Optimal Transport
for
Imaging and Learning

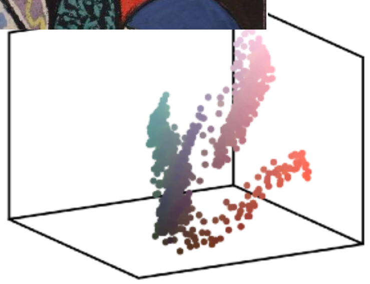
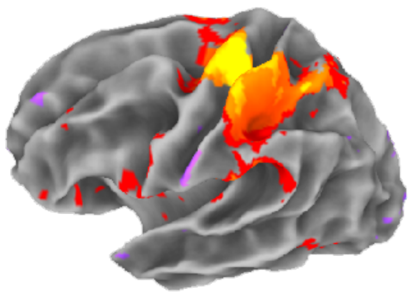
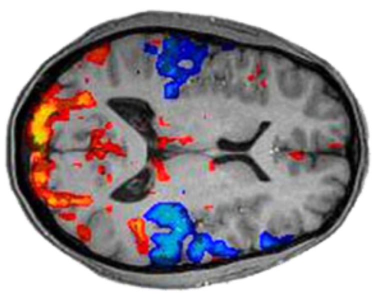
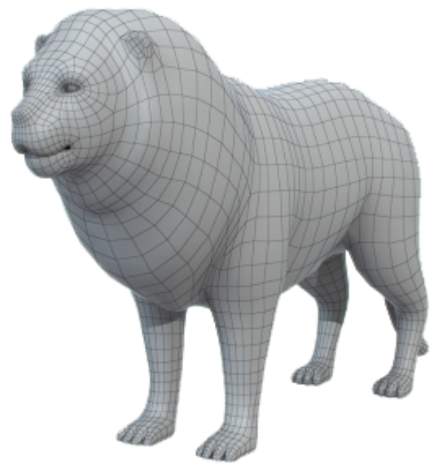
Gabriel Peyré **Marco Cuturi** **Aude Genevay**

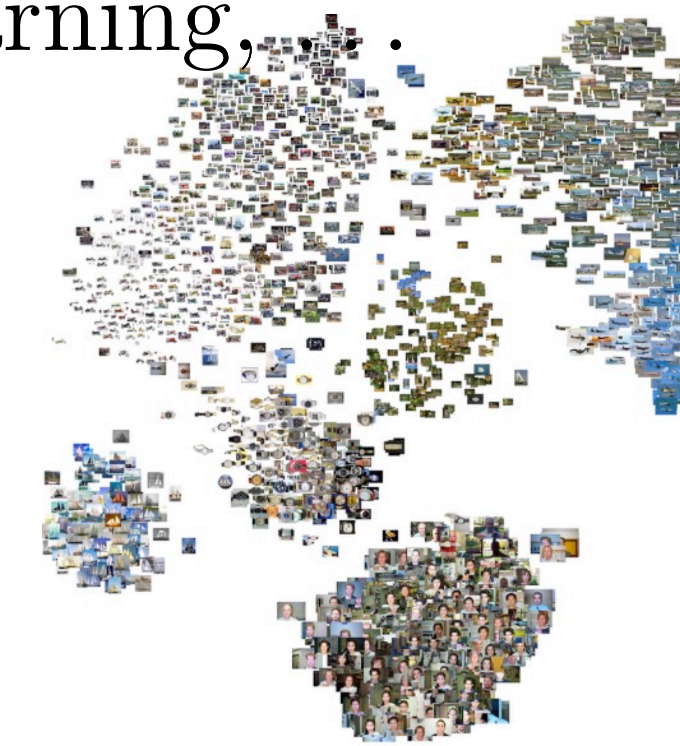
CNRS — ÉCOLE NORMALE SUPÉRIEURE
ENSAE ParisTech — École nationale de la statistique et de l'administration économique — université PARIS-SACLAY
DAUPHINE UNIVERSITÉ PARIS — ÉCOLE NORMALE SUPÉRIEURE

www.numerical-tours.com

Comparing Measures and Spaces

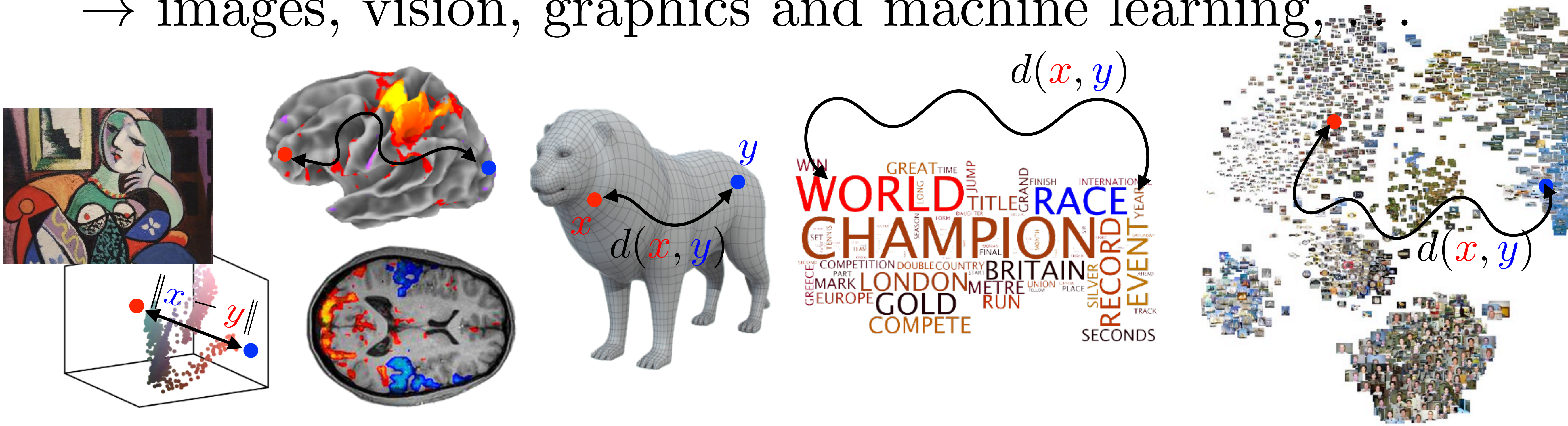
- *Probability distributions and histograms*
  $\rightarrow$ images, vision, graphics and machine learning, . . .

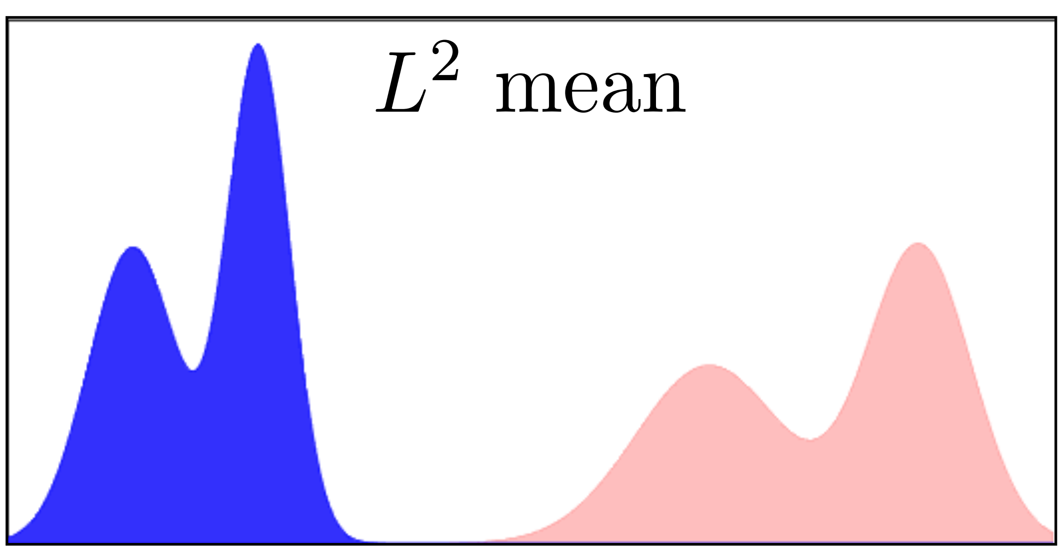



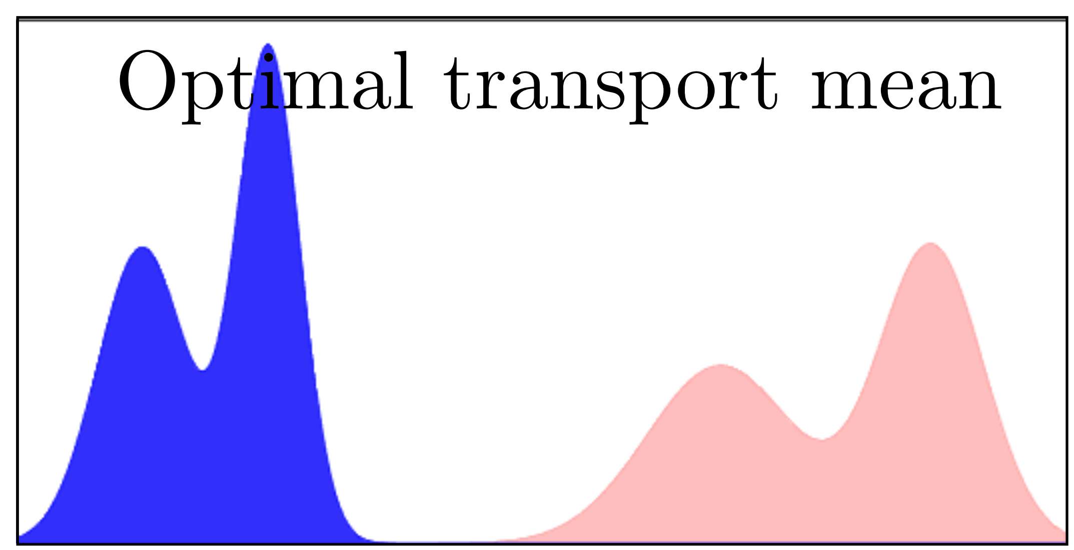
Comparing Measures and Spaces

- *Probability distributions and histograms*
  $\rightarrow$ images, vision, graphics and machine learning, . . .

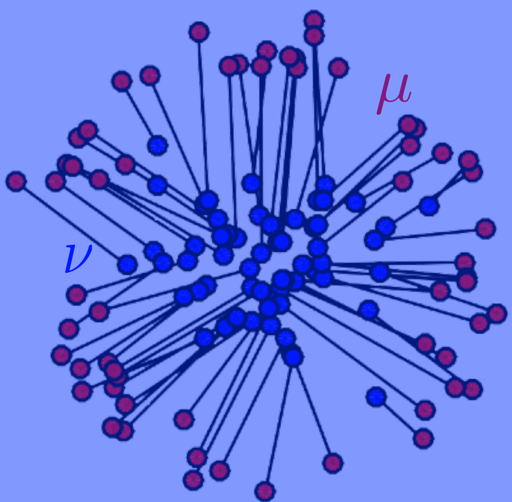


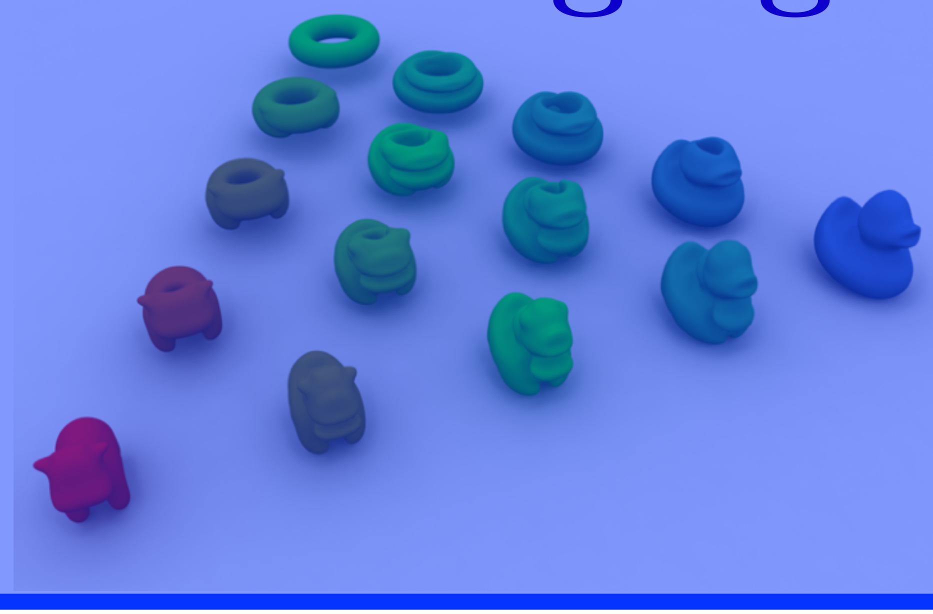
- *Optimal transport*
  $\rightarrow$ takes into account a metric d.

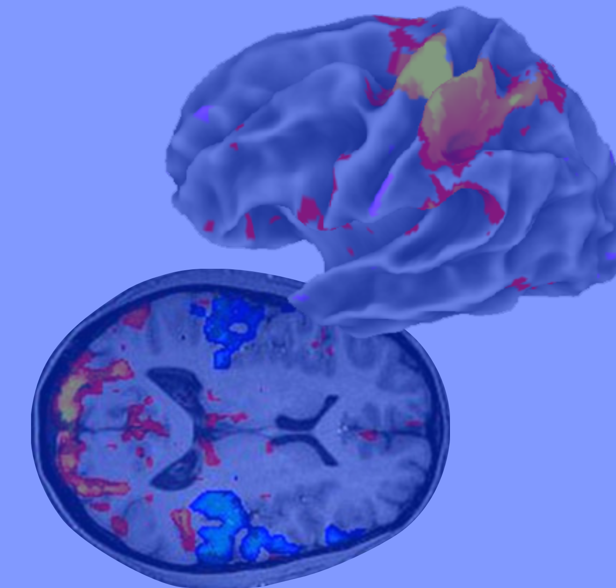




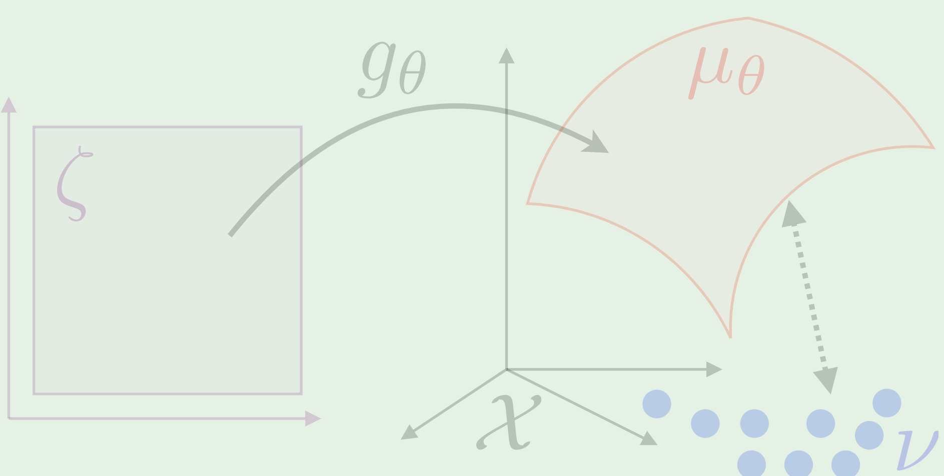
1. OT for Imaging Sciences

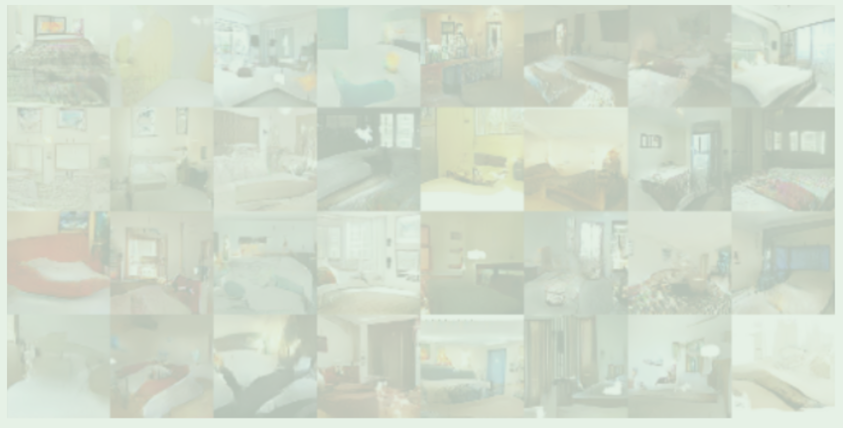


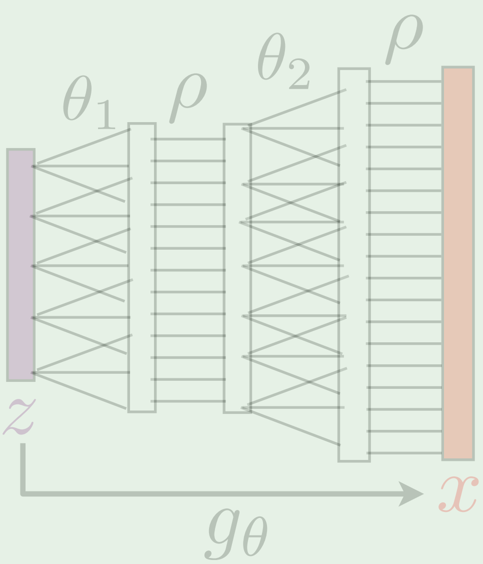




2. OT for Machine Learning



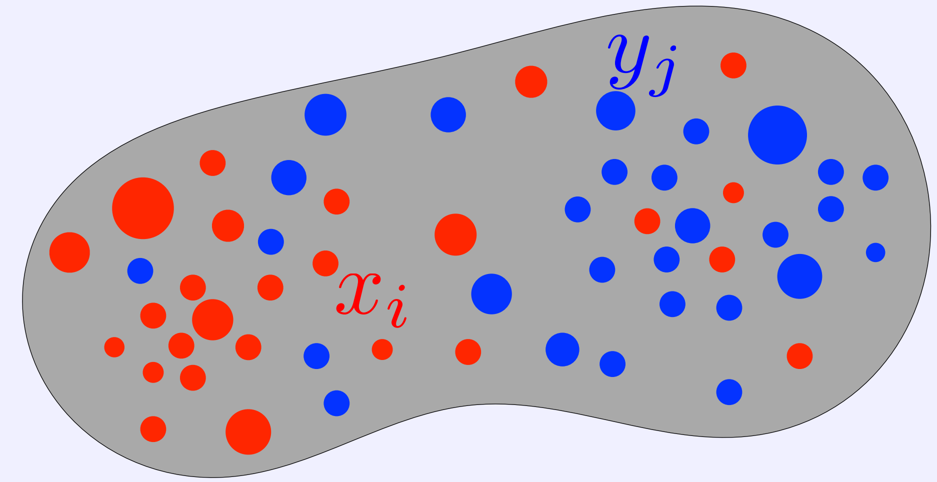




Couplings and Optimal Transport

Input distributions $\mu = \sum_i \mu_i \delta_{x_i}$

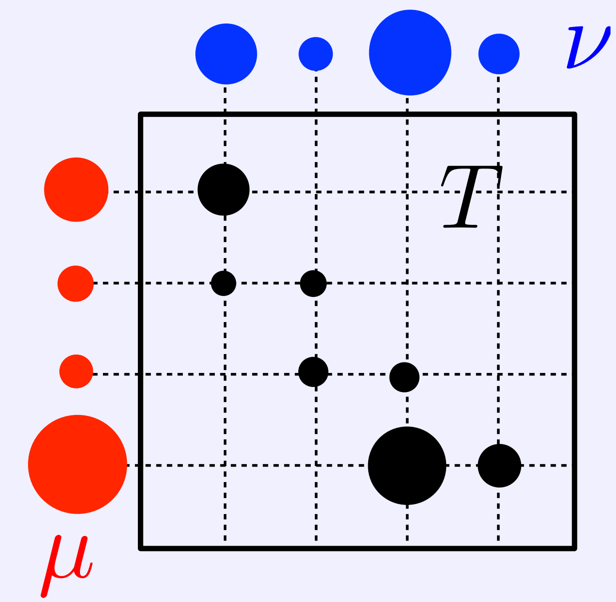
$\nu = \sum_j \nu_j \delta_{y_j}$



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

Weights $\mu_i \geqslant 0$, $\nu_j \geqslant 0$.

$\sum_{i=1}^{N_1} \mu_i = \sum_{j=1}^{N_2} \nu_j = 1$



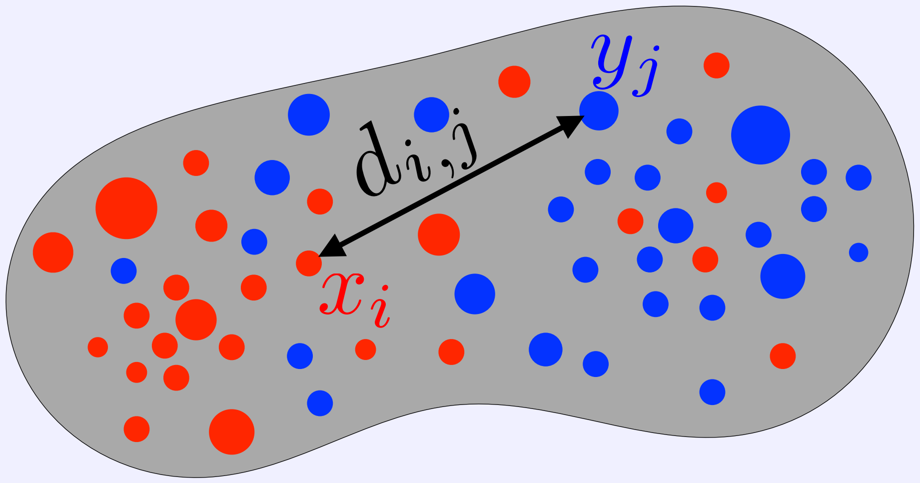
Def. *Couplings*

$$\mathcal{C}_{\mu,\nu} \stackrel{\text{def.}}{=} \left\{ T \in \mathbb{R}_+^{N_1 \times N_2} \; ; \; T \mathbb{1}_{N_1} = \mu, T^\top \mathbb{1}_{N_2} = \nu \right\}$$

Couplings and Optimal Transport

Input distributions $\mu = \sum_i \mu_i \delta_{x_i}$

$\nu = \sum_j \nu_j \delta_{y_j}$

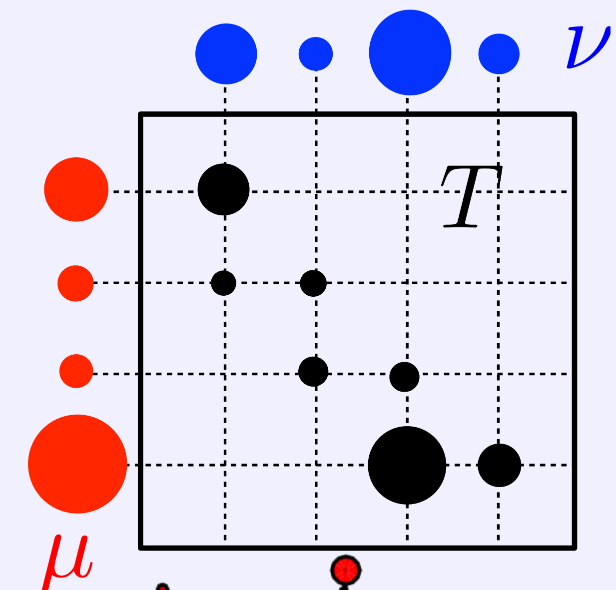


Points $(x_i)_i$, $(y_j)_j$

Weights $\mu_i \geqslant 0$, $\nu_j \geqslant 0$.

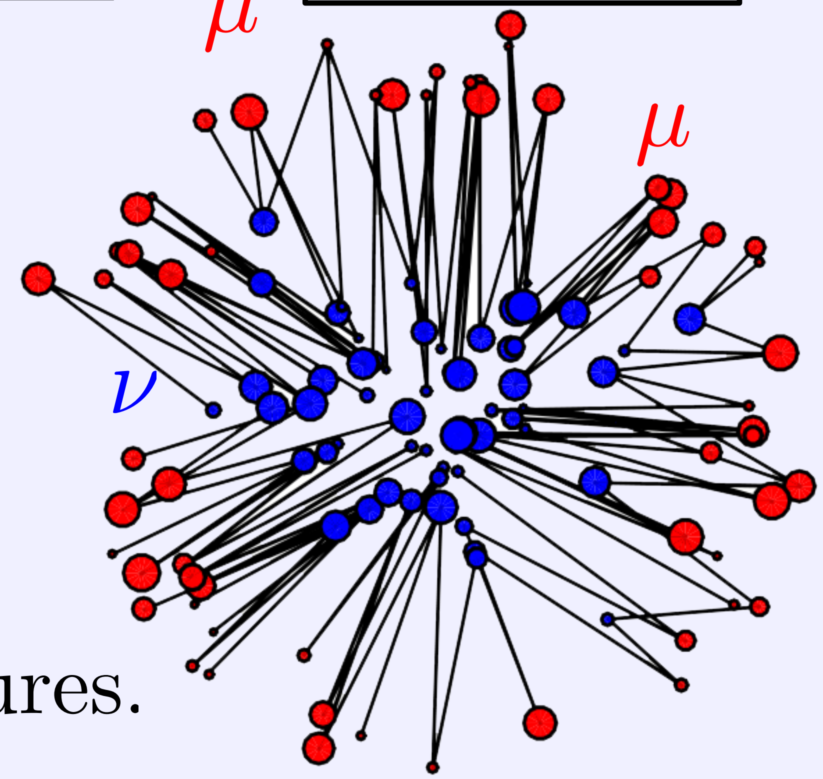
$\sum_{i=1}^{N_1} \mu_i = \sum_{j=1}^{N_2} \nu_j = 1$

$d_{i,j} = d(x_i, y_j)$



Def. *Couplings*

$$\mathcal{C}_{\mu,\nu} \overset{\text{def.}}{=} \left\{ T \in \mathbb{R}_+^{N_1 \times N_2} \;;\; T \mathbb{1}_{N_1} = \mu, T^\top \mathbb{1}_{N_2} = \nu \right\}$$



Def. *Wasserstein Distance / EMD*

$$W_p^p(\mu, \nu) \overset{\text{def.}}{=} \min \left\{ \sum_{i,j} T_{i,j} d_{i,j}^p \;;\; T \in \mathcal{C}_{\mu,\nu} \right\}$$

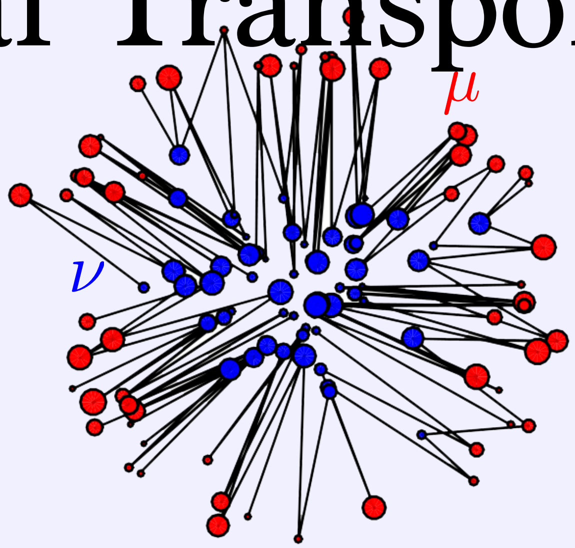
[Kantorovich 1942]

$\rightarrow W_p$ is a distance over Radon probability measures.

Numerical Optimal Transport

Linear programming:

$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$



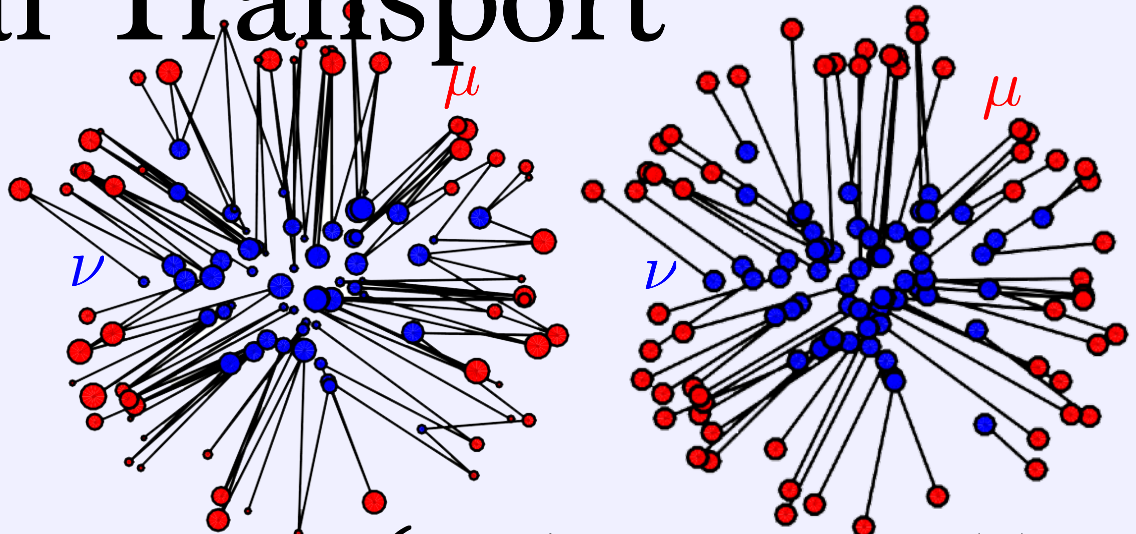
Numerical Optimal Transport

Linear programming:

$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$

Hungarian/Auction: $\sim O(N^3)$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$



$$T_{i,j} = \begin{cases} 1/N \text{ if } j = \sigma(i), \\ 0 \text{ otherwise.} \end{cases}$$

Numerical Optimal Transport

Linear programming:

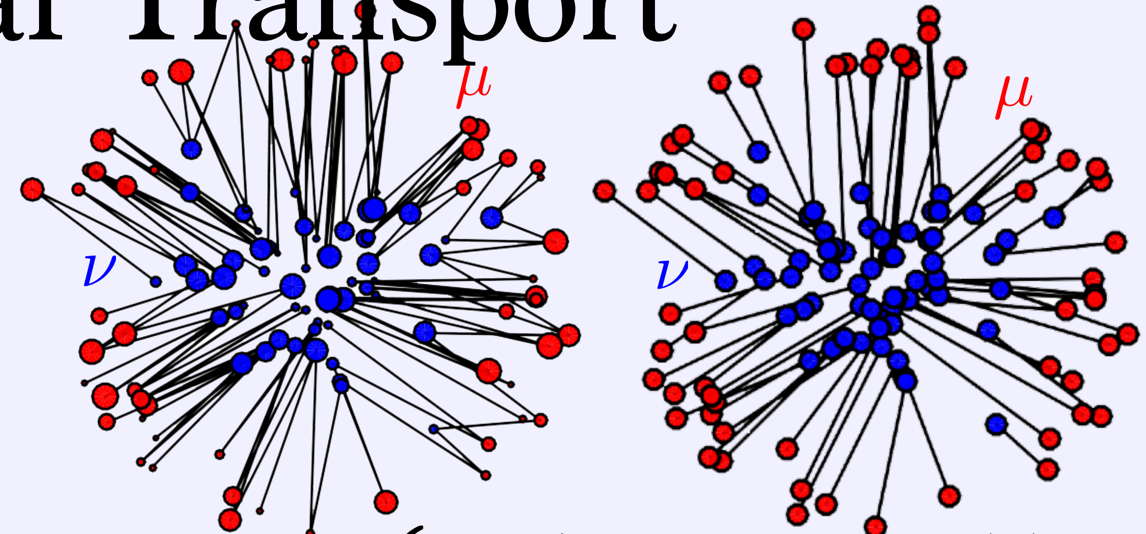
$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$

Hungarian/Auction: $\sim O(N^3)$

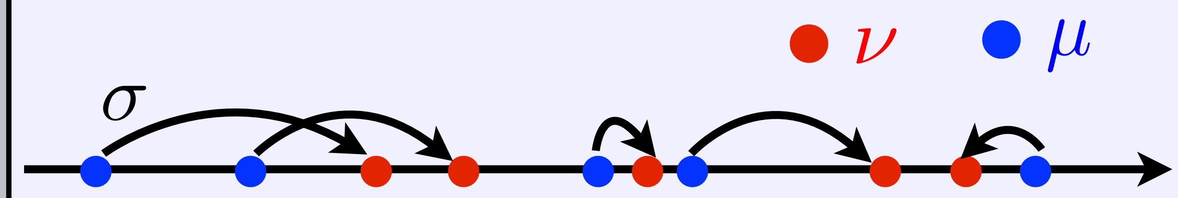
$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$

1-D case, $d = |\cdot|^p, p \geqslant 1$.

$\rightarrow$ sorting, $O(N \log(N))$ operations.



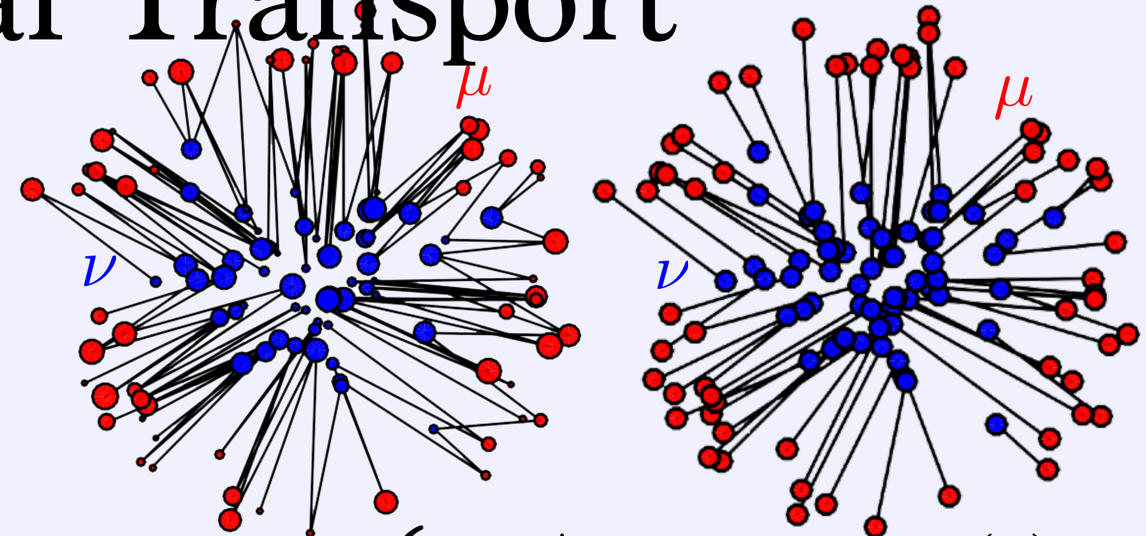
$$T_{i,j} = \begin{cases} 1/N \text{ if } j = \sigma(i), \\ 0 \text{ otherwise.} \end{cases}$$



Numerical Optimal Transport

Linear programming:

$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$



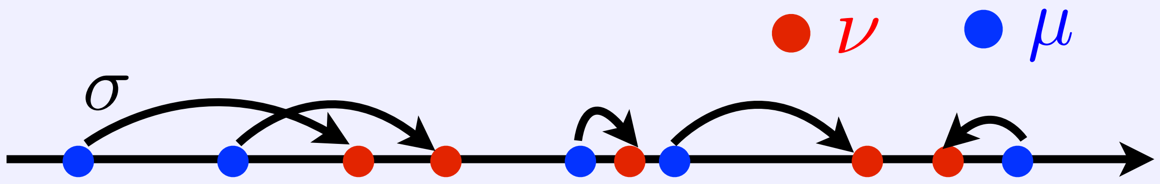
Hungarian/Auction: $\sim O(N^3)$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$

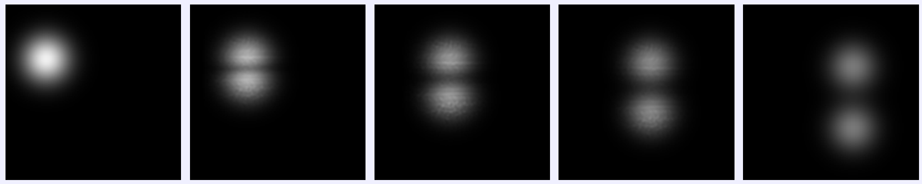
$$T_{i,j} = \begin{cases} 1/N \text{ if } j = \sigma(i), \\ 0 \text{ otherwise.} \end{cases}$$

1-D case, $d = |\cdot|^p, p \geqslant 1$.

$\rightarrow$ sorting, $O(N\log(N))$ operations.



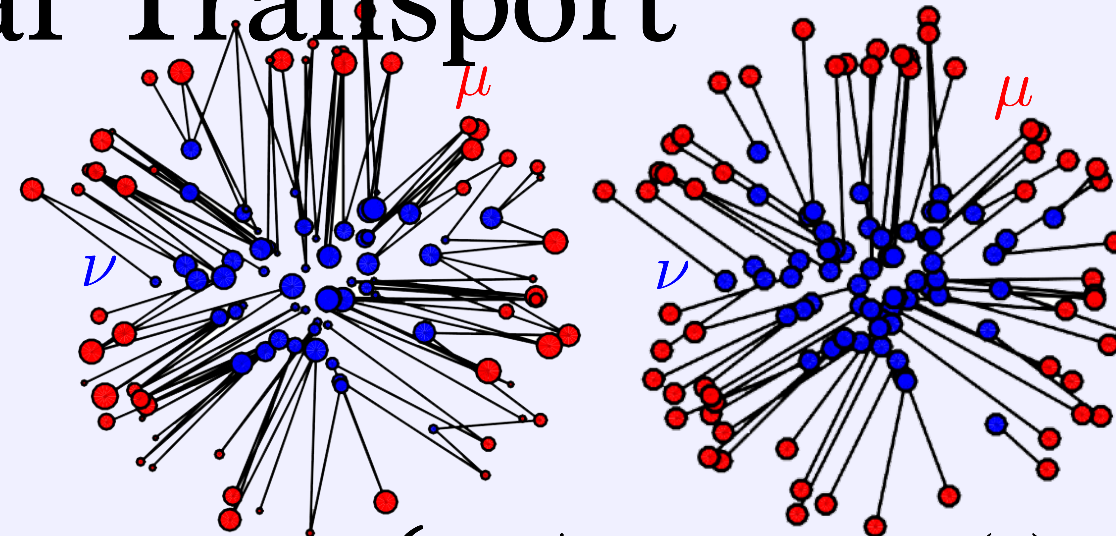
Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.



Numerical Optimal Transport

Linear programming:

$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$

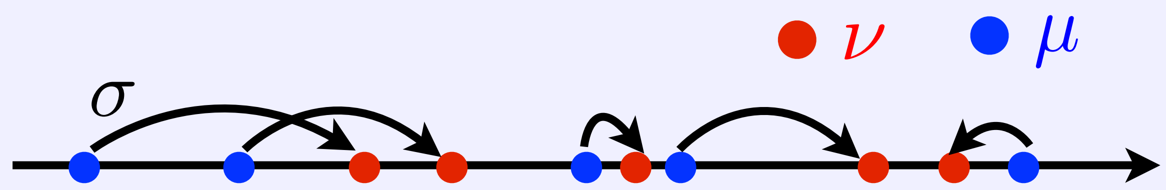


Hungarian/Auction: $\sim O(N^3)$

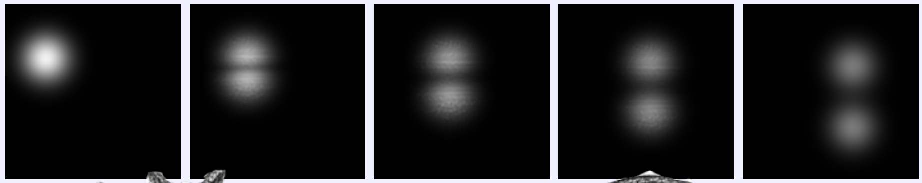
$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$

$$T_{i,j} = \begin{cases} 1/N \text{ if } j = \sigma(i), \\ 0 \text{ otherwise.} \end{cases}$$

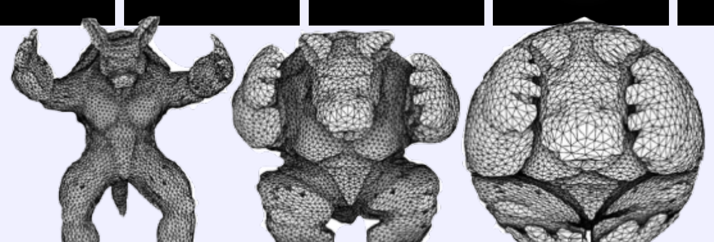
1-D case, $d = |\cdot|^p, p \geqslant 1$.
$\rightarrow$ sorting, $O(N \log(N))$ operations.



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



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

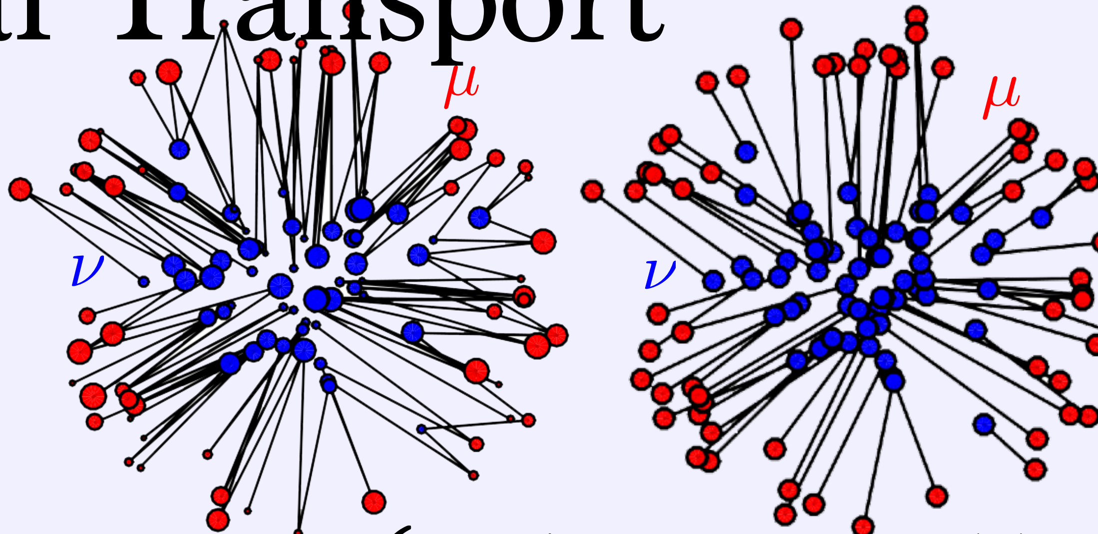


[Levy,'15]

Numerical Optimal Transport

Linear programming:

$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$

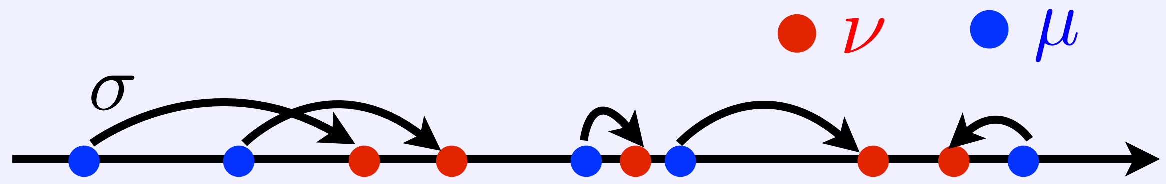


Hungarian/Auction: $\sim O(N^3)$

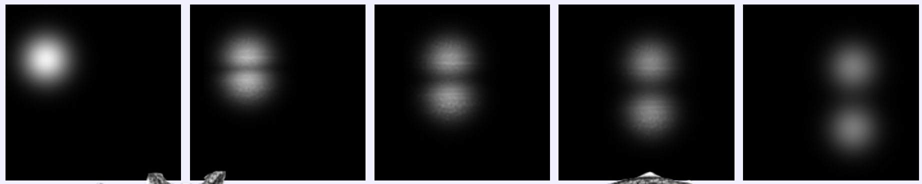
$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$

$$T_{i,j} = \begin{cases} 1/N \text{ if } j = \sigma(i), \\ 0 \text{ otherwise.} \end{cases}$$

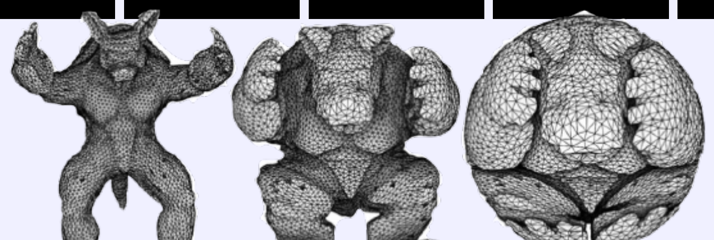
1-D case, $d = |\cdot|^p, p \geqslant 1$.
$\rightarrow$ sorting, $O(N \log(N))$ operations.



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



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



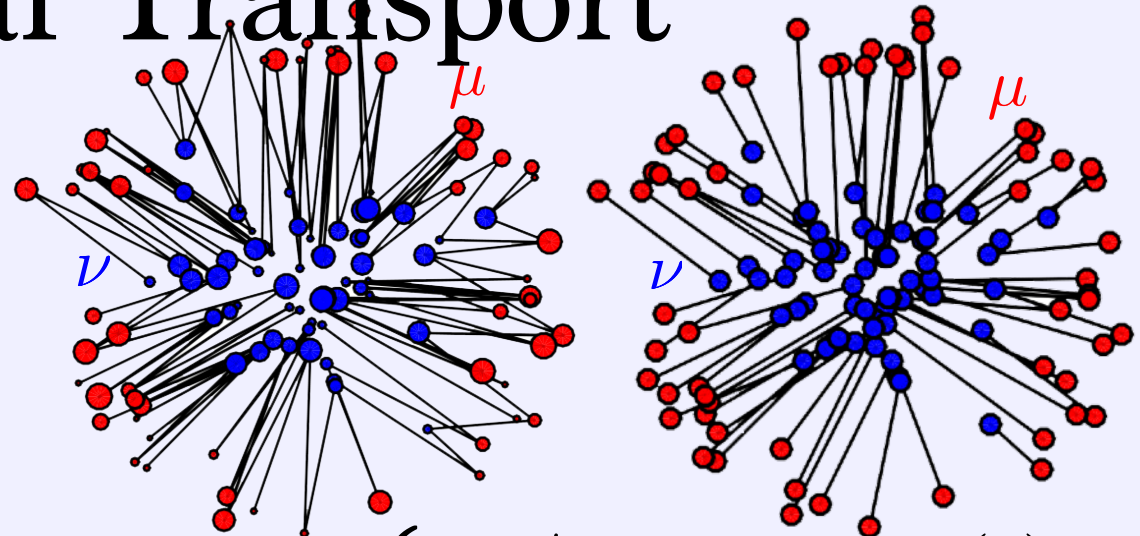
[Levy,'15]

$$d = \|\cdot\|, p = 1 : W_1(\mu, \nu) = \min_{\mathrm{div}(v)=\mu-\nu} \int \|u(x)\| \mathrm{d}x \rightarrow \text{max-flow.}$$

Numerical Optimal Transport

Linear programming:

$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$

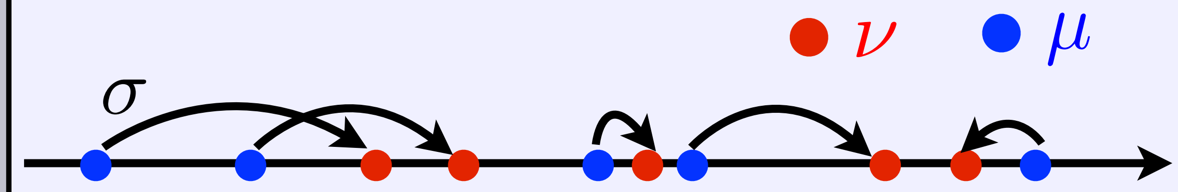


Hungarian/Auction: $\sim O(N^3)$

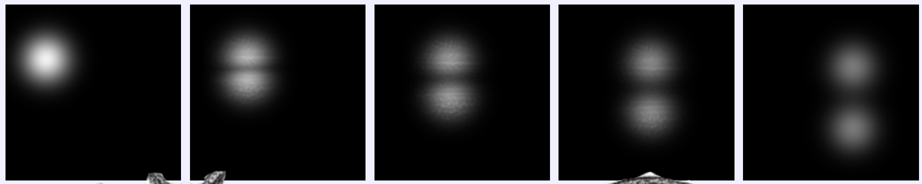
$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$

$$T_{i,j} = \begin{cases} 1/N \text{ if } j = \sigma(i), \\ 0 \text{ otherwise.} \end{cases}$$

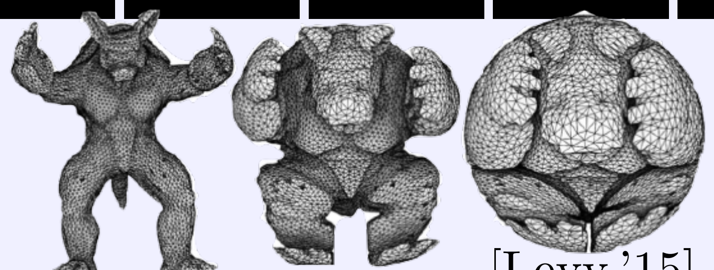
1-D case, $d = |\cdot|^p, p \geqslant 1$.
$\rightarrow$ sorting, $O(N \log(N))$ operations.



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



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



[Levy,'15]

$$d = \|\cdot\|, p = 1 : W_1(\mu, \nu) = \min_{\operatorname{div}(v) = \mu - \nu} \int \|u(x)\| \mathrm{d}x \quad \rightarrow \text{ max-flow.}$$

Need for fast approximate algorithms for generic c.

Entropic Regularization

Entropy: $H(T) \stackrel{\text{def.}}{=} -\sum_{i,j=1}^{N} T_{i,j}(\log(T_{i,j})-1)$

Def. *Regularized OT:* [Cuturi NIPS'13]

$$\min_{T} \left\{ \sum_{i,j} d_{i,j}^{p} T_{i,j} - \varepsilon H(T) \;;\; T \in \mathcal{C}_{\mu,\nu} \right\}$$

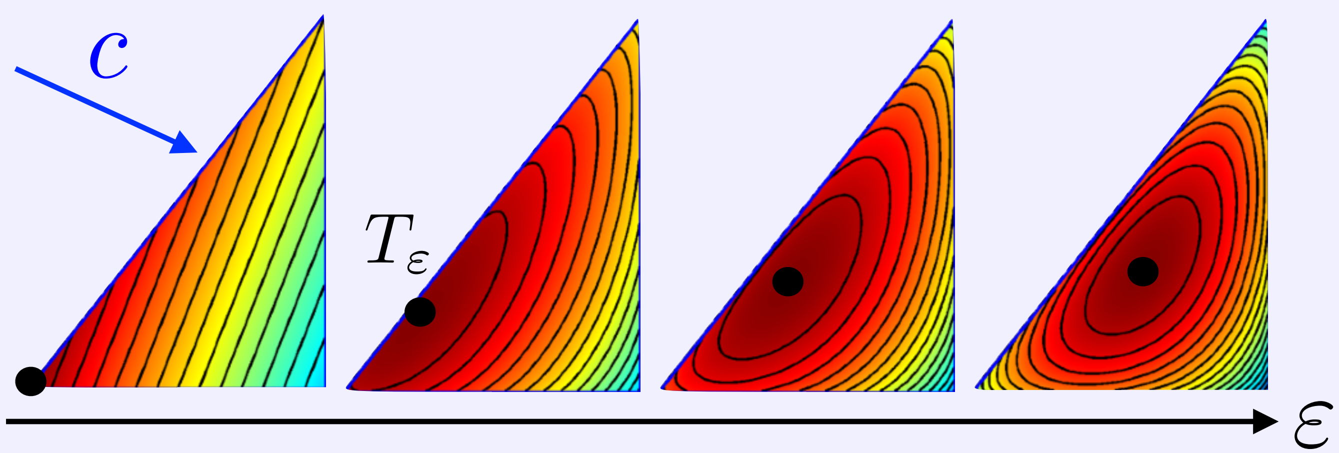
Entropic Regularization

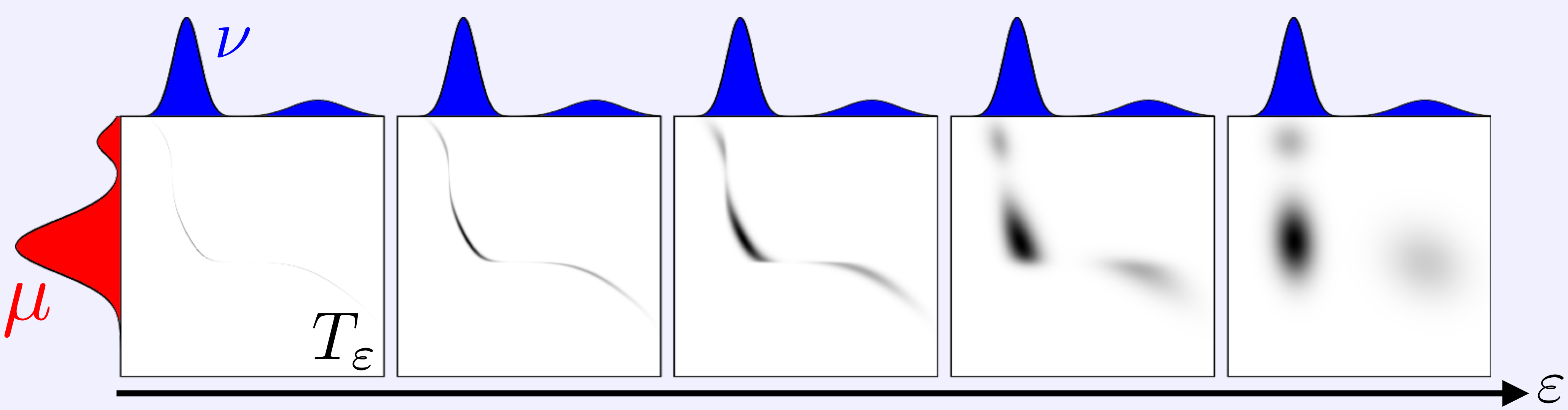
Entropy: $H(T) \stackrel{\text{def.}}{=} -\sum_{i,j=1}^{N} T_{i,j}(\log(T_{i,j}) - 1)$

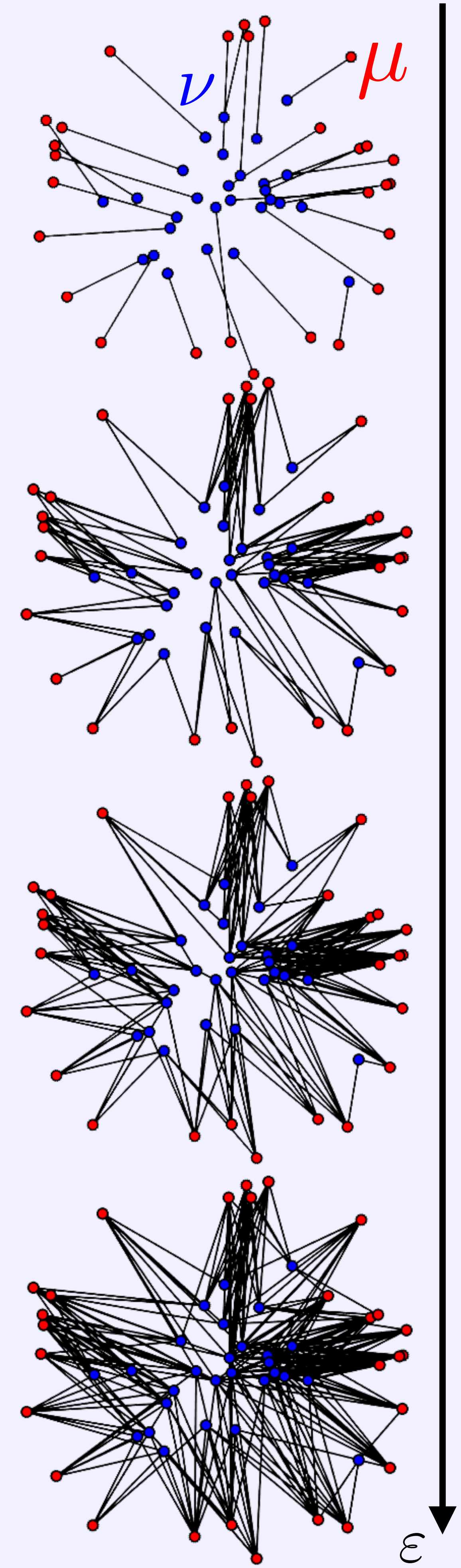
Def. *Regularized OT:* [Cuturi NIPS'13]

$$\min_{T} \left\{ \sum_{i,j} d_{i,j}^{p} T_{i,j} - \varepsilon H(T) \; ; \; T \in \mathcal{C}_{\mu,\nu} \right\}$$

Regularization impact on solution:







Sinkhorn's Algorithm

$$\min_{T} \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) \; ; \; T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\star)$$

Prop. One has $T = \mathrm{diag}(a) K \, \mathrm{diag}(b)$, where $K = e^{-\frac{d^p}{\varepsilon}}$.

Sinkhorn's Algorithm

$$\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) \; ; \; T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\star)$$

Prop. One has $T = \mathrm{diag}(a) K \,\mathrm{diag}(b)$, where $K = e^{-\frac{d^p}{\varepsilon}}$.

Row constraint: $T \mathbb{1}_{N_2} = \mu \iff a \odot (Kb) = \mu$

Col. constraint: $T^\top \mathbb{1}_{N_2} = \nu \iff b \odot (K^\top a) = \nu$

Sinkhorn iterations: $a \leftarrow \dfrac{\mu}{Kb}$ and $b \leftarrow \dfrac{\nu}{K^\top a}$

Sinkhorn's Algorithm

$$\min_{T} \left\{ \textstyle\sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) \; ; \; T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\star)$$

Prop. One has $T = \mathrm{diag}(a) K \,\mathrm{diag}(b)$, where $K = e^{-\frac{d^p}{\varepsilon}}$.

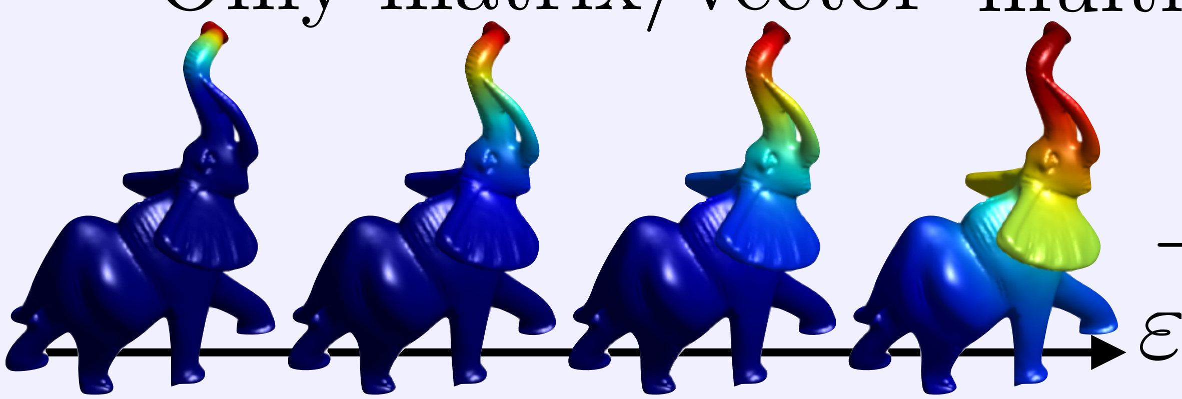
Row constraint: $T \mathbb{1}_{N_2} = \mu \iff a \odot (Kb) = \mu$

Col. constraint: $T^\top \mathbb{1}_{N_2} = \nu \iff b \odot (K^\top a) = \nu$

Sinkhorn iterations: $a \leftarrow \frac{\mu}{Kb}$ and $b \leftarrow \frac{\nu}{K^\top a}$

Only matrix/vector multiplications. $\rightarrow$ Parallelizable.

$\rightarrow$ Streams well on GPU.



$\rightarrow$ convolutive/heat structure for K

Sinkhorn's Algorithm

$$\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) \; ; \; T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\star)$$

Prop. One has $T = \mathrm{diag}(a) K \,\mathrm{diag}(b)$, where $K = e^{-\frac{d^p}{\varepsilon}}$.

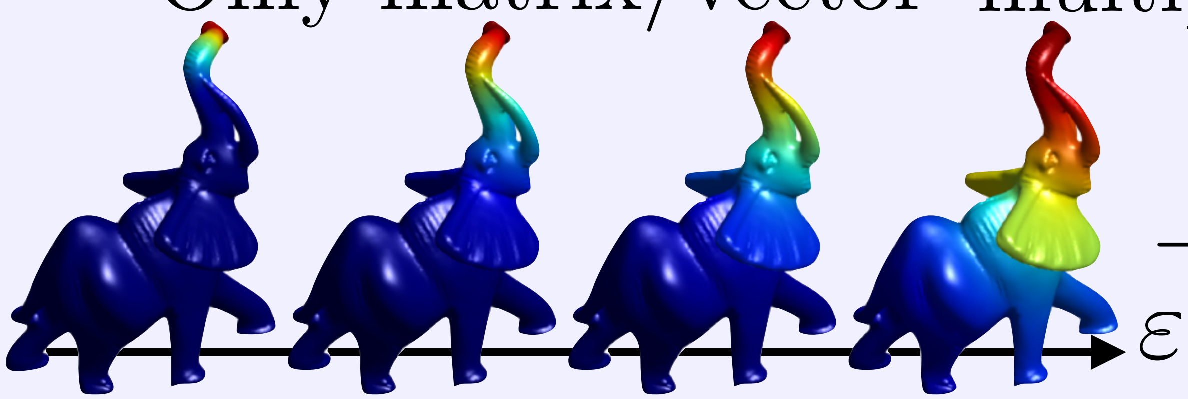
Row constraint: $T \mathbb{1}_{N_2} = \mu \iff a \odot (Kb) = \mu$

Col. constraint: $T^\top \mathbb{1}_{N_2} = \nu \iff b \odot (K^\top a) = \nu$

Sinkhorn iterations: $a \leftarrow \dfrac{\mu}{Kb}$ and $b \leftarrow \dfrac{\nu}{K^\top a}$

Only matrix/vector multiplications. $\rightarrow$ Parallelizable.

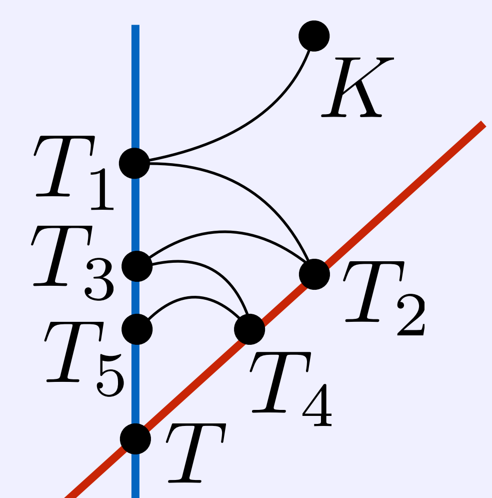
$\rightarrow$ Streams well on GPU.



$\rightarrow$ convolutive/heat structure for K

Prop. $(\star) \iff \min_T \{\mathrm{KL}(T|K) \; ; \; T \in \mathcal{C}_{\mu,\nu}\}$

Sinkhorn $\iff$ iterative projections.

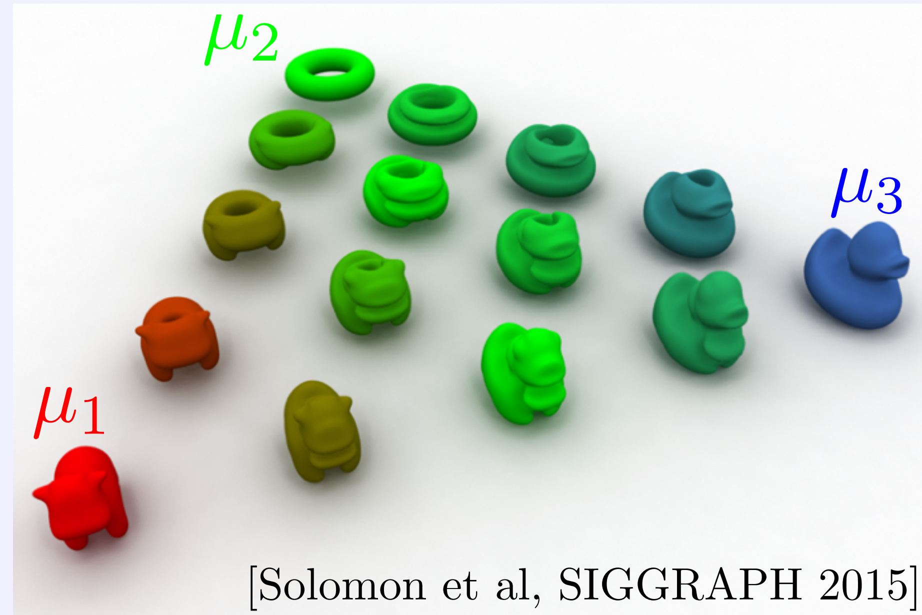


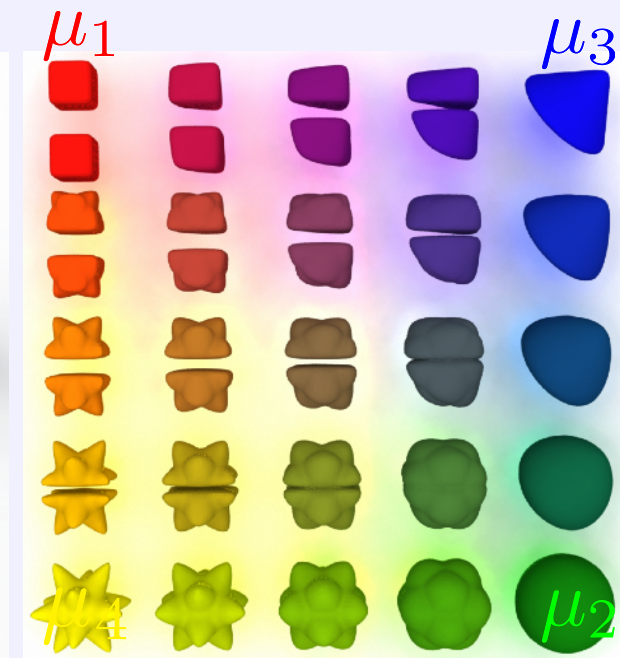
Generalizations

OT barycenters:

$$\min_{\nu} \sum_k \lambda_k W_2^2(\mu_k, \nu)$$

[Agueh, Carlier 2010]



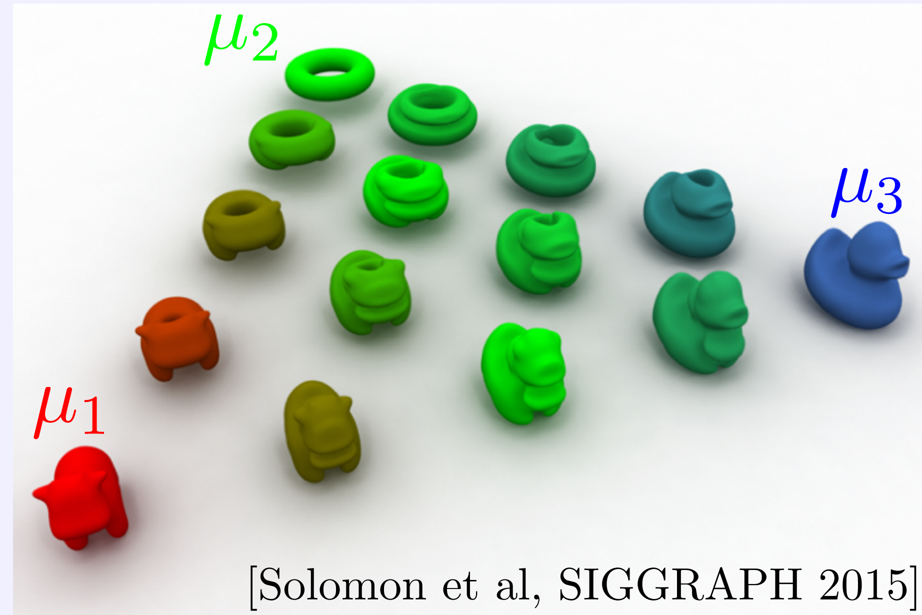


Generalizations

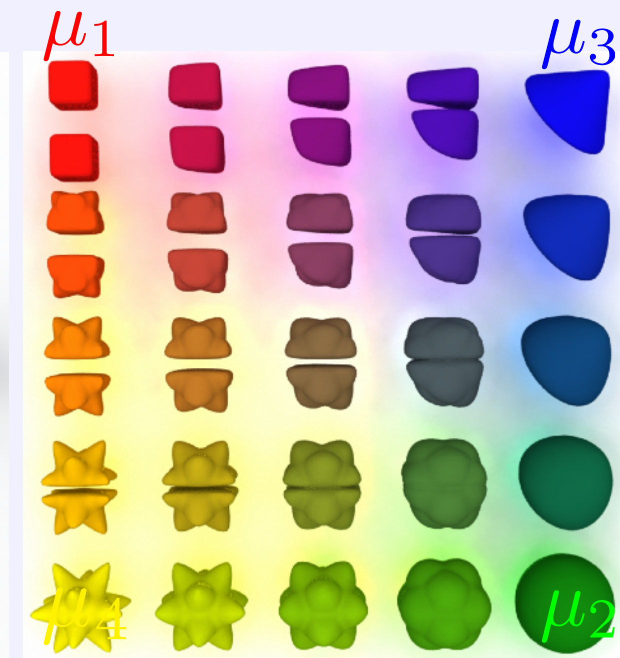
OT barycenters:

$$\min_{\nu} \sum_k \lambda_k W_2^2(\mu_k, \nu)$$

[Agueh, Carlier 2010]



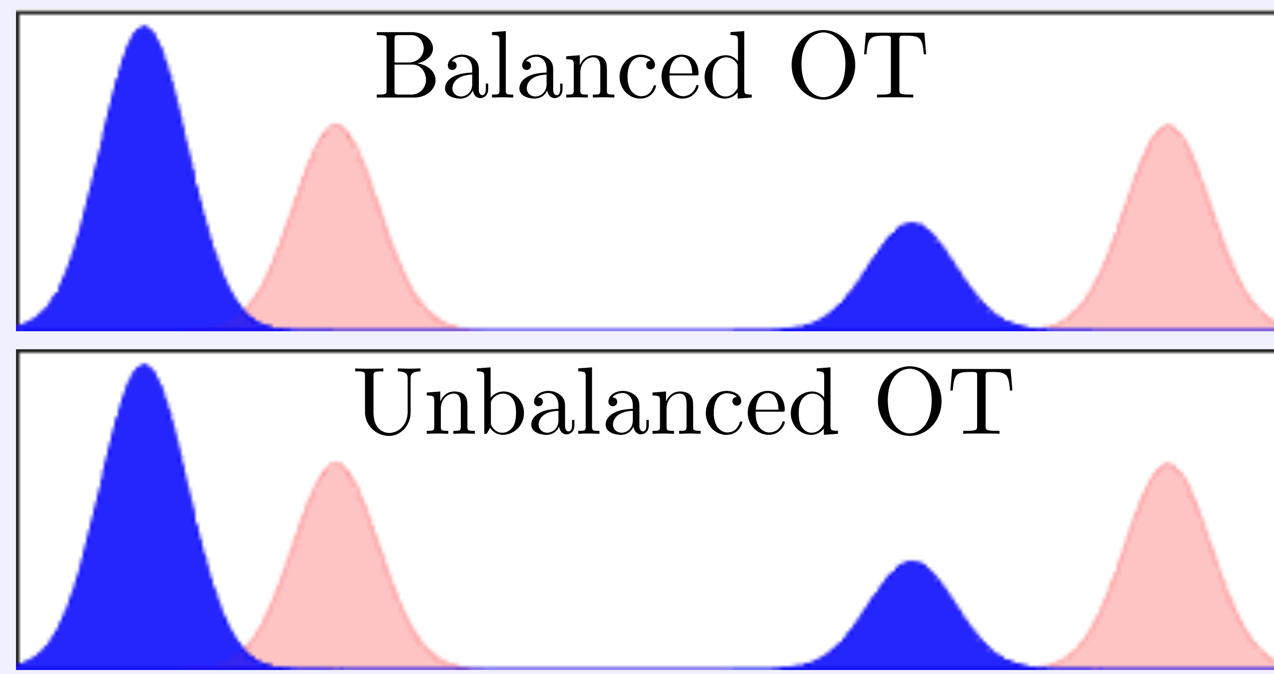
[Solomon et al, SIGGRAPH 2015]



Unbalanced transport:

$$\min_{T} \sum_{i,j} d_{i,j}^p T_{i,j} +$$
$$\rho \mathrm{KL}(T\mathbb{1}_{N_1} | \mu) + \rho \mathrm{KL}(T^\top \mathbb{1}_{N_2} | \nu)$$

[Liereo, Mielke, Savaré 2015]
[Chizat, Schmitzer, Peyré, Vialard 2015]
[Kondratyev, Monsaingeon, Vorotnikov, 2015]

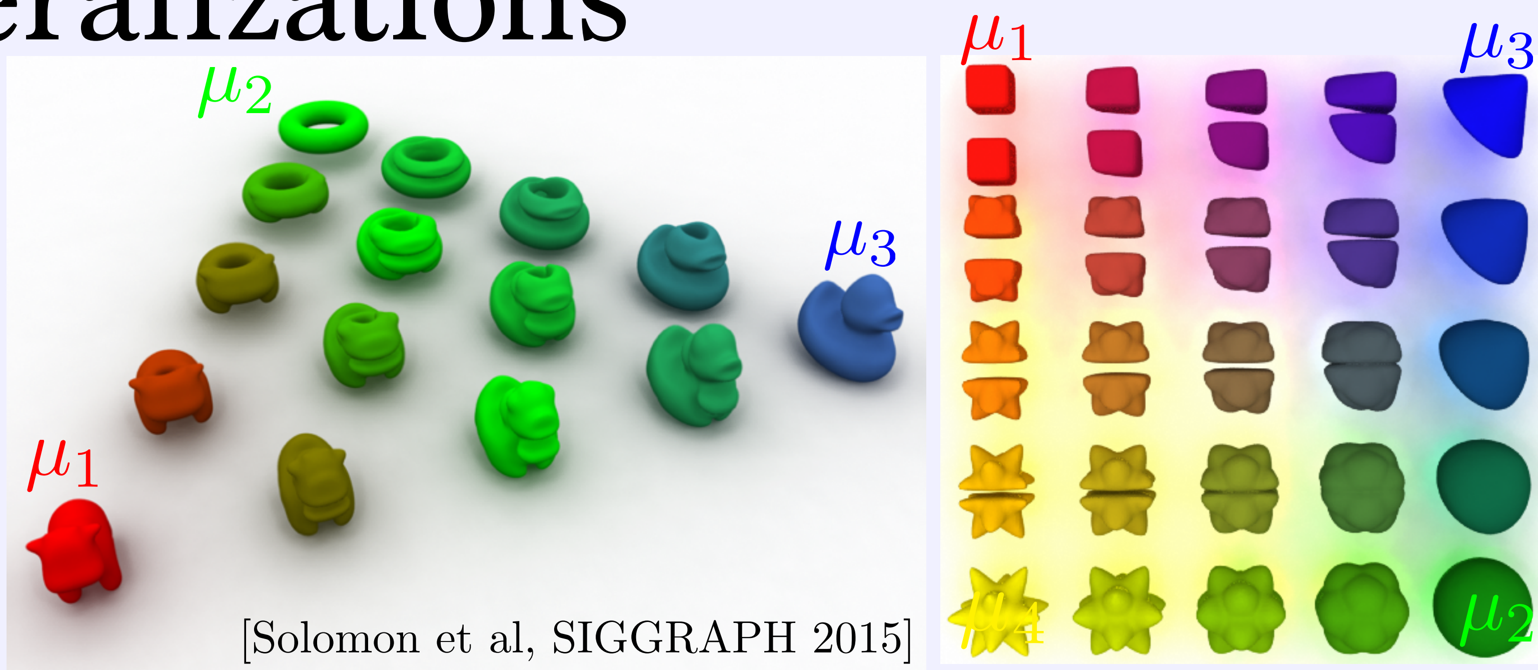


Generalizations

OT barycenters:

$$\min_{\nu} \sum_k \lambda_k W_2^2(\mu_k, \nu)$$

[Agueh, Carlier 2010]

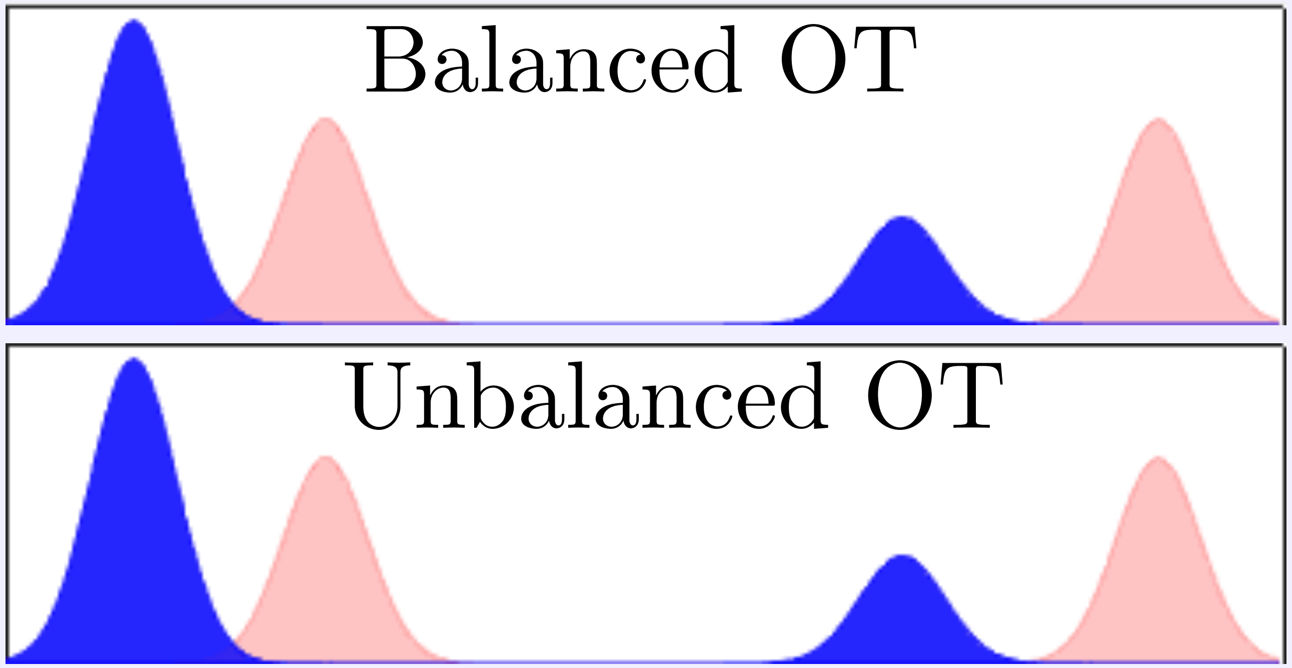


[Solomon et al, SIGGRAPH 2015]

Unbalanced transport:

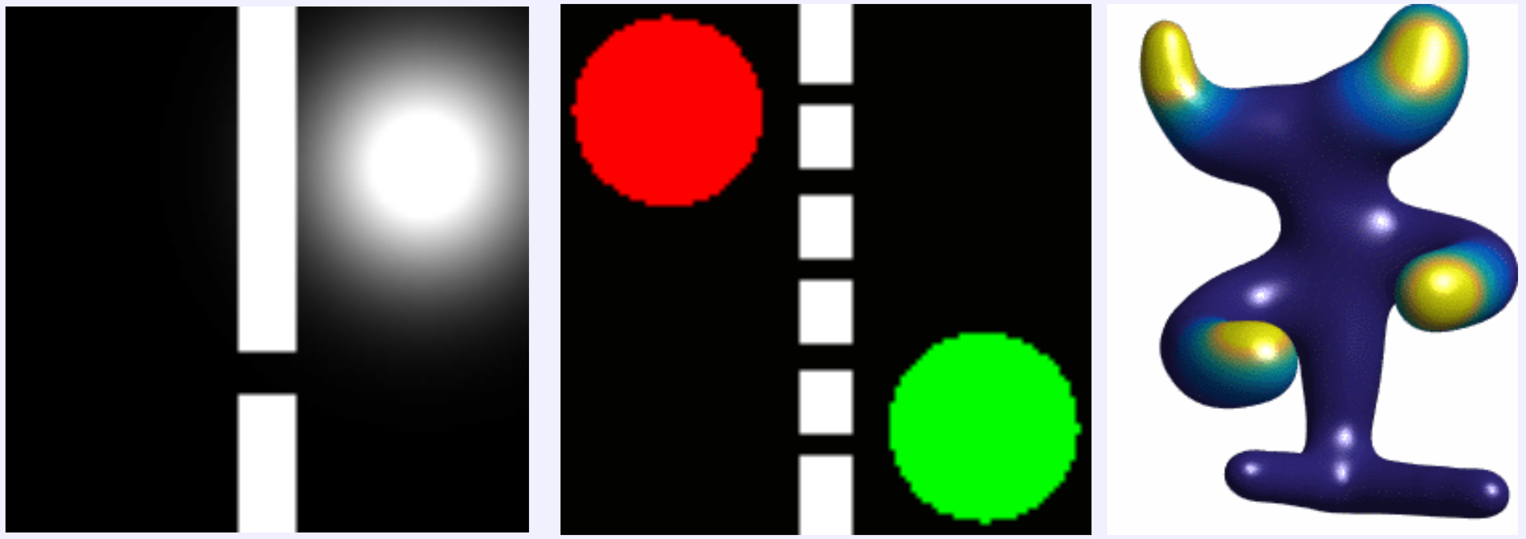
$$\min_{T} \sum_{i,j} d_{i,j}^p T_{i,j} +$$
$$\rho \mathrm{KL}(T \mathbb{1}_{N_1} | \mu) + \rho \mathrm{KL}(T^\top \mathbb{1}_{N_2} | \nu)$$

[Liereo, Mielke, Savaré 2015]
[Chizat, Schmitzer, Peyré, Vialard 2015]
[Kondratyev, Monsaingeon, Vorotnikov, 2015]

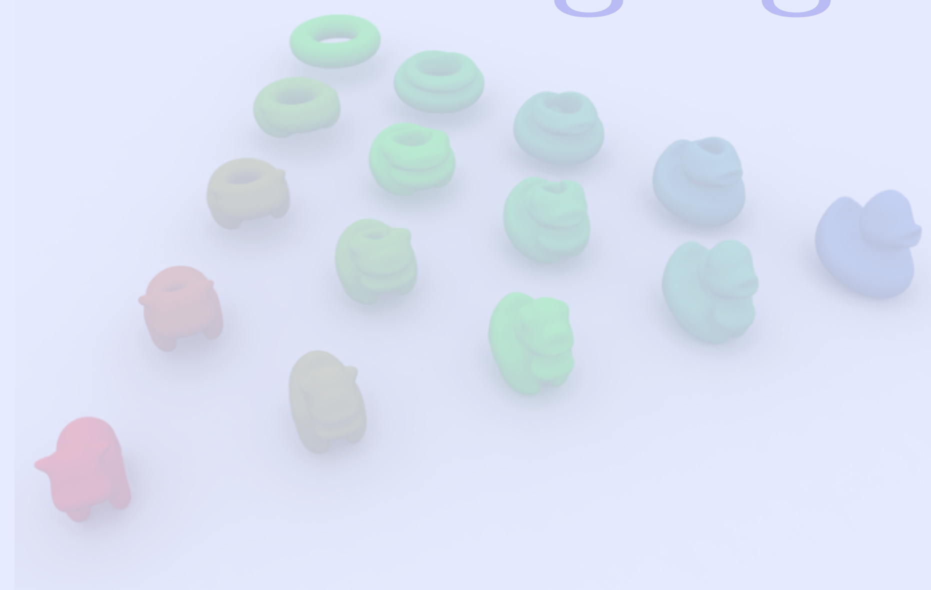


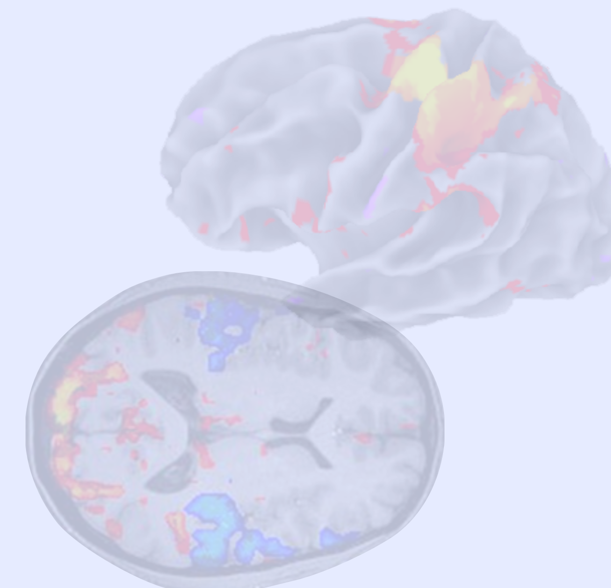
Gradient flows:

$$\mu_{t+1} = \min_{\mu} \frac{1}{2\tau} W(\mu_t, \mu) + f(\mu)$$

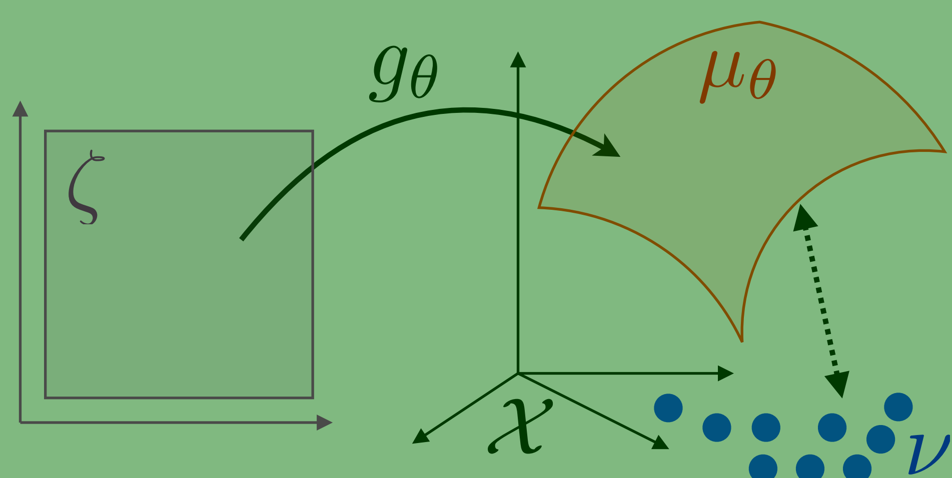


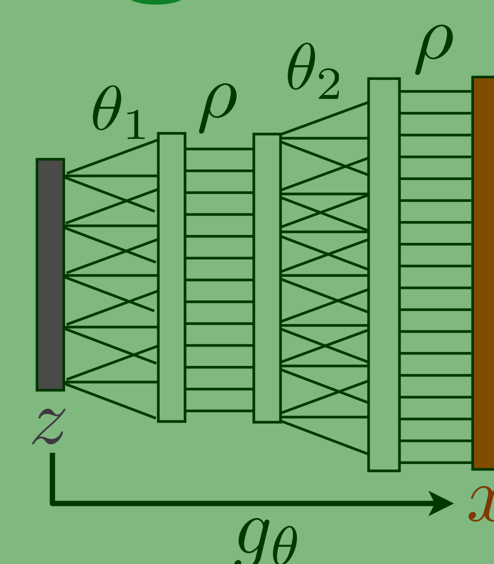
1. OT for Imaging Sciences





2. OT for Machine Learning

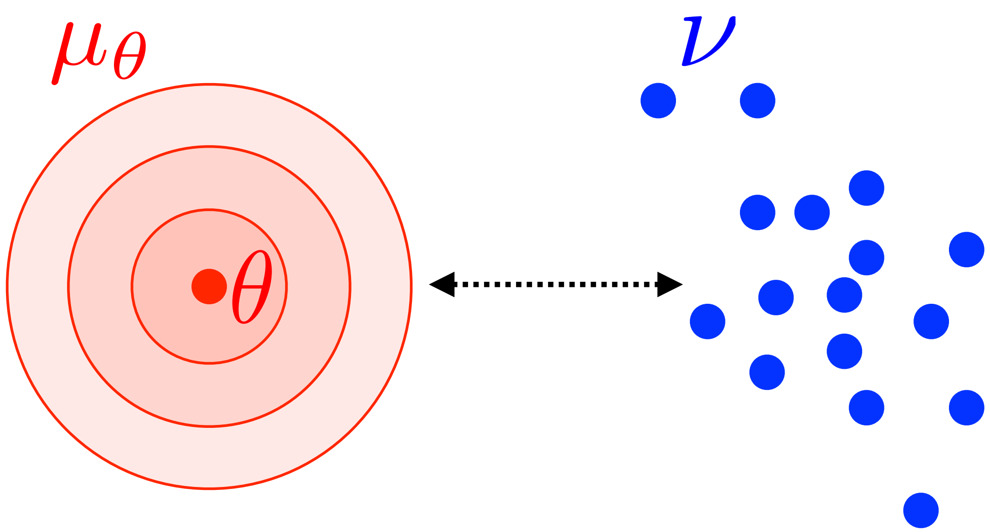




Density Fitting and Generative Models

Observations: $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

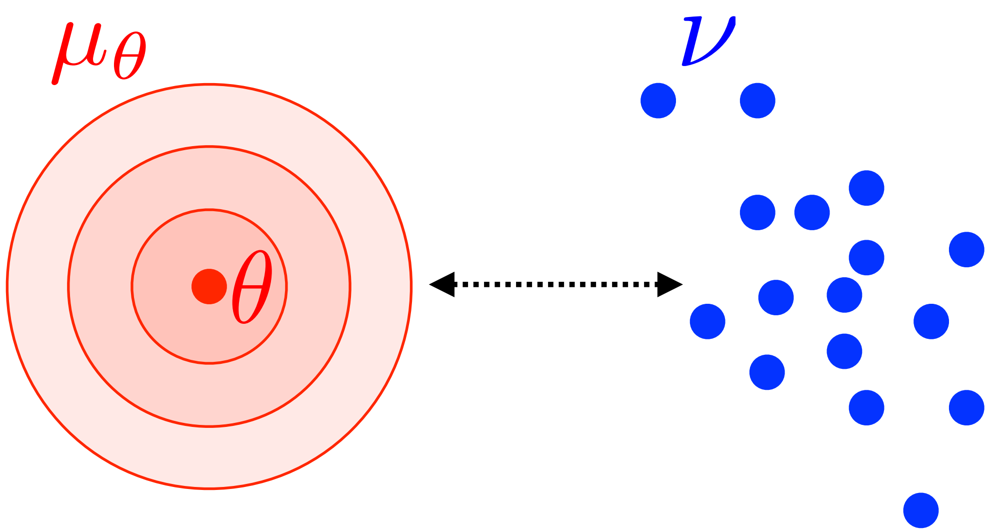
Parametric model: $\theta \mapsto \mu_\theta$



Density Fitting and Generative Models

Observations: $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Parametric model: $\theta \mapsto \mu_\theta$



Density fitting: $\mathrm{d}\mu_\theta(y) = f_\theta(y)\mathrm{d}y$

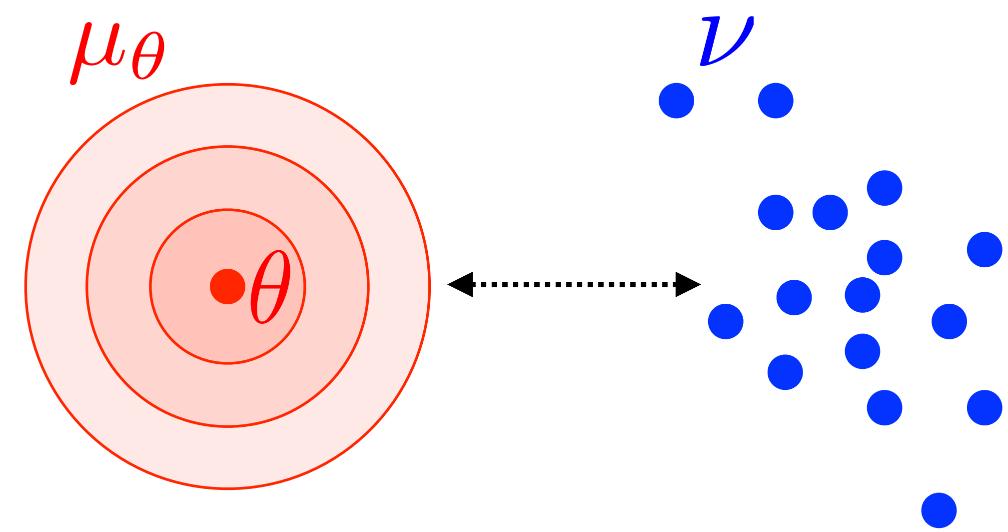
$$\min_\theta \widehat{\mathrm{KL}}(\mu_\theta|\nu) \stackrel{\text{def.}}{=} -\sum_j \log(f_\theta(y_j))$$

Maximum likelihood (MLE)

Density Fitting and Generative Models

Observations: $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Parametric model: $\theta \mapsto \mu_\theta$



Density fitting: $\mathrm{d}\mu_\theta(y) = f_\theta(y)\mathrm{d}y$

$$\min_\theta \widehat{\mathrm{KL}}(\mu_\theta|\nu) \overset{\text{def.}}{=} -\sum_j \log(f_\theta(y_j))$$

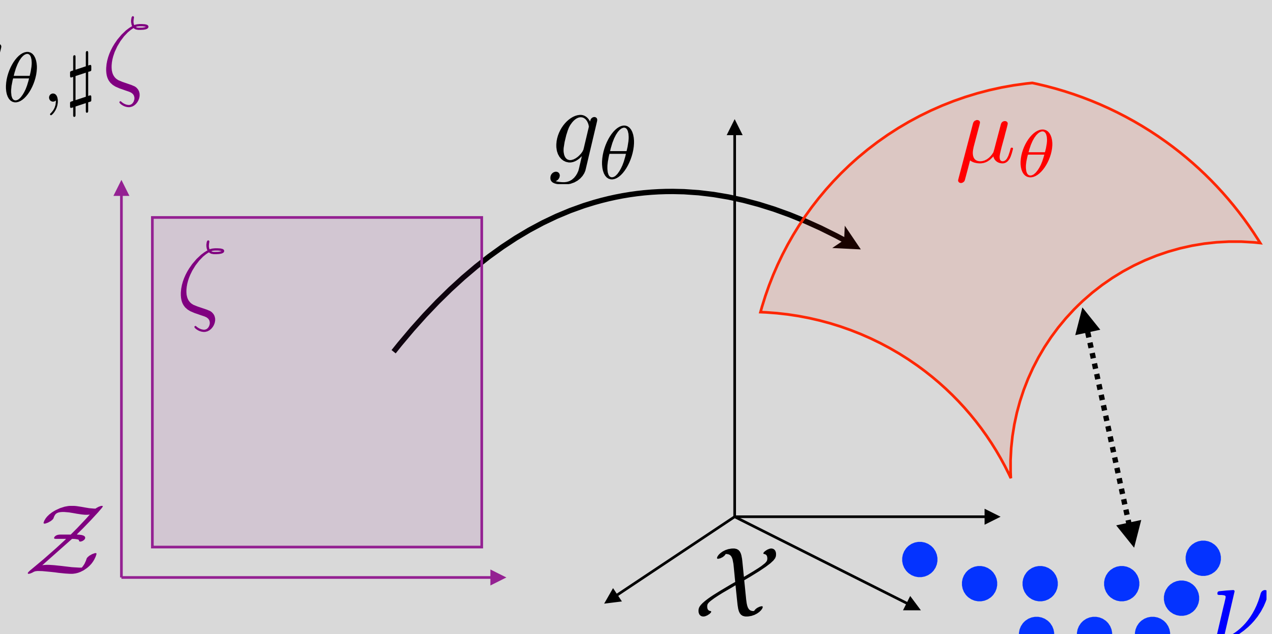
Maximum likelihood (MLE)

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

$$\widehat{\mathrm{KL}}(\mu_\theta|\nu) = +\infty$$

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.



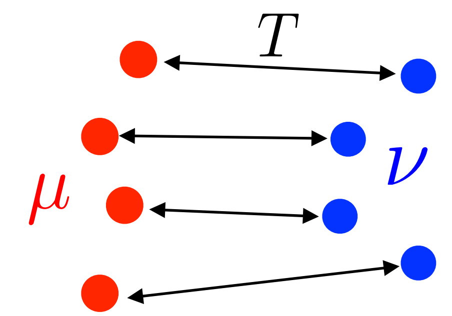
Loss Functions for Measures

Density fitting: $\min_{\theta} D(\mu_\theta, \nu)$

$\nu = \frac{1}{P}\sum_j \delta_{y_j}$ $\quad \mu = \frac{1}{N}\sum_i \delta_{x_i}$

Optimal Transport Distances

$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p$$



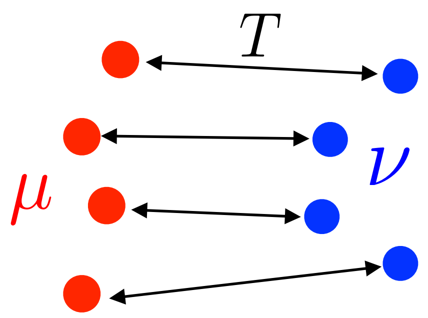
Loss Functions for Measures

Density fitting: $\min_{\theta} D(\mu_\theta, \nu)$

$\nu = \frac{1}{P} \sum_j \delta_{y_j}$ $\mu = \frac{1}{N} \sum_i \delta_{x_i}$

Optimal Transport Distances

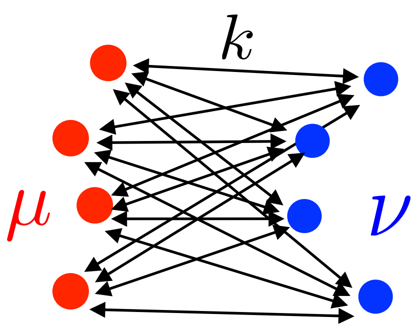
$$W(\mu, \nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

$$\|\mu - \nu\|_k^2 \overset{\text{def.}}{=} \frac{1}{N^2} \sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{P^2} \sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{NP} \sum_{i,j} k(x_i, y_j)$$

Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.

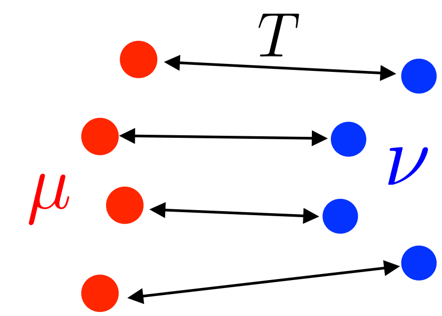


Loss Functions for Measures

Density fitting: $\min_{\theta} D(\mu_\theta, \nu)$ $\qquad \nu = \frac{1}{P}\sum_j \delta_{y_j} \quad \mu = \frac{1}{N}\sum_i \delta_{x_i}$

Optimal Transport Distances

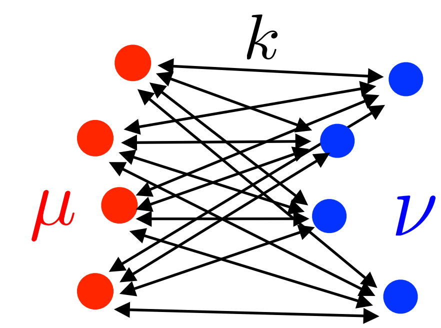
$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

$$\|\mu - \nu\|_k^2 \overset{\text{def.}}{=} \frac{1}{N^2}\sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{P^2}\sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{NP}\sum_{i,j} k(x_i, y_j)$$

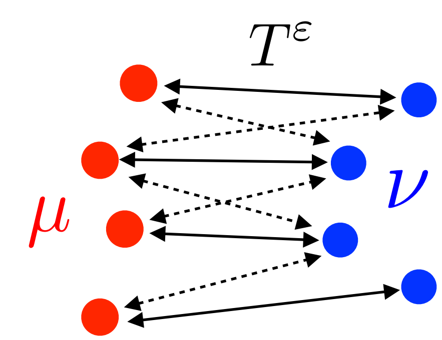
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. $\quad$ Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Cuturi 13]

$$W_\varepsilon(\mu,\nu)^p \overset{\text{def.}}{=} \sum_{i,j} T^\varepsilon_{i,j} \|x_i - y_j\|^p$$

$$\bar{W}_\varepsilon(\mu,\nu)^p \overset{\text{def.}}{=} W_\varepsilon(\mu,\nu)^p - \frac{1}{2} W_\varepsilon(\mu,\mu)^p - \frac{1}{2} W_\varepsilon(\nu,\nu)^p$$

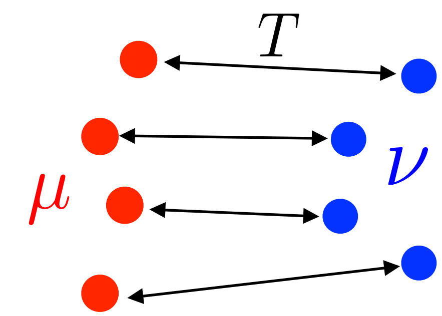


Loss Functions for Measures

Density fitting: $\min_\theta D(\mu_\theta, \nu)$ $\nu = \frac{1}{P}\sum_j \delta_{y_j}$ $\mu = \frac{1}{N}\sum_i \delta_{x_i}$

Optimal Transport Distances

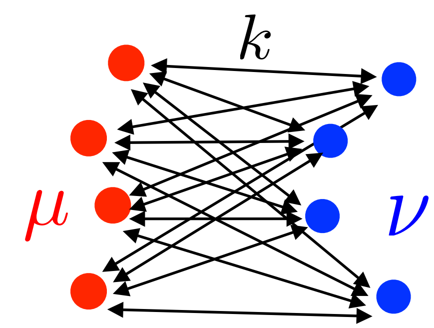
$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

$$\|\mu - \nu\|_k^2 \overset{\text{def.}}{=} \frac{1}{N^2}\sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{P^2}\sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{NP}\sum_{i,j} k(x_i, y_j)$$

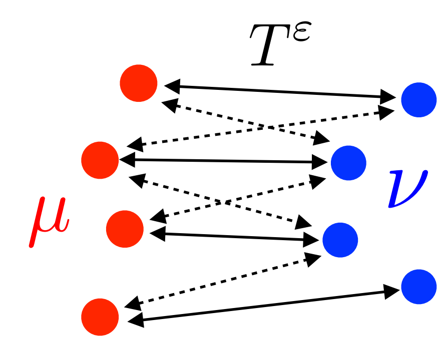
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Cuturi 13]

$$W_\varepsilon(\mu,\nu)^p \overset{\text{def.}}{=} \sum_{i,j} T^\varepsilon_{i,j} \|x_i - y_j\|^p$$

$$\bar{W}_\varepsilon(\mu,\nu)^p \overset{\text{def.}}{=} W_\varepsilon(\mu,\nu)^p - \tfrac{1}{2} W_\varepsilon(\mu,\mu)^p - \tfrac{1}{2} W_\varepsilon(\nu,\nu)^p$$



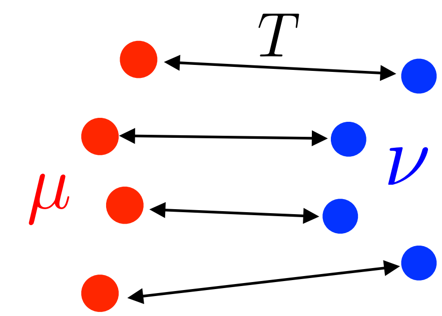
Theorem: [Ramdas, G.Trillos, Cuturi 17] $\bar{W}_\varepsilon(\mu,\nu)^p \xrightarrow{\varepsilon \to 0} W(\mu,\nu)^p$, $\bar{W}_\varepsilon(\mu,\nu)^p \xrightarrow{\varepsilon \to +\infty} \|\mu-\nu\|_k^2$ for $k(x,y) = -\|x-y\|^p$

Loss Functions for Measures

Density fitting: $\min_{\theta} D(\mu_\theta, \nu)$ $\quad \nu = \frac{1}{P}\sum_j \delta_{y_j} \quad \mu = \frac{1}{N}\sum_i \delta_{x_i}$

Optimal Transport Distances

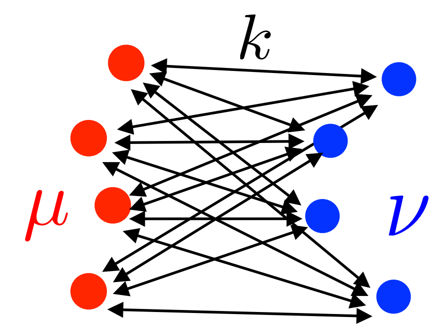
$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

$$\|\mu - \nu\|_k^2 \overset{\text{def.}}{=} \frac{1}{N^2}\sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{P^2}\sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{NP}\sum_{i,j} k(x_i, y_j)$$

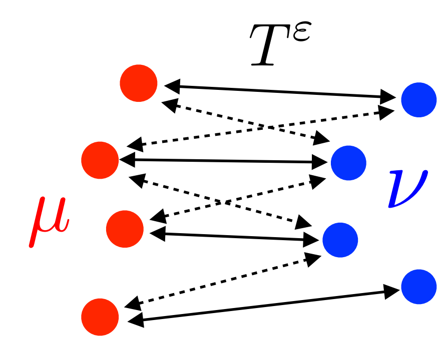
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Cuturi 13]

$$W_\varepsilon(\mu,\nu)^p \overset{\text{def.}}{=} \sum_{i,j} T^\varepsilon_{i,j} \|x_i - y_j\|^p$$

$$\bar{W}_\varepsilon(\mu,\nu)^p \overset{\text{def.}}{=} W_\varepsilon(\mu,\nu)^p - \frac{1}{2} W_\varepsilon(\mu,\mu)^p - \frac{1}{2} W_\varepsilon(\nu,\nu)^p$$



Theorem: [Ramdas, G.Trillos, Cuturi 17] $\bar{W}_\varepsilon(\mu,\nu)^p \xrightarrow{\varepsilon \to 0} W(\mu,\nu)^p$, $\quad \bar{W}_\varepsilon(\mu,\nu)^p \xrightarrow{\varepsilon \to +\infty} \|\mu - \nu\|_k^2$ $\quad$ for $k(x,y) = -\|x-y\|^p$

Best of both worlds:

$\rightarrow$ cross-validate ε

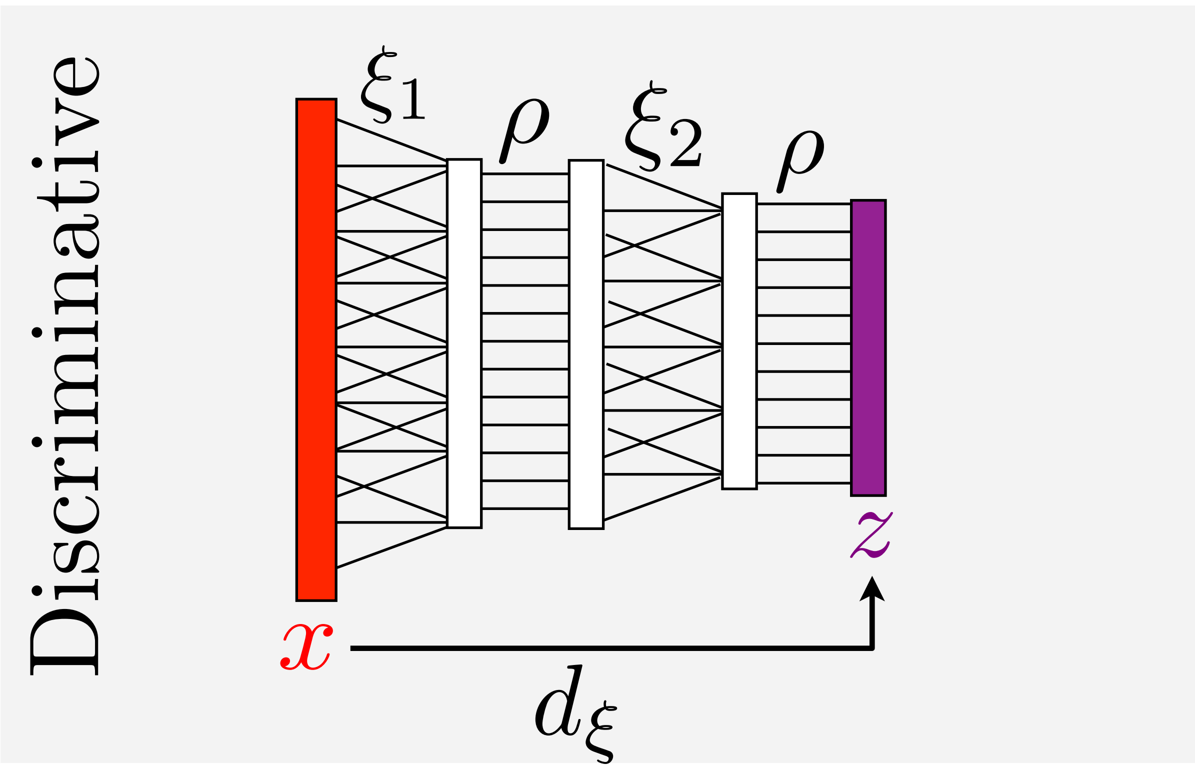
- Scale free (no σ, no heavy tail kernel).
- Non-Euclidean, arbitrary ground distance.
- Less biased gradient.
- No curse of dimension (low sample complexity).

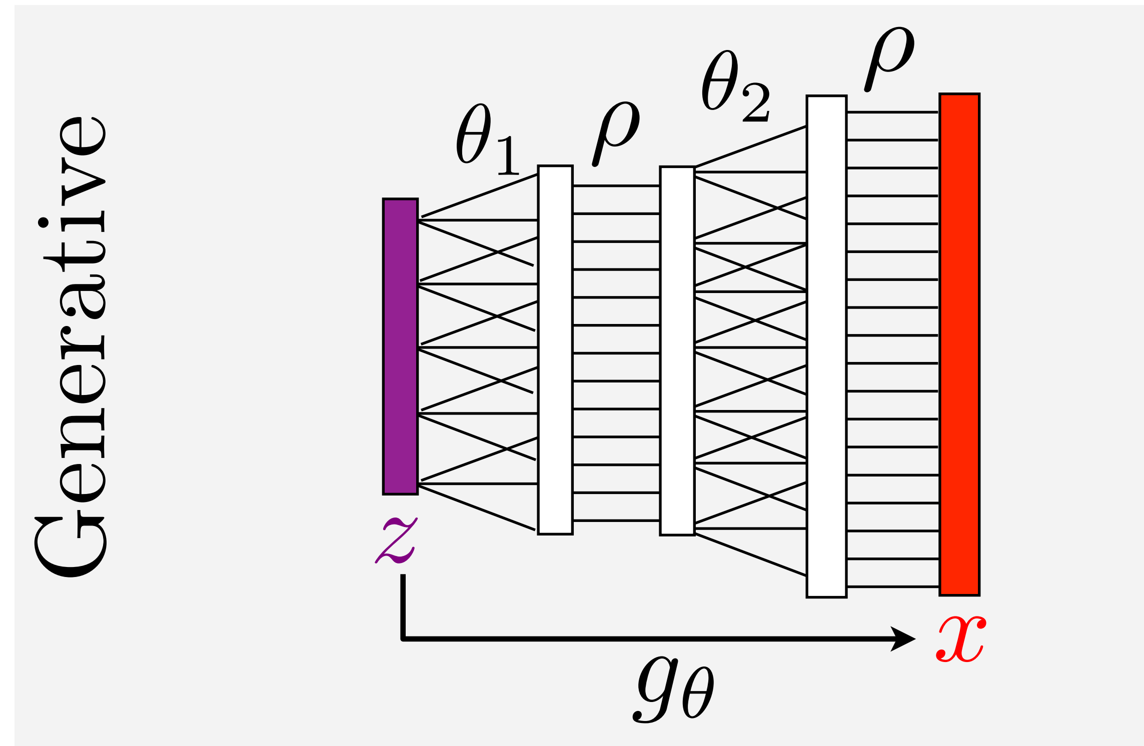
Deep Discriminative vs Generative Models

Deep networks:

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

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



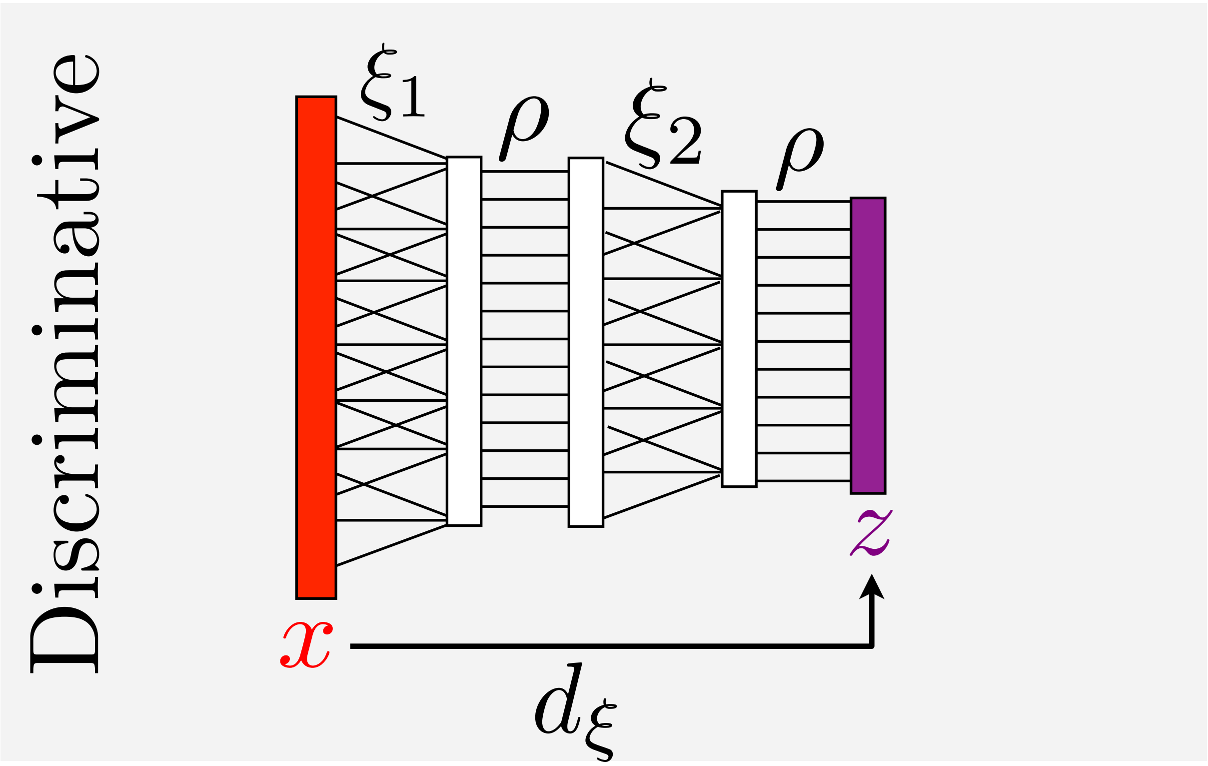


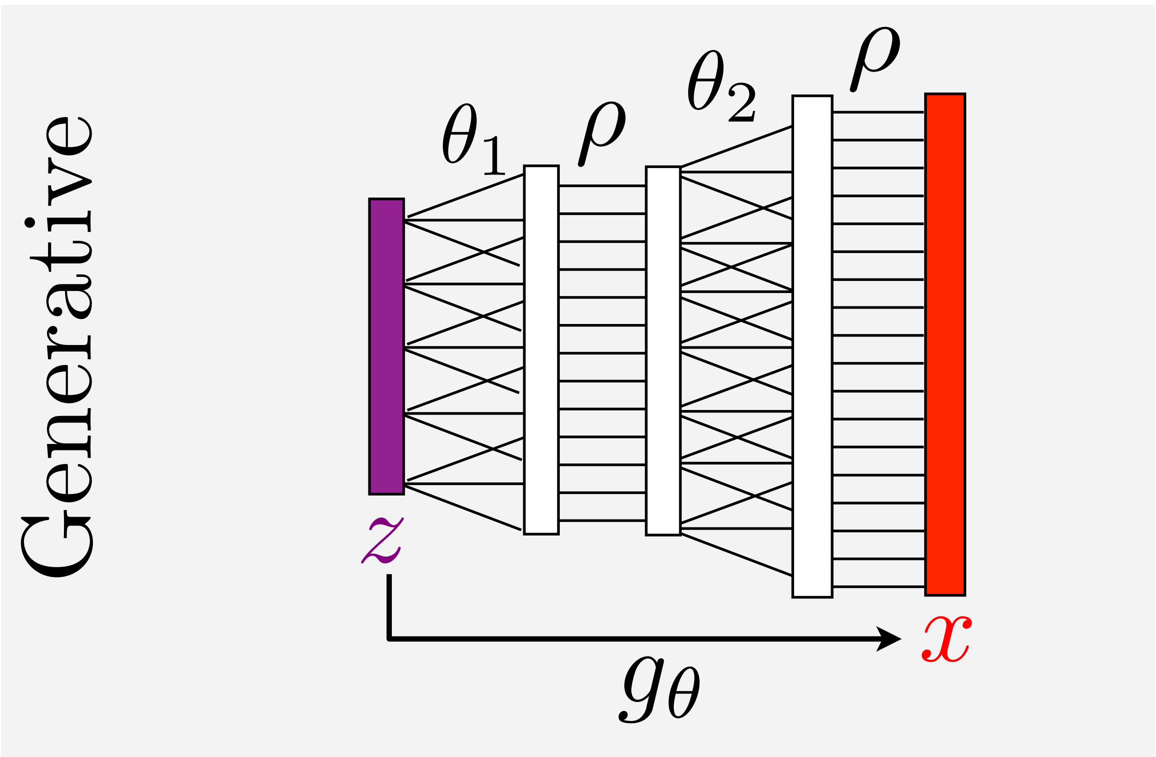
Deep Discriminative vs Generative Models

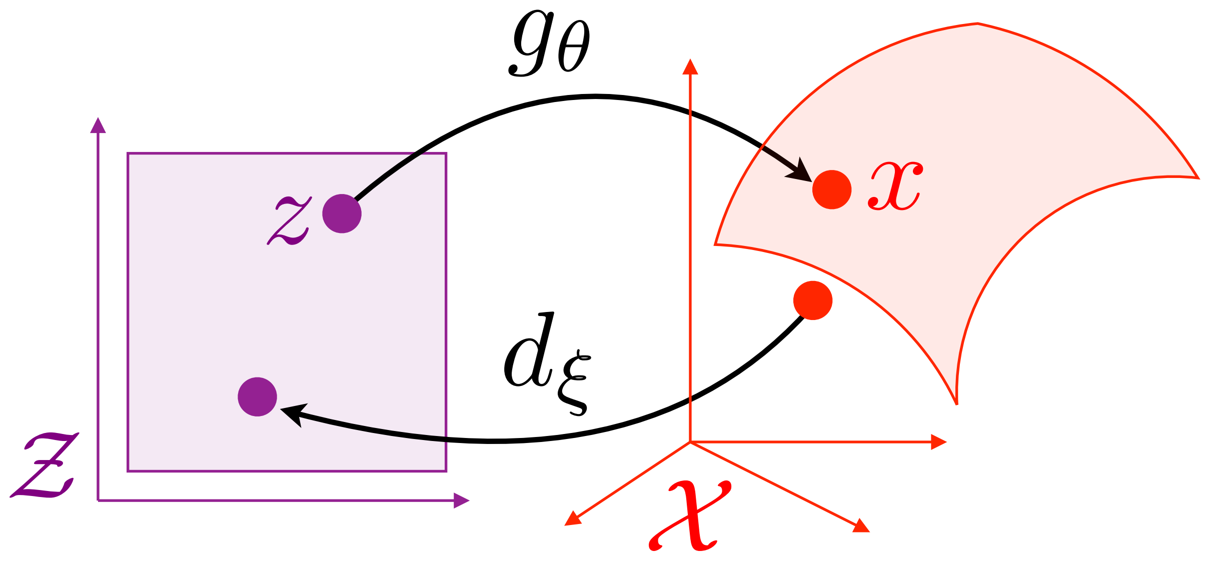
Deep networks:

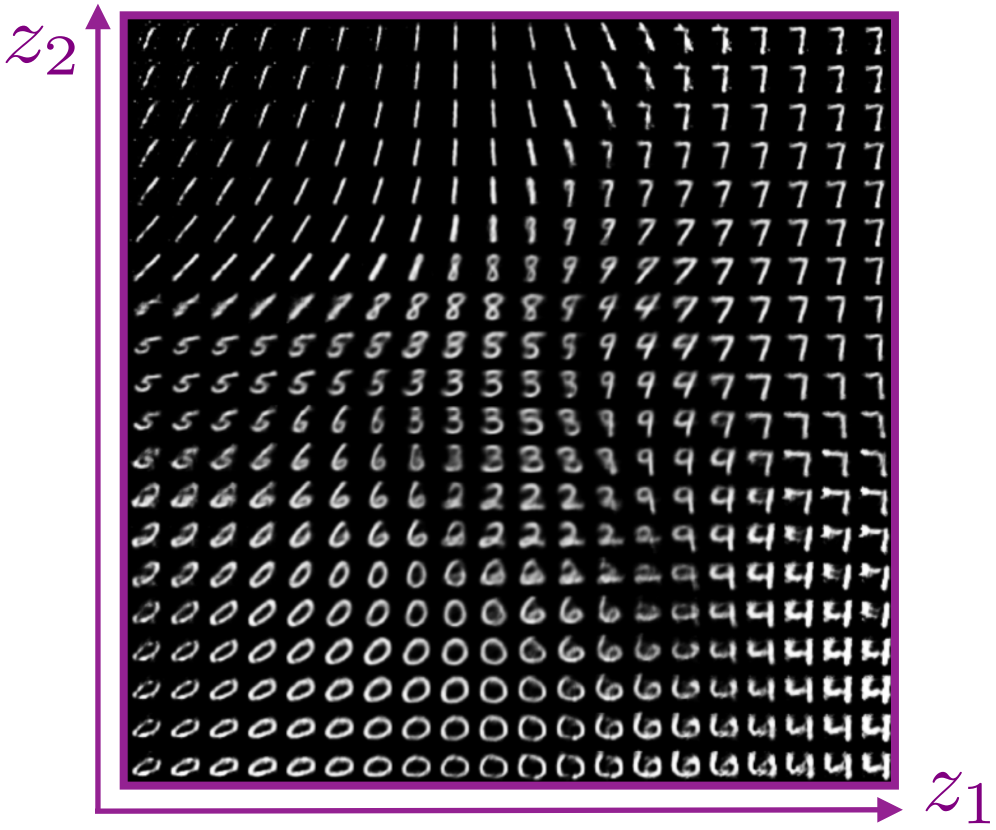
$$d_\xi(x) = \rho(\xi_K(\ldots\rho(\xi_2(\rho(\xi_1(x))\ldots)$$

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

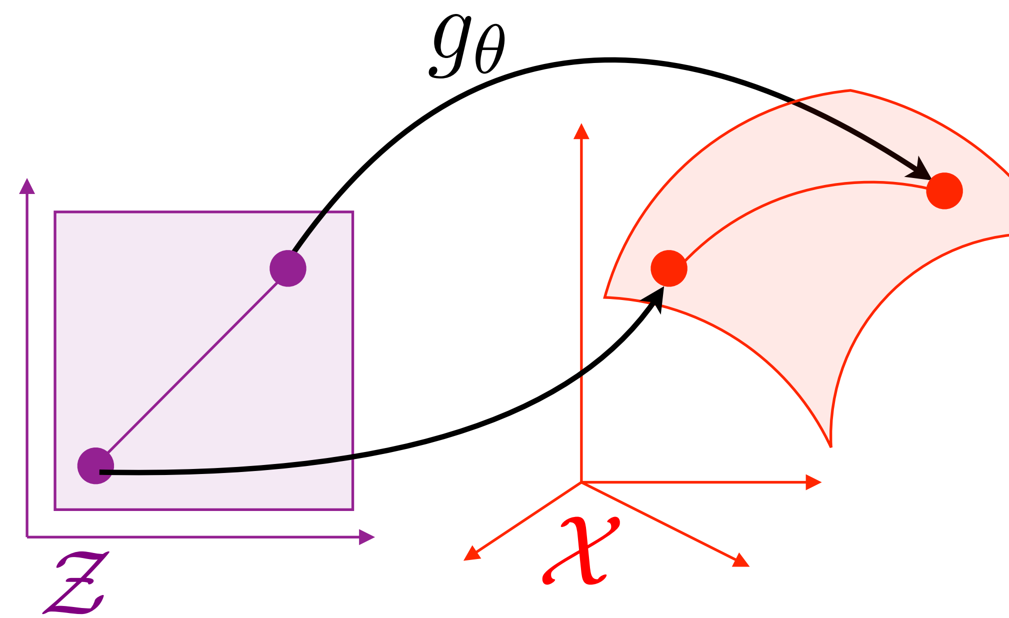








Examples of Image Generation



[Credit ArXiv:1511.06434]

Conclusion: Toward High-dimensional OT

Monge Kantorovich Dantzig Brenier Otto McCann Villani