Computational Optimal Transport

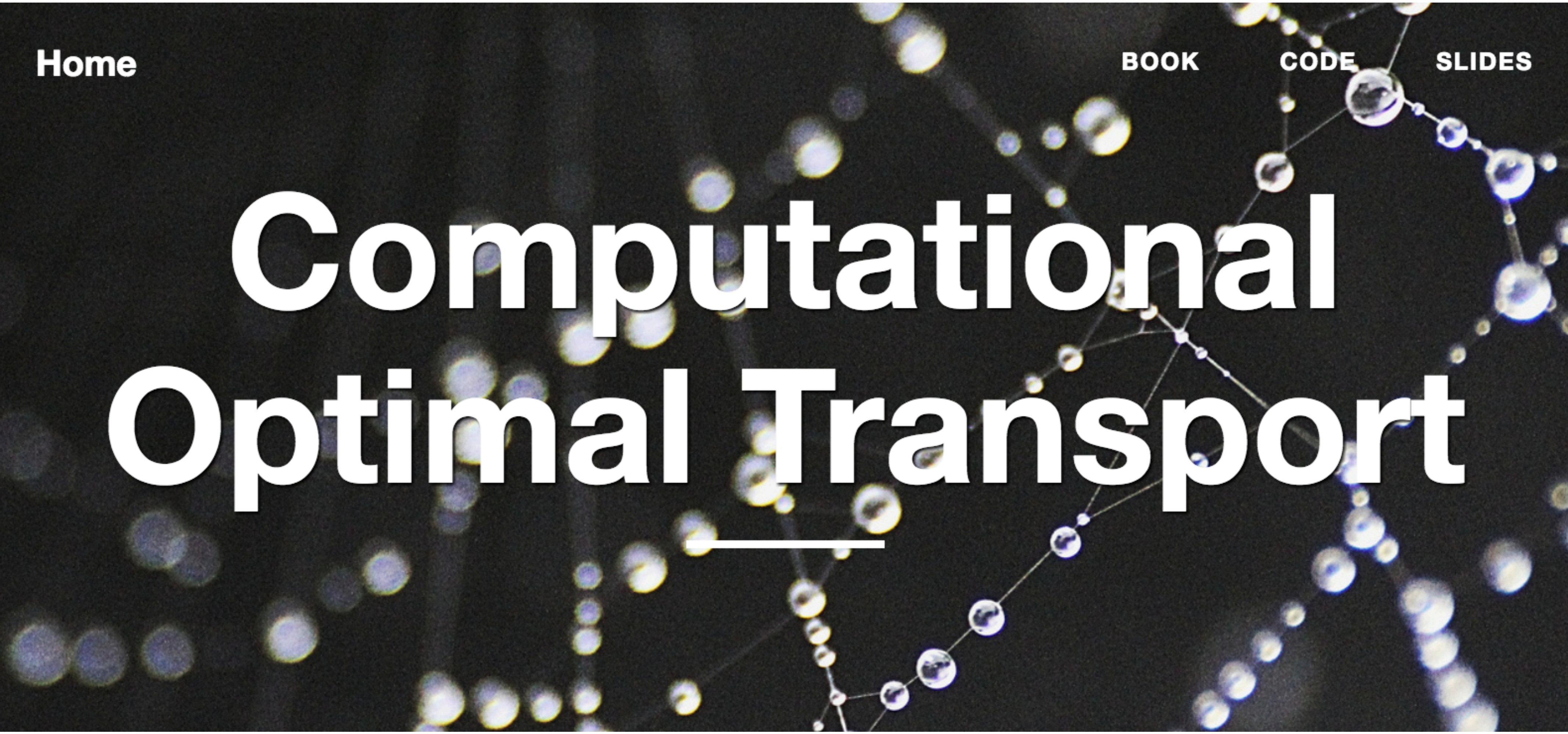
http://optimaltransport.github.io

Introduction

Gabriel Peyré

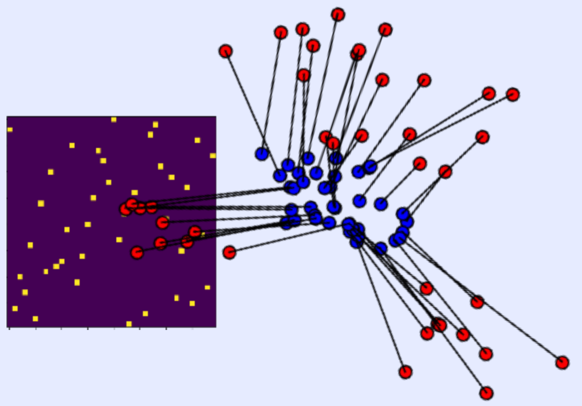
www.numerical-tours.com

https://optimaltransport.github.io

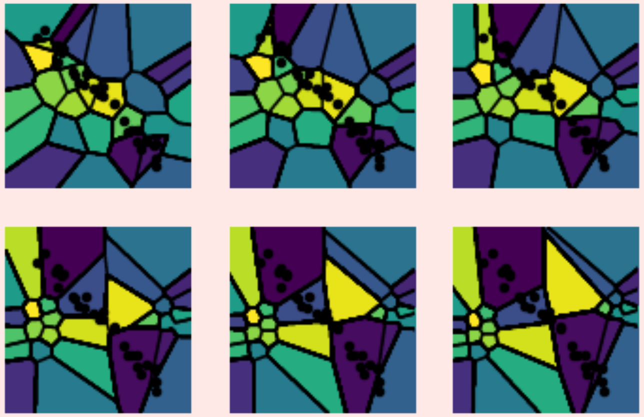


www.numerical-tours.com

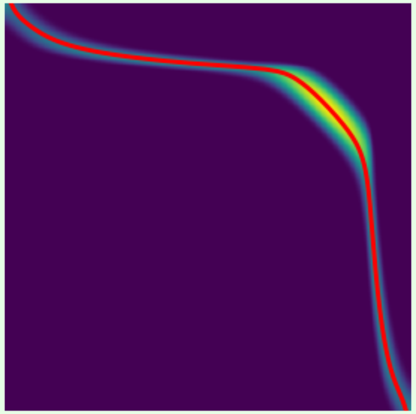
Optimal Transport with Linear Programming



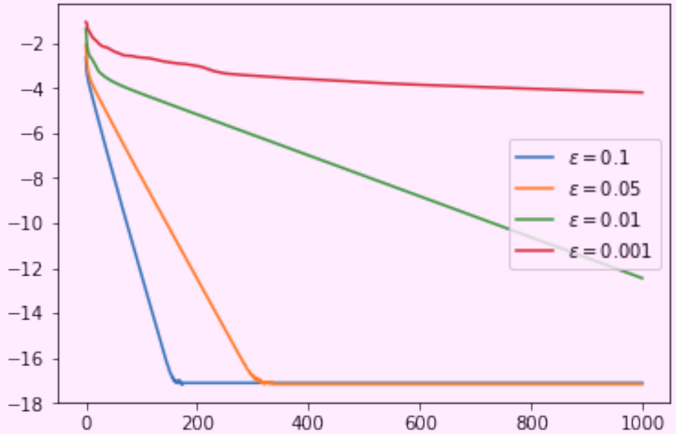
Semi-discrete Optimal Transport



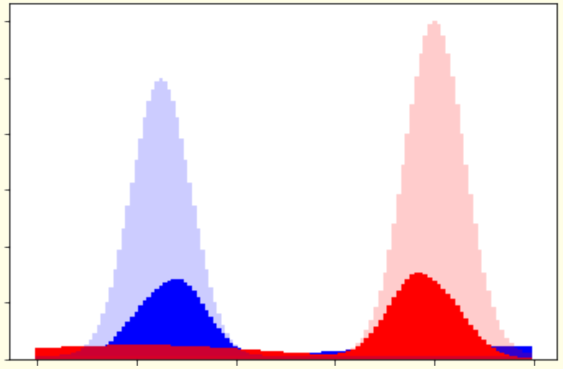
Entropic Regularization of Optimal Transport



Advanced Topics on Sinkhorn Algorithm



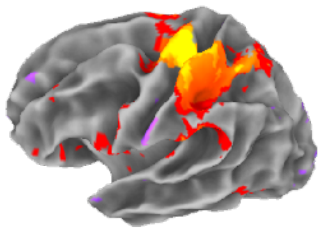
Unbalanced Optimal Transport

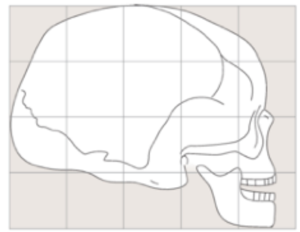


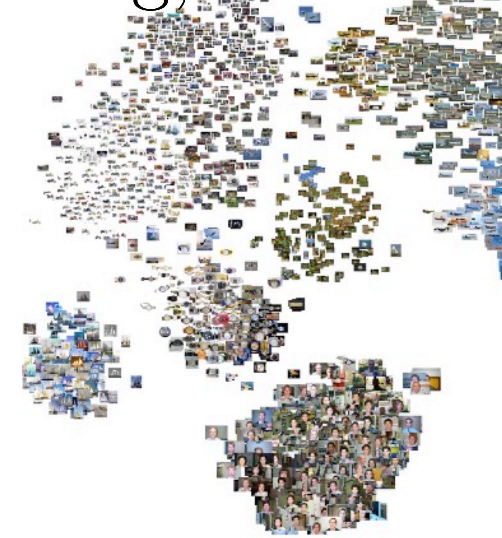
Probability Distributions in Data Sciences

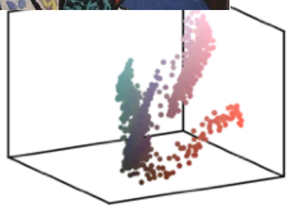
Probability distributions and histograms

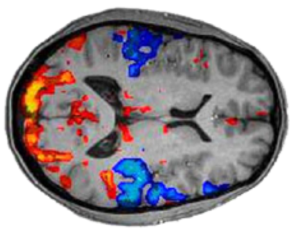
$\rightarrow$ images, vision, graphics and machine learning, . . .

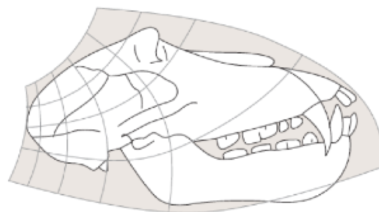








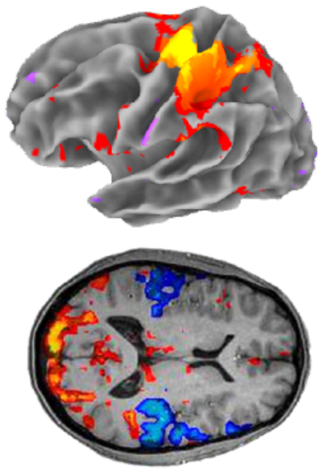


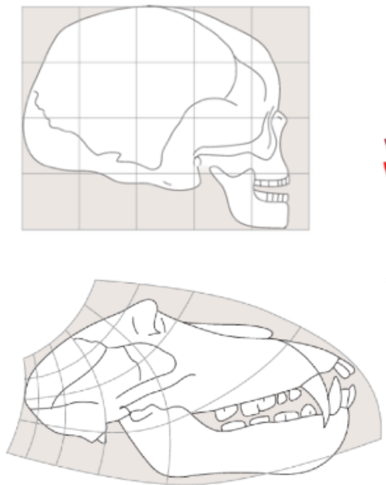


Probability Distributions in Data Sciences

Probability distributions and histograms

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

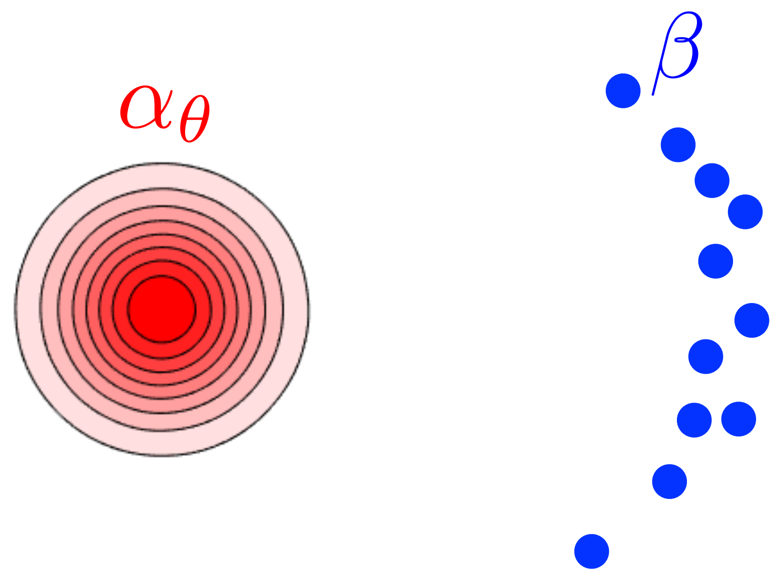




Unsupervised learning

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

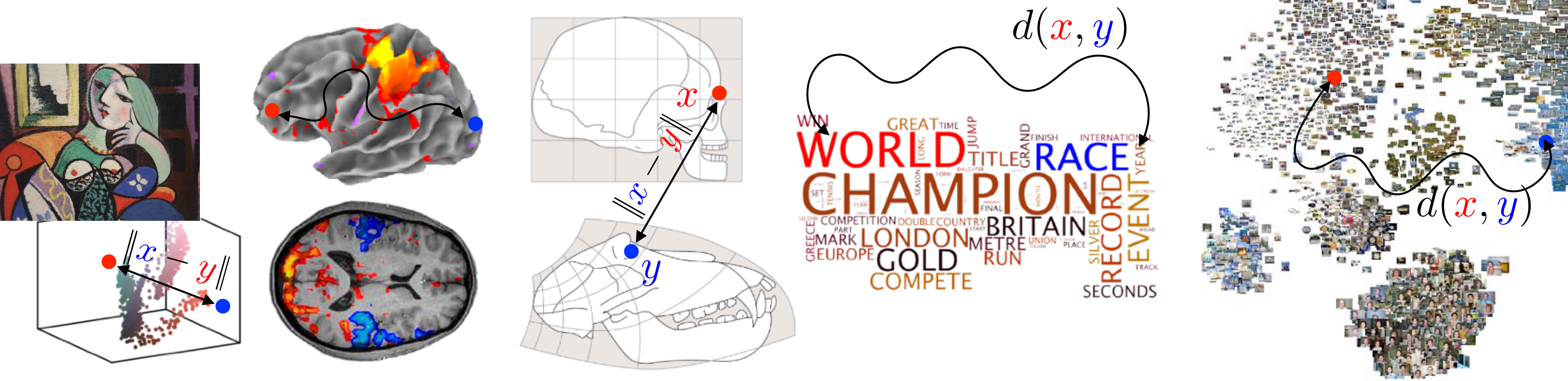
Parametric model: $\theta \mapsto \alpha_\theta$



Probability Distributions in Data Sciences

Probability distributions and histograms

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



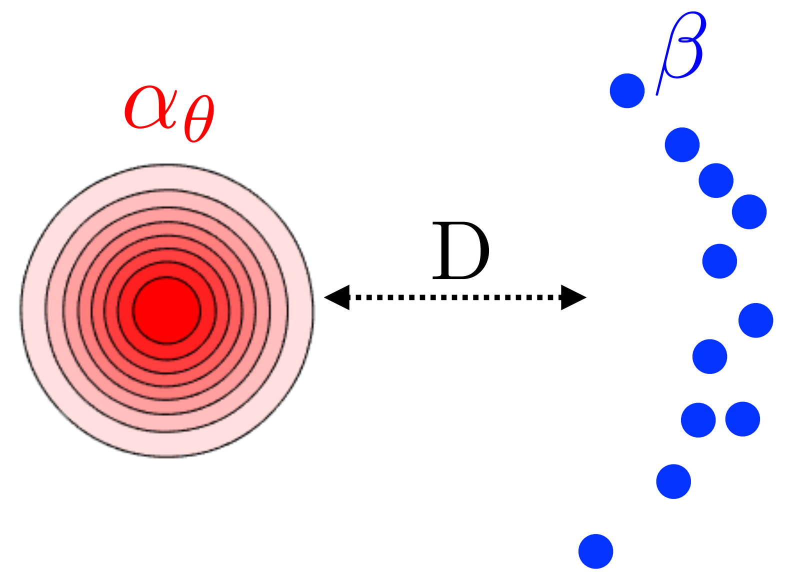
Unsupervised learning

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.



Overview

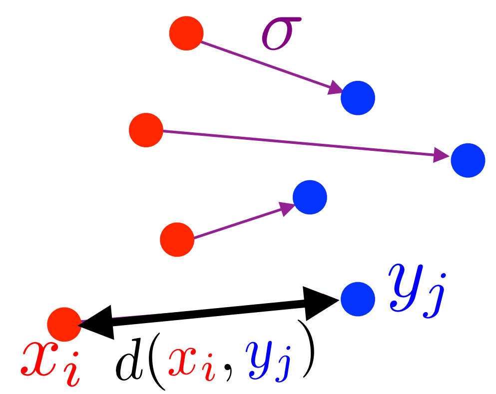
- **Monge Formulation**
- Continuous Optimal Transport
- Kantorovitch Formulation
- Applications

Monge's Problem

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

Permutation:

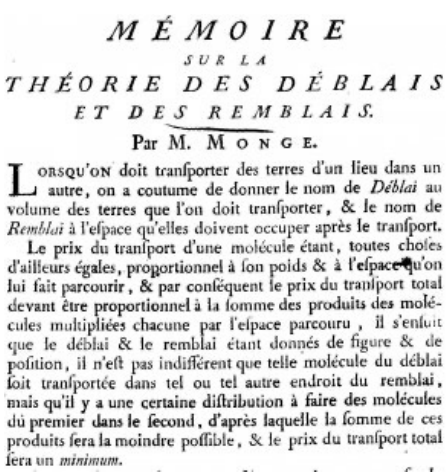
$$\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$$



Monge optimal matching:

$$\mathrm{D}(X, Y) = \min_{\sigma} \sum_{i=1}^{n} d(x_i, y_{\sigma(i)})$$

[Monge 1784]



MÉMOIRE

SUR LA

THÉORIE DES DÉBLAIS ET DES REMBLAIS.

Par M. MONGE.

Lorsqu'on doit tranſporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit tranſporter, & le nom de *Remblai* à l'eſpace qu'elles doivent occuper après le tranſport.

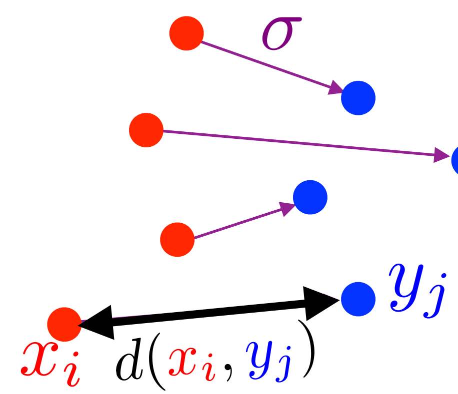
Le prix du tranſport d'une molécule étant, toutes choſes d'ailleurs égales, proportionnel à ſon poids & à l'eſpace qu'on lui fait parcourir, & par conſéquent le prix du tranſport total devant être proportionnel à la ſomme des produits des molécules multipliées chacune par l'eſpace parcouru, il s'enſuit que le déblai & le remblai étant donnés de figure & de poſition, il n'eſt pas indifférent que telle molécule du déblai ſoit tranſportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine diſtribution à faire des molécules du premier dans le ſecond, d'après laquelle la ſomme de ces produits ſera la moindre poſſible, & le prix du tranſport total ſera un *minimum*.

Monge's Problem

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

Permutation:

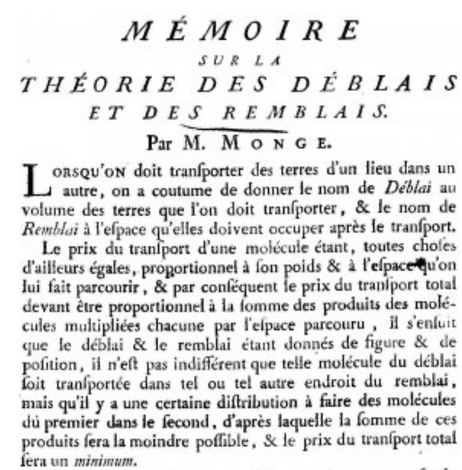
$$\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$$



Monge optimal matching:

$$\mathrm{D}(X, Y) = \min_{\sigma} \sum_{i=1}^{n} d(x_i, y_{\sigma(i)})$$

[Monge 1784]



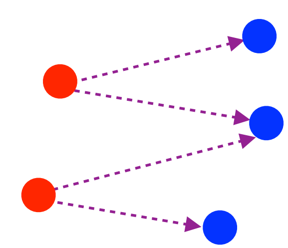
MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. MONGE.

Lorsqu'on doit tranfporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit tranfporter, & le nom de *Remblai* à l'efpace qu'elles doivent occuper après le tranfport.

Le prix du tranfport d'une molécule étant, toutes chofes d'ailleurs égales, proportionnel à fon poids & à l'efpace qu'on lui fait parcourir, & par conféquent le prix du tranfport total devant être proportionnel à la fomme des produits des molécules multipliées chacune par l'efpace parcouru, il s'enfuit que le déblai & le remblai étant donnés de figure & de pofition, il n'eft pas indifférent que telle molécule du déblai foit tranfportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine diftribution à faire des molécules du premier dans le fecond, d'après laquelle la fomme de ces produits fera la moindre poffible, & le prix du tranfport total fera un *minimum*.

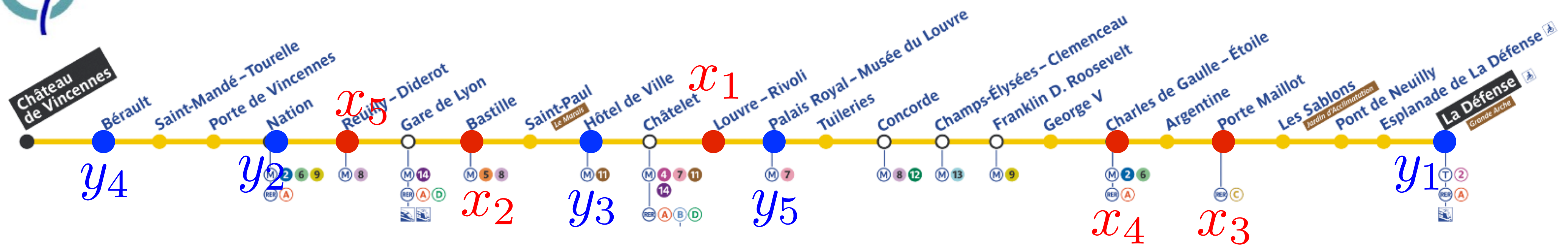
→ Seems intractable: $n!$ possibilities.

→ Different number of points?



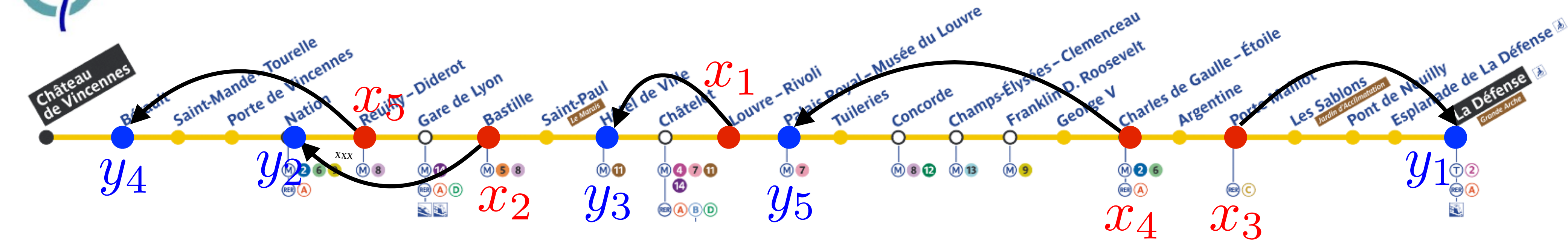
1-D Optimal Transport

$$\min_{\sigma \in \Sigma_n} \sum_{i=1}^{n} |x_i - y_{\sigma(i)}|^p, \quad p \geqslant 1$$



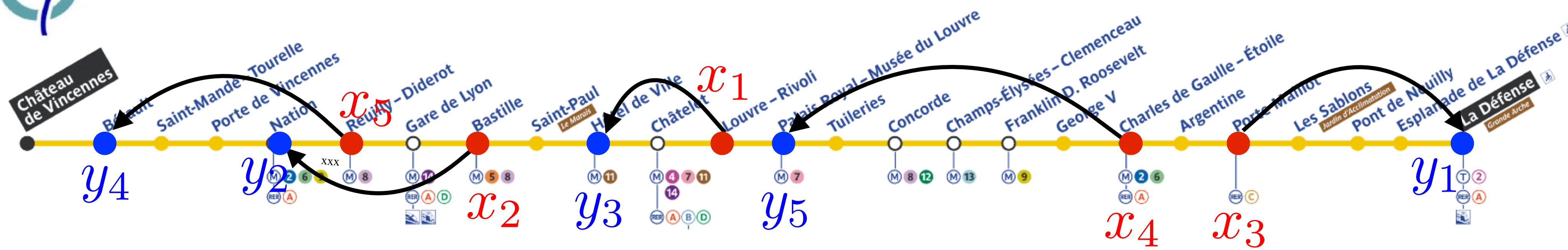
1-D Optimal Transport

$$\min_{\sigma \in \Sigma_n} \sum_{i=1}^{n} |x_i - y_{\sigma(i)}|^p, \quad p \geqslant 1$$



1-D Optimal Transport

$$\min_{\sigma \in \Sigma_n} \sum_{i=1}^{n} |x_i - y_{\sigma(i)}|^p, \quad p \geqslant 1$$

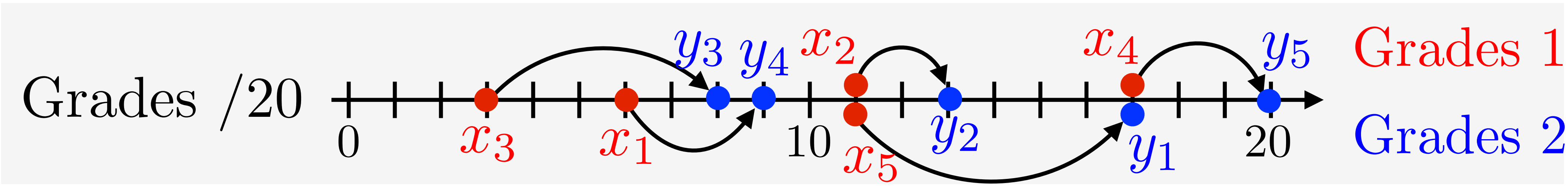


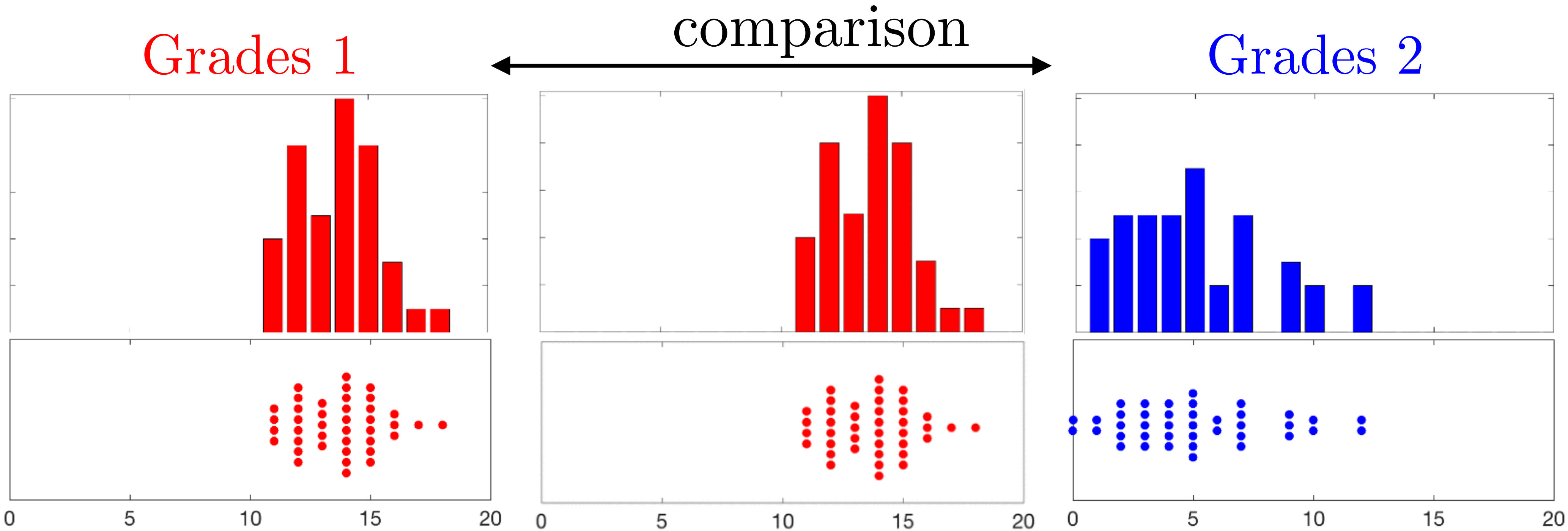
Sorting algorithms: insertion $n(n-1)/2$ worst case.

n	n !	n(n-1)/2	n log(n)
10	3628800	45	23
11	39916800	55	26
12	479001600	66	30
25	$1{,}551\text{x}10^{25}$	300	80
70	$1{,}198\text{x}10^{100}$	21415	297

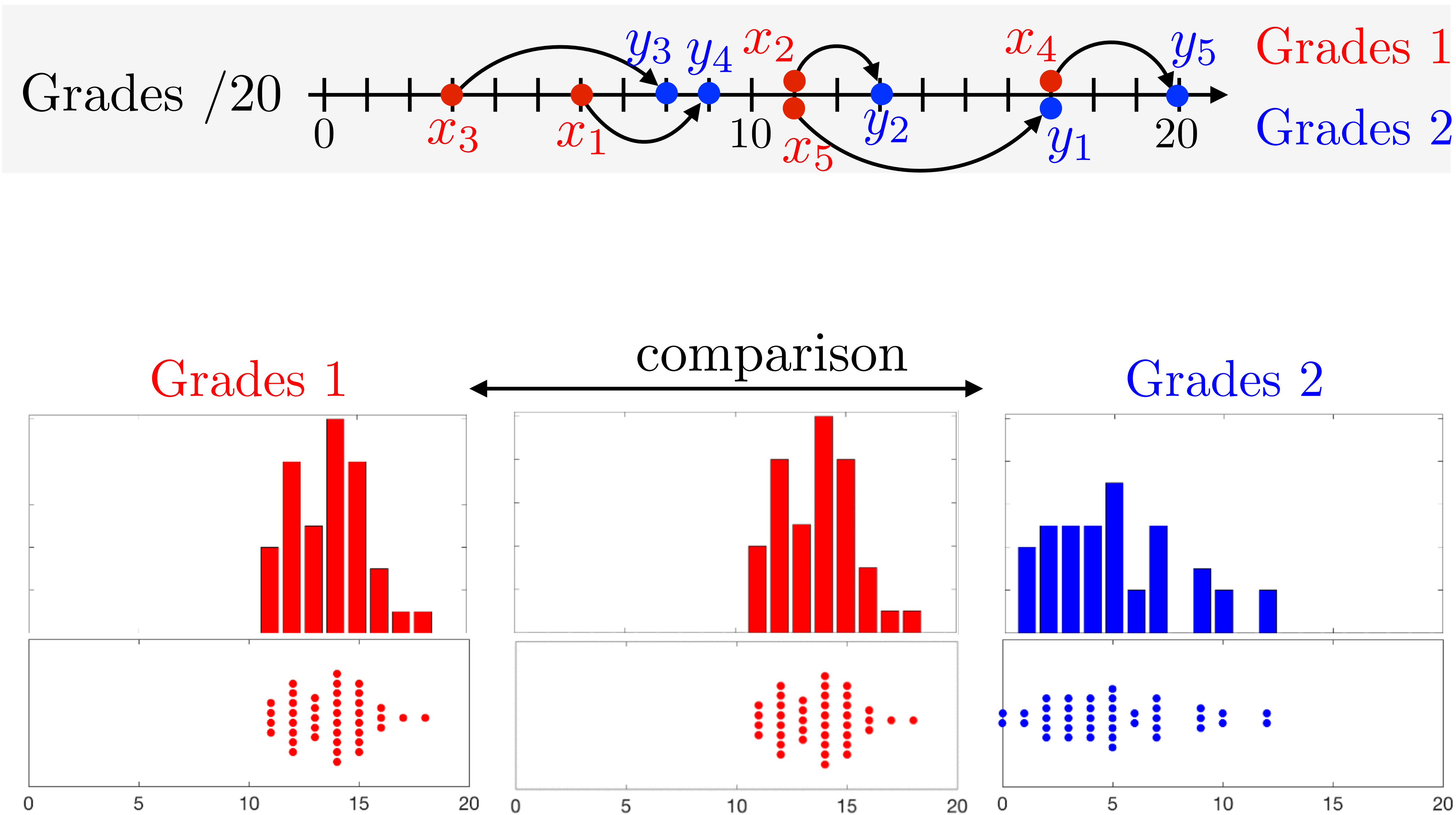
QuickSort: $O(n \log(n))$.

1-D OT Interpolation



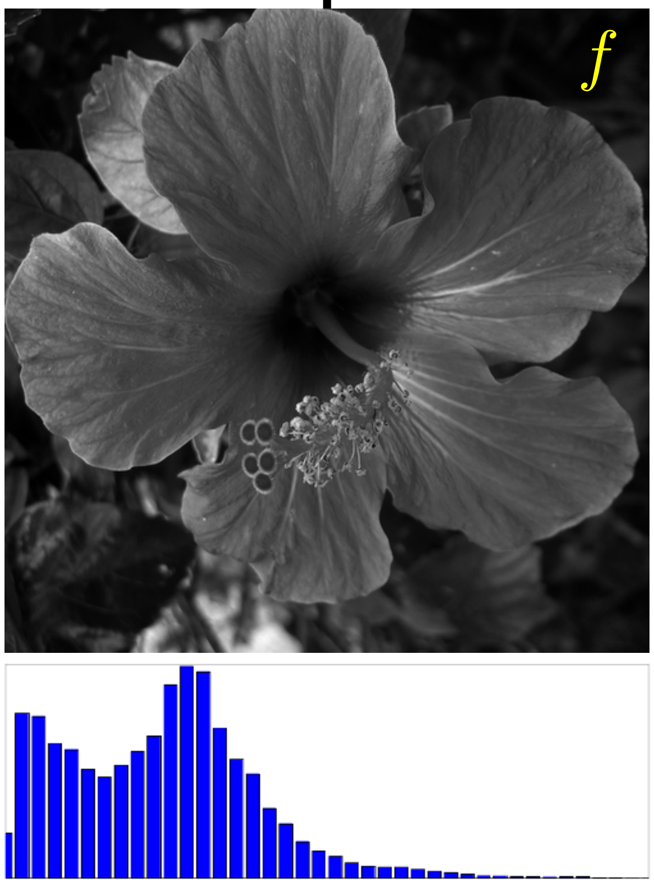


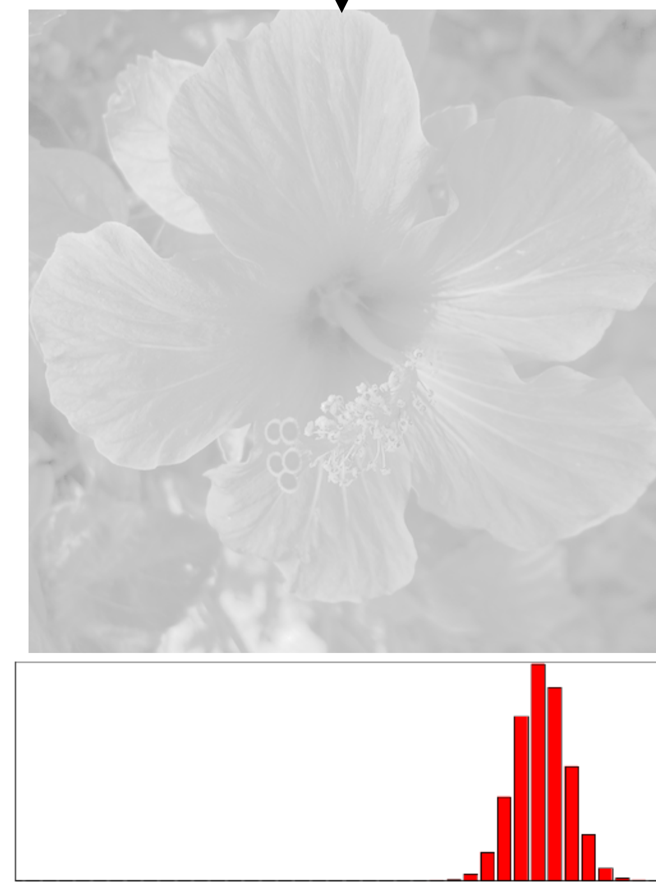
1-D OT Interpolation



Grayscale Histogram Equalization

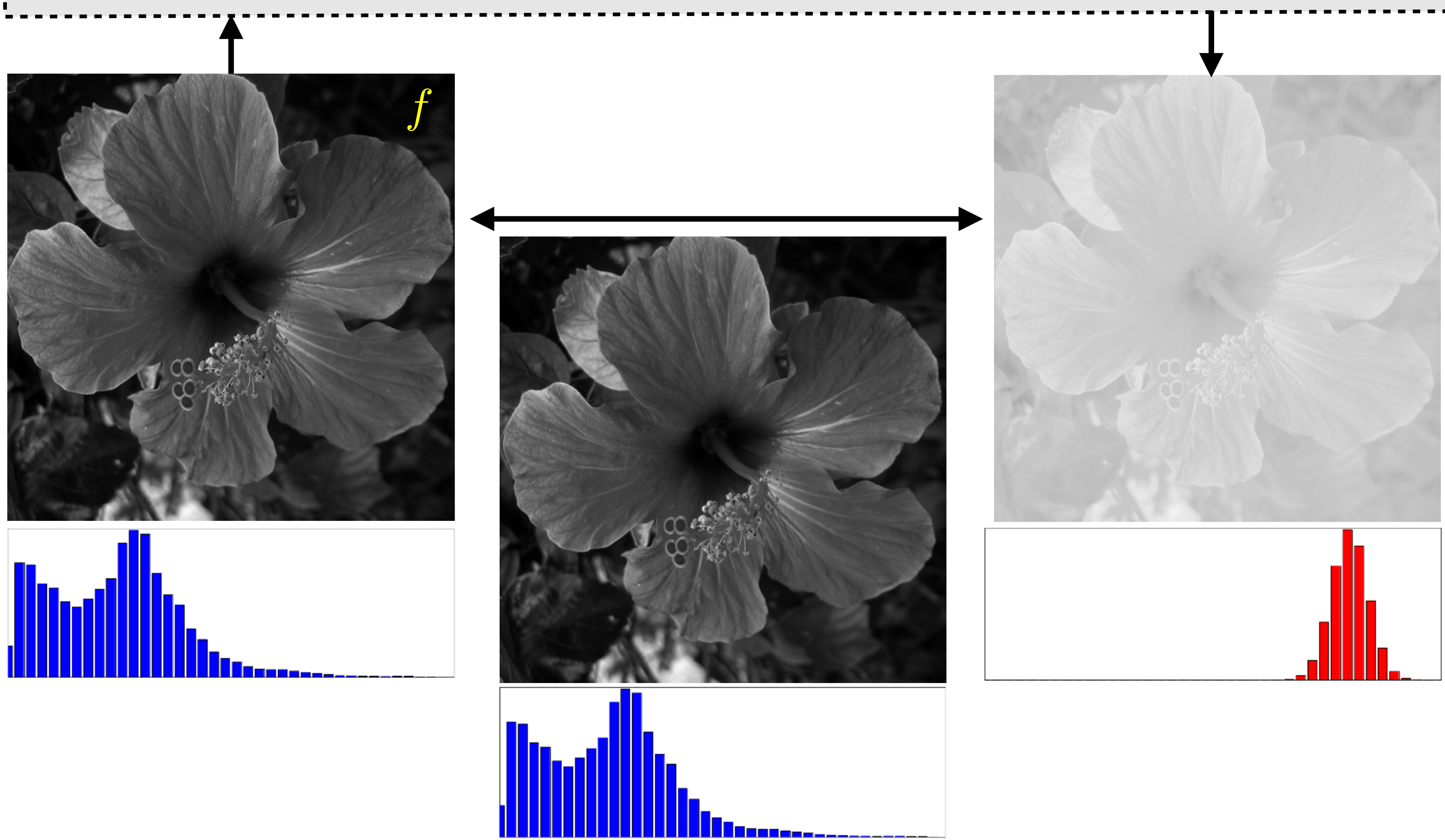
```
f[argsort(f.flatten())] = np.sort(g.flatten())
```





Grayscale Histogram Equalization

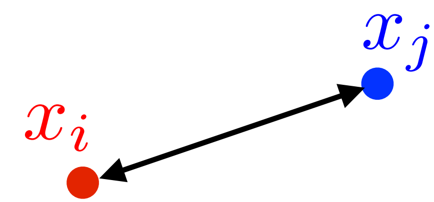
```
f[argsort(f.flatten())] = np.sort(g.flatten())
```

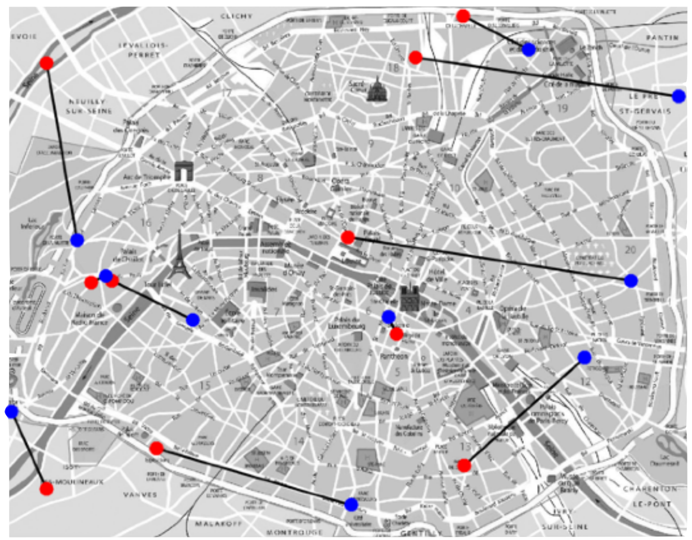


In 2-D

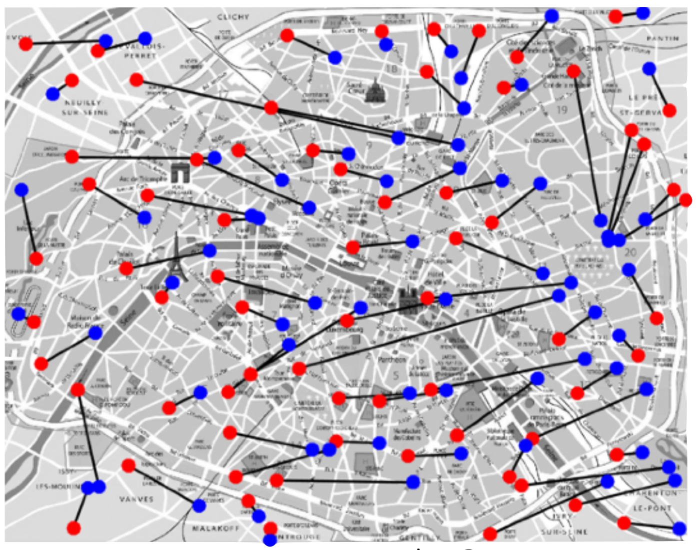
$x_i, y_j \in \mathbb{R}^2$

$$c_{i,j} = \|x_i - y_j\| = \sqrt{(x_i^1 - y_j^1)^2 + (x_i^2 - y_j^2)^2}$$

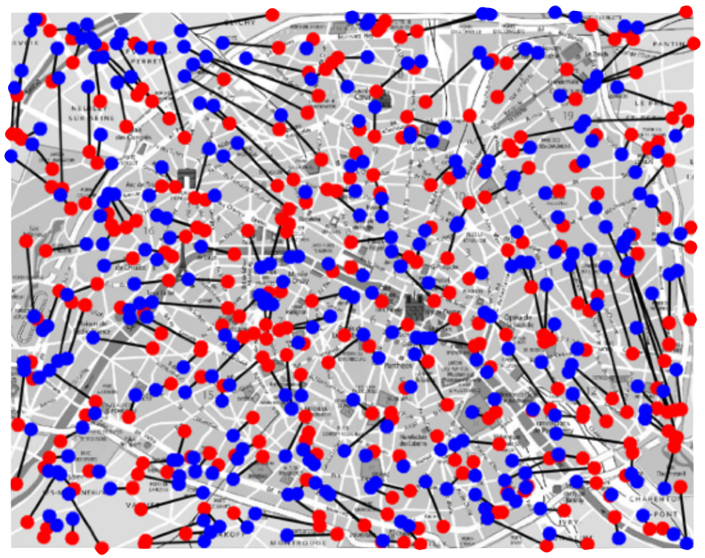




$n = 10$

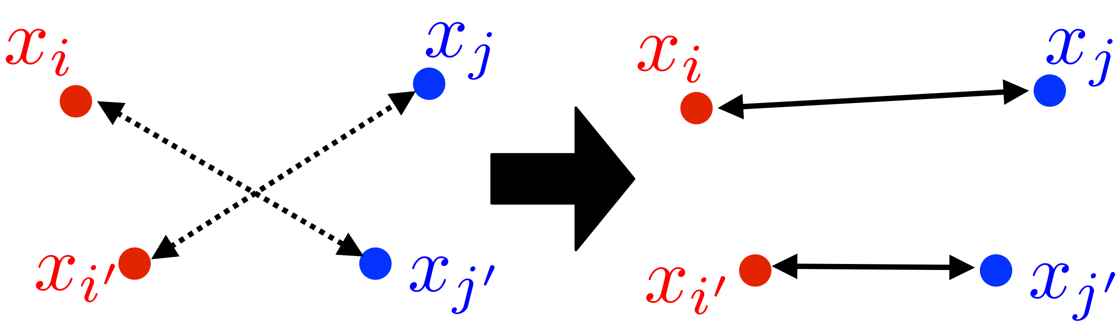


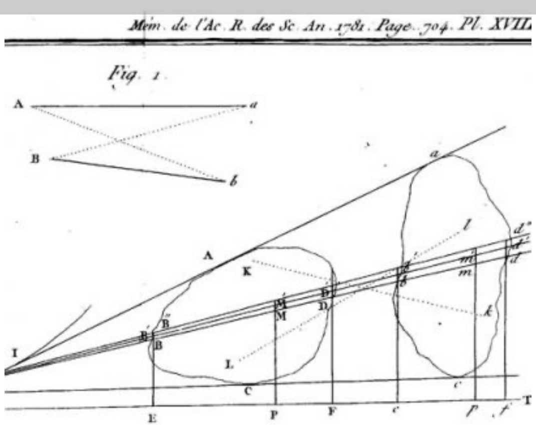
$n = 70$



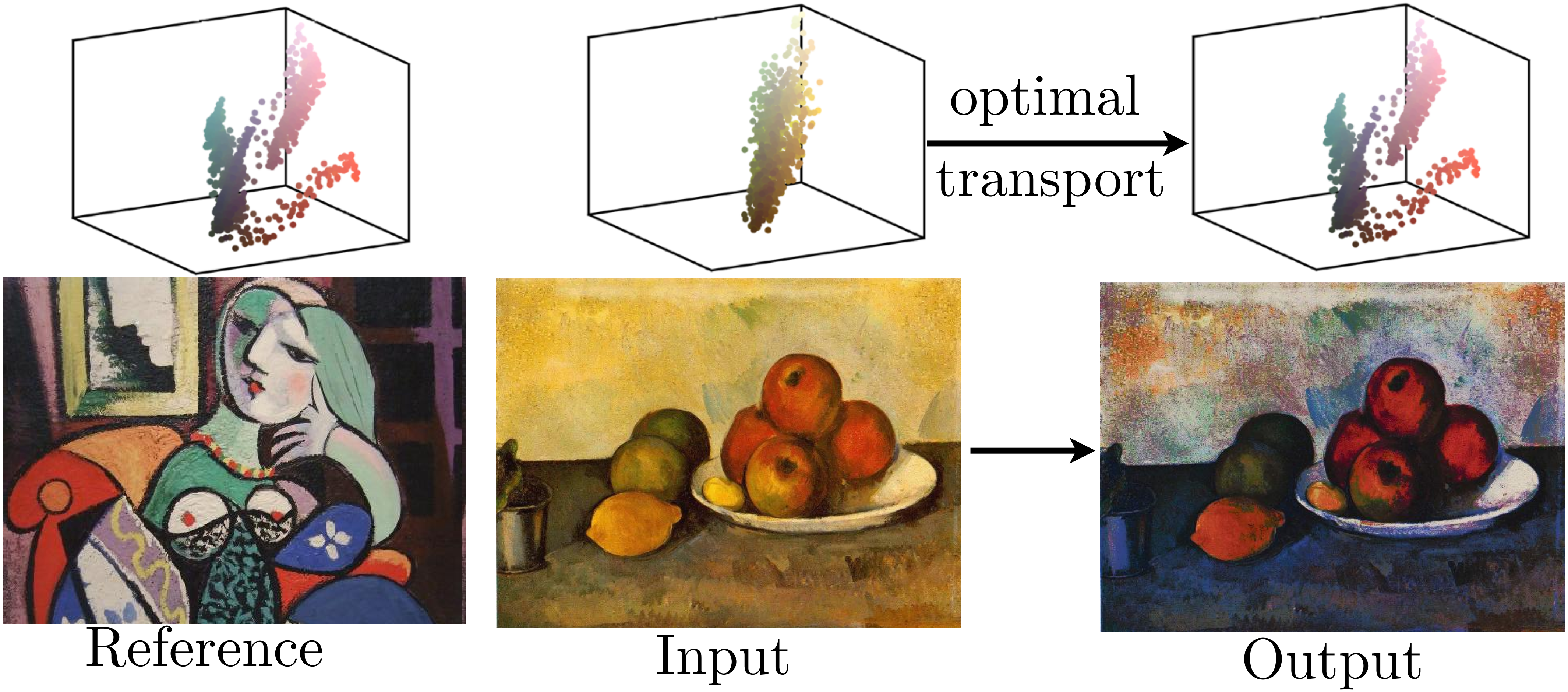
$n = 300$

Proposition: two segments never cross.





In 3-D: Color Image Palette Equalization



Color Image Palette Equalization

Color Image Palette Equalization

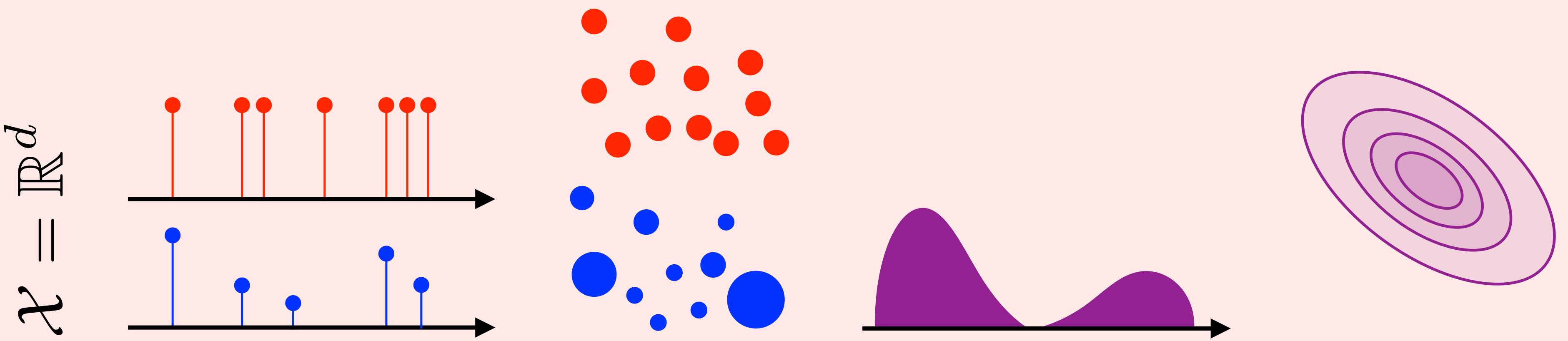
Overview

- Monge Formulation
- **Continuous Optimal Transport**
- Kantorovitch Formulation
- Applications

Probability Measures

Positive Radon measure α on a metric space $\mathcal{X}$.

$$\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x \qquad \alpha = \sum_i \mathbf{a}_i \delta_{x_i}$$

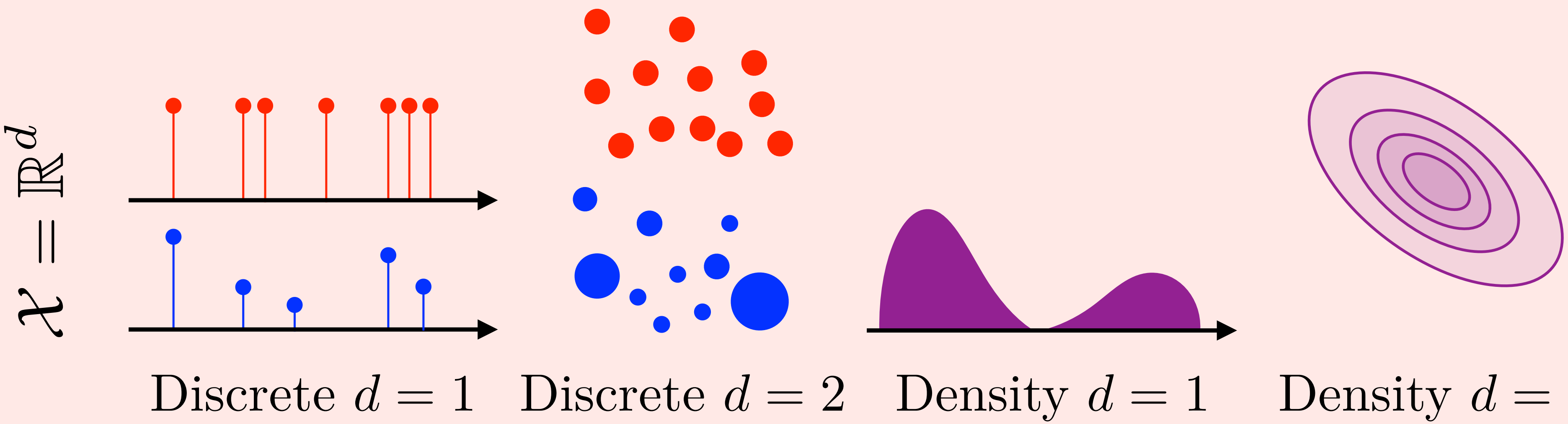


Discrete $d = 1$ Discrete $d = 2$ Density $d = 1$ Density $d = 2$

Probability Measures

Positive Radon measure α on a metric space $\mathcal{X}$.

$$\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x \qquad \alpha = \sum_i \mathbf{a}_i \delta_{x_i}$$



Measure of sets $A \subset \mathcal{X}$: $\alpha(A) = \int_A \mathrm{d}\alpha(x) \geqslant 0$

Integration against continuous functions: $\int_{\mathcal{X}} g(x)\mathrm{d}\alpha(x) \geqslant 0$

$$\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x \longrightarrow \int_{\mathcal{X}} g\mathrm{d}\alpha = \int_{\mathcal{X}} g(x)\rho_\alpha(x)\mathrm{d}x$$

$$\alpha = \sum_i \mathbf{a}_i \delta_{x_i} \longrightarrow \int_{\mathcal{X}} g\mathrm{d}\alpha = \sum_i \mathbf{a}_i g(x_i)$$

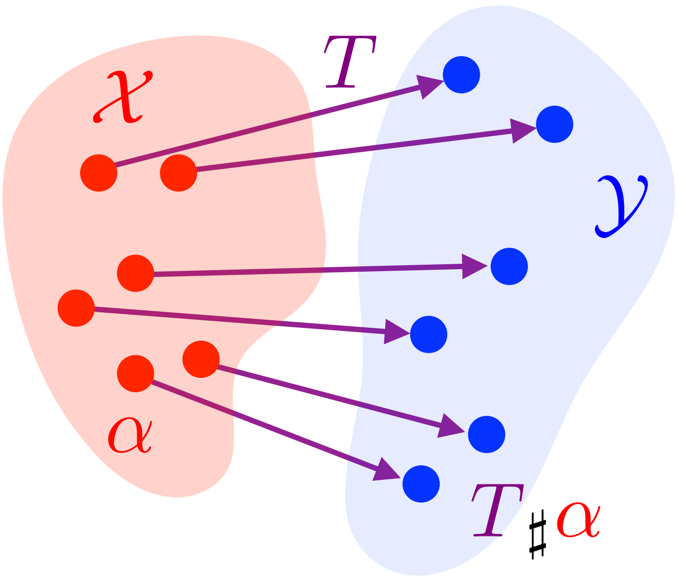
Probability (normalized) measure: $\alpha(\mathcal{X}) = \int_{\mathcal{X}} \mathrm{d}\alpha(x) = 1$

Push Forward

Map: $T : \mathcal{X} \to \mathcal{Y}$

Push-forward:

$$T_\sharp : \begin{cases} \delta_x \longmapsto \delta_{T(x)} \\ \sum_i \delta_{x_i} \longmapsto \sum_i \delta_{T(x_i)} \\ \sum_i \mathbf{a}_i \delta_{x_i} \longmapsto \sum_i \mathbf{a}_i \delta_{T(x_i)} \end{cases}$$

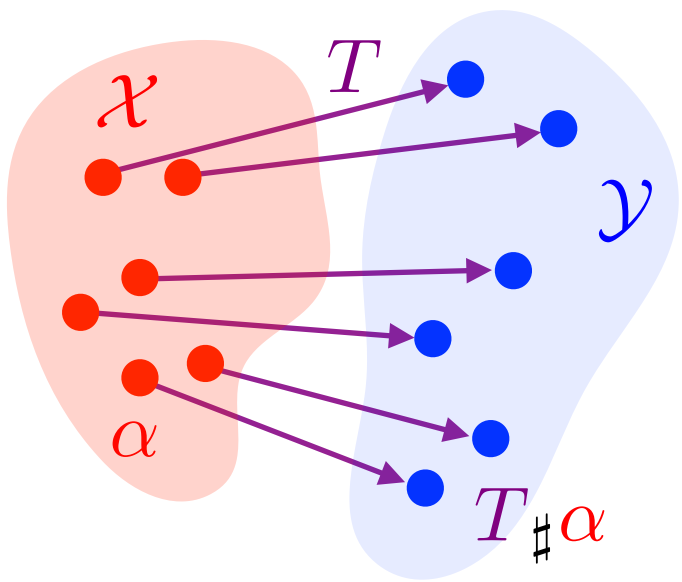


Push Forward

Map: $T : \mathcal{X} \to \mathcal{Y}$

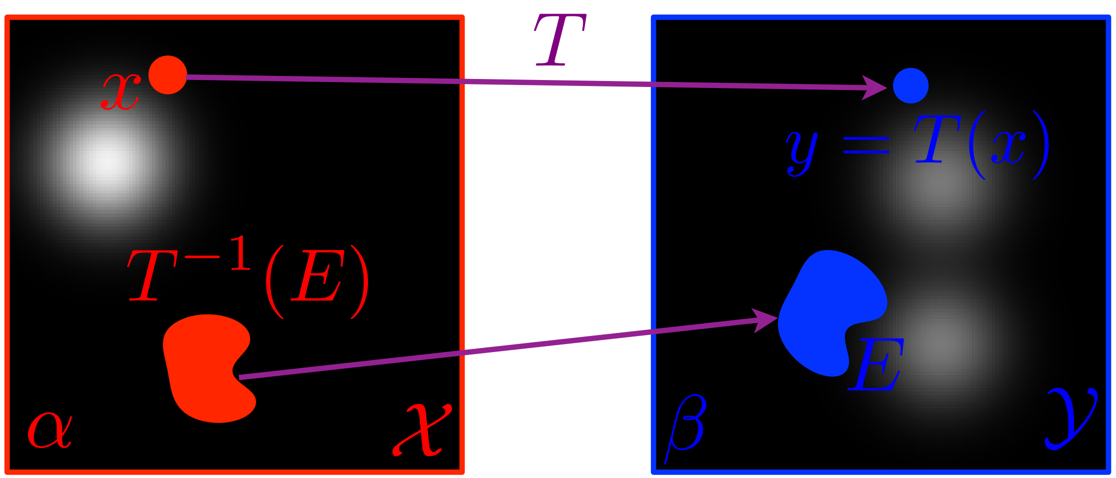
Push-forward:

$$T_\sharp : \begin{cases} \delta_x \longmapsto \delta_{T(x)} \\ \sum_i \delta_{x_i} \longmapsto \sum_i \delta_{T(x_i)} \\ \sum_i \mathbf{a}_i \delta_{x_i} \longmapsto \sum_i \mathbf{a}_i \delta_{T(x_i)} \end{cases}$$



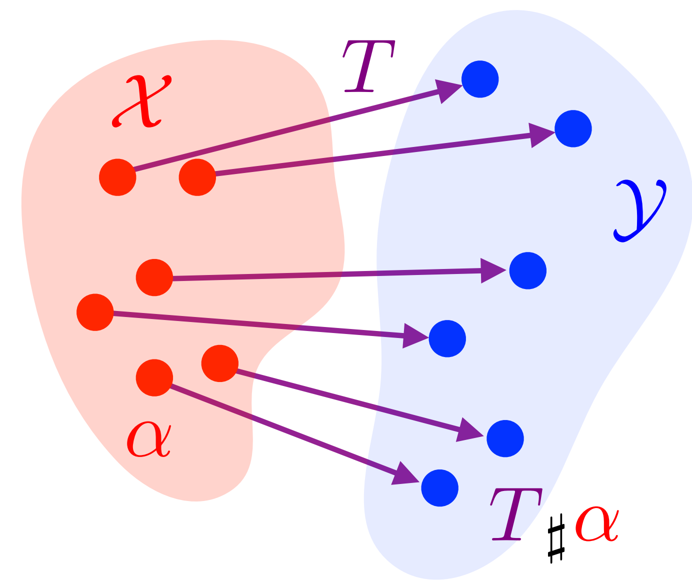
General case:

$$(T_\sharp \alpha)(E) \stackrel{\text{def.}}{=} \alpha(T^{-1}(E))$$



Push Forward

Map: $T : \mathcal{X} \rightarrow \mathcal{Y}$

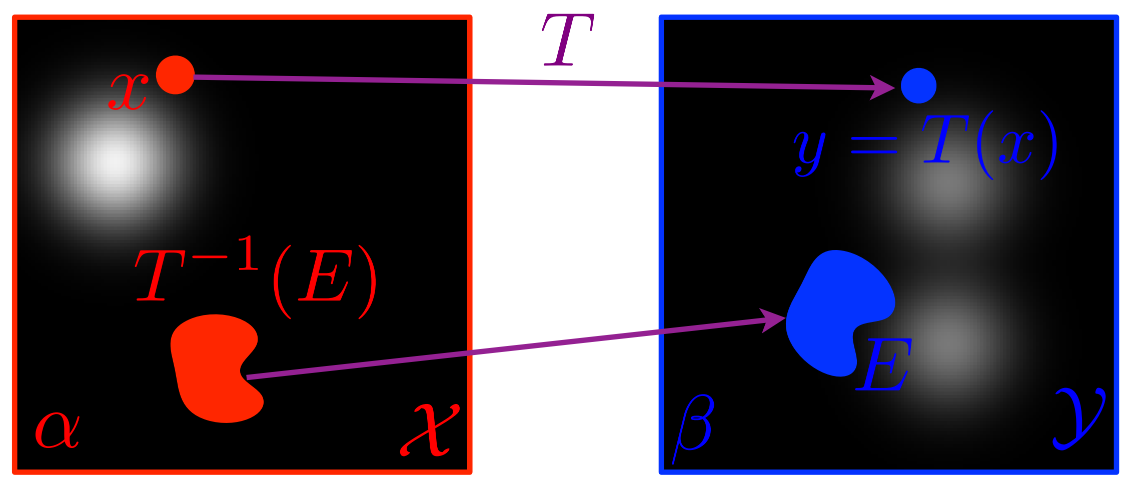


Push-forward:

$$T_\sharp : \left\{ \begin{array}{rcl} \delta_x & \longmapsto & \delta_{T(x)} \\ \sum_i \delta_{x_i} & \longmapsto & \sum_i \delta_{T(x_i)} \\ \sum_i \mathbf{a}_i \delta_{x_i} & \longmapsto & \sum_i \mathbf{a}_i \delta_{T(x_i)} \end{array} \right.$$

General case:

$$(T_\sharp \alpha)(E) \overset{\text{def.}}{=} \alpha(T^{-1}(E))$$



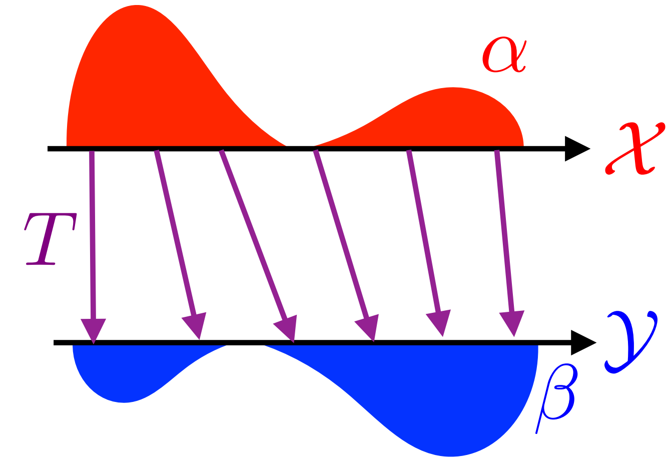
Change of variables:

$$\beta = T_\sharp \alpha \quad \Longleftrightarrow \quad \int_{\mathcal{Y}} g(y) \mathrm{d}\beta(y) = \int_{\mathcal{X}} g(T(x)) \mathrm{d}\alpha(x)$$

Densities $\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \rho_\alpha$*:* $\qquad \rho_\alpha(x) = |\det(\partial T(x))| \rho_\beta(T(x))$

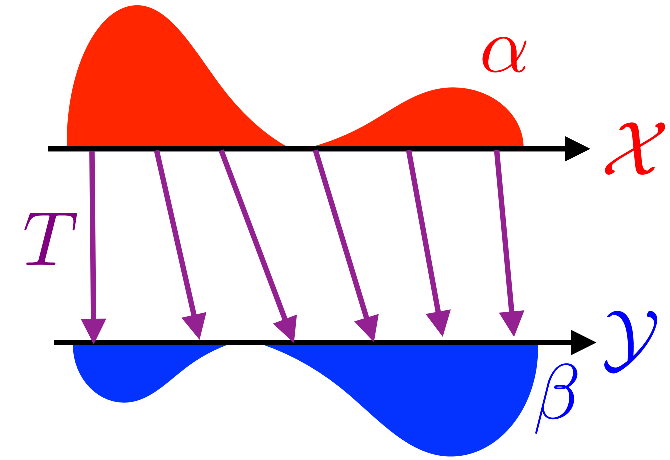
Continuous Monge's Problem

$$\inf_{\beta = T_\sharp \alpha} \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x)$$



Continuous Monge's Problem

$$\inf_{\beta = T_\sharp \alpha} \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x)$$



Discrete case:

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

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

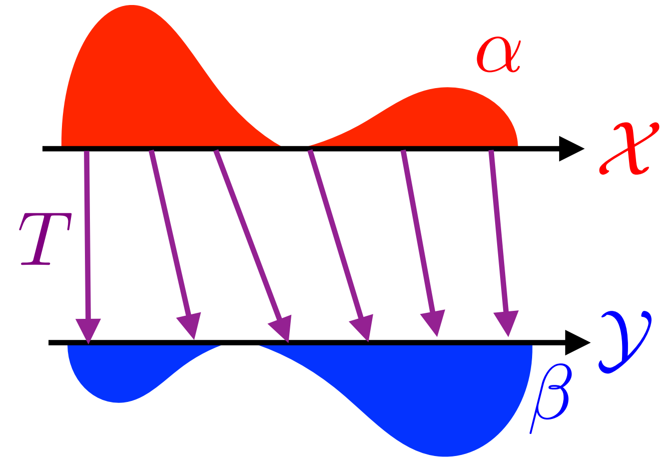
$$\min_{\sigma \in \Sigma_n} \sum_{i=1}^n \mathbf{C}_{i,\sigma(i)}$$

$$T : x_i \longmapsto y_{\sigma(i)}$$

$$\mathbf{C}_{i,j} = c(x_i, y_j)$$

Continuous Monge's Problem

$$\inf_{\beta = T_\sharp \alpha} \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x)$$



Discrete case:

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

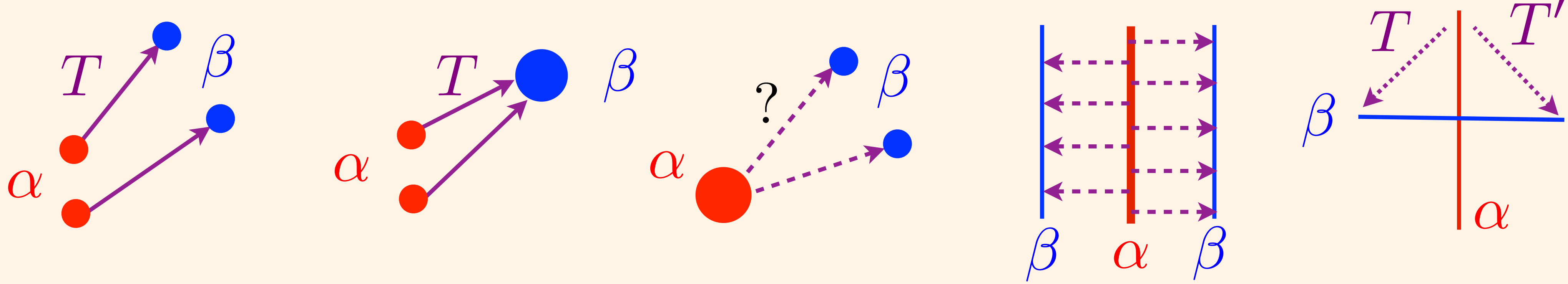
$\beta = \sum_{j=1}^n \delta_{y_j}$

$$\min_{\sigma \in \Sigma_n} \sum_{i=1}^n \mathbf{C}_{i,\sigma(i)}$$

$T : x_i \longmapsto y_{\sigma(i)}$

$\mathbf{C}_{i,j} = c(x_i, y_j)$

Non-symmetry, non-existence, non-uniqueness:

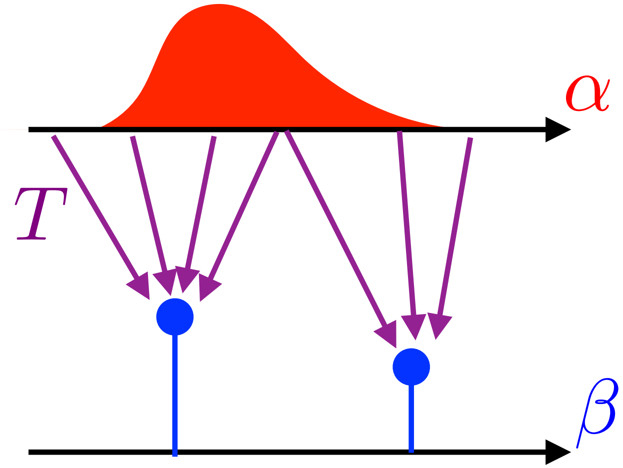


Brenier's Theorem

Hypotheses: $c(x, y) = \|x - y\|^2$

$\mathcal{X} = \mathbb{R}^d$ $\qquad \frac{\mathrm{d}\alpha}{\mathrm{d}x} = \rho_\alpha$ density.

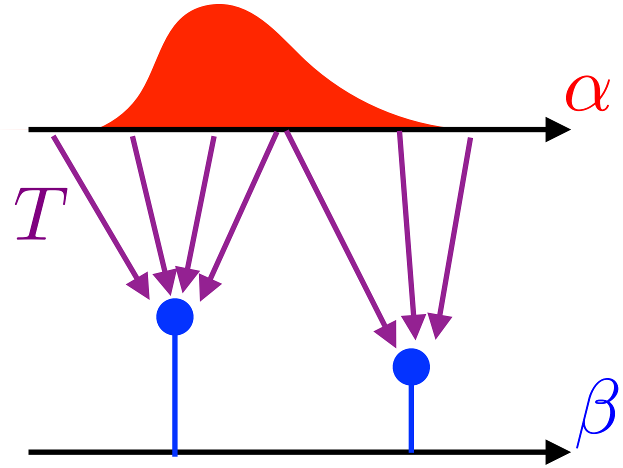
$$W_2^2(\alpha, \beta) \overset{\text{def.}}{=} \inf_{\beta = T_\sharp \alpha} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \mathrm{d}\alpha(x)$$



Brenier's Theorem

Hypotheses: $c(x,y) = \|x-y\|^2$

$\mathcal{X} = \mathbb{R}^d$ $\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \rho_\alpha$ density.



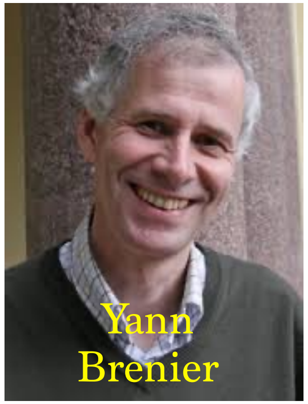
$$W_2^2(\alpha,\beta) \overset{\text{def.}}{=} \inf_{\beta = T_\sharp \alpha} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \mathrm{d}\alpha(x)$$

Theorem: [Brenier, 1991]

There exists a unique Monge map T.

It is the unique $T = \nabla\varphi$ such that

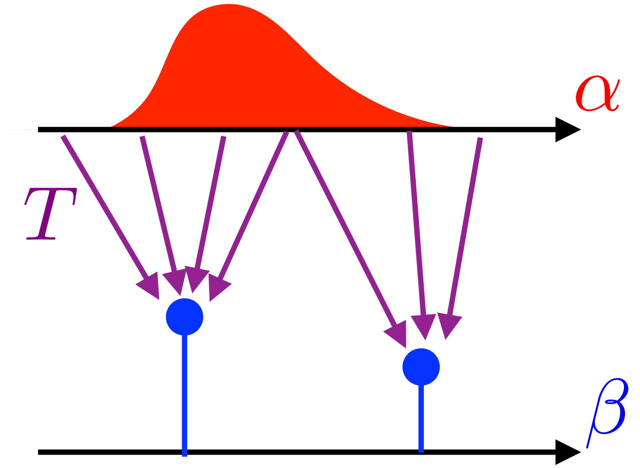
φ is convex and $(\nabla\varphi)_\sharp \alpha = \beta$.



Brenier's Theorem

Hypotheses: $c(x, y) = \|x - y\|^2$

$\mathcal{X} = \mathbb{R}^d$ $\quad \frac{\mathrm{d}\alpha}{\mathrm{d}x} = \rho_\alpha$ density.



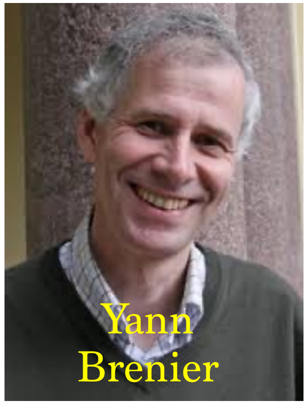
$$W_2^2(\alpha, \beta) \stackrel{\text{def.}}{=} \inf_{\beta = T_\sharp \alpha} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \mathrm{d}\alpha(x)$$

Theorem: [Brenier, 1991]

There exists a unique Monge map T.

It is the unique $T = \nabla\varphi$ such that

φ is convex and $(\nabla\varphi)_\sharp \alpha = \beta$.

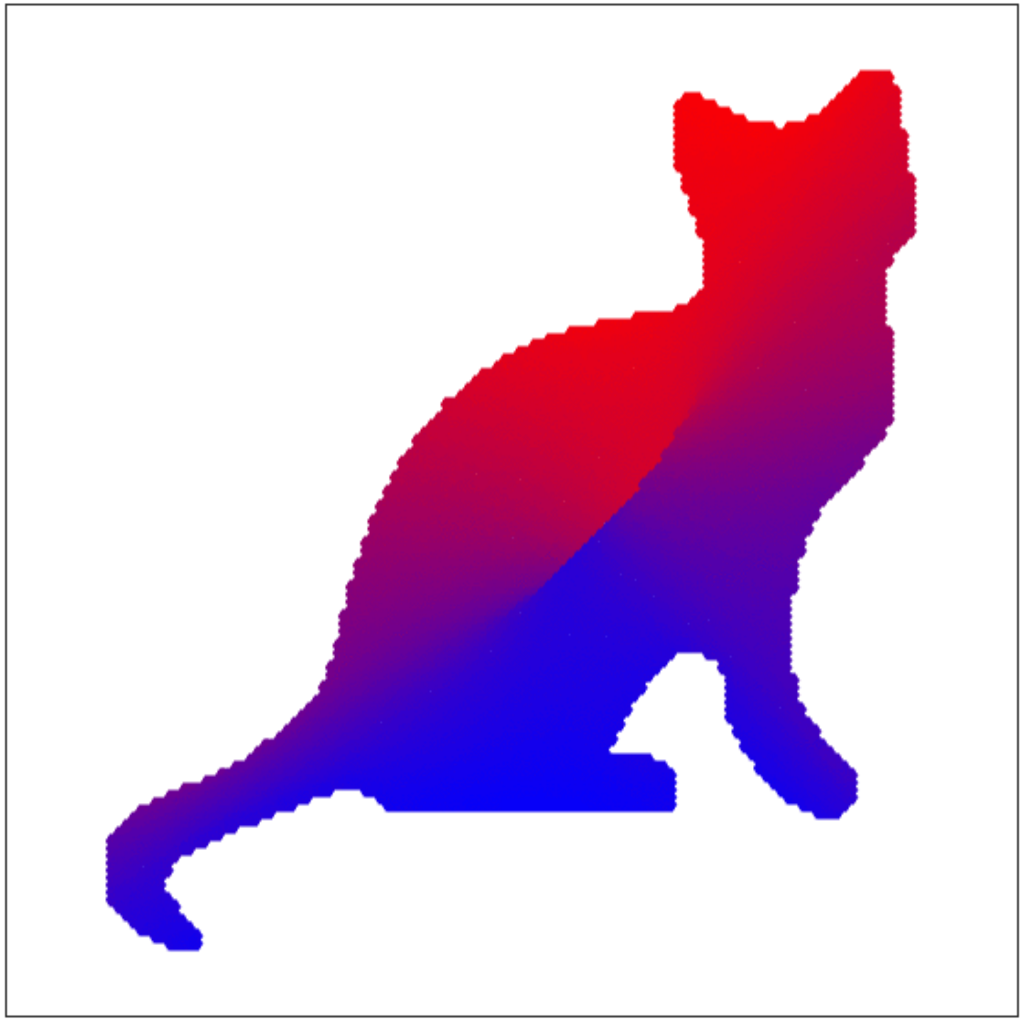


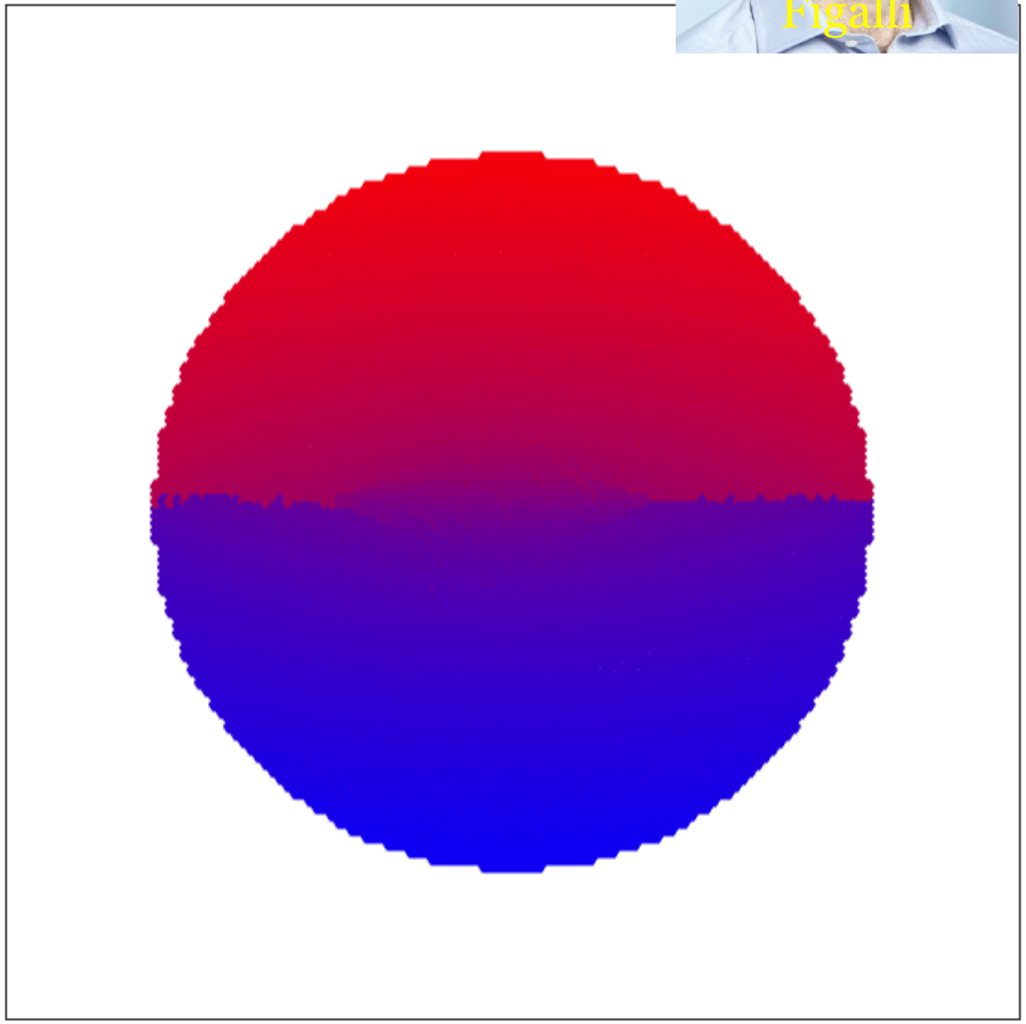
$\rightarrow$ Monge-Ampère equation (non-linear, degenerate elliptic).

$$\rho_\alpha(x) = |\det(\partial^2 \varphi(x))| \rho_\beta(T(x)) \quad \text{s.t. } \varphi \text{ convex.}$$

Regularity Theory

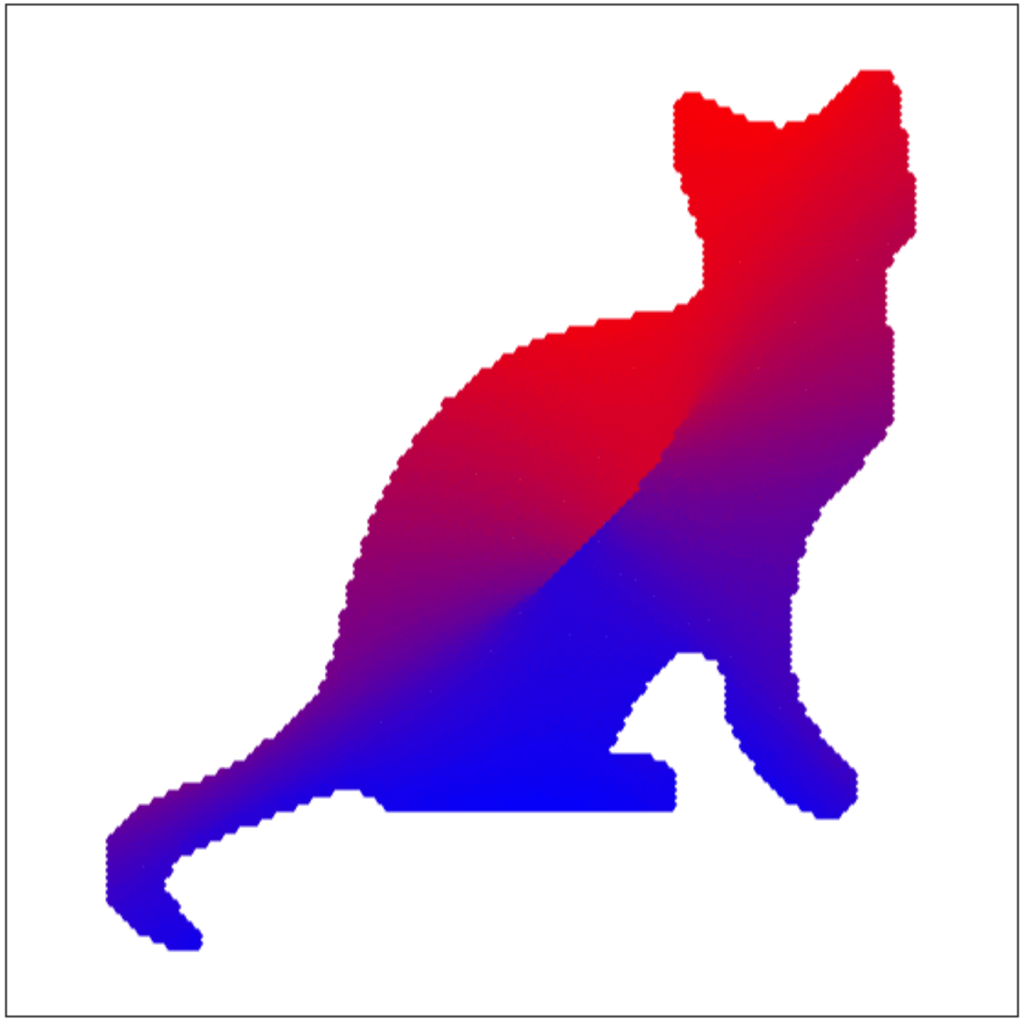
$\rightarrow$ Regularity of T requires convex target.

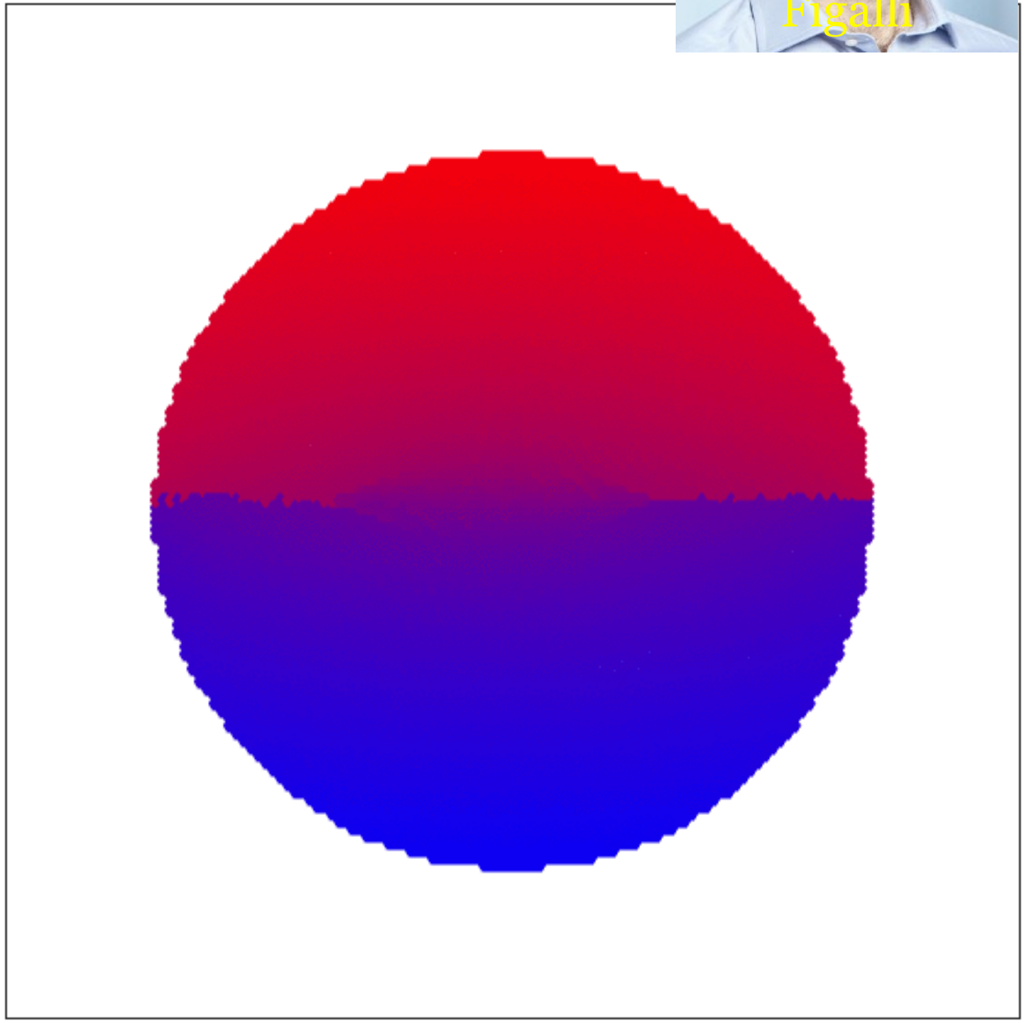




Regularity Theory

→ Regularity of T requires convex target.

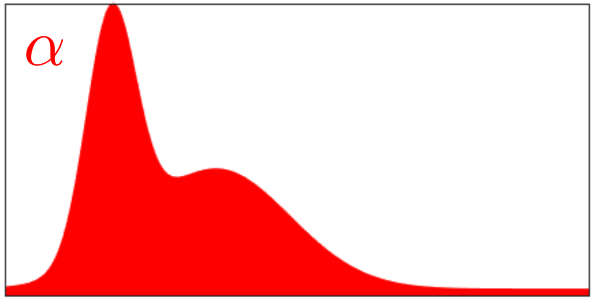


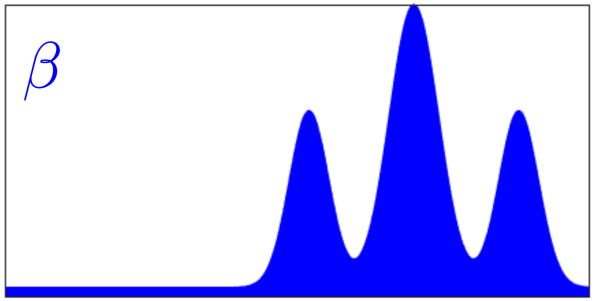


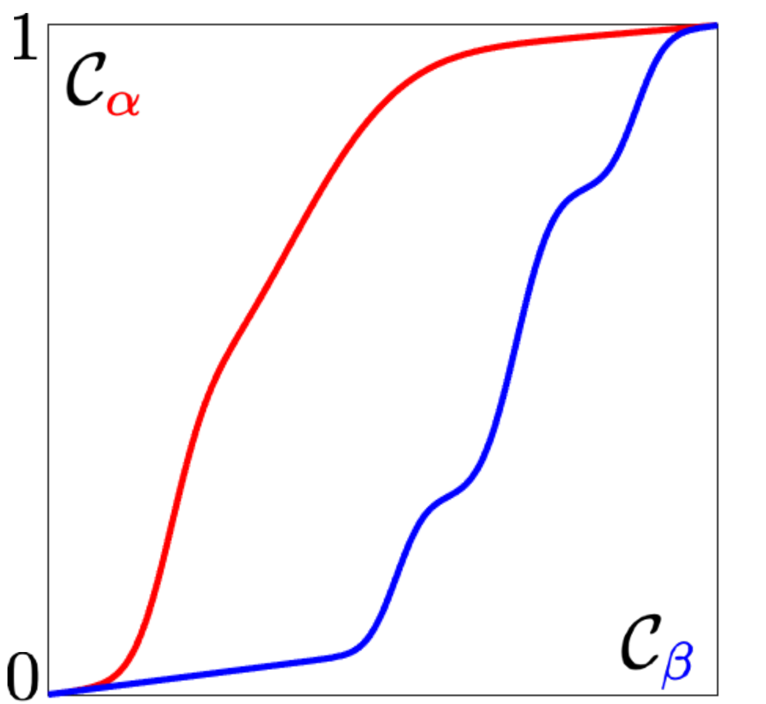
1-D Optimal Transport

Cumulative function: $\mathcal{C}_\alpha(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$

Cumulative function: $C_{\alpha\sharp} : \alpha \longmapsto \mathcal{U}_{[0,1]}$





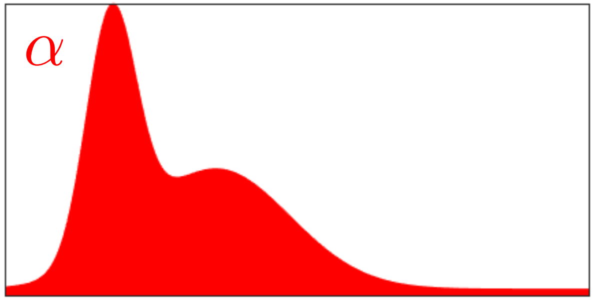


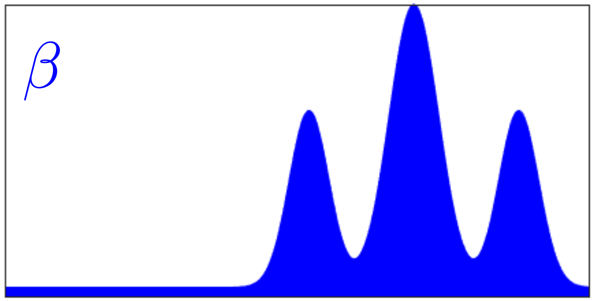
0 1

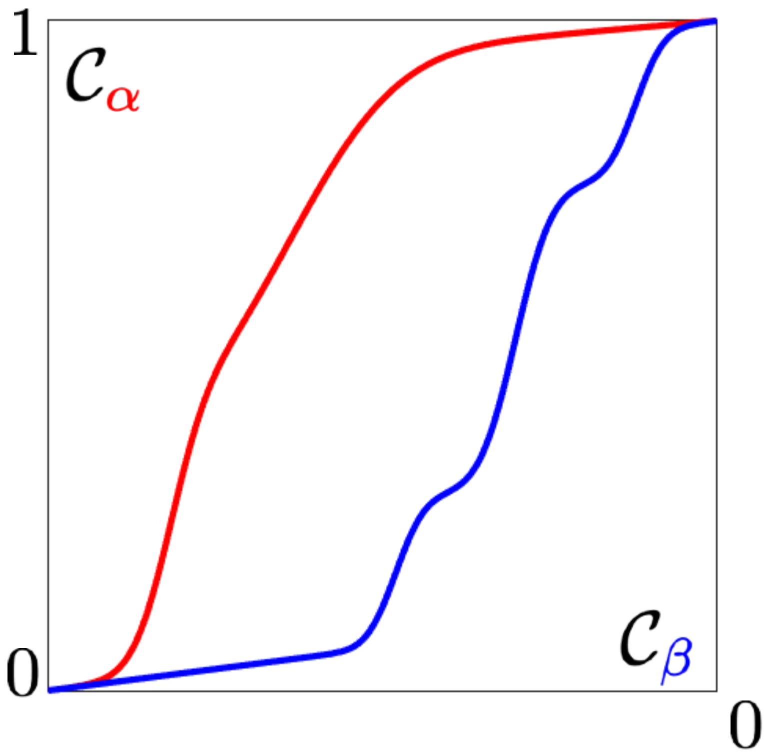
1-D Optimal Transport

Cumulative function: $\mathcal{C}_\alpha(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$

Cumulative function: $C_{\alpha\sharp} : \alpha \longmapsto \mathcal{U}_{[0,1]}$





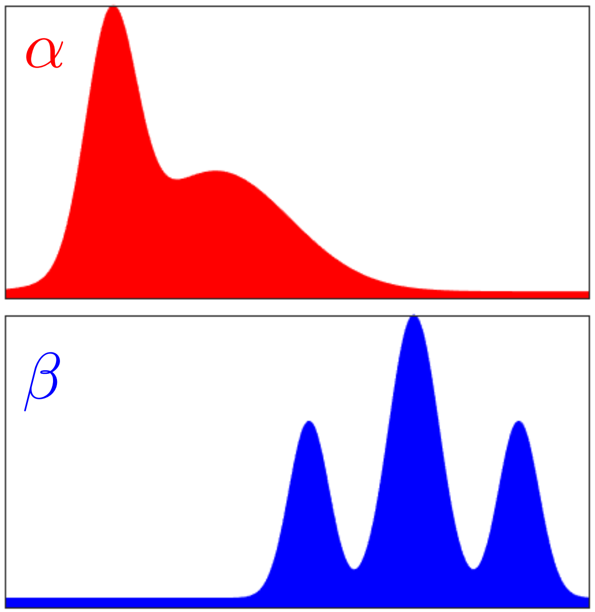


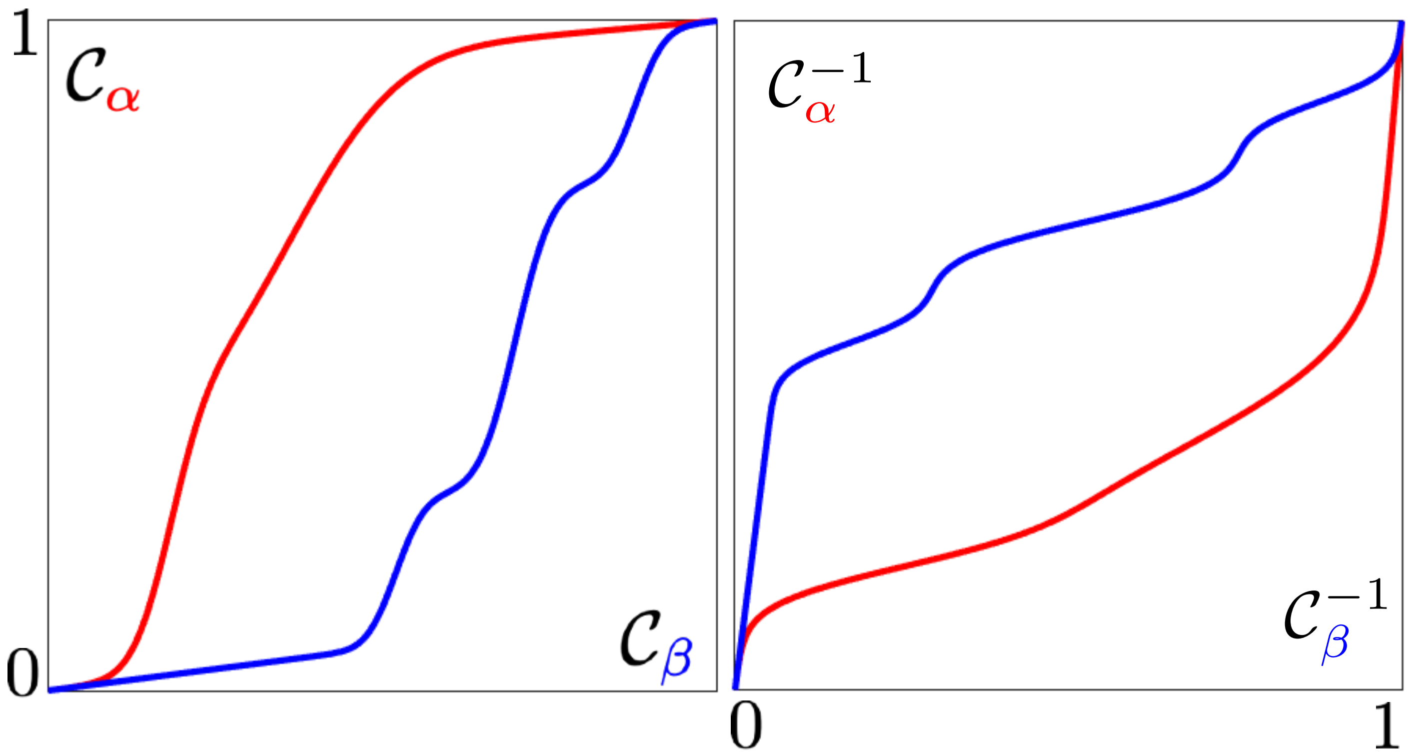
1-D Optimal Transport

Cumulative function: $\mathcal{C}_{\alpha}(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$

Cumulative function: $C_{\alpha\sharp} : \alpha \longmapsto \mathcal{U}_{[0,1]}$

Quantile function: $C_{\beta}^{-1}{}_{\sharp} : \mathcal{U}_{[0,1]} \longmapsto \beta$





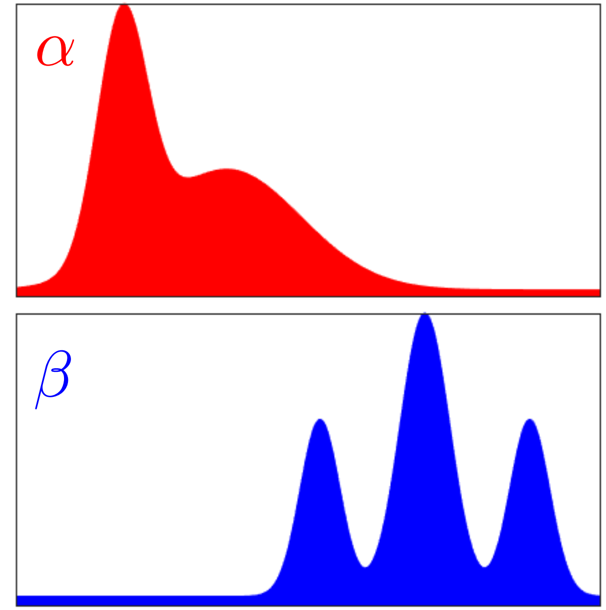
1-D Optimal Transport

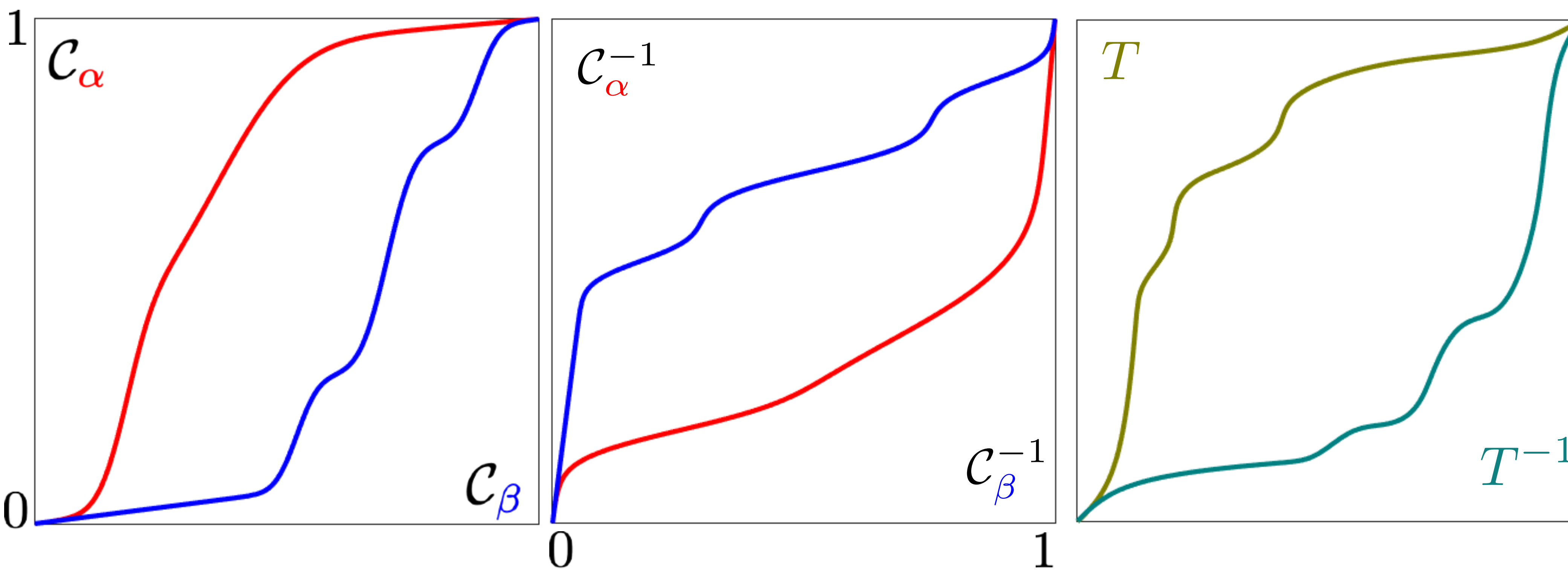
Cumulative function: $\mathcal{C}_\alpha(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$

Cumulative function: $C_{\alpha\sharp} : \alpha \longmapsto \mathcal{U}_{[0,1]}$

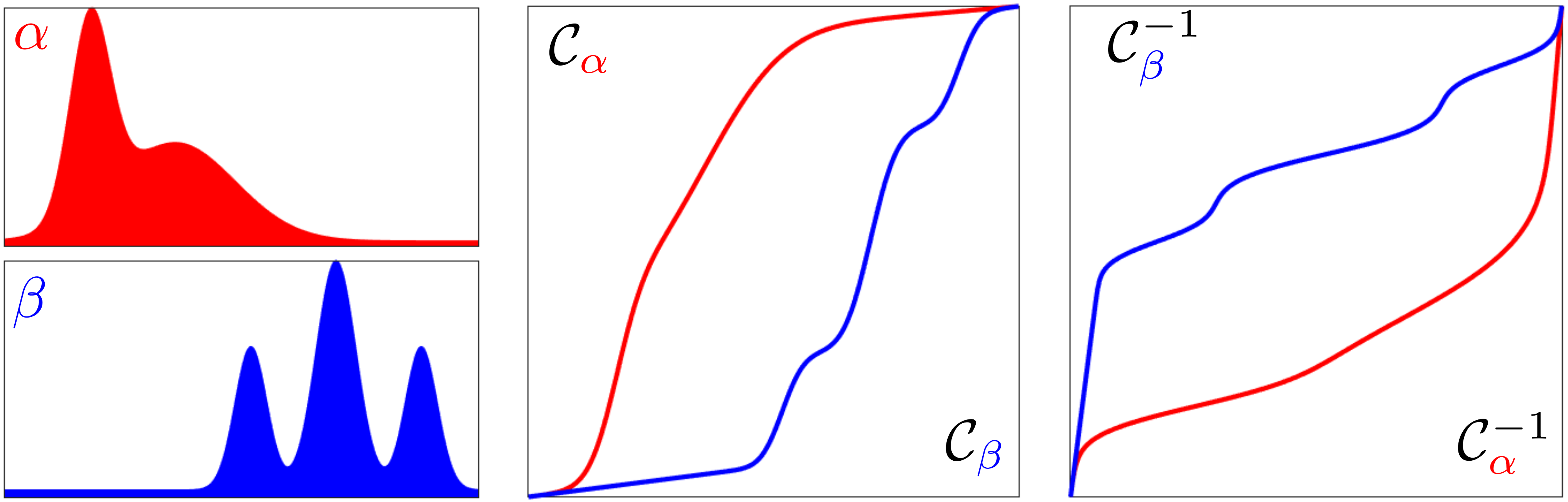
Quantile function: $C_{\beta}^{-1}{}_{\sharp} : \mathcal{U}_{[0,1]} \longmapsto \beta$

Optimal transport $\alpha \mapsto \beta$: $T = C_\beta^{-1} \circ C_\alpha$





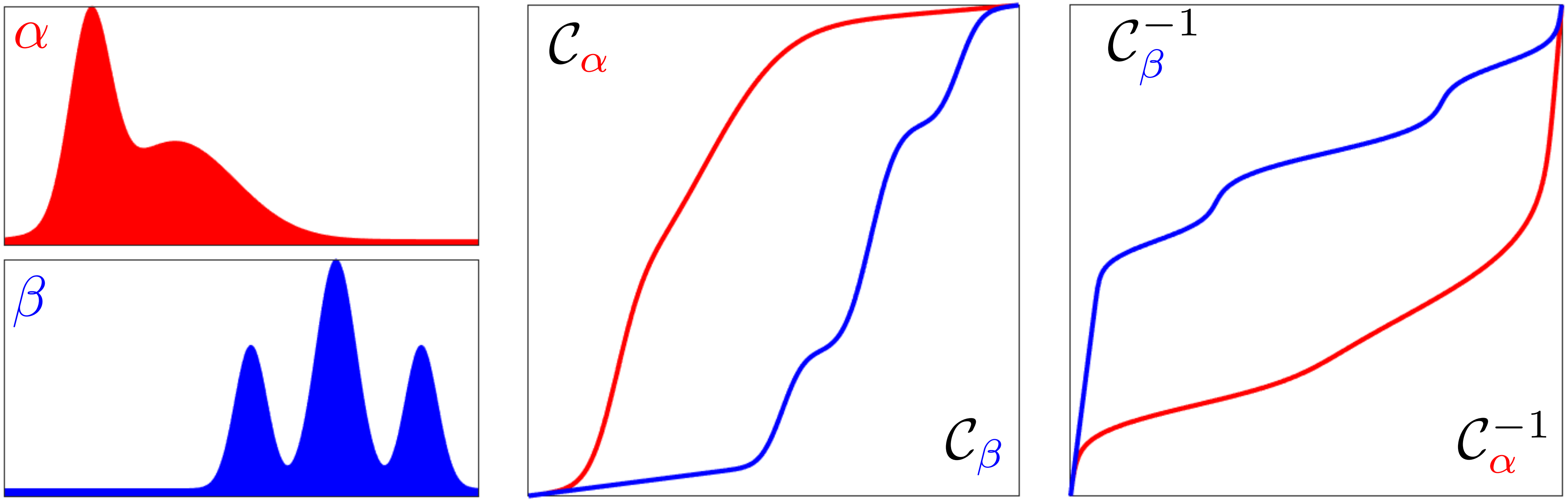
Cumulative function: $\mathcal{C}_\alpha(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$



$$W_p(\alpha,\beta)^p \overset{\text{def.}}{=} \int \|T(x)-x\|^p \mathrm{d}\alpha(x) = \int_0^1 |\mathcal{C}_\alpha^{-1}(t) - \mathcal{C}_\beta^{-1}(t)|^p \mathrm{d}t$$

$$W_1(\alpha,\beta) = \|\alpha-\beta\|_{W_1} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| \mathrm{d}x$$

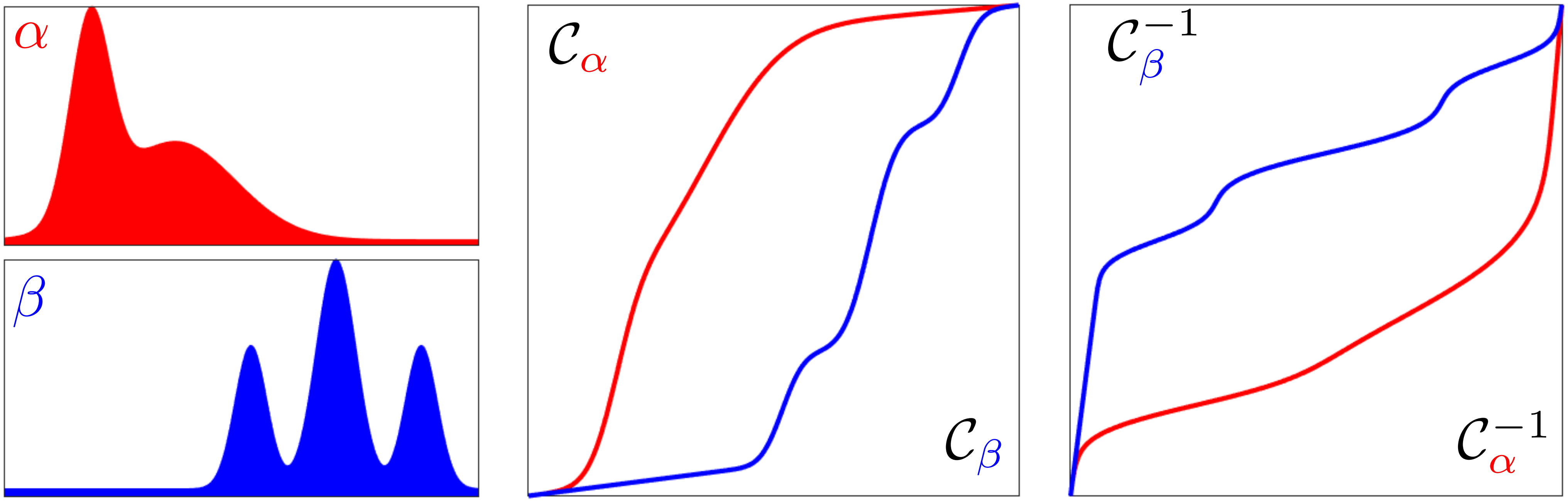
Cumulative function: $\mathcal{C}_{\alpha}(x) \stackrel{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$



$$W_p(\alpha, \beta)^p \stackrel{\text{def.}}{=} \int \|T(x) - x\|^p \mathrm{d}\alpha(x) = \int_0^1 |\mathcal{C}_{\alpha}^{-1}(t) - \mathcal{C}_{\beta}^{-1}(t)|^p \mathrm{d}t$$

$$W_1(\alpha, \beta) = \|\alpha - \beta\|_{W_1} = \int_{\mathbb{R}} |\mathcal{C}_{\alpha}(x) - \mathcal{C}_{\beta}(x)| \mathrm{d}x$$

Cumulative function: $\mathcal{C}_\alpha(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$



$$W_p(\alpha,\beta)^p \overset{\text{def.}}{=} \int \|T(x)-x\|^p \mathrm{d}\alpha(x) = \int_0^1 |\mathcal{C}_\alpha^{-1}(t) - \mathcal{C}_\beta^{-1}(t)|^p \mathrm{d}t$$

$$W_1(\alpha,\beta) = \|\alpha-\beta\|_{W_1} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| \mathrm{d}x$$

Kramer (Sobolev) norm: $\|\alpha-\beta\|_K^2 = \int_0^1 |\mathcal{C}_\alpha(t) - \mathcal{C}_\beta(t)|^2 \mathrm{d}t$

Kolmogorov-Smirnov norm: $\|\alpha-\beta\|_{KS} = \sup_x |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)|$

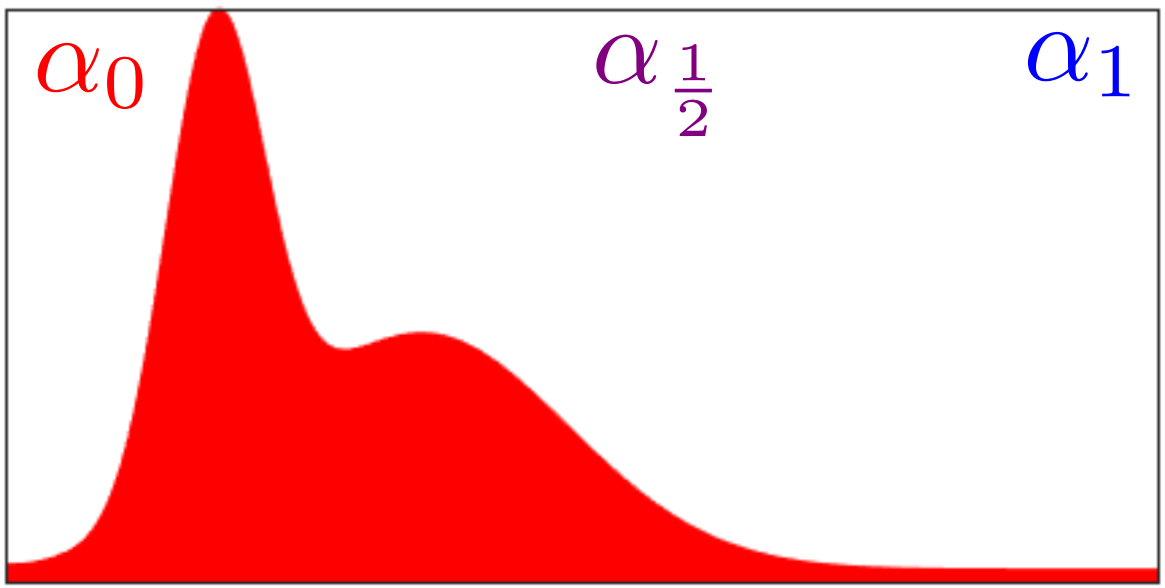
Area under the curve: $\mathrm{AUC}(\alpha,\beta) = 1 - \int_0^1 \mathcal{C}_\alpha \circ \mathcal{C}_\beta{}^{-1}(x)$

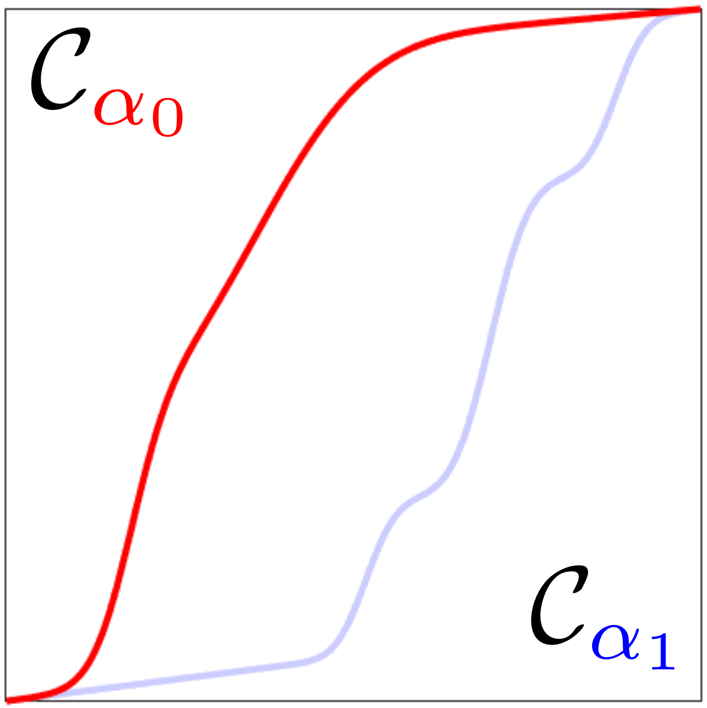
1-D Optimal Transport Interpolation

Cumulative function: $\mathcal{C}_{\alpha}(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$

Optimal transport interpolation $\alpha_0 \leftrightarrow \alpha_1$

$$\forall\, t \in [0,1], \mathcal{C}_{\alpha_t}^{-1} = (1-t)\mathcal{C}_{\alpha_0}^{-1} + t\mathcal{C}_{\alpha_1}^{-1}$$



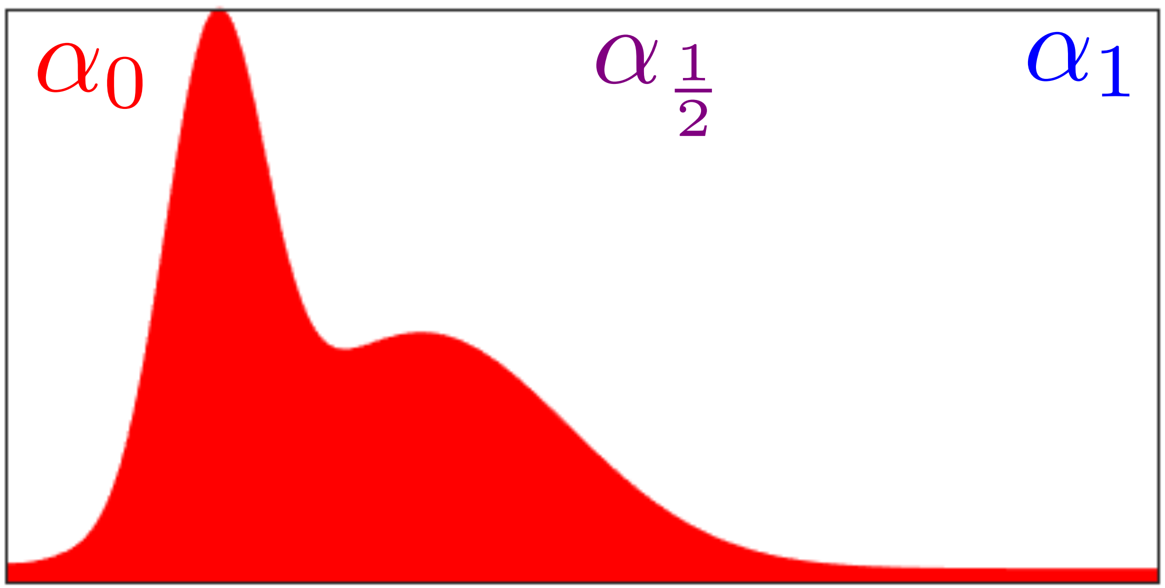


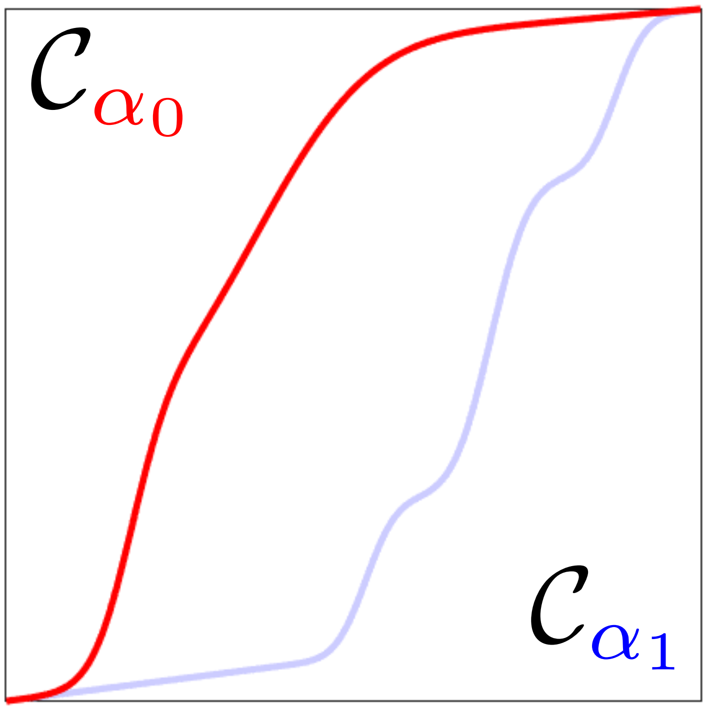
1-D Optimal Transport Interpolation

Cumulative function: $\quad \mathcal{C}_{\alpha}(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha$

Optimal transport interpolation $\alpha_0 \leftrightarrow \alpha_1$

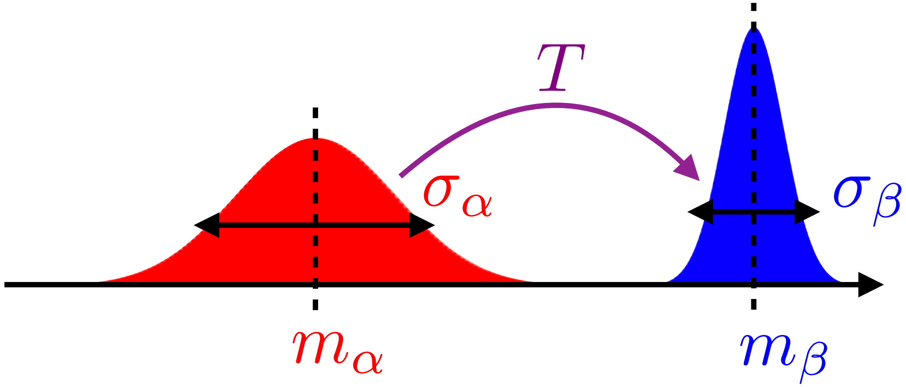
$$\forall\, t \in [0,1], \mathcal{C}_{\alpha_t}^{-1} = (1-t)\mathcal{C}_{\alpha_0}^{-1} + t\mathcal{C}_{\alpha_1}^{-1}$$





OT Between 1D Gaussians

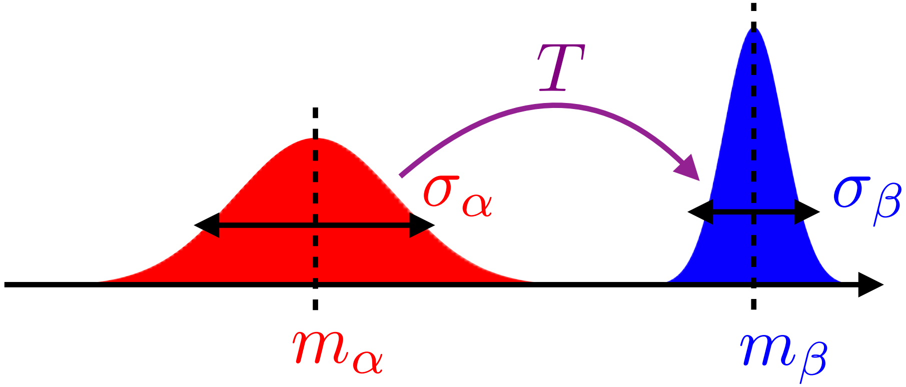
$$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \frac{1}{\sigma_\alpha\sqrt{2\pi}} e^{-\frac{(x-m_\alpha)^2}{2\sigma_\alpha^2}}$$



$$T(x) = \frac{\sigma_\beta}{\sigma_\alpha}(x - m_\alpha) + m_\beta$$

OT Between 1D Gaussians

$$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \frac{1}{\sigma_\alpha\sqrt{2\pi}} e^{-\frac{(x-m_\alpha)^2}{2\sigma_\alpha^2}}$$



$$T(x) = \frac{\sigma_\beta}{\sigma_\alpha}(x - m_\alpha) + m_\beta$$

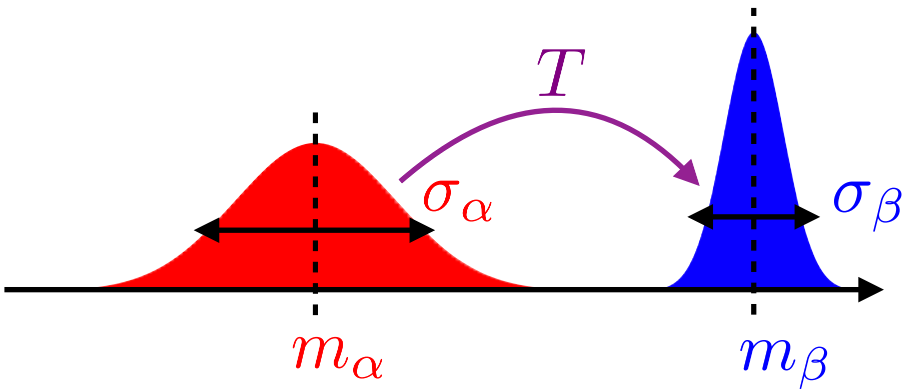
$$\boxed{T = \nabla\varphi \qquad \varphi \text{ is convex.}}$$

$$\overset{\text{Brenier}}{\Longrightarrow} \quad \boxed{T \equiv \mathrm{OT}}$$

$$\varphi(x) = \frac{\sigma_\beta}{2\sigma_\alpha}(x - m_\alpha)^2 + m_\beta x$$

OT Between 1D Gaussians

$$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \frac{1}{\sigma_\alpha\sqrt{2\pi}} e^{-\frac{(x-m_\alpha)^2}{2\sigma_\alpha^2}}$$



$$T(x) = \frac{\sigma_\beta}{\sigma_\alpha}(x - m_\alpha) + m_\beta$$

$$\boxed{T = \nabla\varphi \qquad \varphi \text{ is convex.}}$$

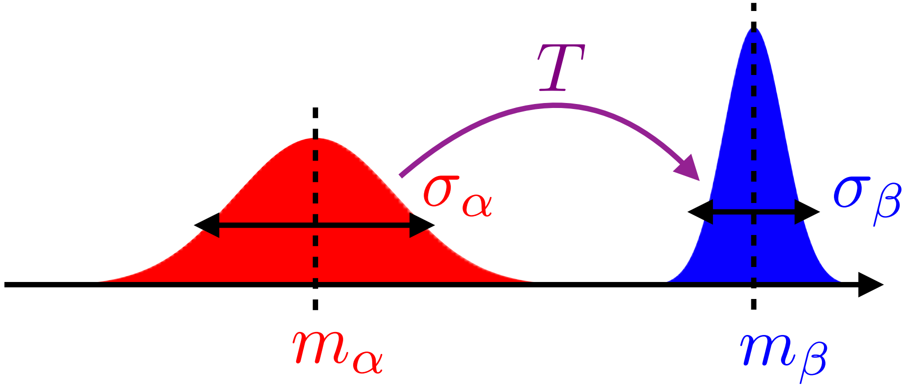
$$\overset{\text{Brenier}}{\Longrightarrow} \quad \boxed{T \equiv \mathrm{OT}}$$

$$\varphi(x) = \frac{\sigma_\beta}{2\sigma_\alpha}(x - m_\alpha)^2 + m_\beta x$$

$$\boxed{\mathrm{W}_2^2(\alpha, \beta) = (m_\alpha - m_\beta)^2 + (\sigma_\alpha - \sigma_\beta)^2}$$

OT Between 1D Gaussians

$$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \frac{1}{\sigma_\alpha\sqrt{2\pi}} e^{-\frac{(x-m_\alpha)^2}{2\sigma_\alpha^2}}$$

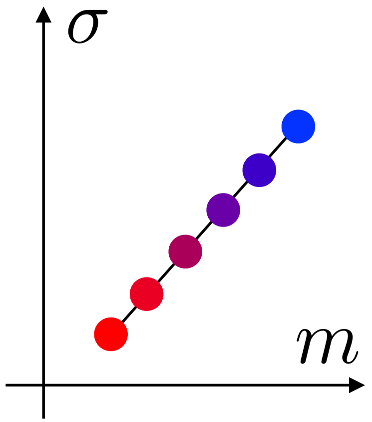


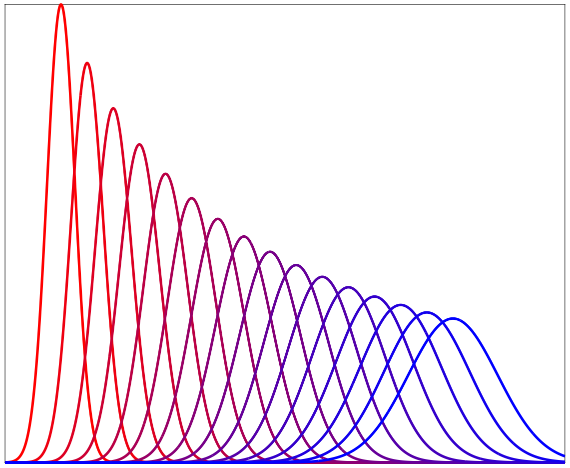
$$T(x) = \frac{\sigma_\beta}{\sigma_\alpha}(x - m_\alpha) + m_\beta$$

$T = \nabla\varphi$ φ is convex.

$\overset{\text{Brenier}}{\Longrightarrow}$ $T \equiv \mathrm{OT}$

$$\varphi(x) = \frac{\sigma_\beta}{2\sigma_\alpha}(x - m_\alpha)^2 + m_\beta x$$



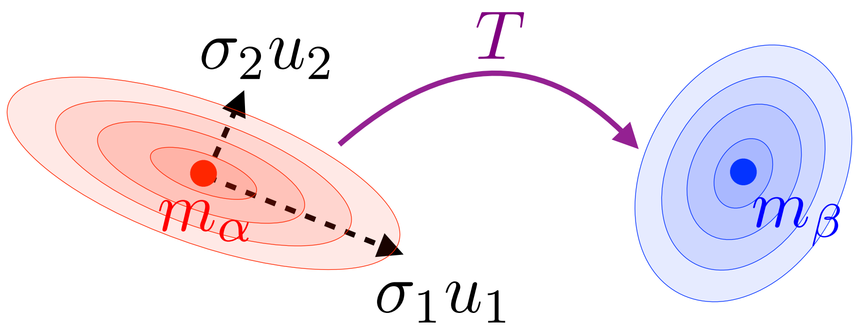


$$\mathrm{W}_2^2(\alpha, \beta) = (m_\alpha - m_\beta)^2 + (\sigma_\alpha - \sigma_\beta)^2$$

OT Between Gaussians

$$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \frac{1}{(2\pi)^{d/2}|\Sigma_\alpha|} e^{-\frac{\|x-m_\alpha\|^2_{\Sigma_\alpha^{-1}}}{2}}$$

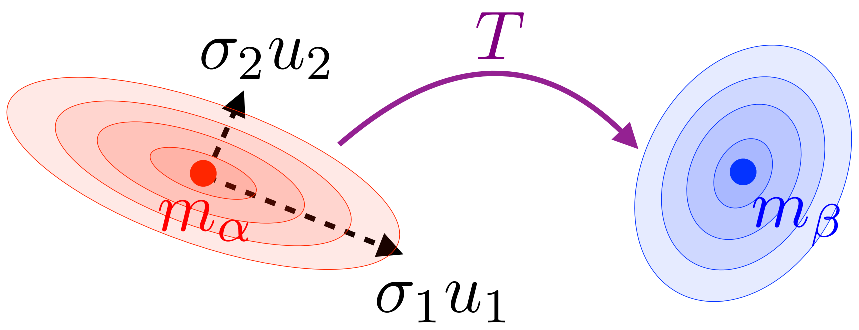
$$\Sigma_\alpha = U_\alpha \operatorname{diag}(\sigma_\alpha) U_\alpha^\top$$



OT Between Gaussians

$$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \frac{1}{(2\pi)^{d/2}|\Sigma_\alpha|} e^{-\frac{\|x-m_\alpha\|^2_{\Sigma_\alpha^{-1}}}{2}}$$

$$\Sigma_\alpha = U_\alpha \,\mathrm{diag}(\sigma_\alpha) U_\alpha^\top$$



$$\text{Ansatz:} \qquad T(x) = A(x - m_\alpha) + m_\beta$$

Proposition: T is the optimal transport if

$$T = \nabla\varphi \;\Leftrightarrow\; \varphi(x) = \frac{1}{2}\langle A(x-m_\alpha), x - m_\alpha\rangle + \langle x, m_\beta\rangle + \mathrm{cst}$$

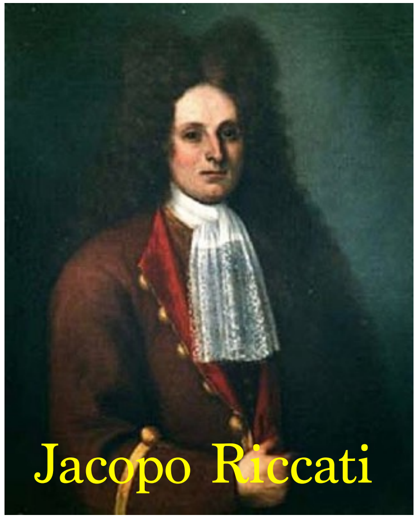
$$\varphi \text{ convex} \;\Leftrightarrow\; A \in \mathcal{S}_+^n$$

$$T_\sharp\alpha = \beta \;\Leftrightarrow\; A\Sigma_\alpha A = \Sigma_\beta$$

Resolution of Algebraic Riccati equations

$$A\Sigma_\alpha A = \Sigma_\beta$$

$$(\Sigma_\alpha^{\frac{1}{2}} A \Sigma_\alpha^{\frac{1}{2}})(\Sigma_\alpha^{\frac{1}{2}} A \Sigma_\alpha^{\frac{1}{2}}) = \Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}$$



Proposition: If $\Sigma \in \mathcal{S}_+$, $\exists! \sqrt{\Sigma} \in \mathcal{S}_+$ s.t. $(\sqrt{\Sigma})^2 = \Sigma$.

Proof: eigen-decomposition $\Sigma = U \operatorname{diag}(\sigma_i) U^\top$, take $\sqrt{\Sigma} = U \operatorname{diag}(\sqrt{\sigma_i}) U^\top$.

Uniqueness: one has $\sqrt{\Sigma}\Sigma = \sqrt{\Sigma}^3 = \Sigma\sqrt{\Sigma}$, they co-diagonalize.

Resolution of Algebraic Riccati equations

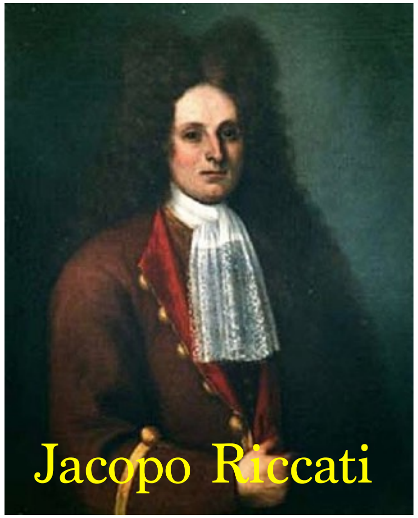
$$A\Sigma_\alpha A = \Sigma_\beta$$

$$(\Sigma_\alpha^{\frac{1}{2}} A \Sigma_\alpha^{\frac{1}{2}})(\Sigma_\alpha^{\frac{1}{2}} A \Sigma_\alpha^{\frac{1}{2}}) = \Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}$$

$$(\Sigma_\alpha^{\frac{1}{2}} A \Sigma_\alpha^{\frac{1}{2}})^2 = \Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}$$

$$\Sigma_\alpha^{\frac{1}{2}} A \Sigma_\alpha^{\frac{1}{2}} = \sqrt{\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}}$$

$$A = \Sigma_\alpha^{-\frac{1}{2}} \sqrt{\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}} \Sigma_\alpha^{-\frac{1}{2}}$$



Proposition: If $\Sigma \in \mathcal{S}_+$, $\exists! \sqrt{\Sigma} \in \mathcal{S}_+$ s.t. $(\sqrt{\Sigma})^2 = \Sigma$.

Proof: eigen-decomposition $\Sigma = U \operatorname{diag}(\sigma_i) U^\top$, take $\sqrt{\Sigma} = U \operatorname{diag}(\sqrt{\sigma_i}) U^\top$.

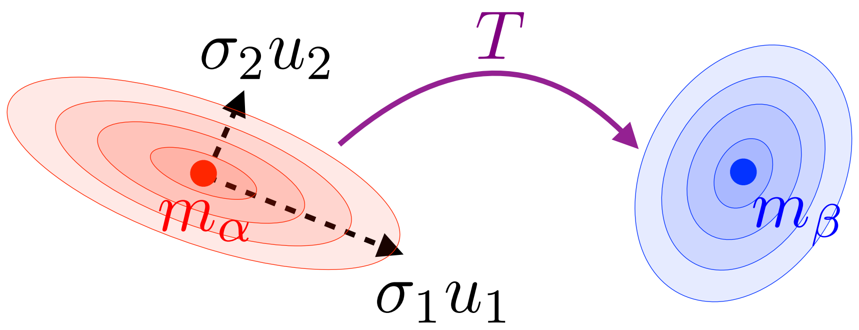
Uniqueness: one has $\sqrt{\Sigma}\Sigma = \sqrt{\Sigma}^3 = \Sigma\sqrt{\Sigma}$, they co-diagonalize.

OT Between Gaussians: Bures Distance

Optimal transport:

$$T(x) = A(x - m_\alpha) + m_\beta$$

$$A = \Sigma_\alpha^{-\frac{1}{2}} \sqrt{\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}} \Sigma_\alpha^{-\frac{1}{2}}$$



$$\mathrm{W}_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2$$

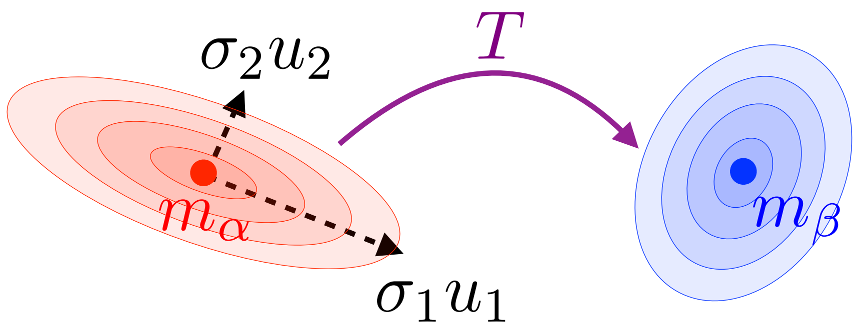
Bures distance: $\mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2 \overset{\text{def.}}{=} \mathrm{tr}\left(\Sigma_\alpha + \Sigma_\beta - 2\sqrt{\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}}\right)$

OT Between Gaussians: Bures Distance

Optimal transport:

$$T(x) = A(x - m_\alpha) + m_\beta$$

$$A = \Sigma_\alpha^{-\frac{1}{2}} \sqrt{\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}} \Sigma_\alpha^{-\frac{1}{2}}$$

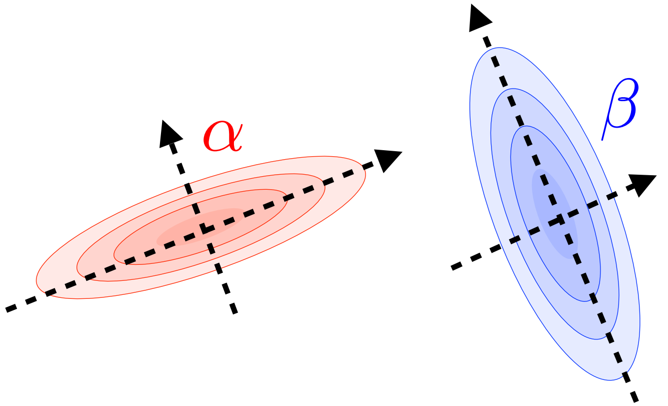


$$\mathrm{W}_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2$$

Bures distance: $\mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2 \stackrel{\text{def.}}{=} \mathrm{tr}\left(\Sigma_\alpha + \Sigma_\beta - 2\sqrt{\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}}\right)$

If $\Sigma_\alpha \Sigma_\beta = \Sigma_\beta \Sigma_\alpha$:

$$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2 = \|\sqrt{\Sigma_\alpha} - \sqrt{\Sigma_\beta}\|^2$$

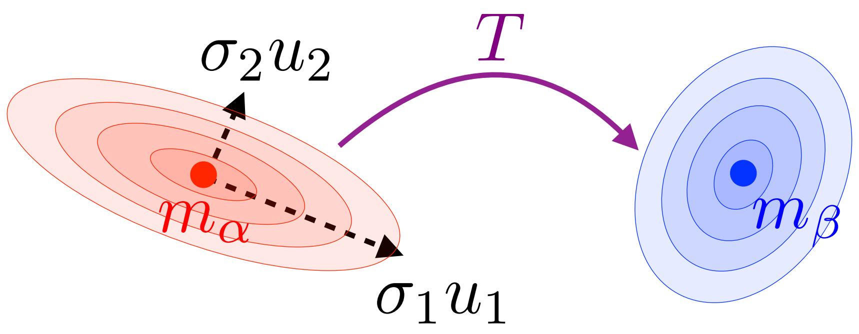


Interpolation Between Gaussians

Optimal transport map $T_\sharp \alpha = \beta$.

$$T(x) = A(x - m_\alpha) + m_\beta$$

$$A = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}}$$



Displacement interpolation: $\alpha_t \overset{\text{def.}}{=} ((1-t)\mathrm{Id} + tT)_\sharp \alpha = \mathcal{N}(m_t, \Sigma_t)$

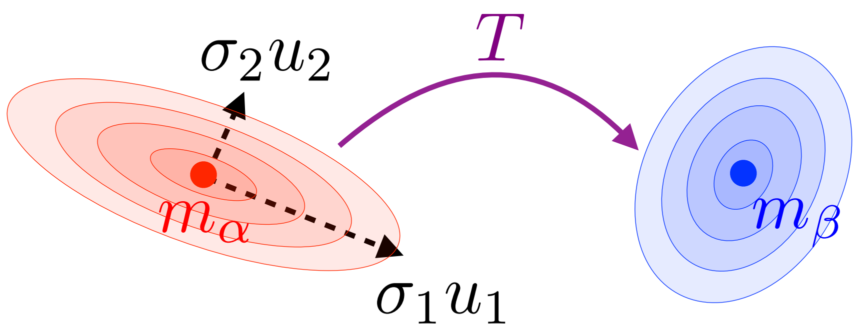
$$m_t = (1-t) m_\alpha + t m_\beta \qquad \Sigma_t = [(1-t)\mathrm{Id} + tA] \Sigma_\alpha [(1-t)\mathrm{Id} + tA]$$

Interpolation Between Gaussians

Optimal transport map $T_\sharp \alpha = \beta$.

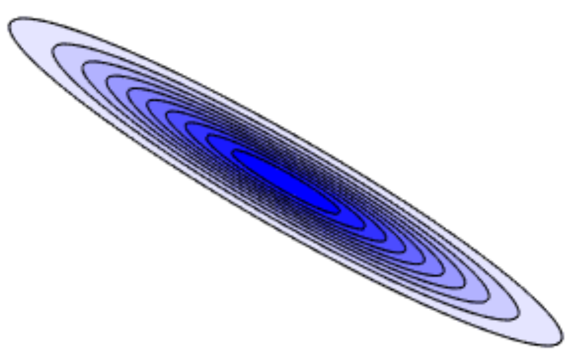
$$T(x) = A(x - m_\alpha) + m_\beta$$

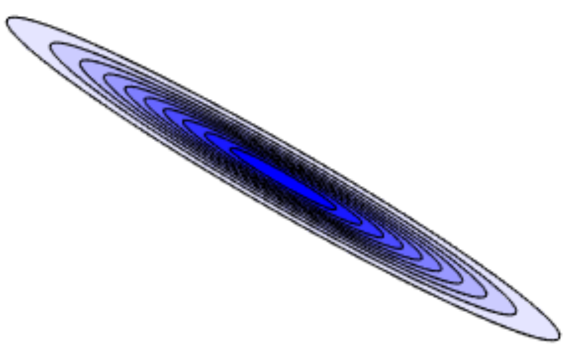
$$A = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}}$$



Displacement interpolation: $\alpha_t \stackrel{\text{def.}}{=} ((1-t)\mathrm{Id} + tT)_\sharp \alpha = \mathcal{N}(m_t, \Sigma_t)$

$$m_t = (1-t) m_\alpha + t m_\beta \qquad \Sigma_t = [(1-t)\mathrm{Id} + tA] \Sigma_\alpha [(1-t)\mathrm{Id} + tA]$$





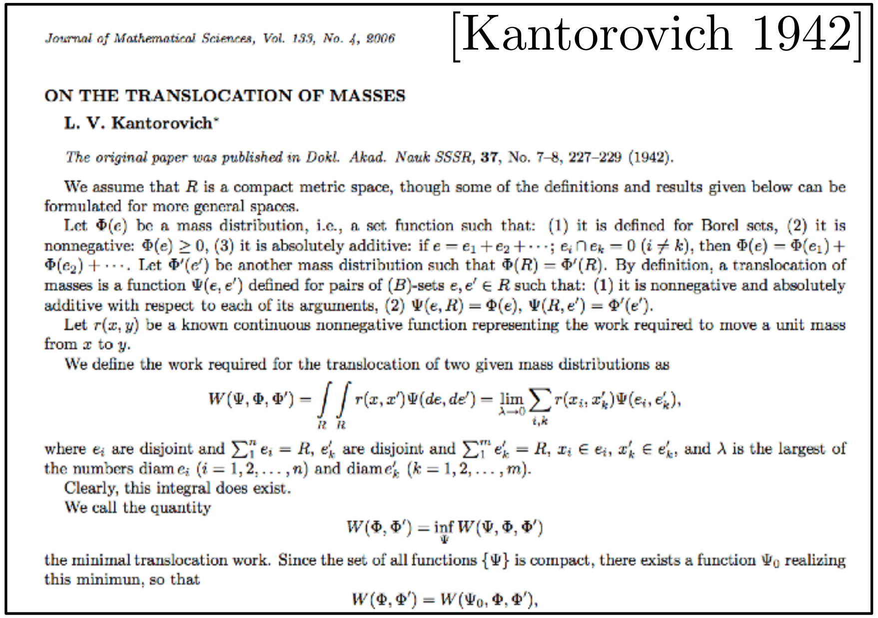
Overview

- Monge Formulation
- Continuous Optimal Transport
- **Kantorovitch Formulation**
- Applications

Leonid Kantorovich (1912-1986)

Леонид Витальевич Канторович

[Kantorovich 1942]



Journal of Mathematical Sciences, Vol. 133, No. 4, 2006

ON THE TRANSLOCATION OF MASSES

L. V. Kantorovich*

The original paper was published in Dokl. Akad. Nauk SSSR, **37**, No. 7–8, 227–229 (1942).

We assume that R is a compact metric space, though some of the definitions and results given below can be formulated for more general spaces.

Let $\Phi(e)$ be a mass distribution, i.e., a set function such that: (1) it is defined for Borel sets, (2) it is nonnegative: $\Phi(e) \geq 0$, (3) it is absolutely additive: if $e = e_1 + e_2 + \cdots$; $e_i \cap e_k = 0$ $(i \neq k)$, then $\Phi(e) = \Phi(e_1) + \Phi(e_2) + \cdots$. Let $\Phi'(e')$ be another mass distribution such that $\Phi(R) = \Phi'(R)$. By definition, a translocation of masses is a function $\Psi(e, e')$ defined for pairs of (B)-sets $e, e' \in R$ such that: (1) it is nonnegative and absolutely additive with respect to each of its arguments, (2) $\Psi(e, R) = \Phi(e)$, $\Psi(R, e') = \Phi'(e')$.

Let $r(x, y)$ be a known continuous nonnegative function representing the work required to move a unit mass from x to y.

We define the work required for the translocation of two given mass distributions as

$$W(\Psi, \Phi, \Phi') = \int_R \int_R r(x, x')\Psi(de, de') = \lim_{\lambda \to 0} \sum_{i,k} r(x_i, x'_k)\Psi(e_i, e'_k),$$

where e_i are disjoint and $\sum_1^n e_i = R$, e'_k are disjoint and $\sum_1^m e'_k = R$, $x_i \in e_i$, $x'_k \in e'_k$, and λ is the largest of the numbers $\operatorname{diam} e_i$ $(i = 1, 2, \ldots, n)$ and $\operatorname{diam} e'_k$ $(k = 1, 2, \ldots, m)$.

Clearly, this integral does exist.

We call the quantity

$$W(\Phi, \Phi') = \inf_\Psi W(\Psi, \Phi, \Phi')$$

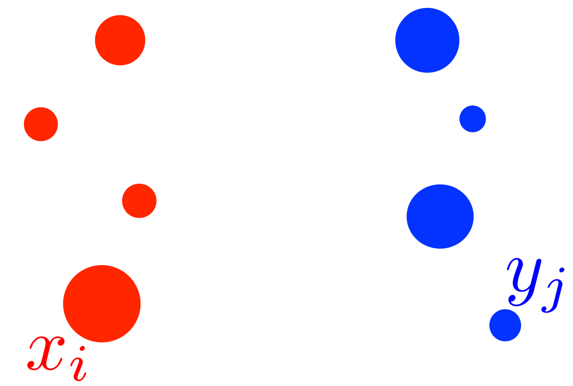
the minimal translocation work. Since the set of all functions $\{\Psi\}$ is compact, there exists a function Ψ_0 realizing this minimun, so that

$$W(\Phi, \Phi') = W(\Psi_0, \Phi, \Phi'),$$

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

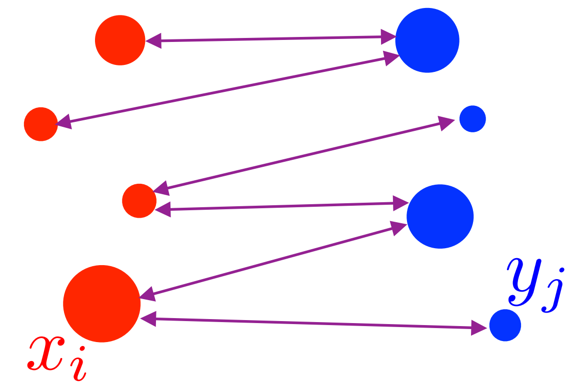
Weights $\mathbf{a}_i \geqslant 0$, $\mathbf{b}_j \geqslant 0$.

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

Kantorovitch's Formulation

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

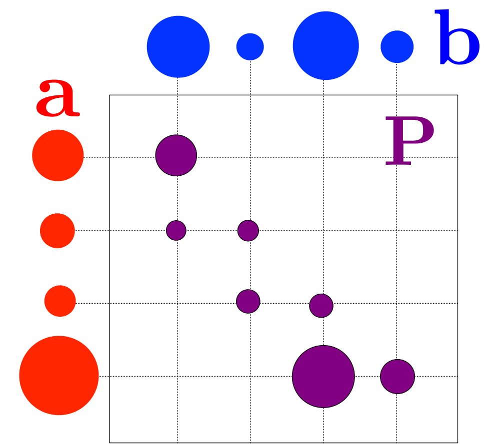
$\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$



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

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

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



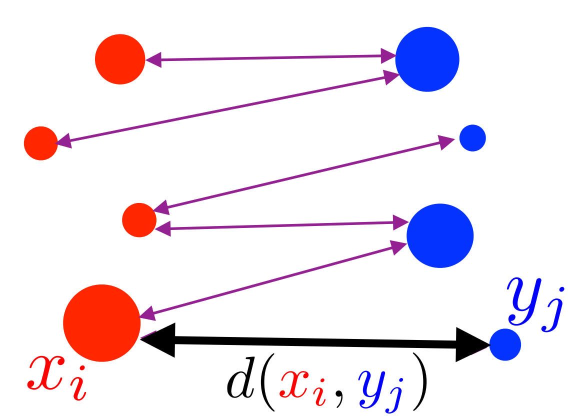
Couplings: $\sum_j \mathbf{P}_{i,j} = \mathbf{a}_i$ $\quad$ $\sum_i \mathbf{P}_{i,j} = \mathbf{b}_j$

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

Kantorovitch's Formulation

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

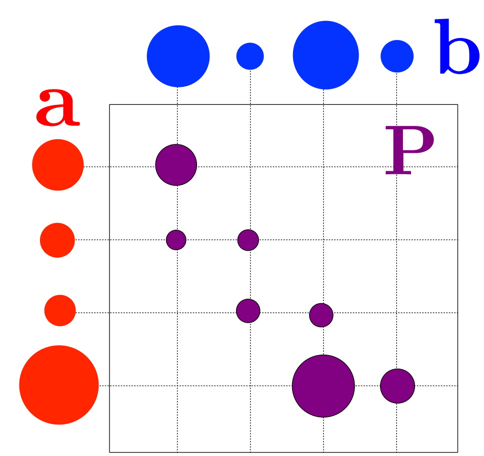
$\beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$



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

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

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



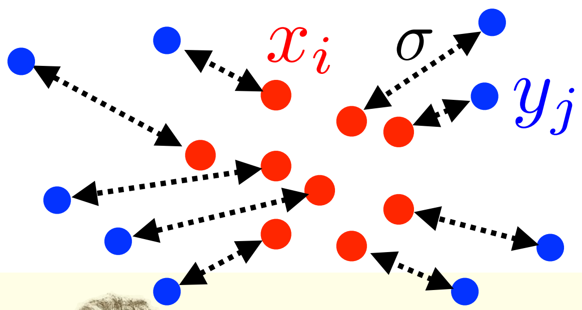
Couplings: $\sum_j \mathbf{P}_{i,j} = \mathbf{a}_i$ $\quad \sum_i \mathbf{P}_{i,j} = \mathbf{b}_j$

[Kantorovich 1942]

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

Kantorovitch's Exact Relaxation

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

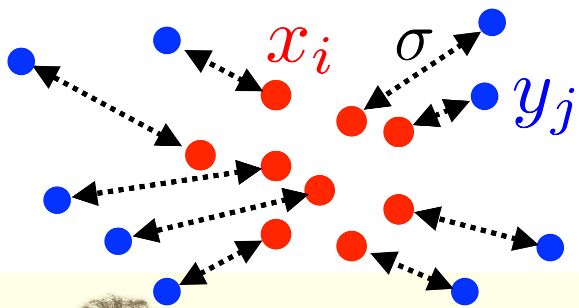


Monge (1784):

$$\min_{\sigma \in \mathrm{Perm}_n} \sum_{i=1}^{n} \mathbf{C}_{i,\sigma(i)}$$

Kantorovitch's Exact Relaxation

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



Permutations "$\subset$" Bi-stochastic matrices:

$$\text{Bist}_n \overset{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times n} \; ; \; \mathbf{P}\mathbb{1} = \mathbb{1}, \mathbf{P}^\top \mathbb{1} = \mathbb{1} \right\}$$

Monge (1784):

$$\min_{\sigma \in \text{Perm}_n} \sum_{i=1}^n \mathbf{C}_{i,\sigma(i)}$$

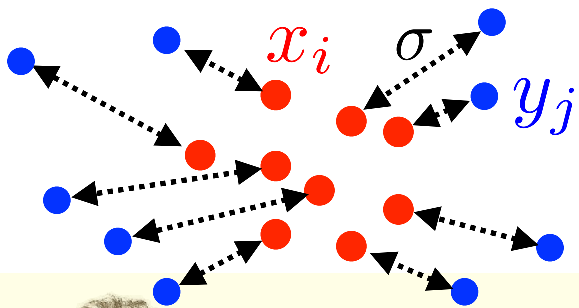
$\geqslant$ (relaxation)

Kantorovitch (1942):

$$\min_{\mathbf{P} \in \text{Bist}_n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{P}_{i,j} \mathbf{C}_{i,j}$$

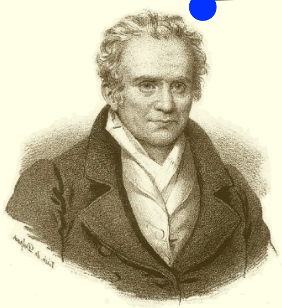
Kantorovitch's Exact Relaxation

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



Permutations "$\subset$" Bi-stochastic matrices:

$$\mathrm{Bist}_n \overset{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times n} \;;\; \mathbf{P}\mathbb{1} = \mathbb{1}, \mathbf{P}^\top \mathbb{1} = \mathbb{1} \right\}$$



Monge (1784):

$$\min_{\sigma \in \mathrm{Perm}_n} \sum_{i=1}^n \mathbf{C}_{i,\sigma(i)}$$

$\geqslant$ (relaxation)

Kantorovitch (1942):

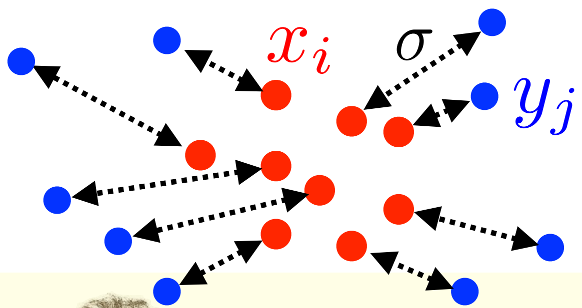
$$\min_{\mathbf{P} \in \mathrm{Bist}_n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{P}_{i,j} \mathbf{C}_{i,j}$$

$n!$ permutations

$O(n^3)$ algorithm

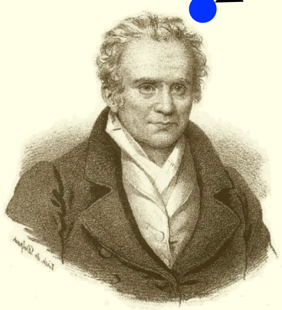
Kantorovitch's Exact Relaxation

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



Permutations "$\subset$" Bi-stochastic matrices:

$$\mathrm{Bist}_n \overset{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times n} \; ; \; \mathbf{P}\mathbb{1} = \mathbb{1}, \mathbf{P}^\top \mathbb{1} = \mathbb{1} \right\}$$



Monge (1784):

$$\min_{\sigma \in \mathrm{Perm}_n} \sum_{i=1}^n \mathbf{C}_{i,\sigma(i)}$$

$\geqslant$ (relaxation)

Kantorovitch (1942):

$$\min_{\mathbf{P} \in \mathrm{Bist}_n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{P}_{i,j} \mathbf{C}_{i,j}$$

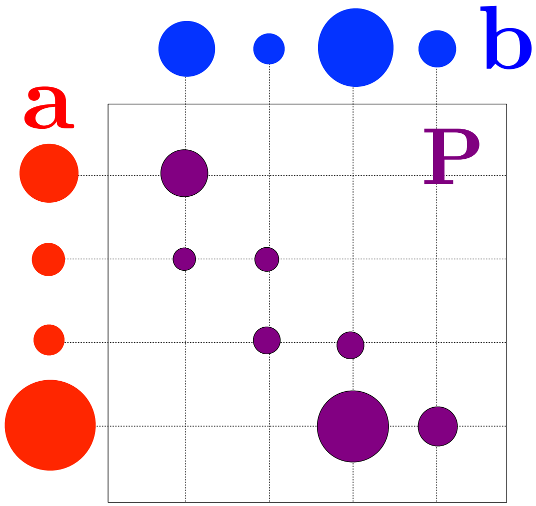
$n!$ permutations

$O(n^3)$ algorithm

Theorem: [Birkhoff-von Neumann]

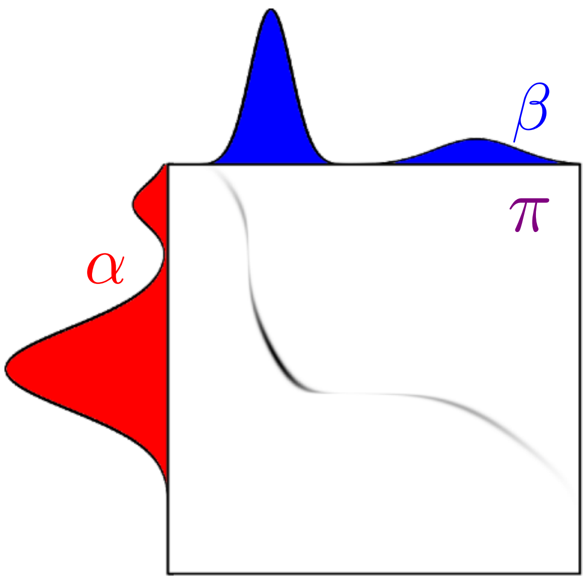
"Monge $\Leftrightarrow$ Kantorovitch"

General Formulation



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

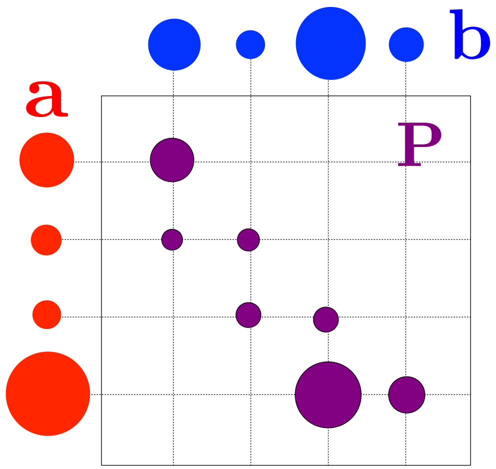
$$c(x,y) = d(x,y)^p$$



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

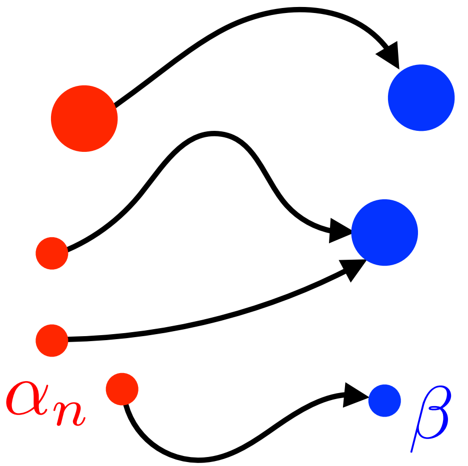
Optimal Transport Distances

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



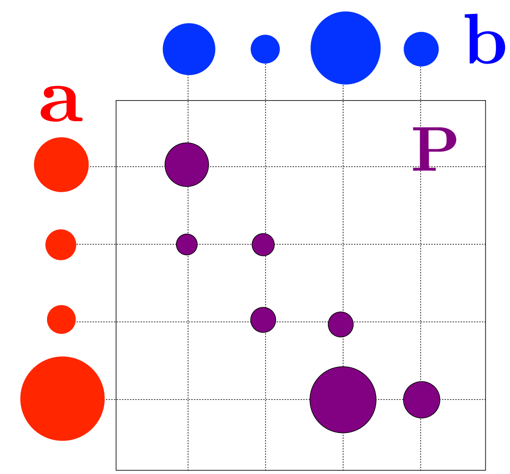
Convergence in law: $\alpha_n \rightharpoonup^* \beta$

$$\Leftrightarrow \; \forall f \in \mathcal{C}_c(\mathcal{X}),\; \int_{\mathcal{X}} f \mathrm{d}\alpha_n \to \int_{\mathcal{X}} f \mathrm{d}\beta$$



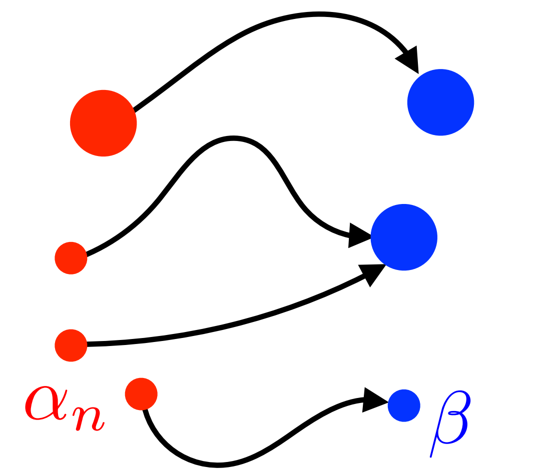
Optimal Transport Distances

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



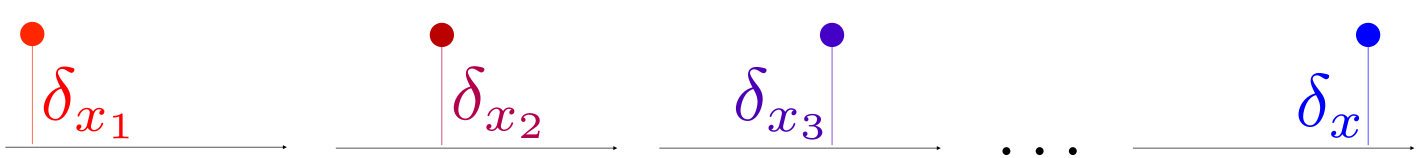
Convergence in law: $\alpha_n \rightharpoonup^* \beta$

$$\Leftrightarrow \quad \forall f \in \mathcal{C}_c(\mathcal{X}), \ \int_{\mathcal{X}} f \mathrm{d}\alpha_n \to \int_{\mathcal{X}} f \mathrm{d}\beta$$



Theorem: W_p is a distance and

$$\left.\begin{array}{c} \alpha_n \rightharpoonup^* \beta \\ \int \|\cdot\|^p \mathrm{d}\alpha_n \to \int \|\cdot\|^p \mathrm{d}\beta \end{array}\right. \quad \Leftrightarrow \quad \mathrm{W}_p(\alpha_n, \beta) \to 0$$



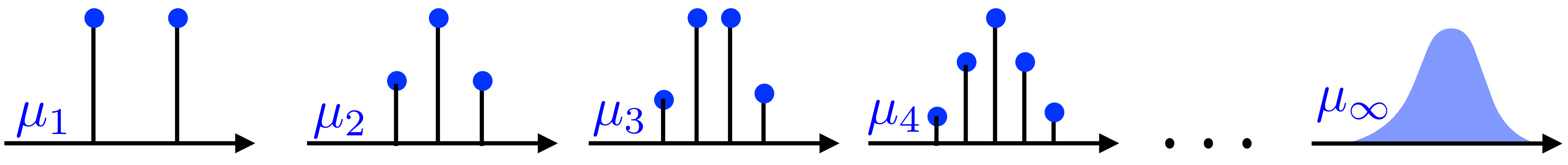
$$\|\delta_{x_n} - \delta_x\|_{\mathrm{TV}} = 2 \quad \textbf{vs.} \quad \mathrm{W}_p(\delta_{x_n}, \delta_x) = d(x_n, x)$$

Application: Central Limit Theorem

Central limit theorem: If $\mathbb{E}(X) = 0, \mathbb{E}(X^2) = 1$ and $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$

$$Y_n \overset{\text{def.}}{=} \frac{X_1 + \ldots + X_n}{\sqrt{n}} \xrightarrow{\text{law}} \mathcal{N}(0,1)$$

$$\mu_n = \text{Law}(Y_n) = (\mu_1 \star \ldots \star \mu_1)(\cdot/\sqrt{n}) \;\; \rightharpoonup^* \mu_\infty = \text{Law}(\mathcal{N}(0,1))$$

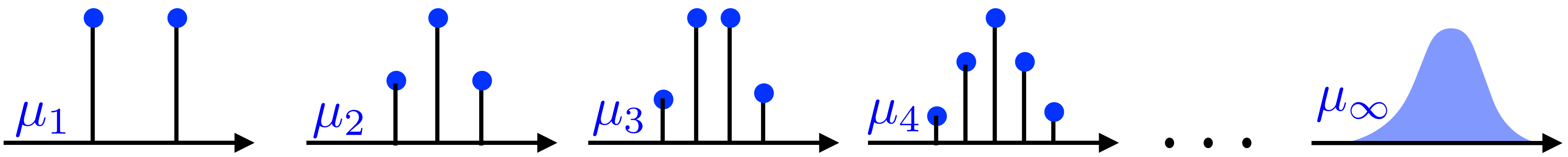


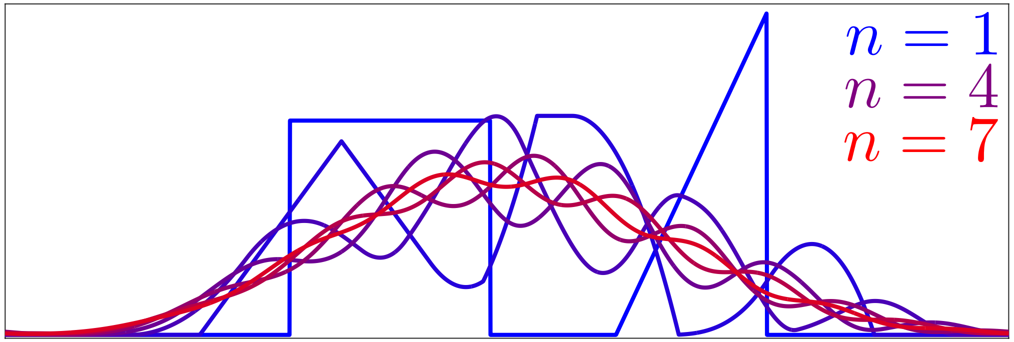
Application: Central Limit Theorem

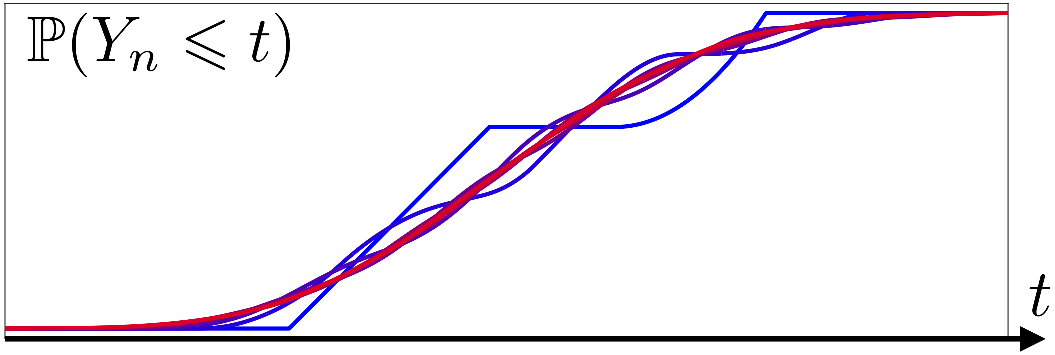
Central limit theorem: If $\mathbb{E}(X) = 0, \mathbb{E}(X^2) = 1$ and $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$

$$Y_n \overset{\text{def.}}{=} \frac{X_1 + \ldots + X_n}{\sqrt{n}} \xrightarrow{\text{law}} \mathcal{N}(0,1)$$

$$\mu_n = \text{Law}(Y_n) = (\mu_1 \star \ldots \star \mu_1)(\cdot/\sqrt{n}) \rightharpoonup^* \mu_\infty = \text{Law}(\mathcal{N}(0,1))$$





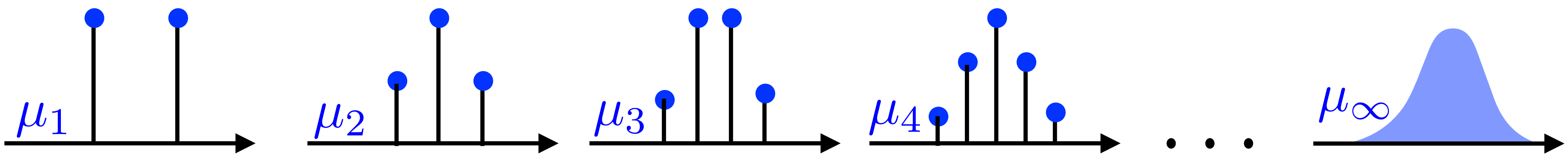


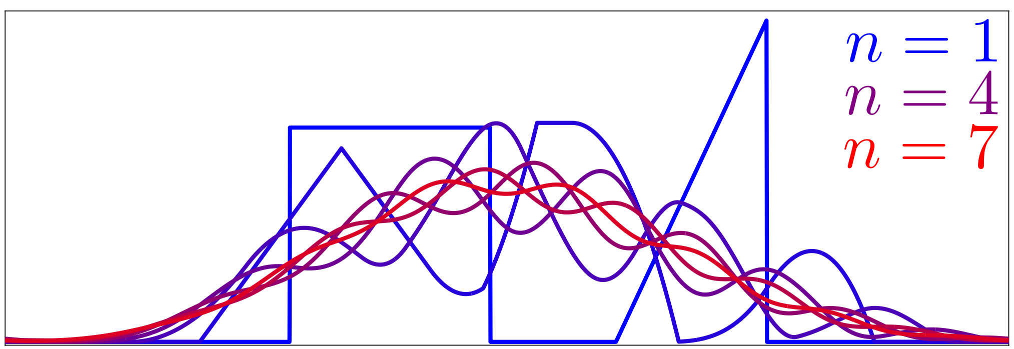
Application: Central Limit Theorem

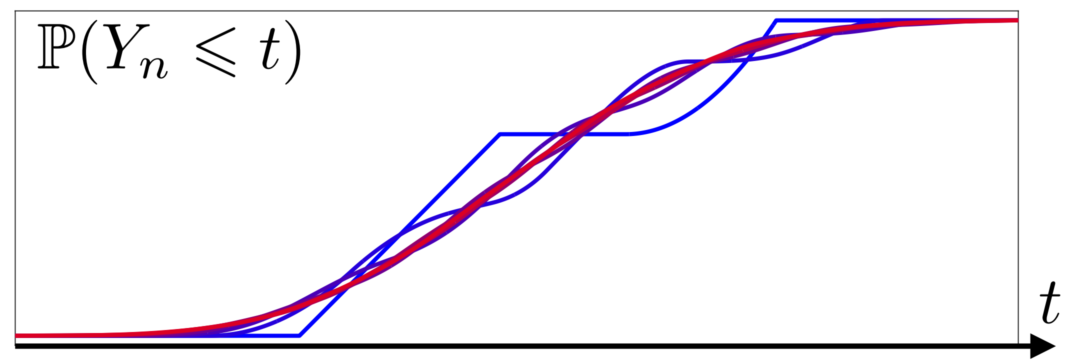
Central limit theorem: If $\mathbb{E}(X) = 0, \mathbb{E}(X^2) = 1$ and $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$

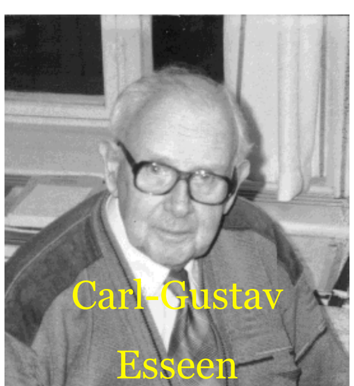
$$Y_n \overset{\text{def.}}{=} \frac{X_1 + \ldots + X_n}{\sqrt{n}} \xrightarrow{\text{law}} \mathcal{N}(0,1)$$

$$\mu_n = \mathrm{Law}(Y_n) = (\mu_1 \star \ldots \star \mu_1)(\cdot/\sqrt{n}) \rightharpoonup^* \mu_\infty = \mathrm{Law}(\mathcal{N}(0,1))$$









Theorem:
[Berry 1941]
[Esseen, 1942]

$$\mathrm{W}_1(\mu_n, \mathcal{N}(0,1)) \leqslant \frac{C\mathbb{E}(|X|^3)}{\sqrt{n}} \qquad C \leqslant 1/2$$

$\longrightarrow$ Generalizes to higher dimensions.

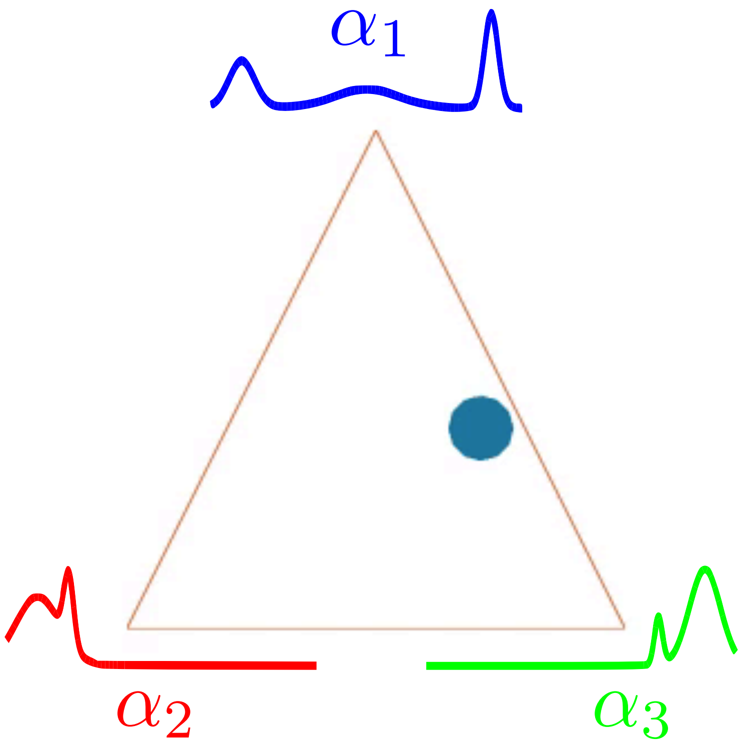
Overview

- Monge Formulation
- Continuous Optimal Transport
- Kantorovitch Formulation
- Applications

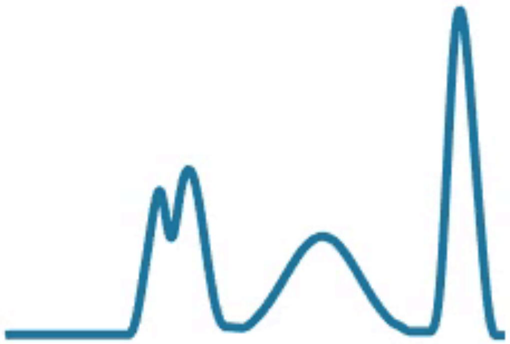
Wasserstein Barycenters

Barycenters of measures $(\alpha_s)_{s=1}^S$: $\quad \sum_s \lambda_s = 1$

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



$\lambda \in \Sigma_3$



$\min_{\alpha} \ \sum_s \lambda_s W_p^p(\alpha, \alpha_s)$

Wasserstein

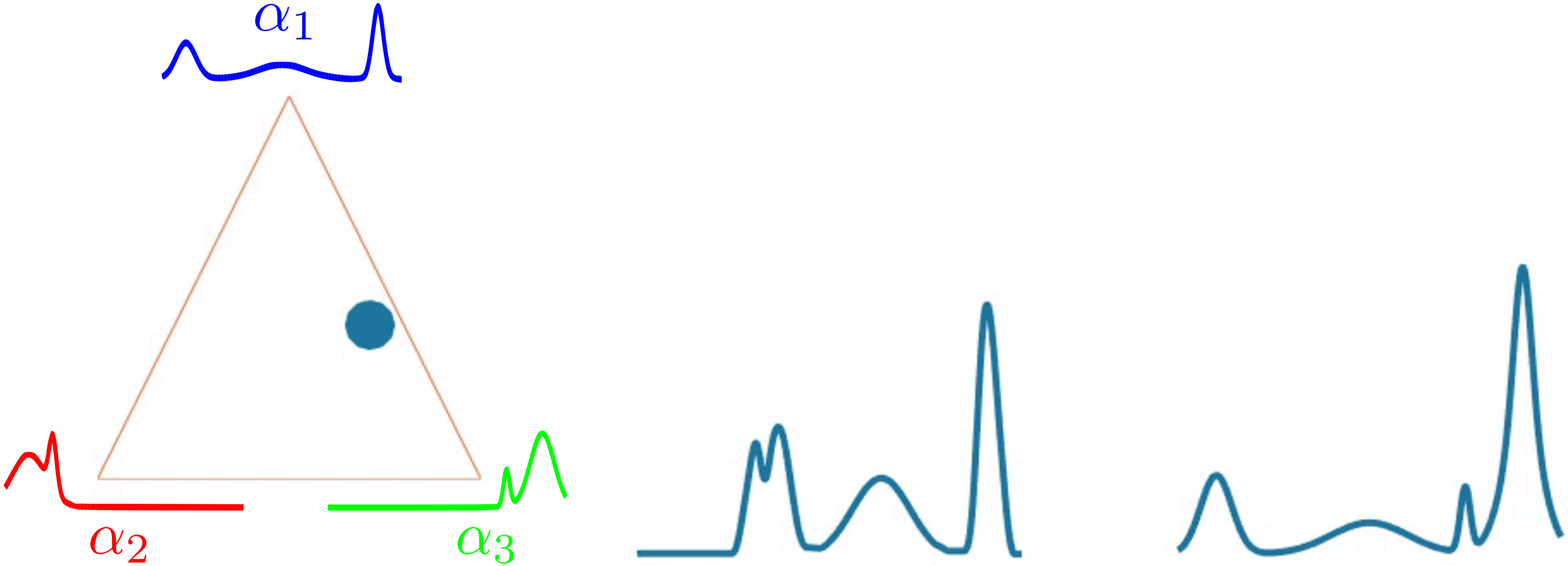
$\sum_s \lambda_s \alpha_s$

Euclidean

Wasserstein Barycenters

Barycenters of measures $(\alpha_s)_{s=1}^{S}$: $\quad \sum_s \lambda_s = 1$

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

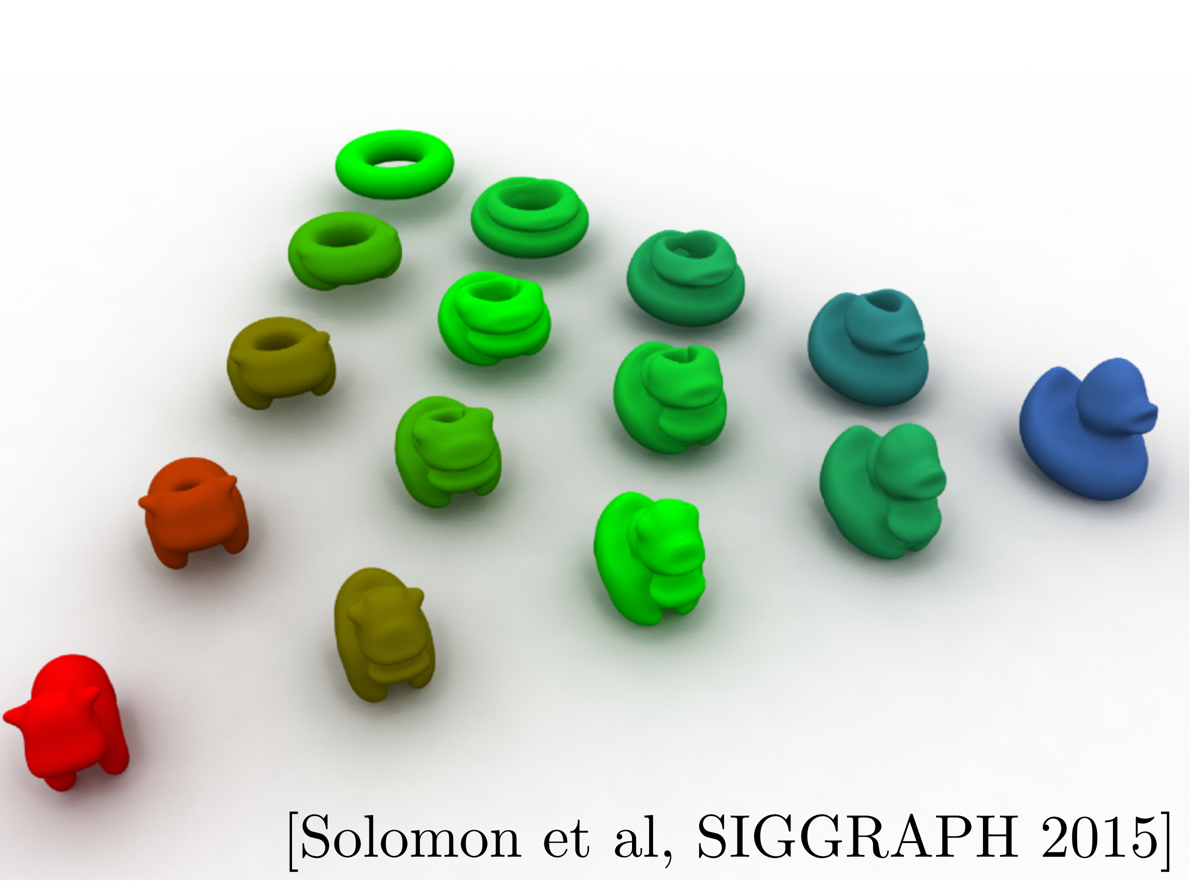


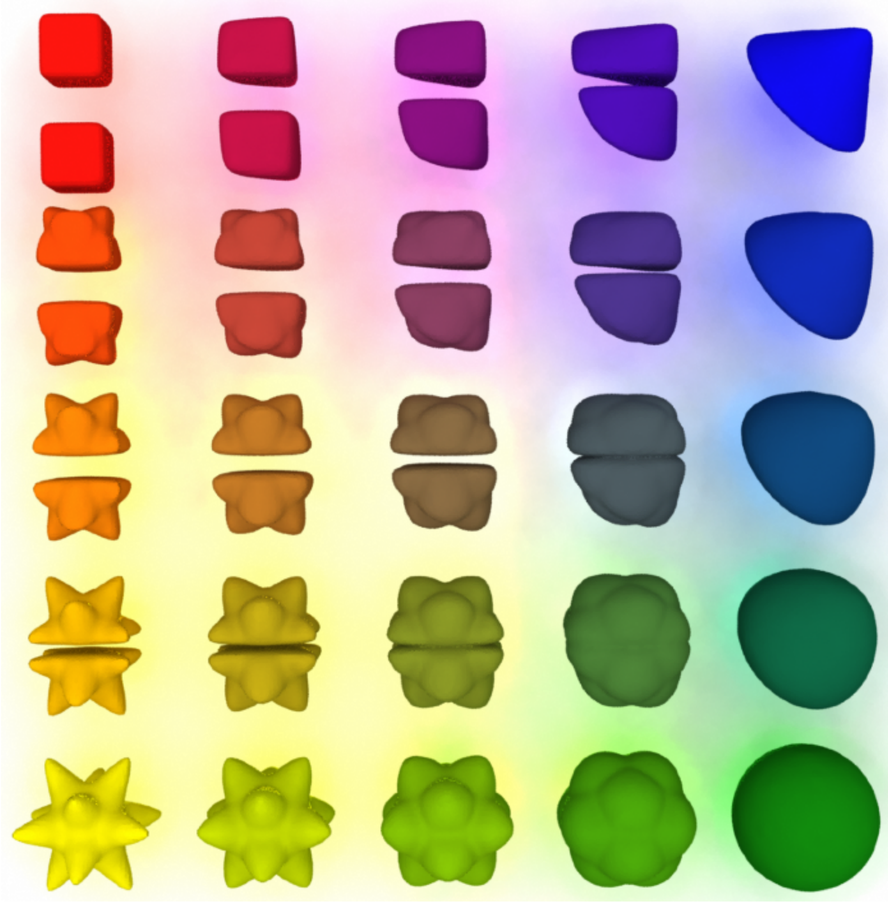
$\lambda \in \Sigma_3$

$\min_\alpha \sum_s \lambda_s W_p^p(\alpha, \alpha_s)$
Wasserstein

$\sum_s \lambda_s \alpha_s$
Euclidean

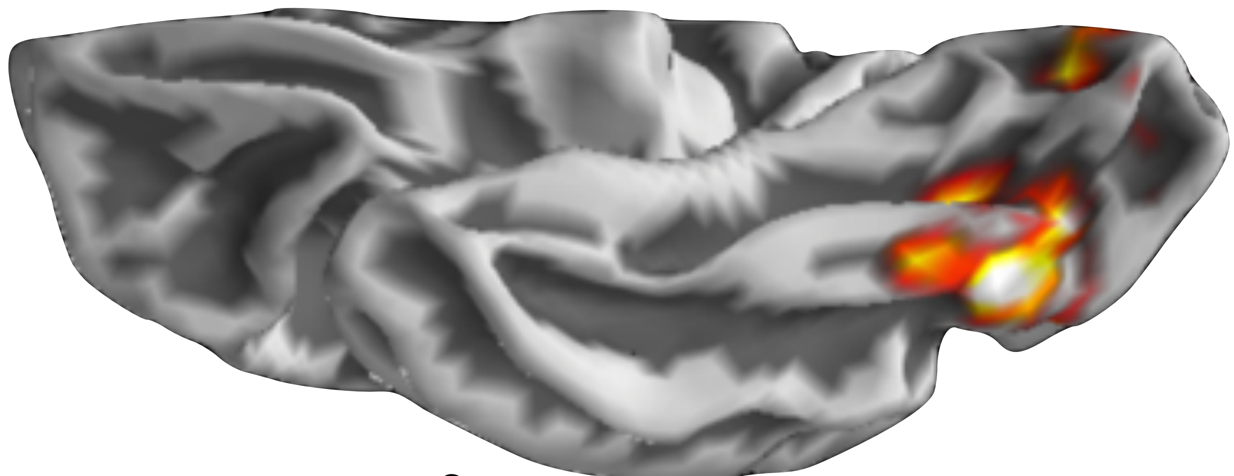
Examples



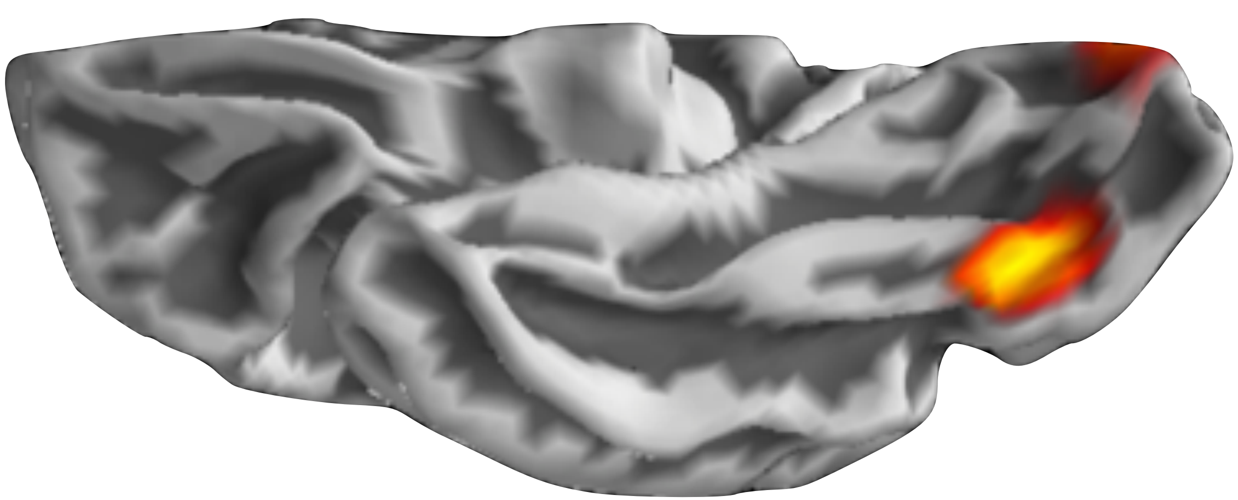


MRI Data Procesing [with A. Gramfort]

Ground cost $c = d_M$: geodesic on cortical surface M.



L^2 barycenter

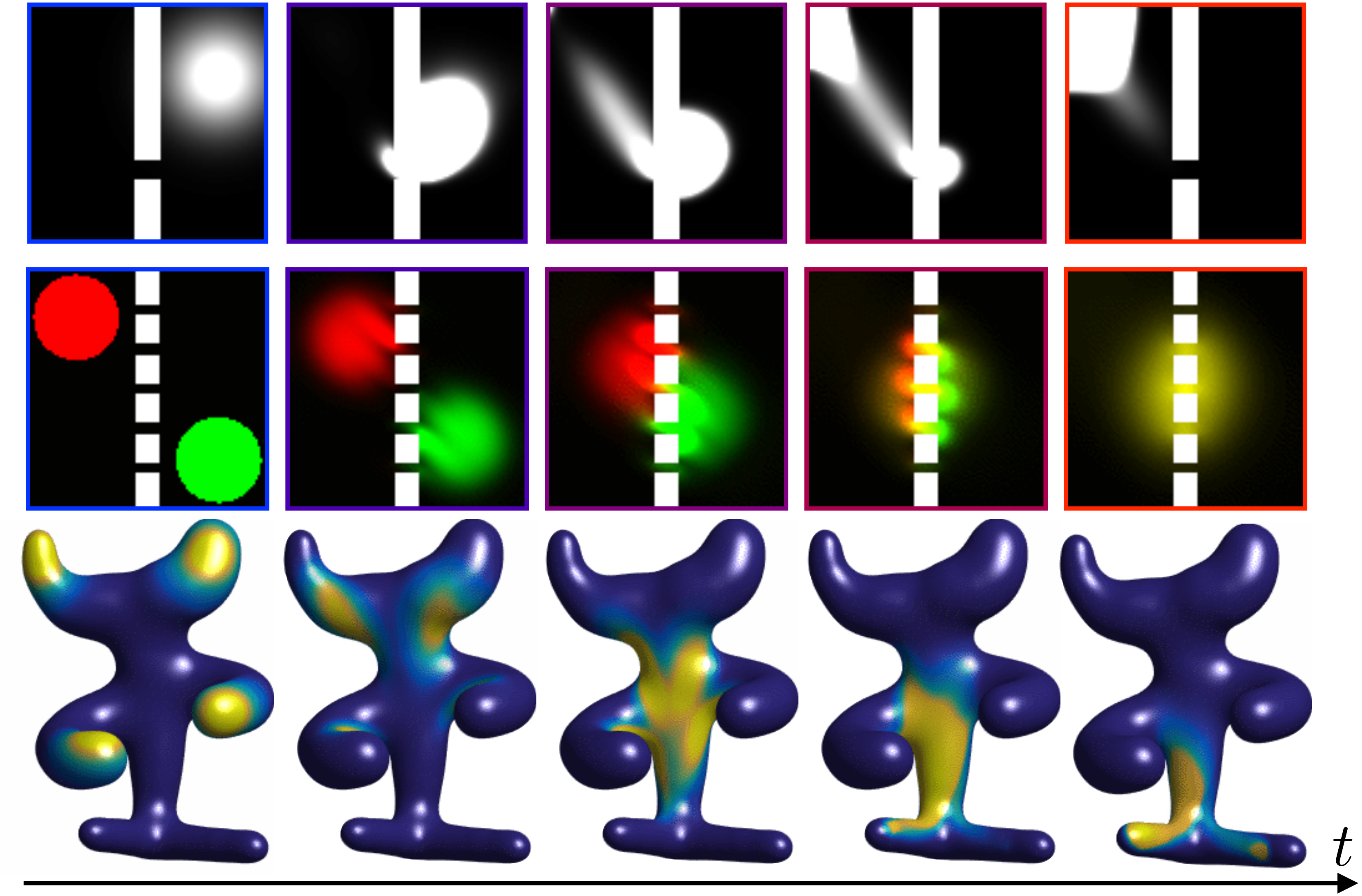


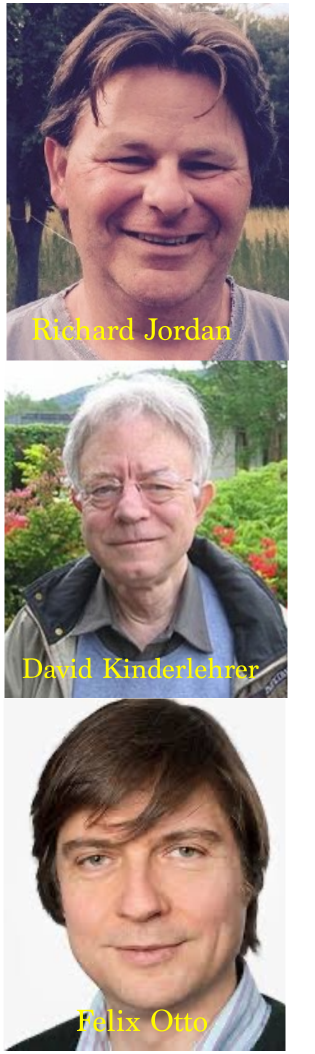
W_2^2 barycenter

Generalizations: Gradient Flows

Implicit stepping: $\alpha_{t+\tau} = \underset{\alpha}{\text{argmin}}\ \mathrm{W}_p^p(\alpha_t, \alpha) + \tau f(\alpha)$

Limit $\tau \to 0$: $\dfrac{\partial \alpha}{\partial t} = \mathrm{div}(\alpha \nabla (f'(\alpha)))$





Also: mean field analysis of 1-hidden layer neural networks.

Gradient Flows Simulation

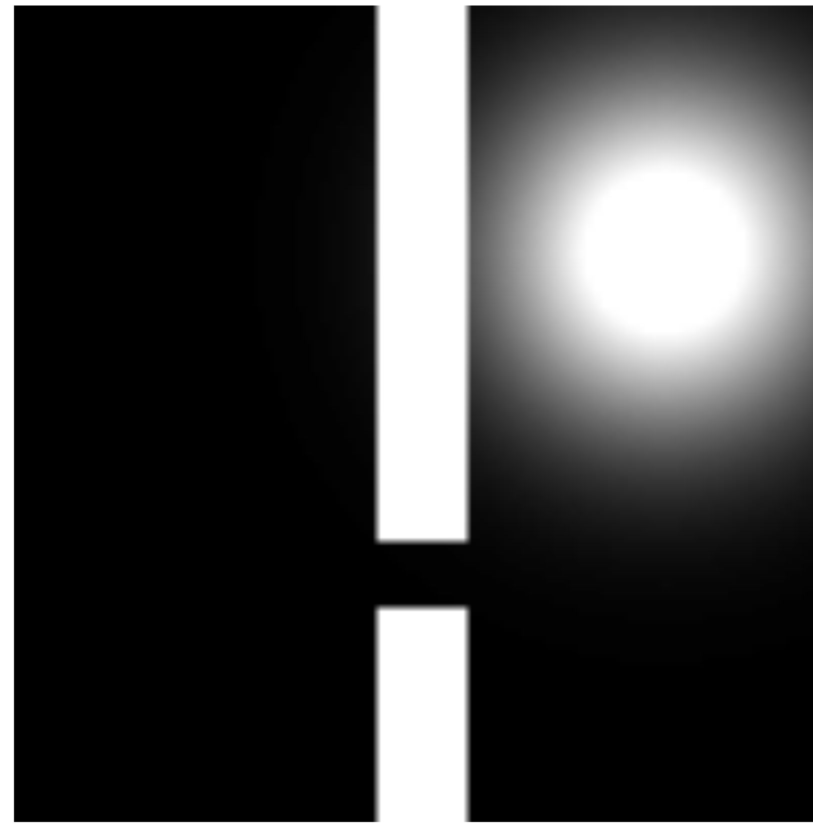
https://www.youtube.com/watch?v=tDQw21ntR64

Tim Whittaker (New Zealand)

Gradient Flows Simulation

https://www.youtube.com/watch?v=tDQw21ntR64

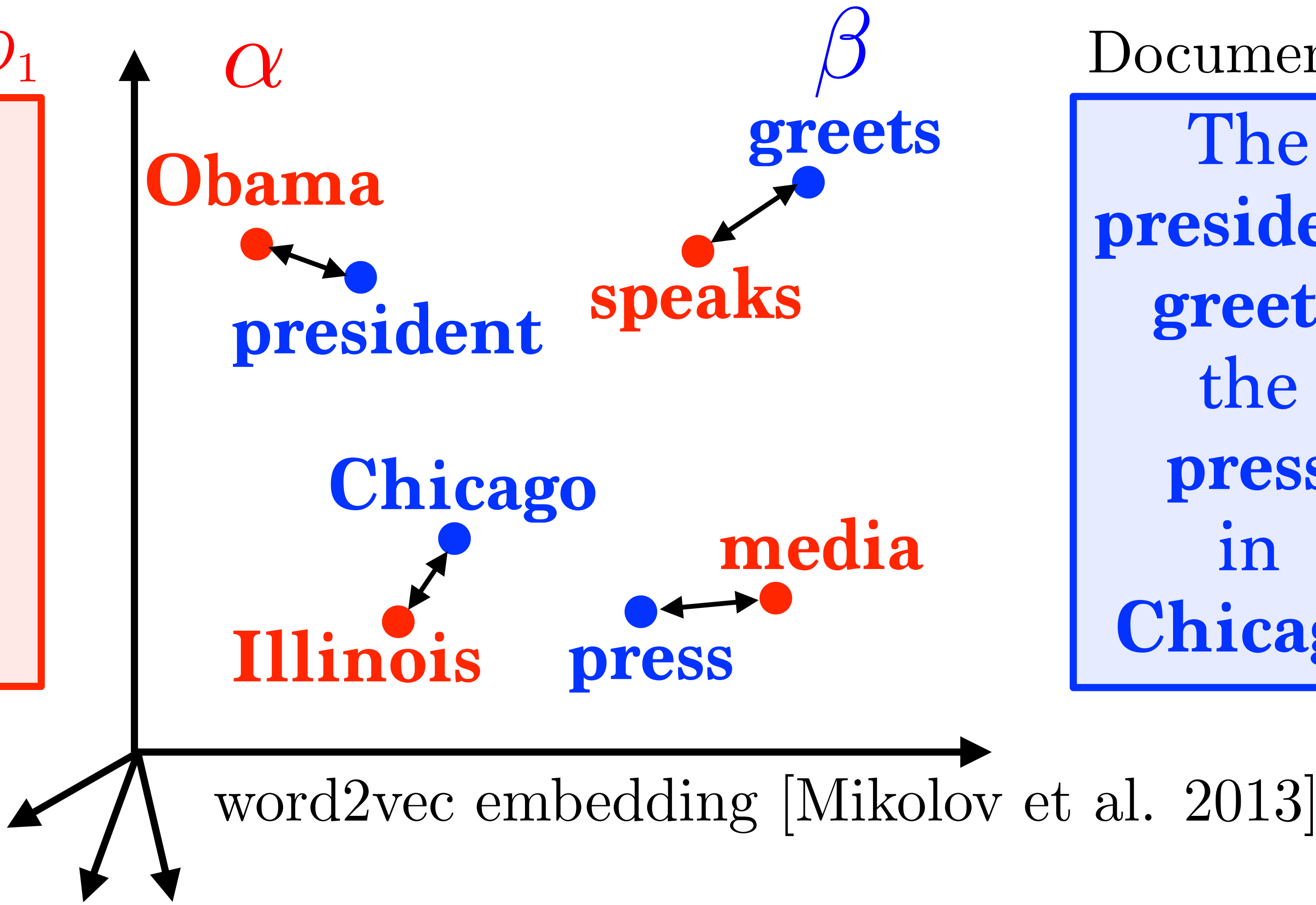
Tim Whittaker (New Zealand)



Bag of Words

Document D_1

Obama speaks to the **media** in **Illinois**



Document D_2

The **president greets** the **press** in **Chicago**

Word mover's distance: [Kusner et al 2015]

$$\mathrm{Dist}(D_1, D_2) = \mathrm{W}_2(\alpha, \beta)$$

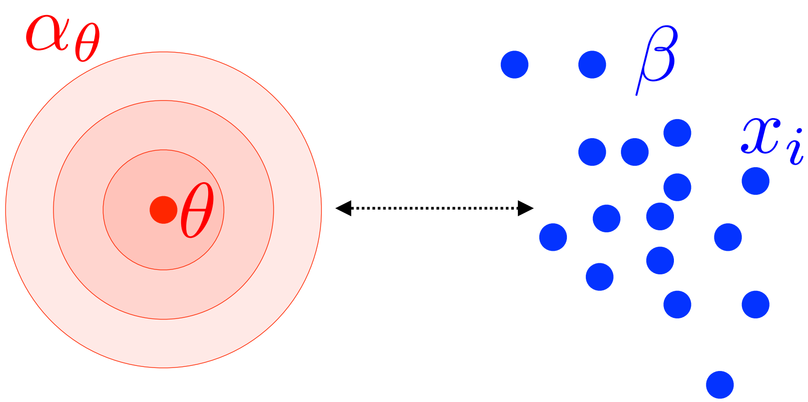
Overview

- Monge Formulation
- Continuous Optimal Transport
- Kantorovitch Formulation
- Applications
- **Generative models**

Density Fitting and Generative Models

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

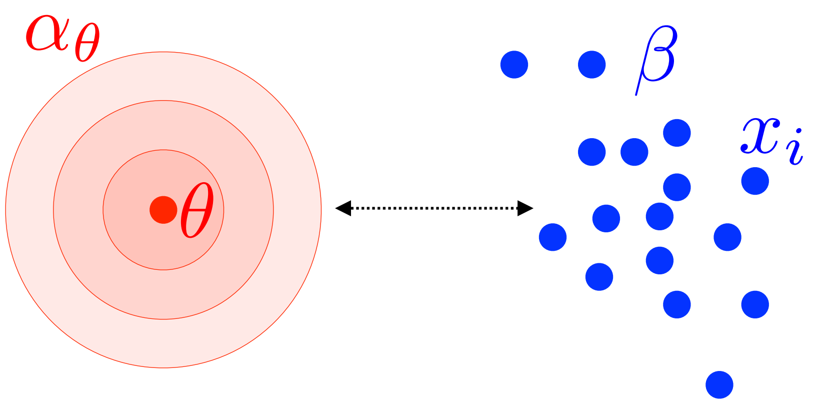
Parametric model: $\theta \mapsto \alpha_\theta$



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

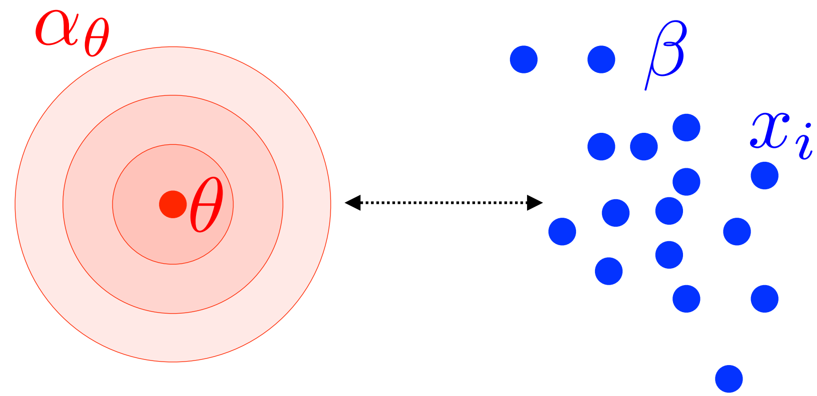
$$\min_\theta \; -\sum_i \log(\rho_\theta(x_i)) \xrightarrow{n \to +\infty} \mathrm{KL}(\beta|\alpha_\theta)$$

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

$$\min_\theta \; -\sum_i \log(\rho_\theta(x_i)) \overset{n\rightarrow+\infty}{\longrightarrow} \mathrm{KL}(\beta|\alpha_\theta)$$

Maximum likelihood (MLE)

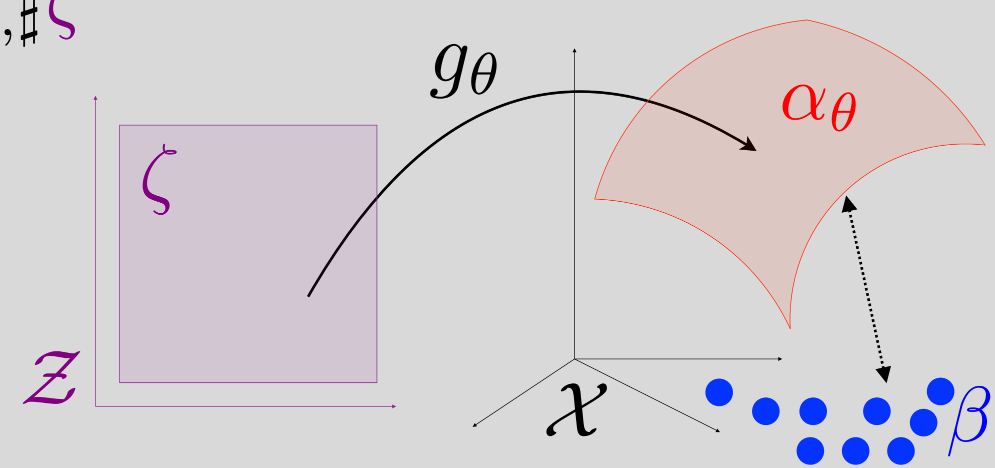
Generative model fit: $\alpha_\theta = g_{\theta,\sharp}\zeta$

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

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.

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

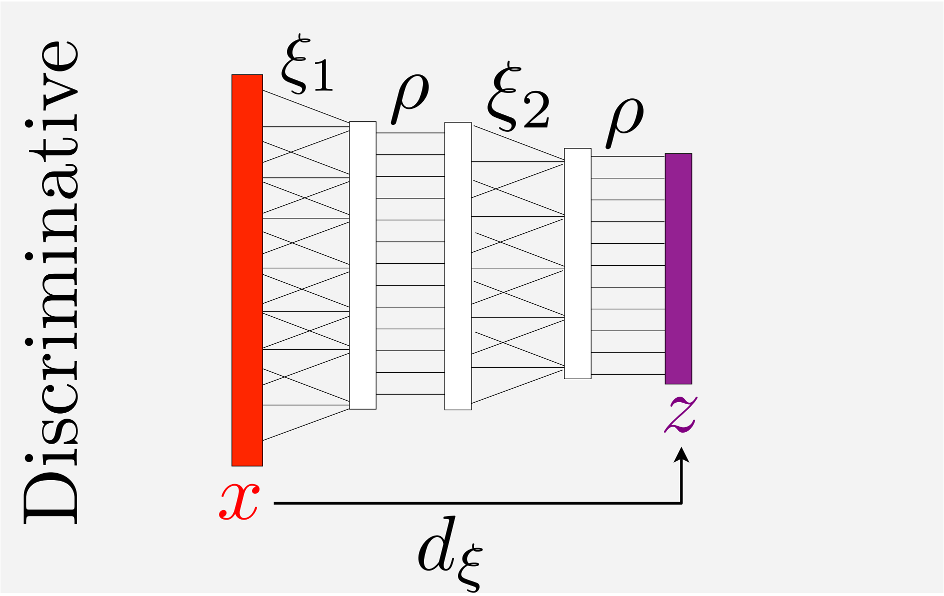


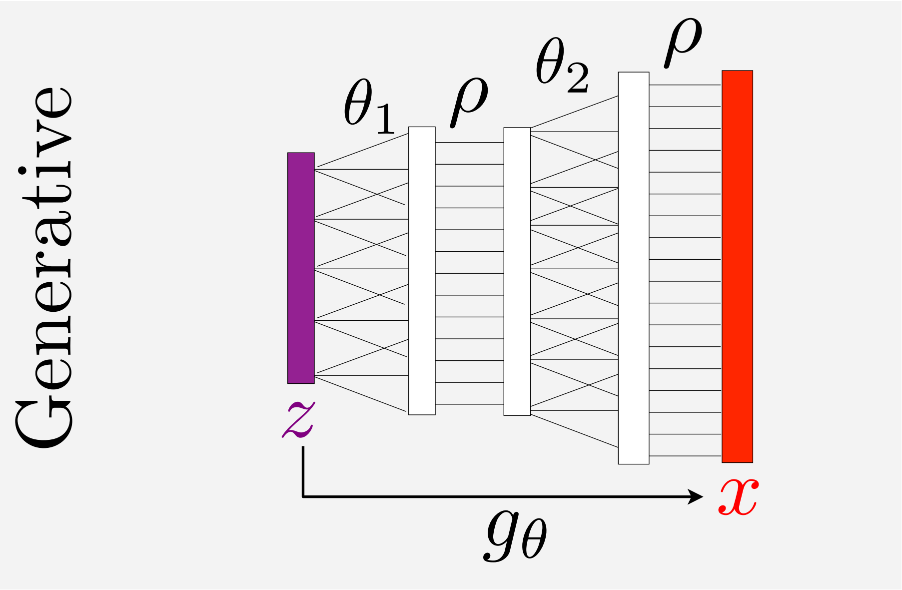
Deep Discriminative vs Generative Models

Deep networks:

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

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



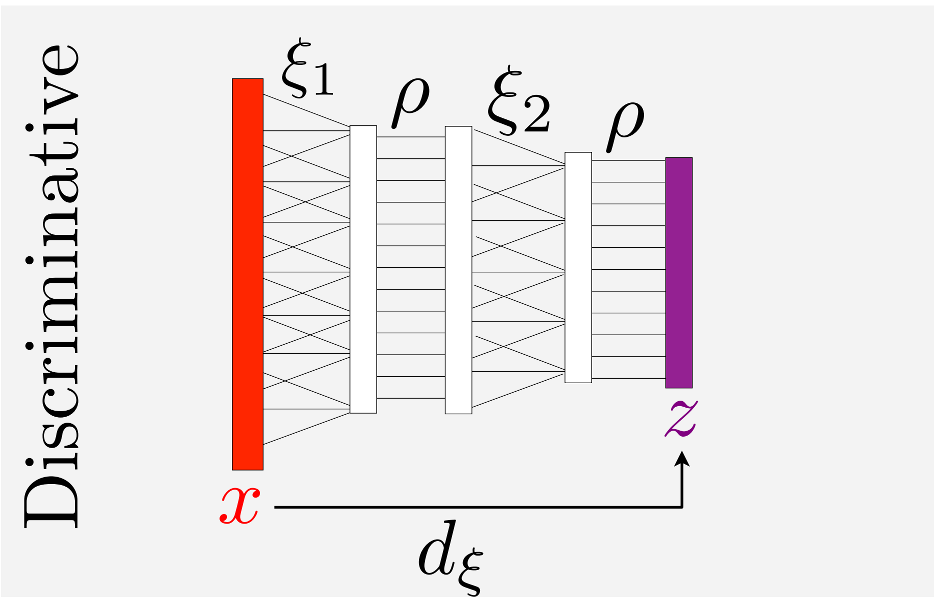


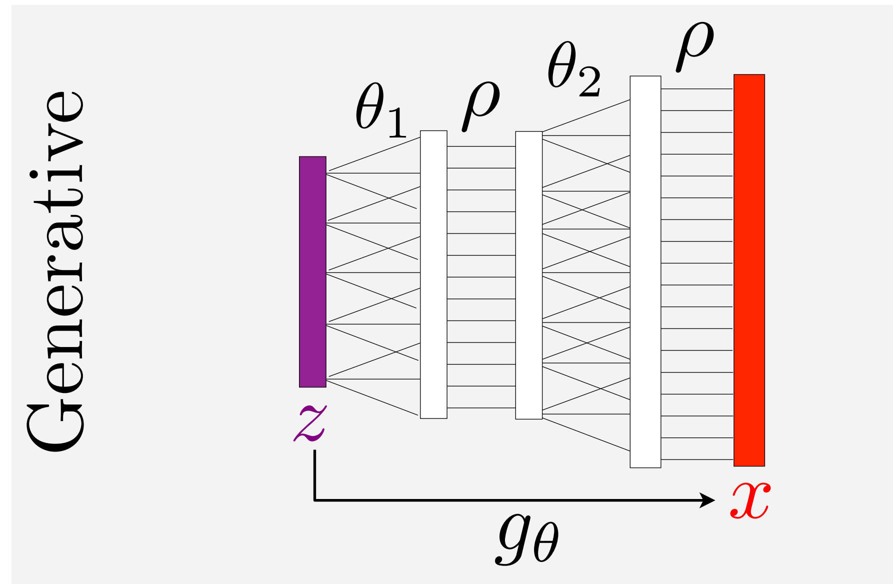
Deep Discriminative vs Generative Models

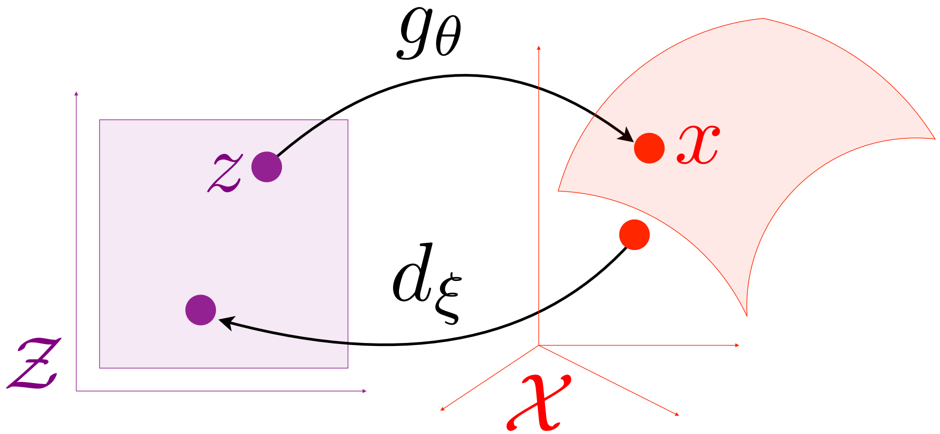
Deep networks:

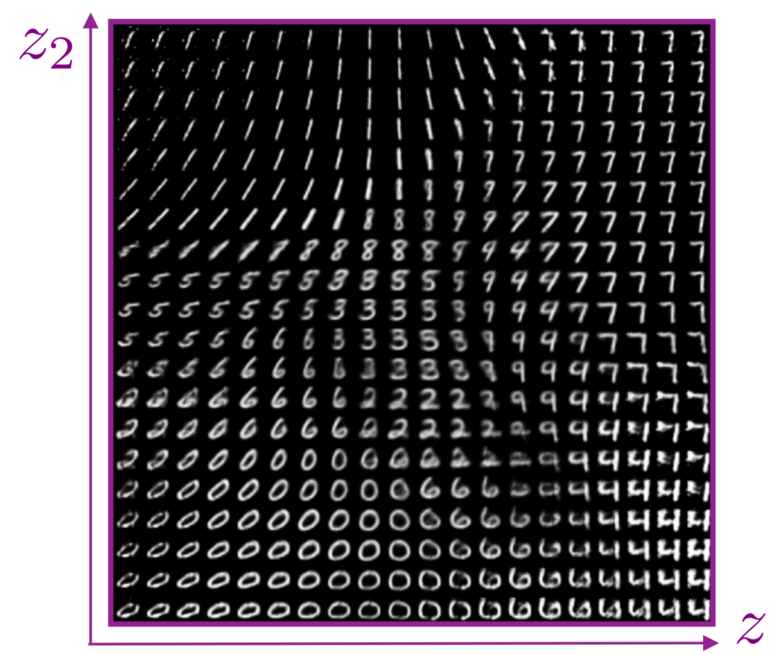
$$d_\xi(x) = \rho(\xi_K(\dots\rho(\xi_2(\rho(\xi_1(x))\dots)$$

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

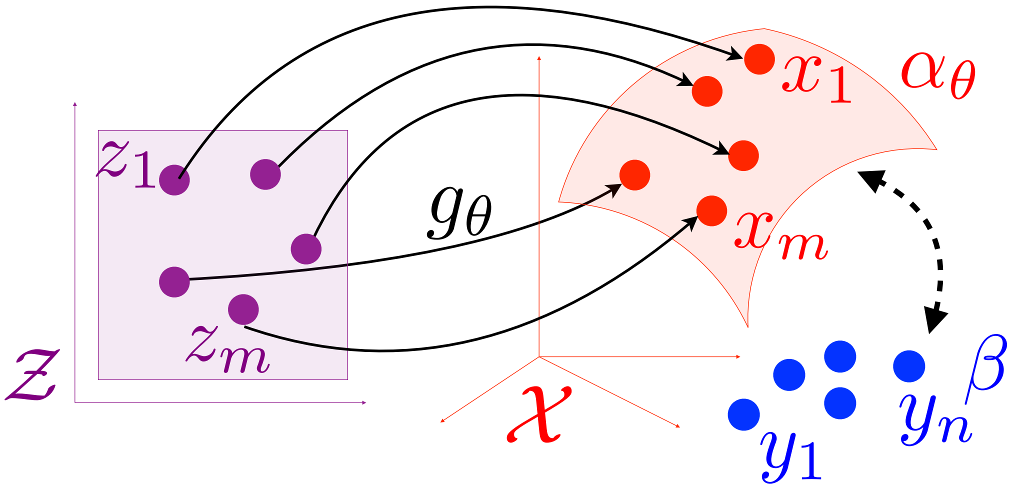








Training Architecture



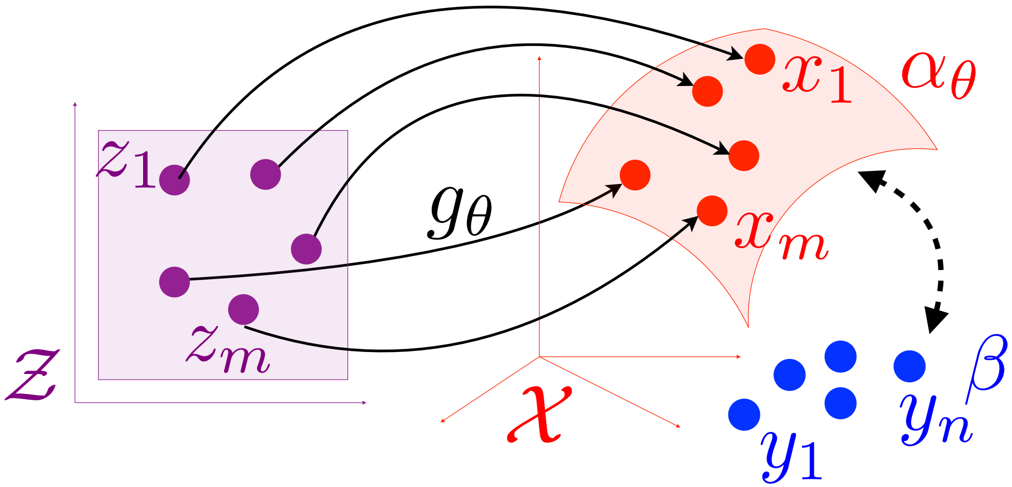
$$\min_\theta \mathcal{E}(\theta) \overset{\text{def.}}{=} \overline{\mathrm{W}}^p_{\varepsilon,p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

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

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

Training Architecture

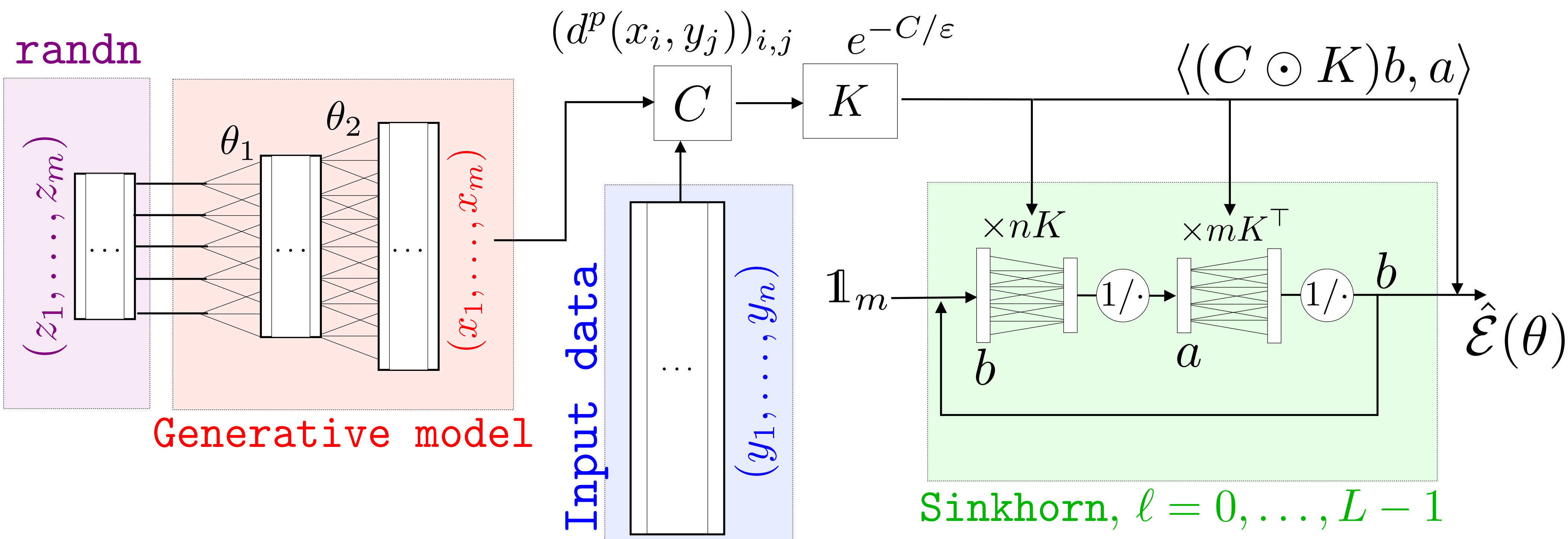


$$\min_\theta \mathcal{E}(\theta) \stackrel{\text{def.}}{=} \overline{\mathrm{W}}_{\varepsilon,p}^{p}(\alpha_\theta, \beta)$$

Stochastic gradient descent

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

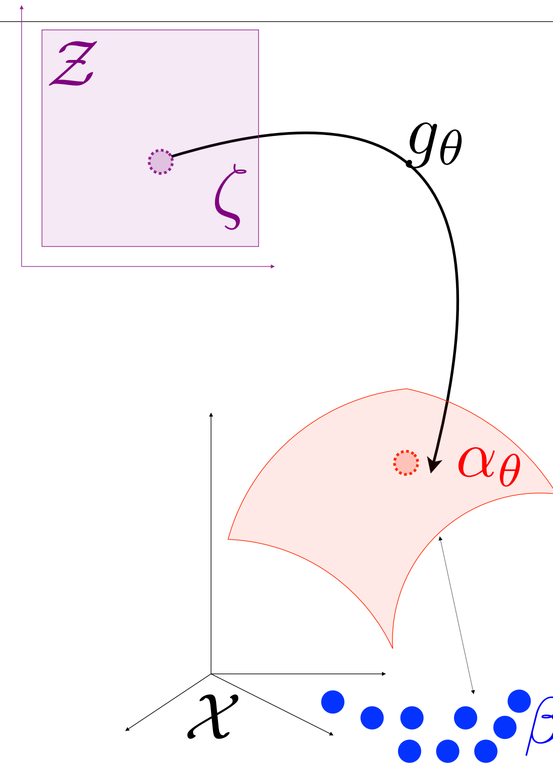
$$\hat{\mathcal{E}}(\theta) \stackrel{\text{def.}}{=} \overline{\mathrm{W}}_{\varepsilon,p}^{p}\left(\tfrac{1}{m}\sum_i \delta_{g_\theta(z_i)}, \beta\right)$$



Examples of Images Generation

Inputs β

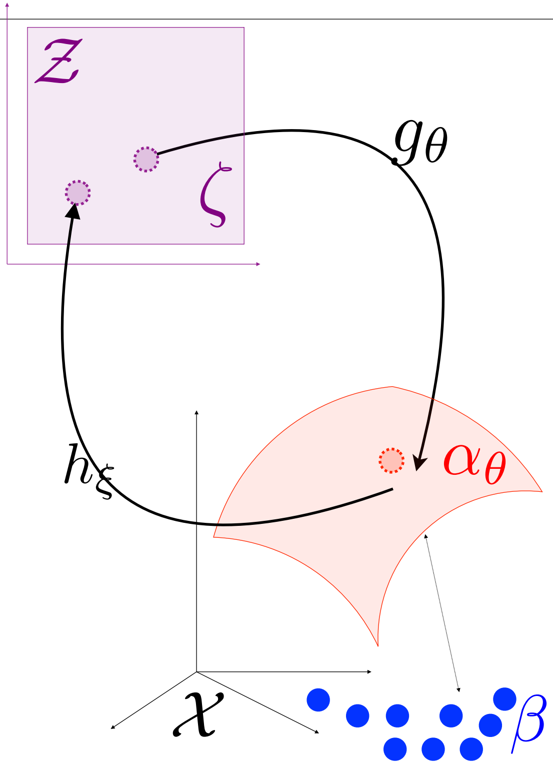
Generated α_θ



Examples of Images Generation

Inputs β

Generated α_θ

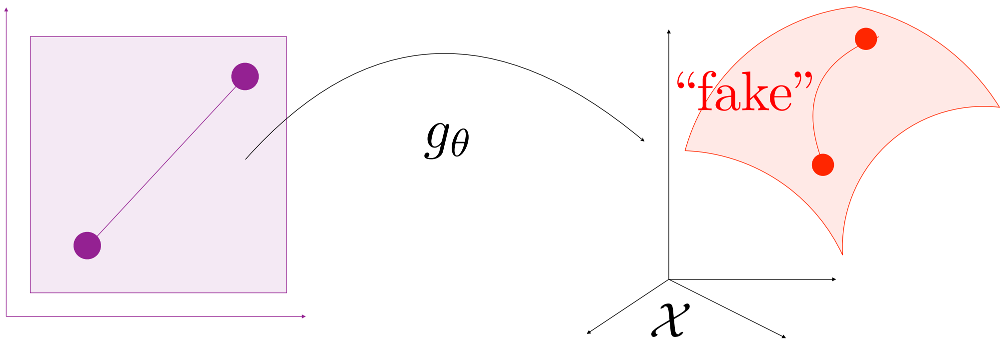


$\rightarrow$ Need to learn the metric $d(x, y) = \|h_\xi(x) - h_\xi(y)\|$ (GANs)

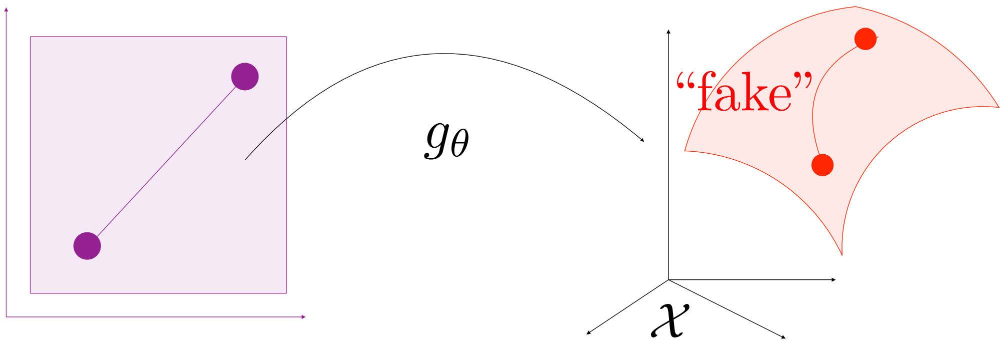
$\rightarrow$ Influence of ε?

$\rightarrow$ 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

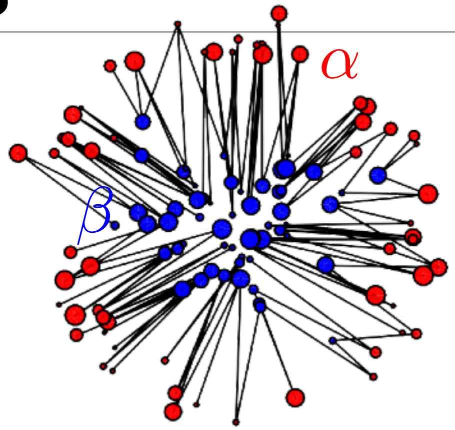


Progressive Growing of GANs for Improved Quality, Stability, and Variation
Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, ICLR 2018



A Glimpse at Algorithms

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

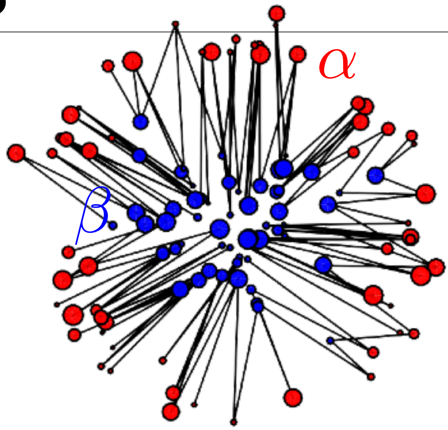


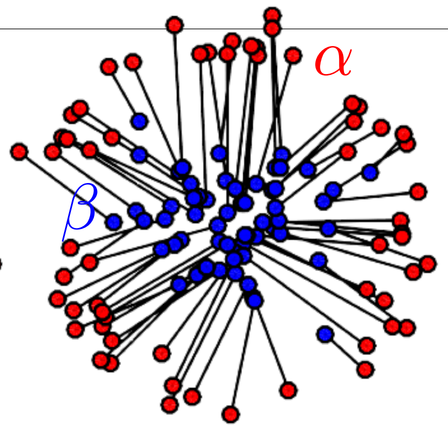
A Glimpse at Algorithms

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

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

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





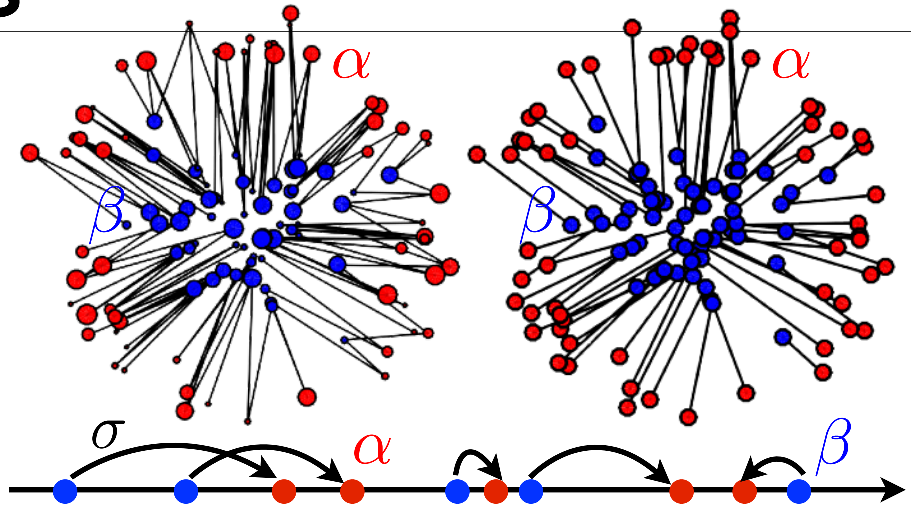
A Glimpse at Algorithms

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

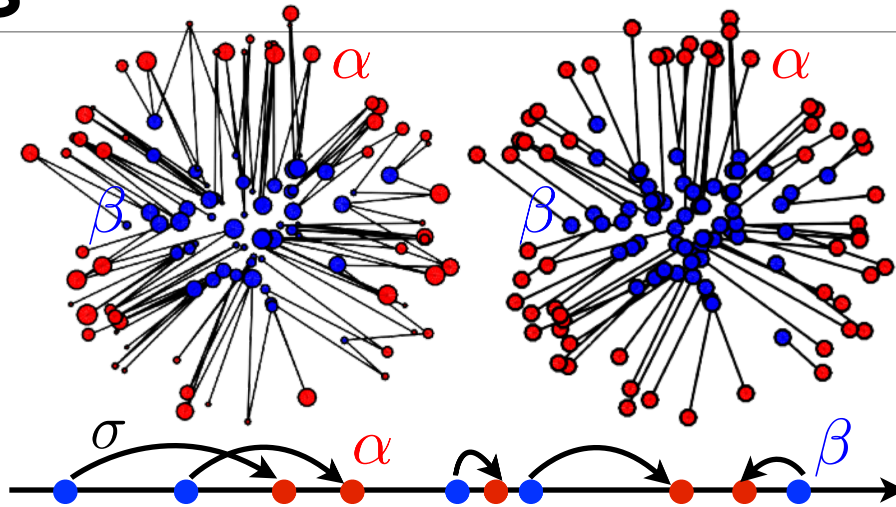
Hungarian/Auction: $O(n^3)$

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

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



A Glimpse at Algorithms



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

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

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

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

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

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

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

A Glimpse at Algorithms

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

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

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

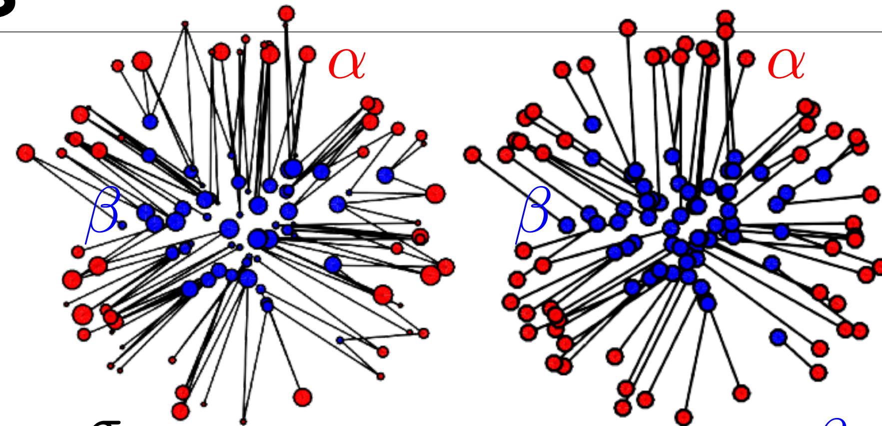
1-D case: sorting $O(n \log(n))$.

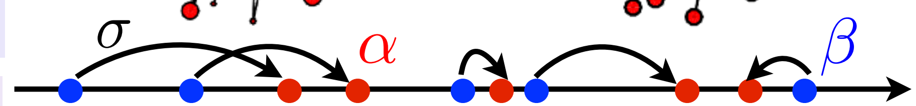
$p = 1$
$d = \|\cdot\|$

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

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

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



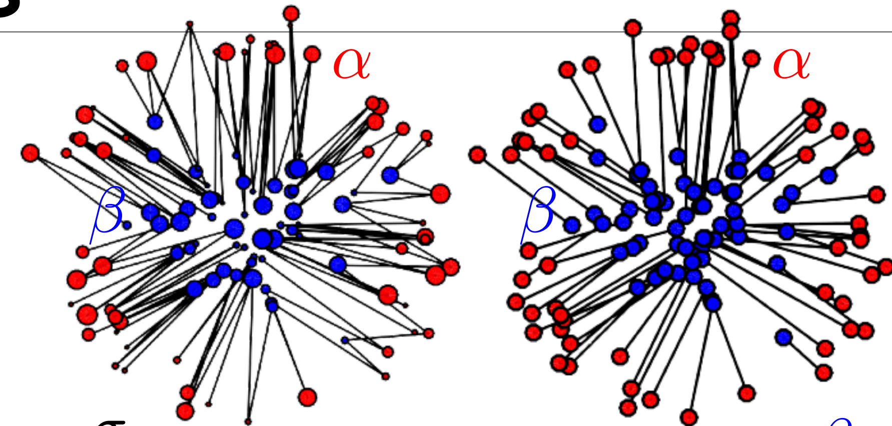


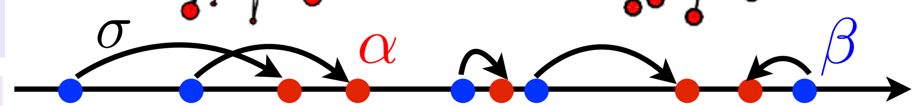
A Glimpse at Algorithms

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

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

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





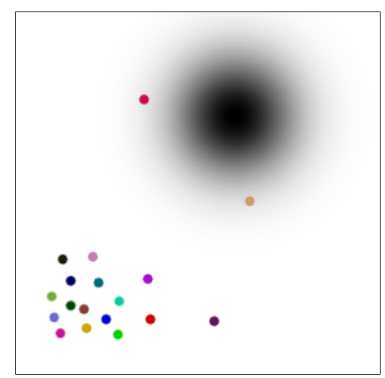
1-D case: sorting $O(n\log(n))$.

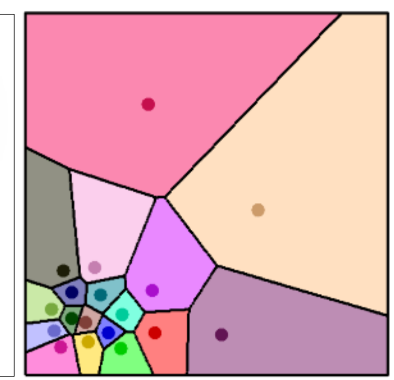
$p = 1$
$d = \|\cdot\|$

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

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

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



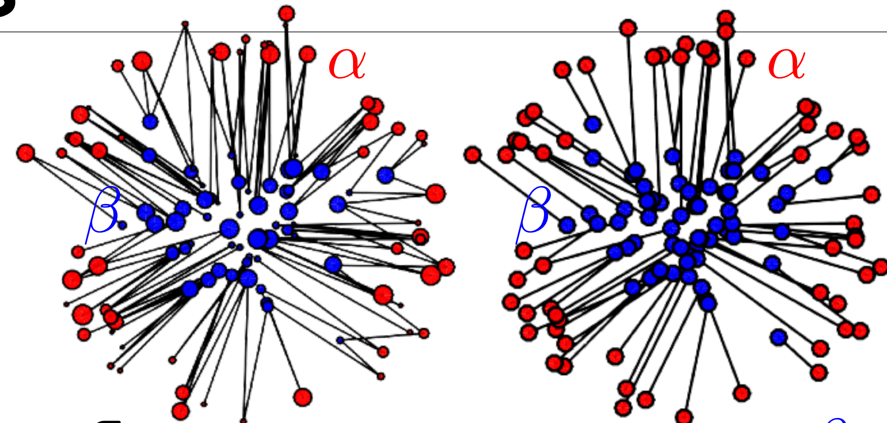


Semi-discrete: Laguerre cells, $d = \|\cdot\|_2^2$.

[Merigot 2013]

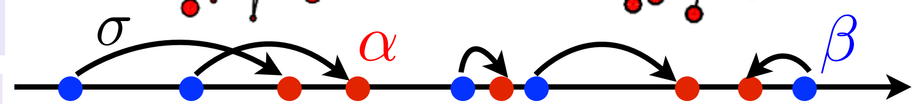
A Glimpse at Algorithms

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



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

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



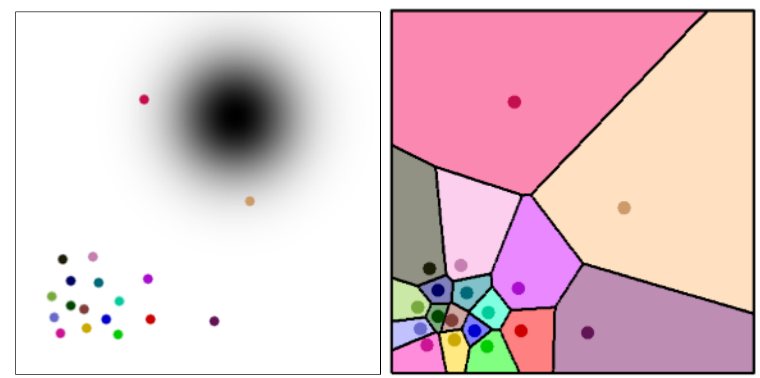
1-D case: sorting $O(n \log(n))$.

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

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

$\rightarrow$ 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]

Entropic regularization: generic d.