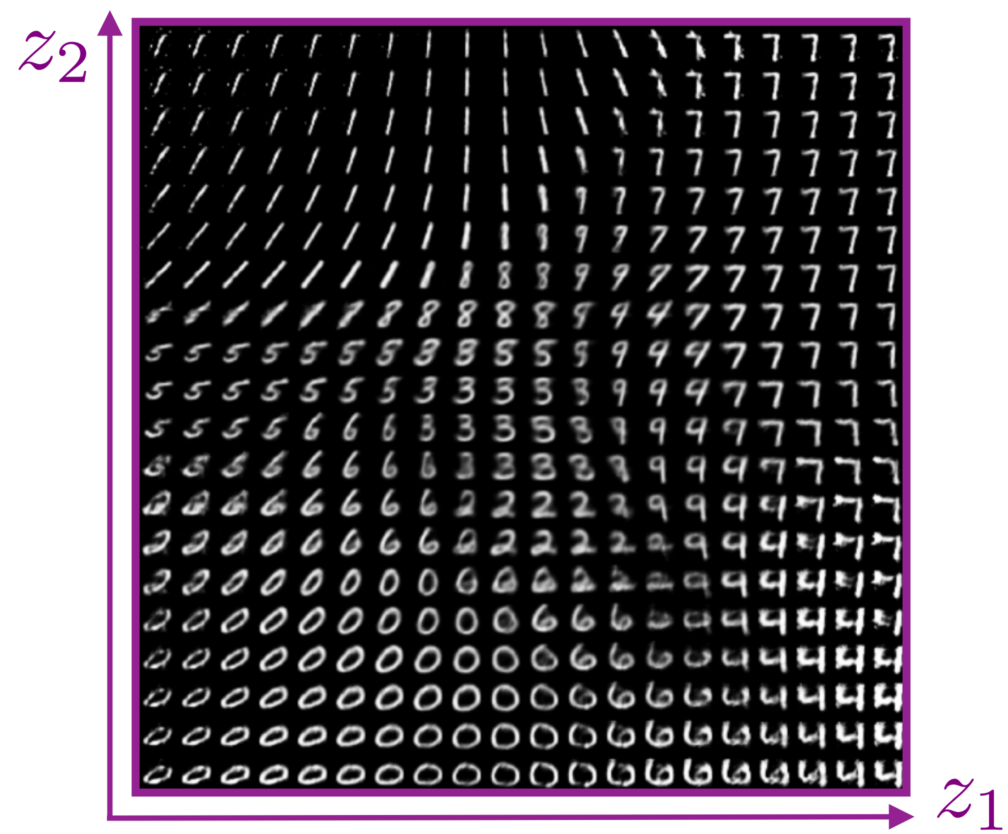
$$d_\xi(x) = \rho(\xi_K(\dots\rho(\xi_2(\rho(\xi_1(x)\dots)$$

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

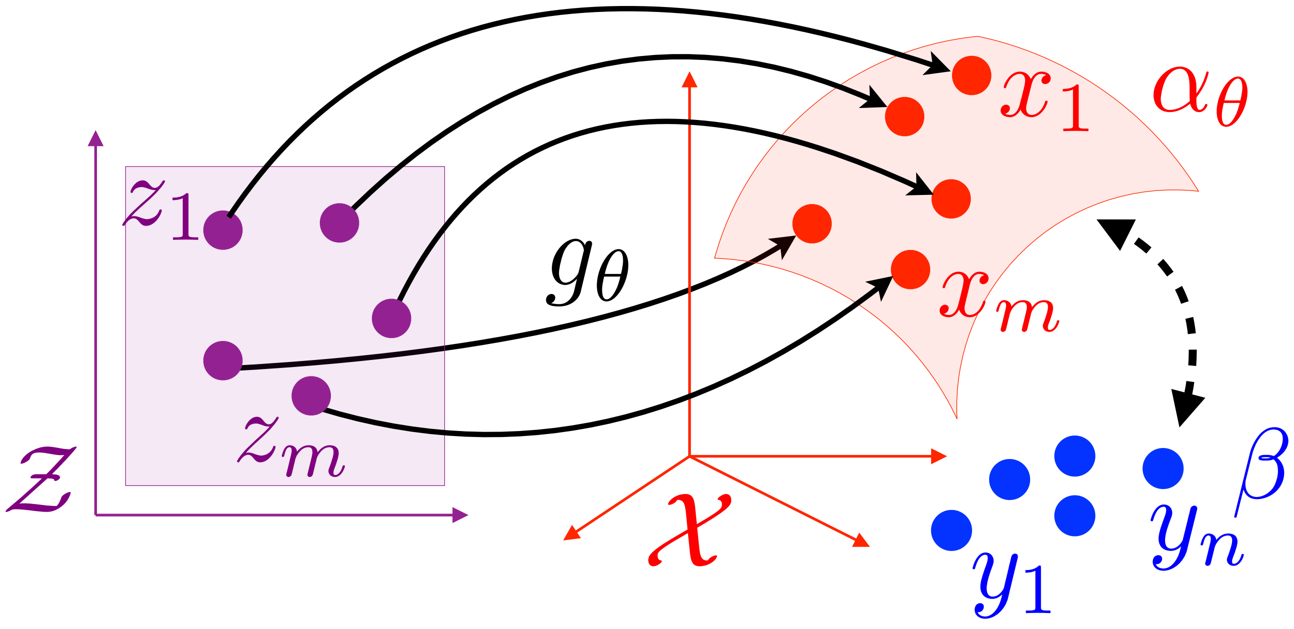








Training Architecture



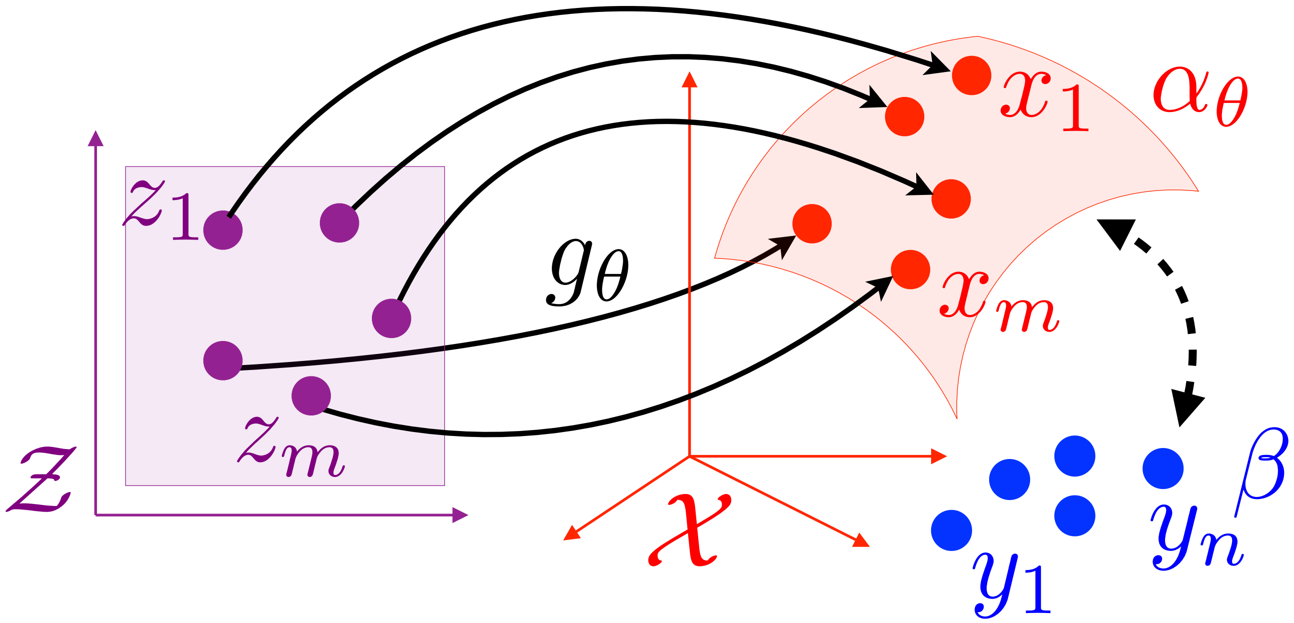
$$\min_\theta \mathcal{E}(\theta) \stackrel{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

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

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

Training Architecture

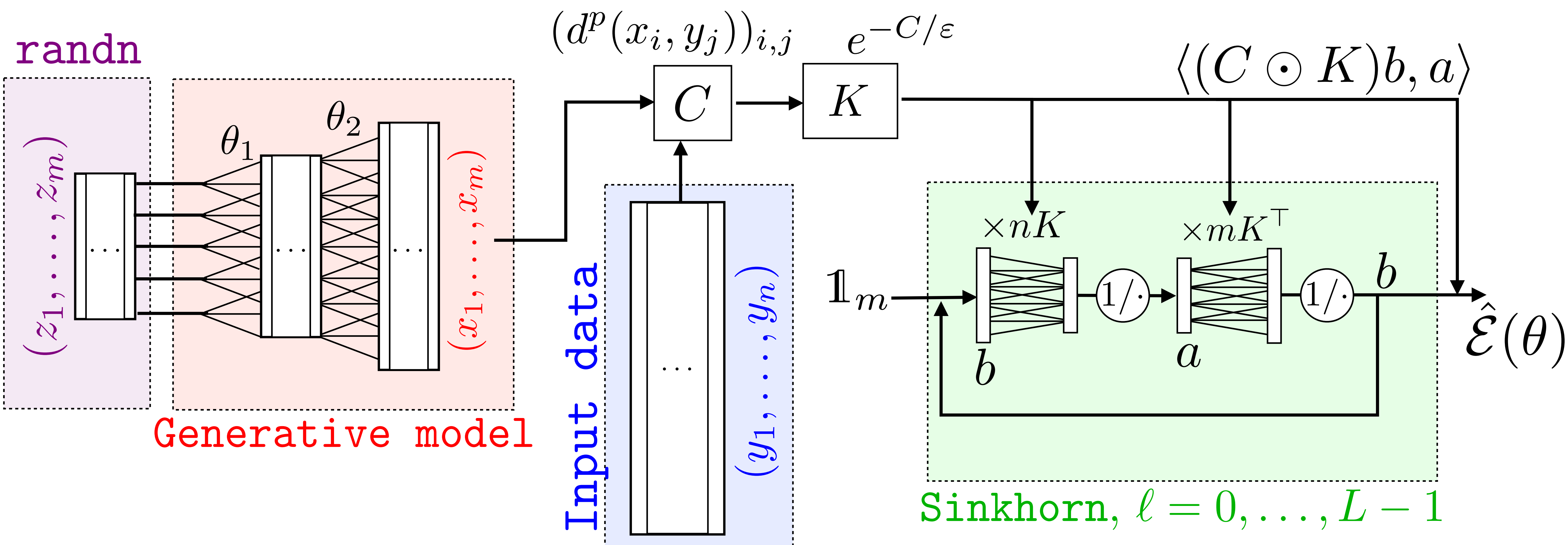


$$\min_\theta \mathcal{E}(\theta) \stackrel{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

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

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



Automatic Differentiation

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

```
def ForwardNN(A,b,Z):
    X = []
    X.append(Z)
    for r in arange(0,R):
        X.append( rhoF( A[r].dot(X[r]) + tile(b[r],[1,Z.shape[1]]) ) )
    return X
```

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

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

Automatic Differentiation

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

```
def ForwardNN(A,b,Z):
    X = []
    X.append(Z)
    for r in arange(0,R):
        X.append( rhoF( A[r].dot(X[r]) + tile(b[r],[1,Z.shape[1]]) ) )
    return X
```

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

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

Finite differences: $$\nabla \mathcal{E}(\theta) \approx \frac{1}{\varepsilon}(\mathcal{E}(\theta + \varepsilon\delta_1) - \mathcal{E}(\theta), \ldots \mathcal{E}(\theta + \varepsilon\delta_1) - \mathcal{E}(\theta))$$

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

Automatic Differentiation

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

```
def ForwardNN(A,b,Z):
    X = []
    X.append(Z)
    for r in arange(0,R):
        X.append( rhoF( A[r].dot(X[r]) + tile(b[r],[1,Z.shape[1]]) ) )
    return X
```

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

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

Finite differences: $$\nabla \mathcal{E}(\theta) \approx \frac{1}{\varepsilon}(\mathcal{E}(\theta + \varepsilon \delta_1) - \mathcal{E}(\theta), \ldots \mathcal{E}(\theta + \varepsilon \delta_1) - \mathcal{E}(\theta))$$

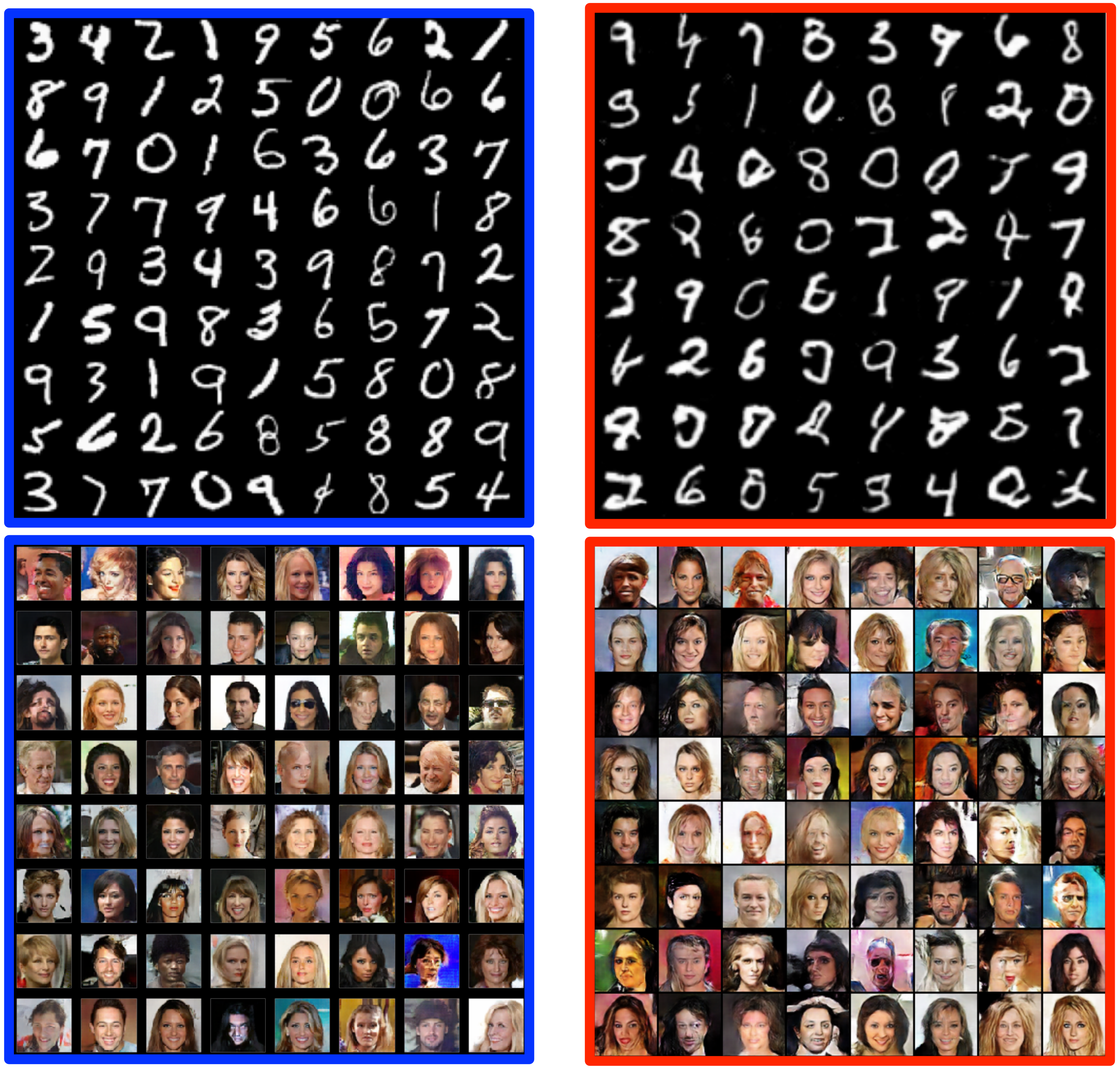
$K(n+1)$ operations, intractable for large n.

Theorem: there is an algorithm to compute $\nabla \mathcal{E}$ in $O(K)$ operations. [Seppo Linnainmaa, 1970]

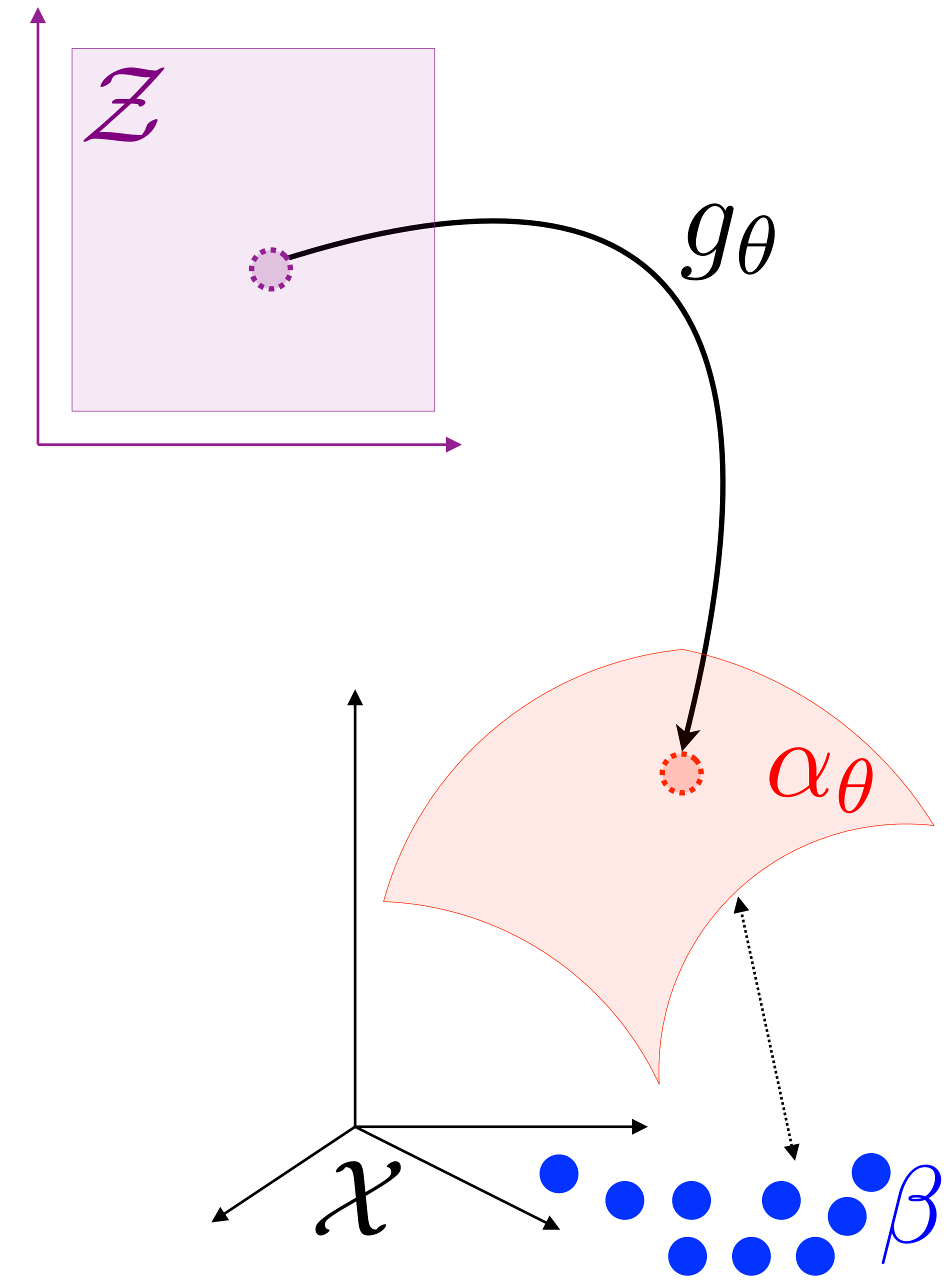
This algorithm is reverse mode automatic differentiation

```
def BackwardNN(A,b,X):
    gx = lossG(X[R],Y) # initialize the gradient
    for r in arange(R-1,-1,-1):
        M = rhoG( A[r].dot(X[r]) + tile(b[r],[1,n]) ) * gx
        gx = A[r].transpose().dot(M)
        gA[r] =  M.dot(X[r].transpose())
        gb[r] =  MakeCol(M.sum(axis=1))
    return [gA,gb]
```

Examples of Images Generation



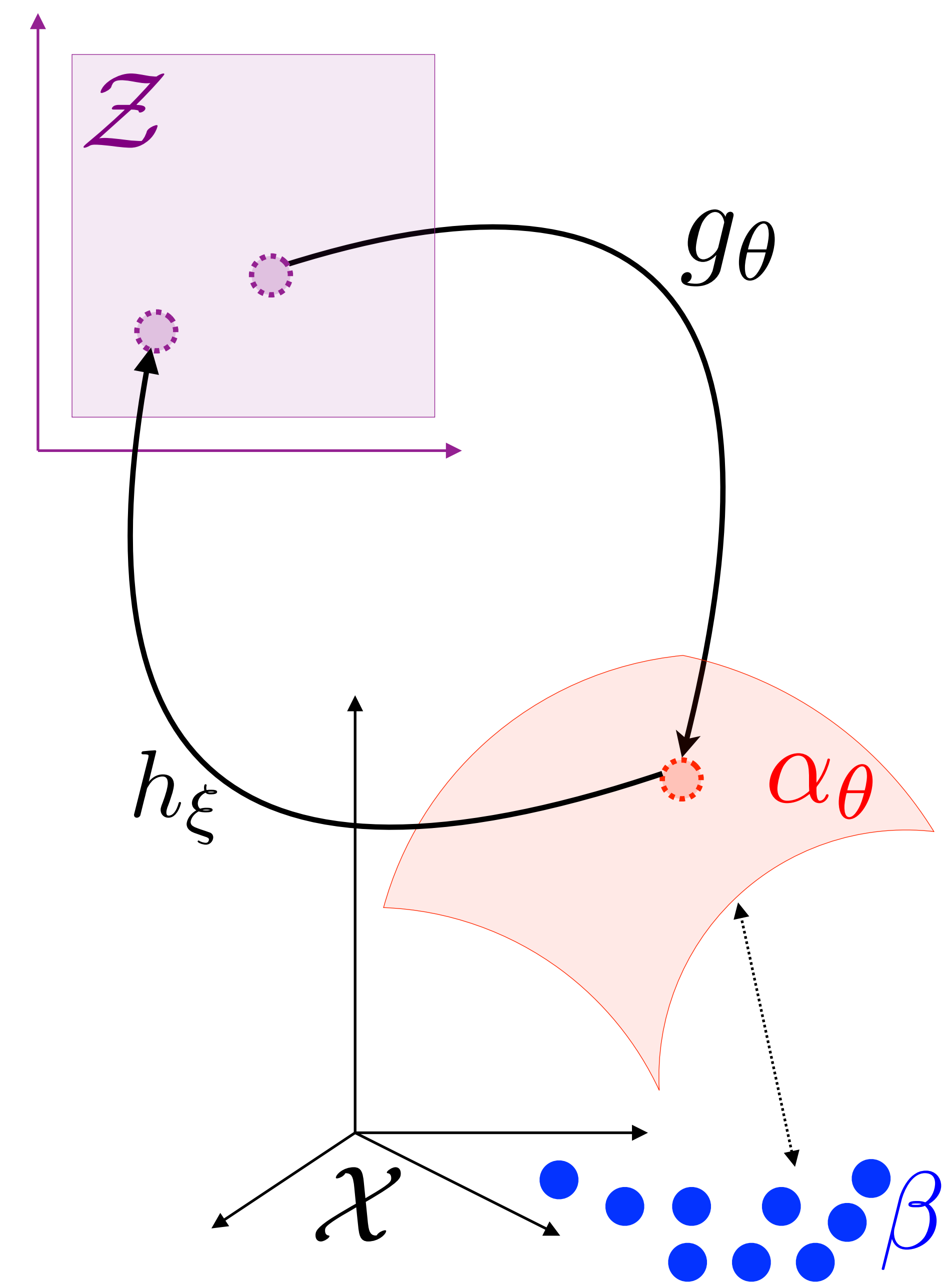
Inputs β Generated α_θ



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