Computational Optimal Transport

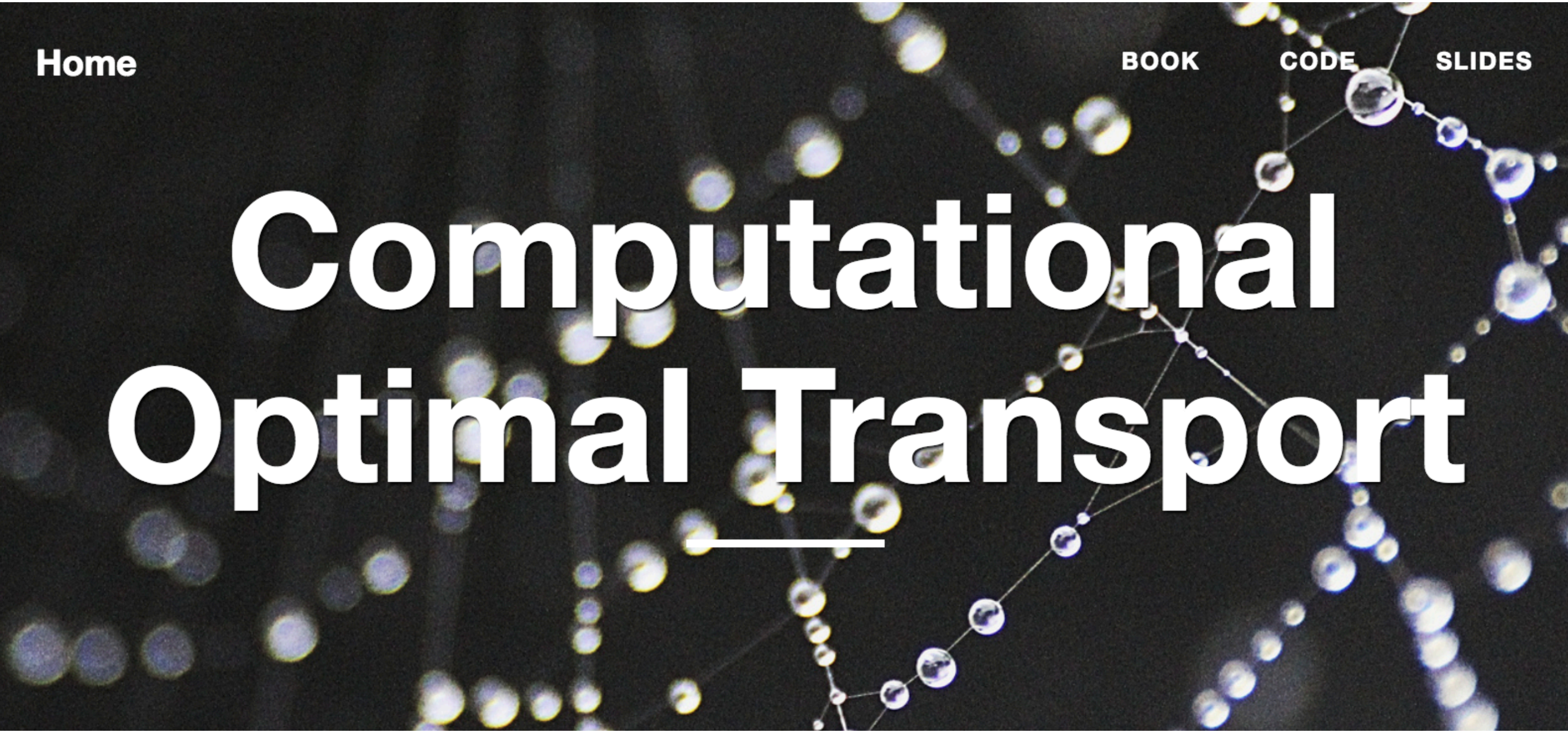
http://optimaltransport.github.io

Introduction

Gabriel Peyré

www.numerical-tours.com

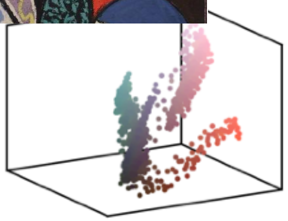
https://optimaltransport.github.io

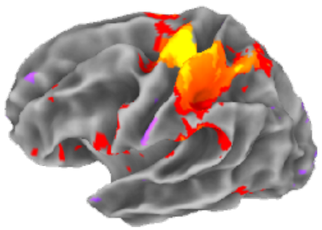


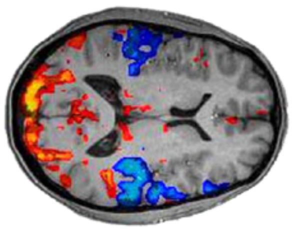
Probability Distributions in Data Sciences

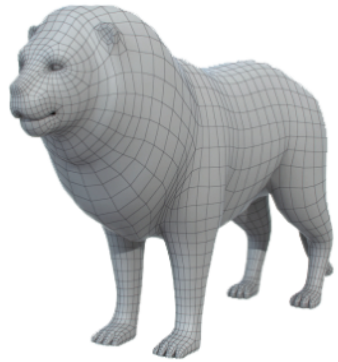
Probability distributions and histograms

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





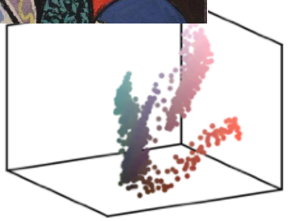


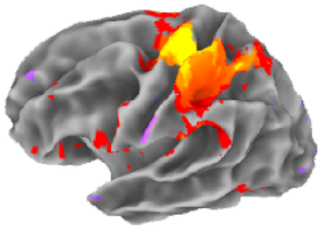


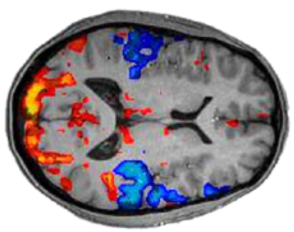
Probability Distributions in Data Sciences

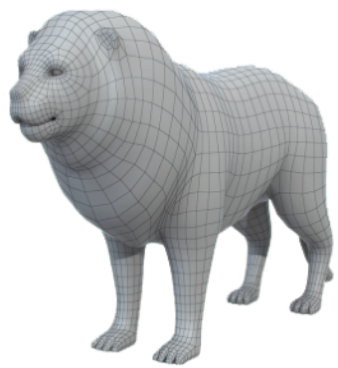
Probability distributions and histograms

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





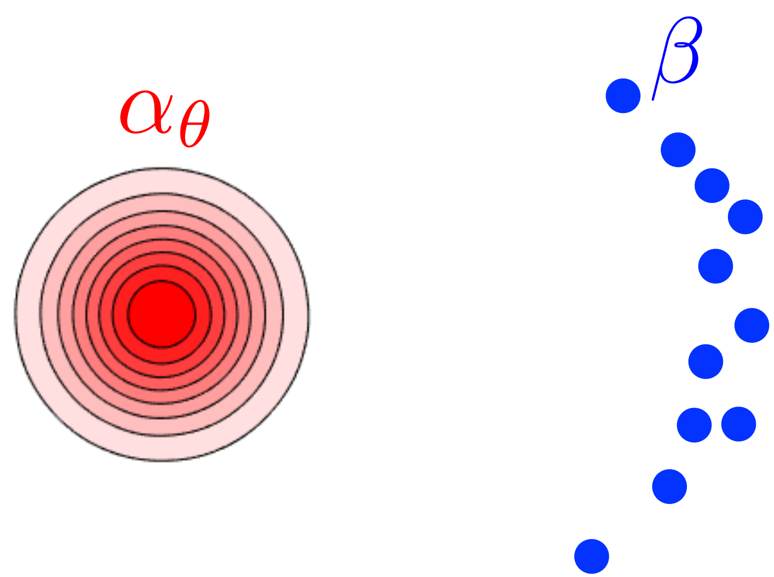




Unsupervised learning

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

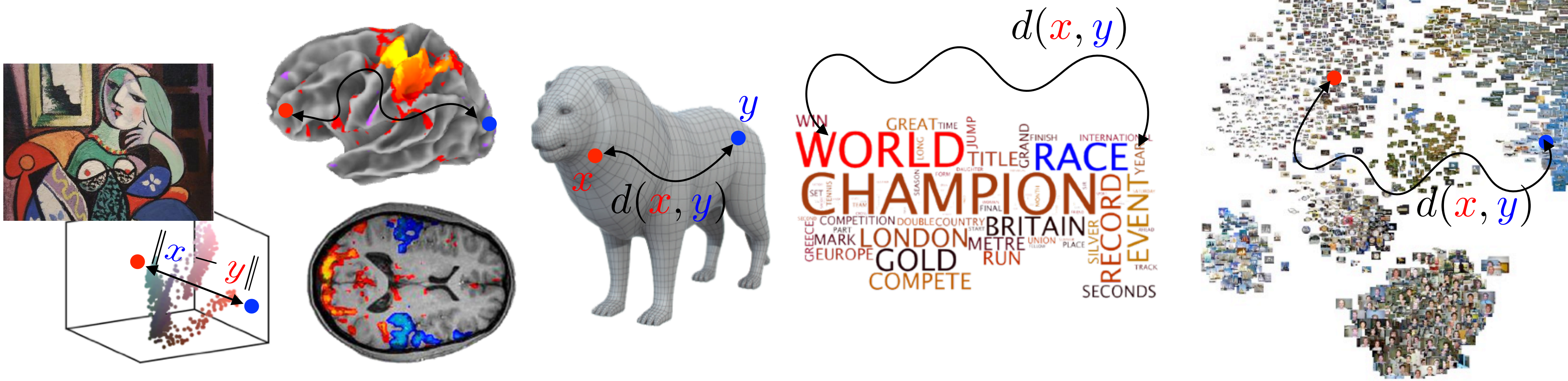
Parametric model: $\theta \mapsto \alpha_\theta$



Probability Distributions in Data Sciences

Probability distributions and histograms

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



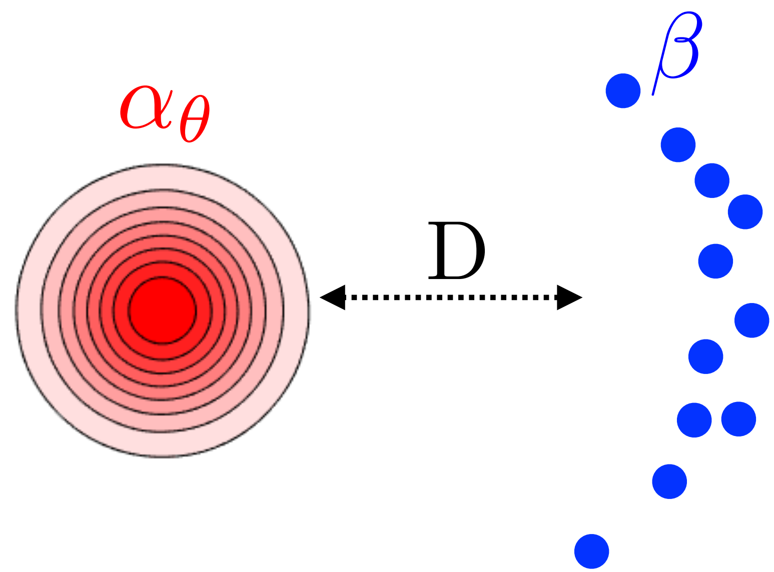
Unsupervised learning

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

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

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

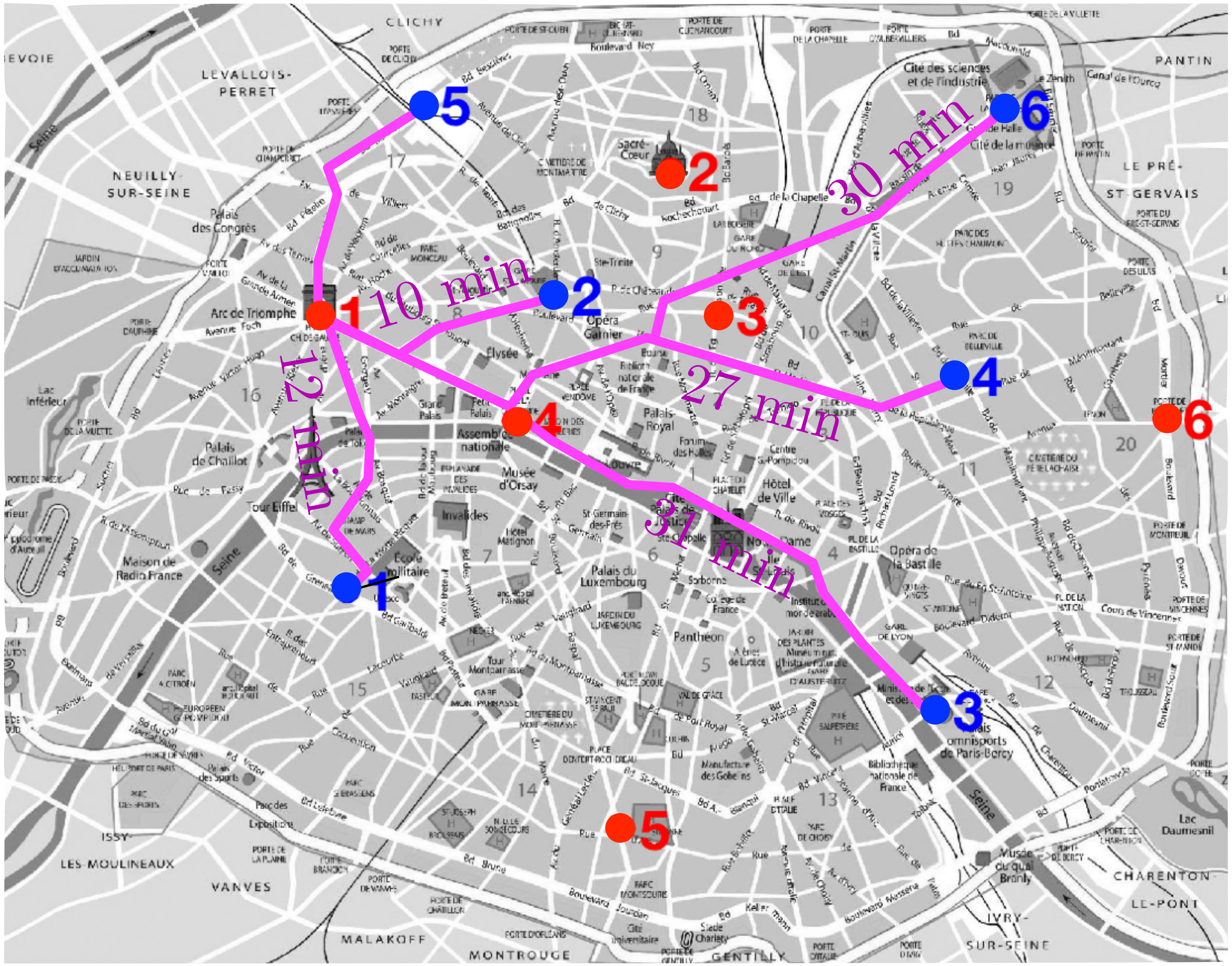
$\rightarrow$ takes into account a metric d.



Overview

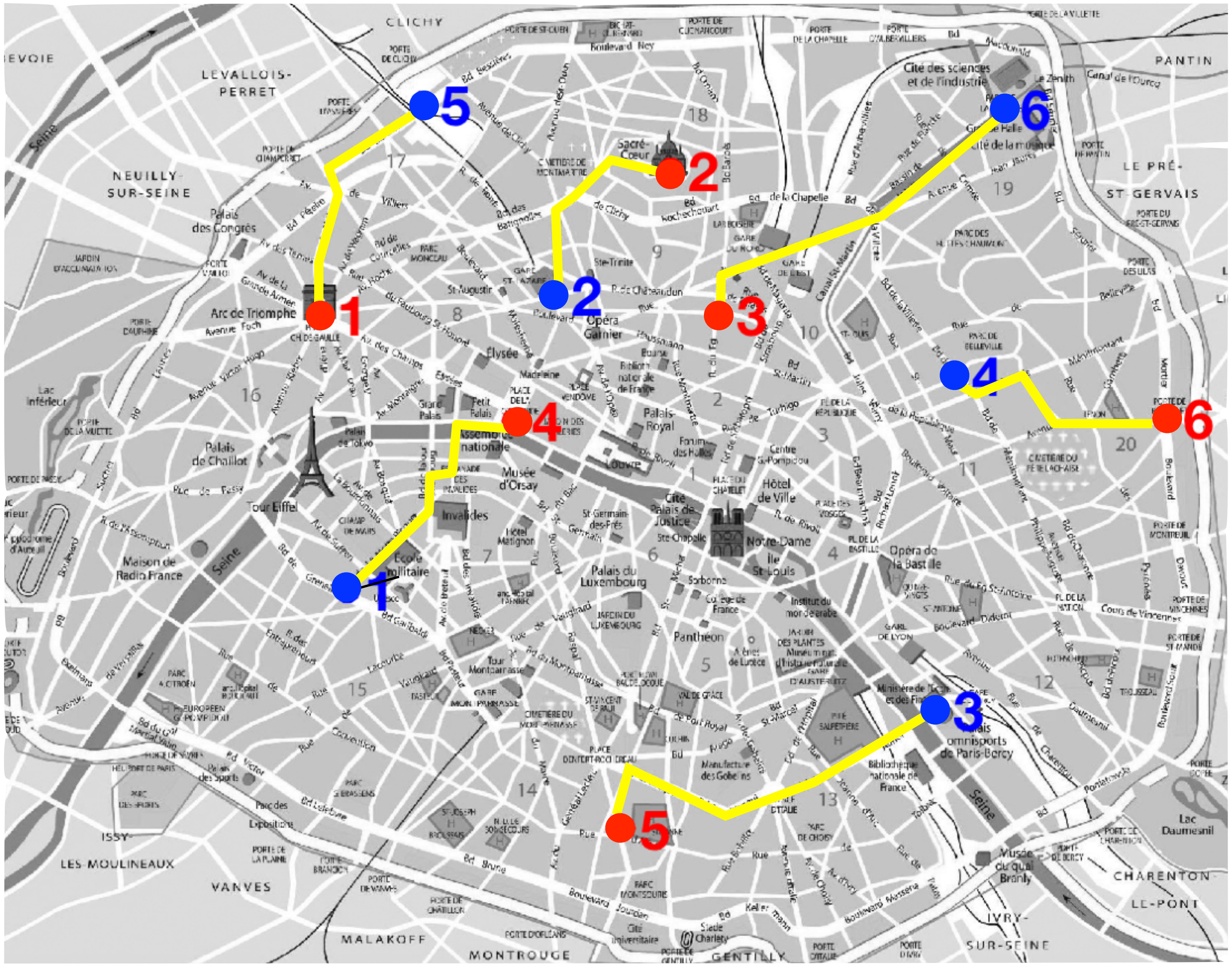
- **Monge Formulation**
- Continuous Optimal Transport
- Kantorovitch Formulation
- Applications

Fom bakeries to cafés



c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15

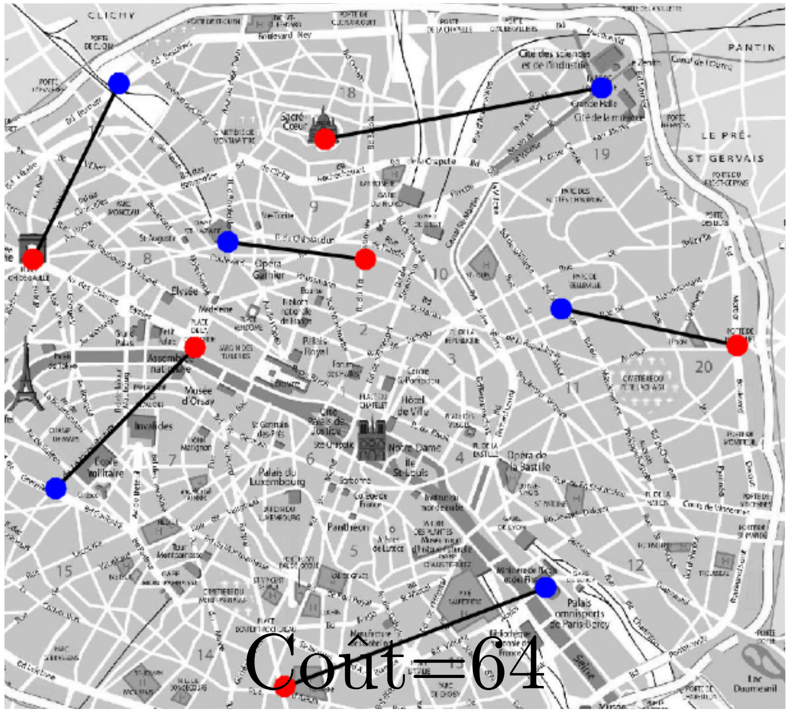
Fom bakeries to cafés

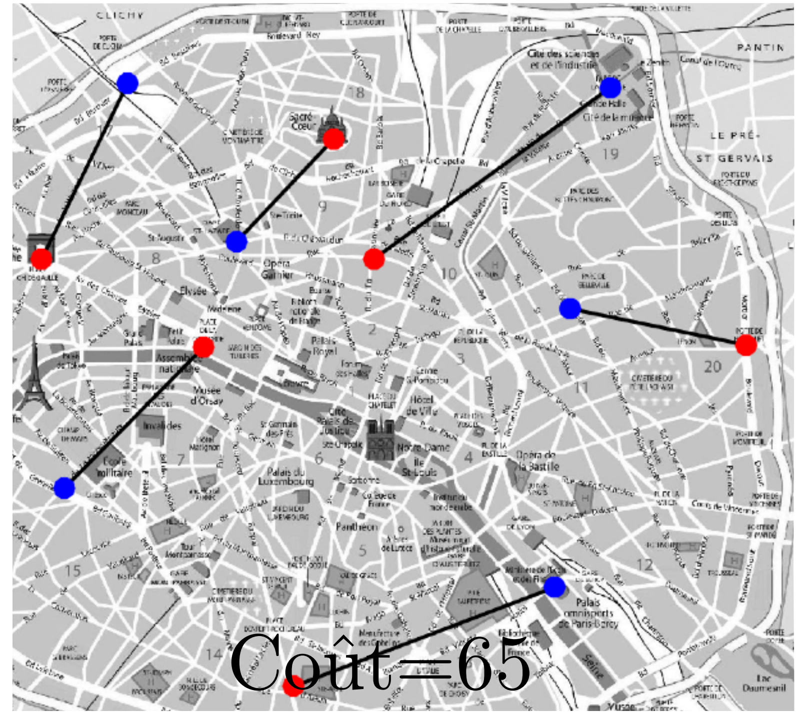


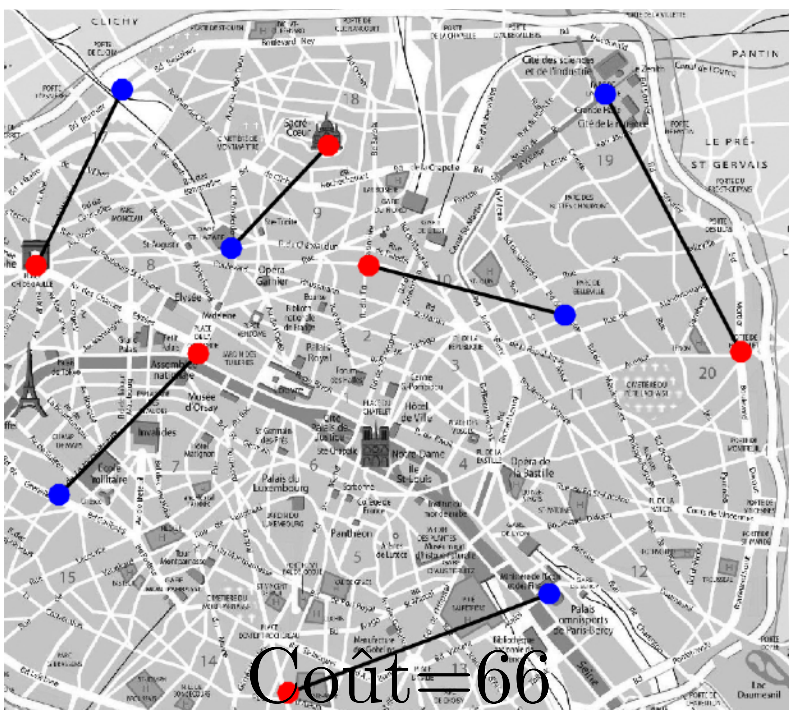
c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15

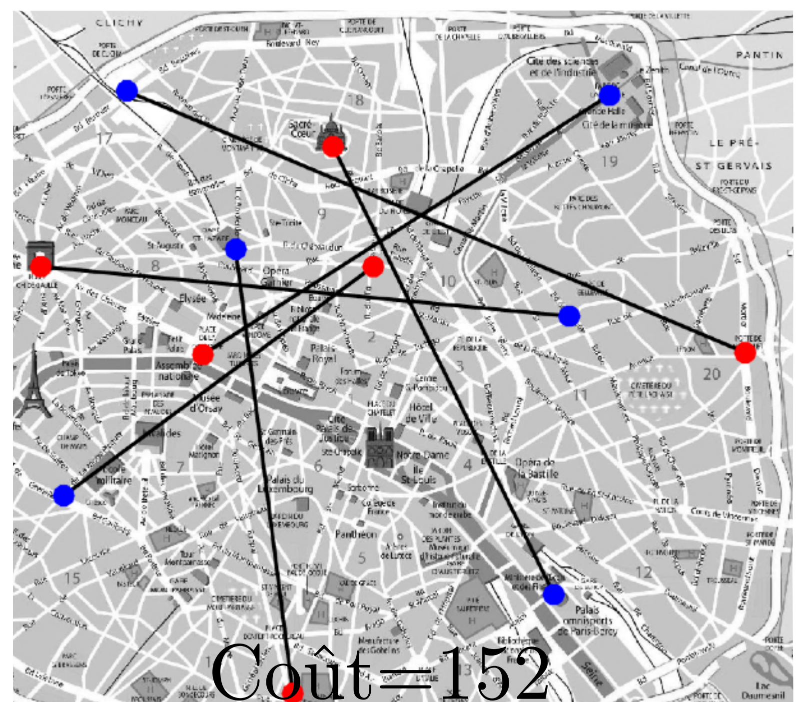
$$\text{Cout:} \quad 10+7+15+10+14+9 \ = 65 \text{ min}$$

From best to worst







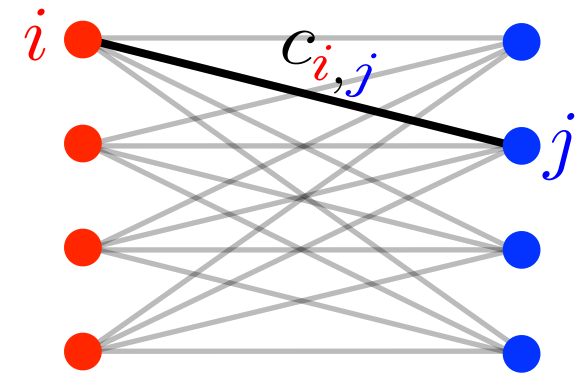


Combinatorial Search

Cost $(c_{i,j})_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$

Permutations:

$$\Sigma_n \overset{\text{def.}}{=} \{\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}\}$$



Monge's problem:

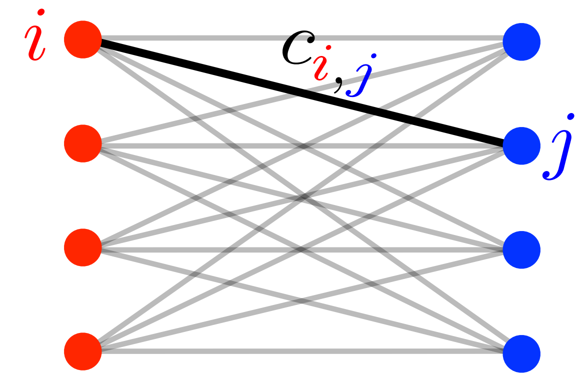
$$\min_{\sigma\in\Sigma_n} \sum_{i=1}^{n} c_{i,\sigma(i)}$$

Combinatorial Search

Cost $(c_{i,j})_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$

Permutations:

$$\Sigma_n \overset{\text{def.}}{=} \{\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}\}$$



Monge's problem:

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

n	$n\,!$
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5 040
8	40 320

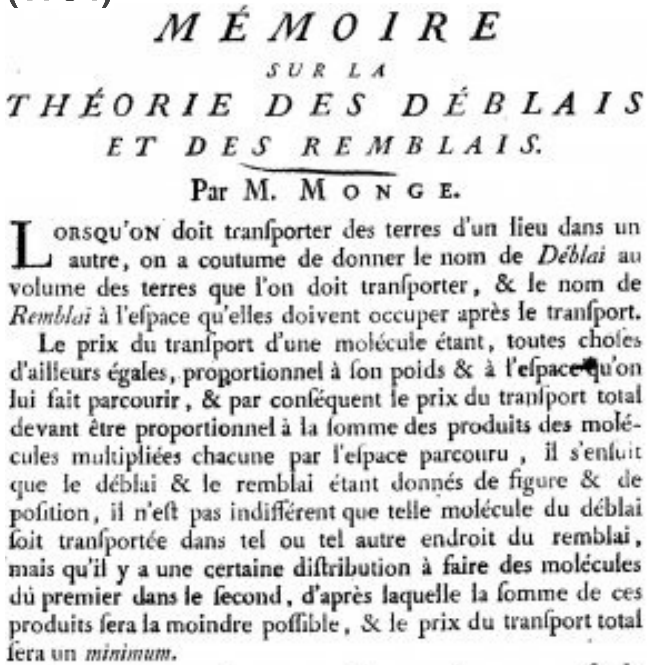
n	$n\,!$
9	362880
10	3628800
11	39916800
12	479001600
...	...
25	$1{,}551\text{x}10^{25}$
...	...
70	$1{,}198\text{x}10^{100}$

Atoms in the universe: 10^{79}

Neurons in the brain: 10^{11}.

Gaspard Monge (1746-1818)

(1784)



MÉMOIRE

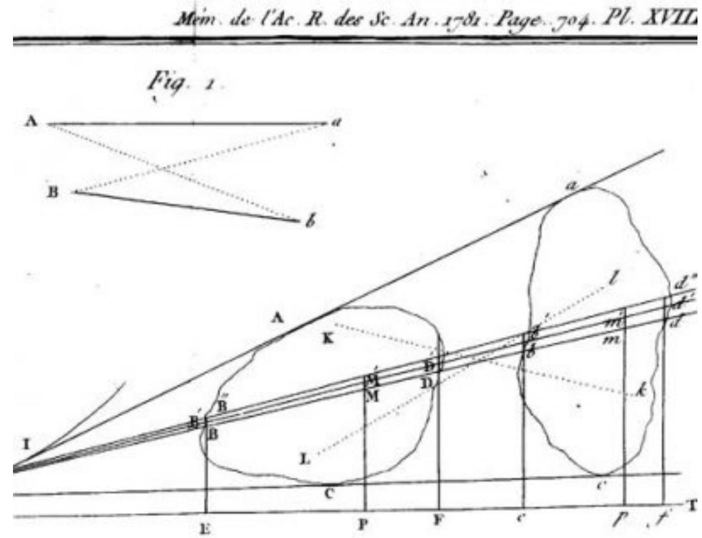
SUR LA

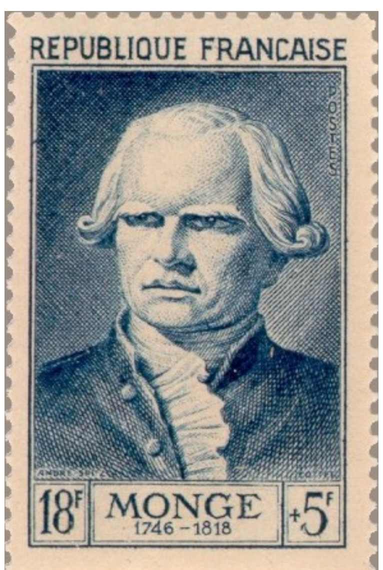
THÉORIE DES DÉBLAIS ET DES REMBLAIS.

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*.



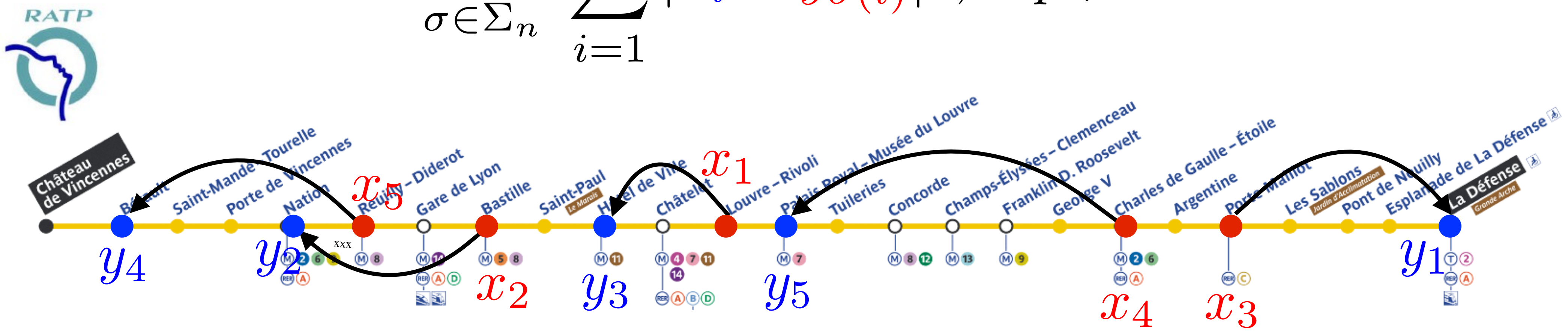


1-D Optimal Transport

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

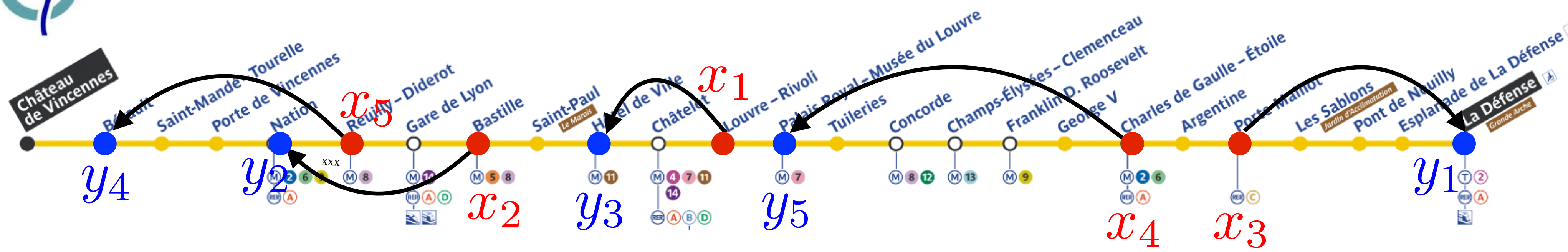
1-D Optimal Transport

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



1-D Optimal Transport

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

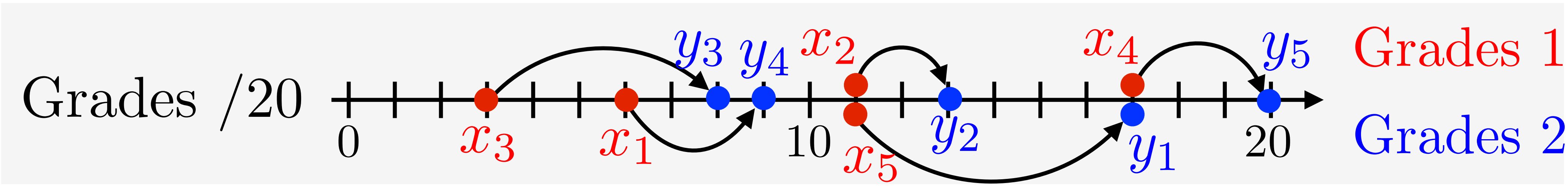


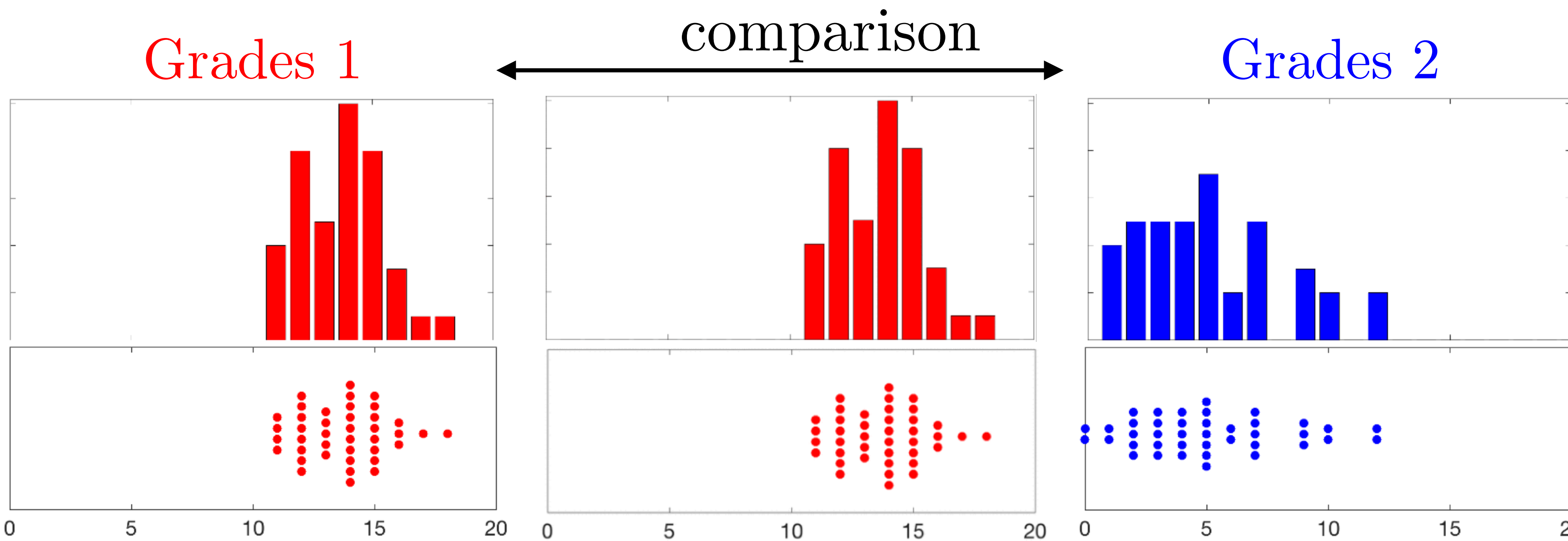
Sorting algorithms: insertion $n(n-1)/2$ worst case.

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

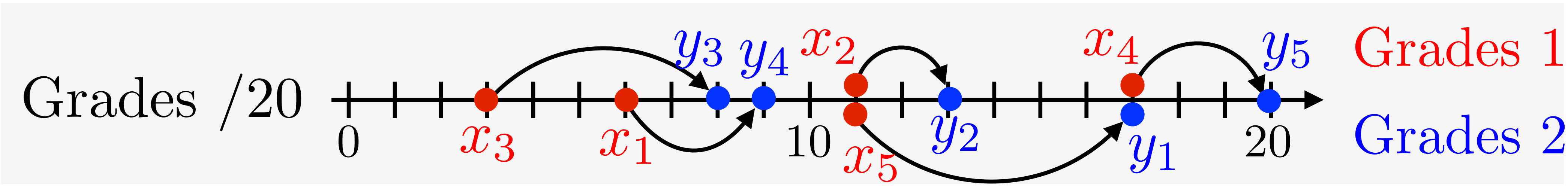
QuickSort: $O(n \log(n))$.

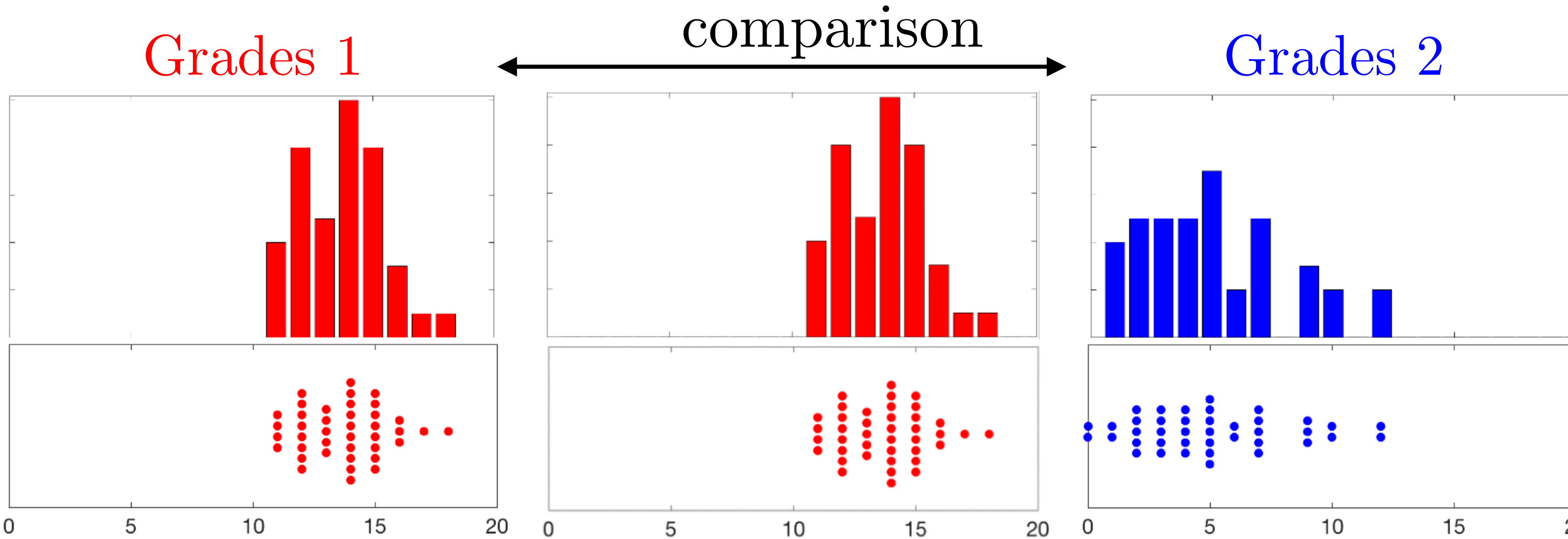
1-D OT Interpolation



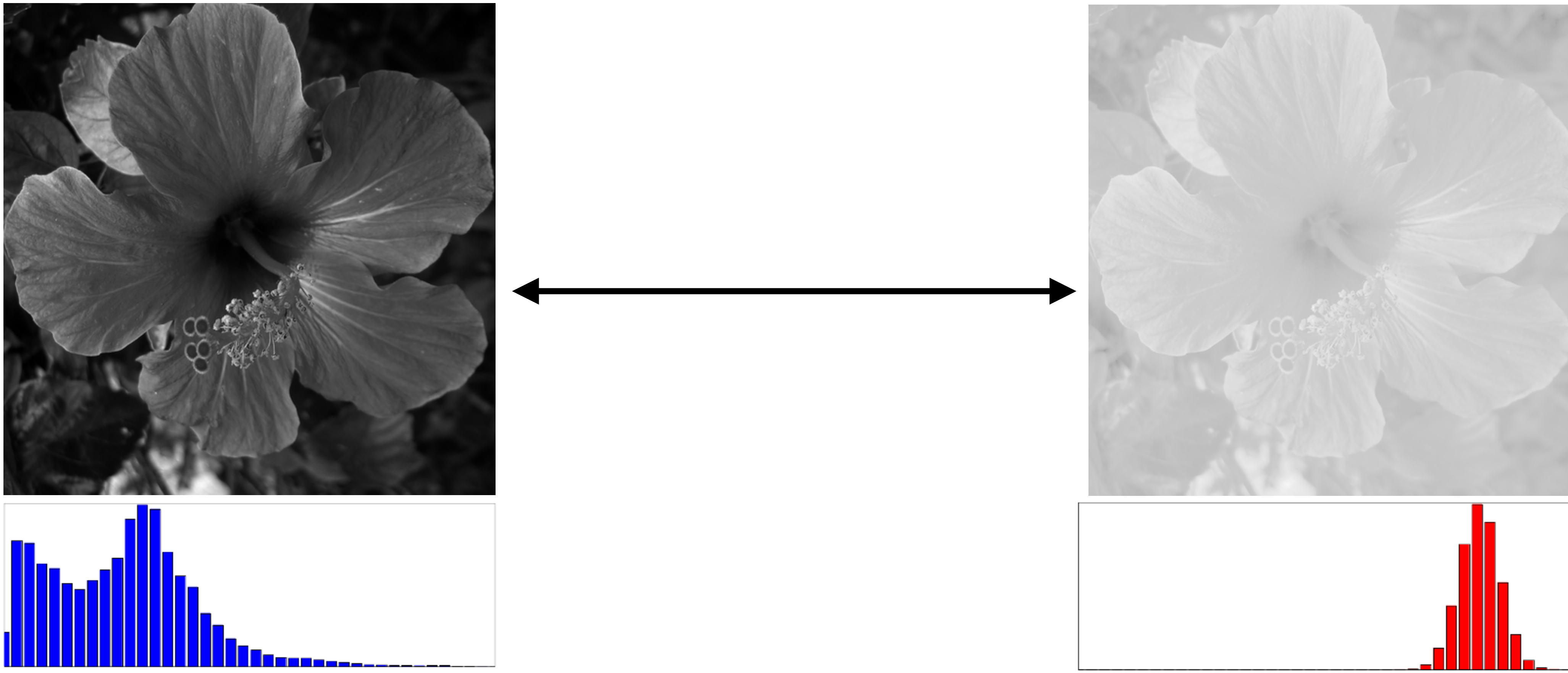


1-D OT Interpolation

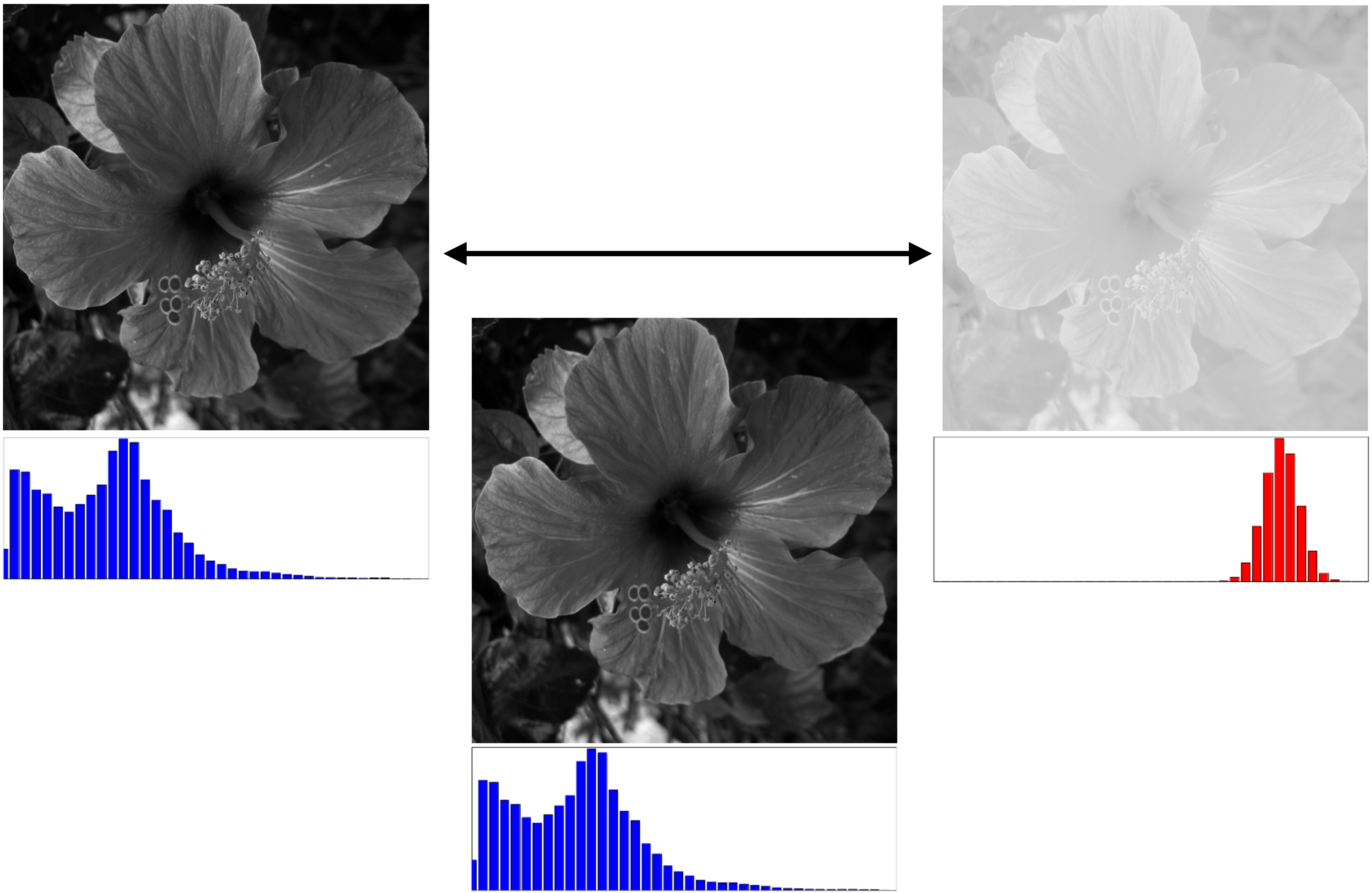




Grayscale Histogram Equalization



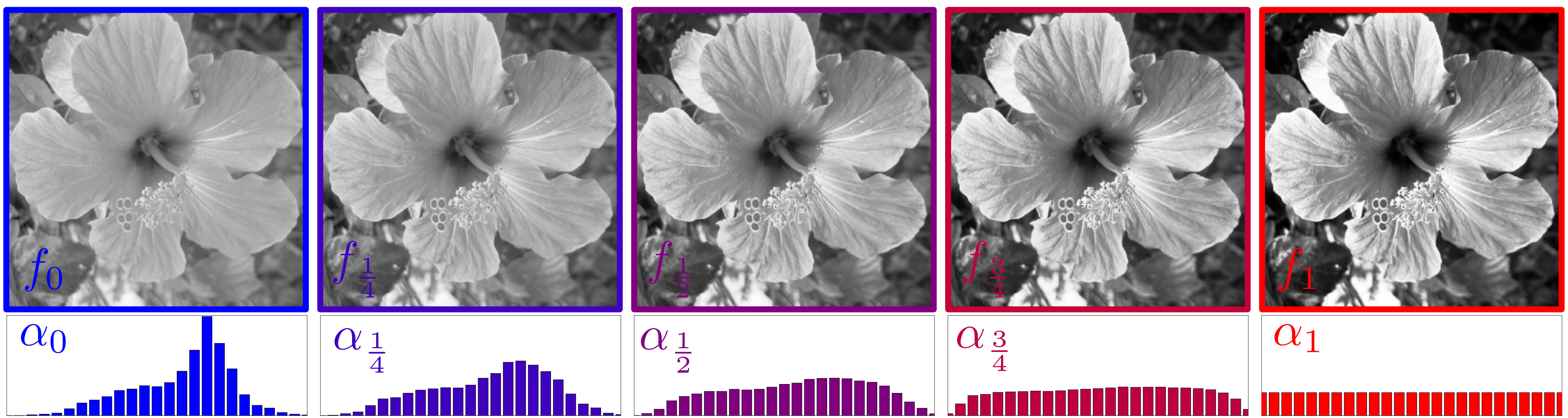
Grayscale Histogram Equalization



Application: Grayscale Histogram Equalization

```
[~,I] = sort(f(:));
f(I) = linspace(0,1,length(f(:)));
```

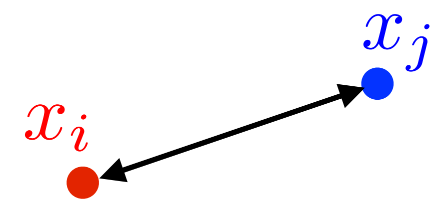
```
f[argsort(f.flatten())] = np.linspace(0,1,n*n)
```

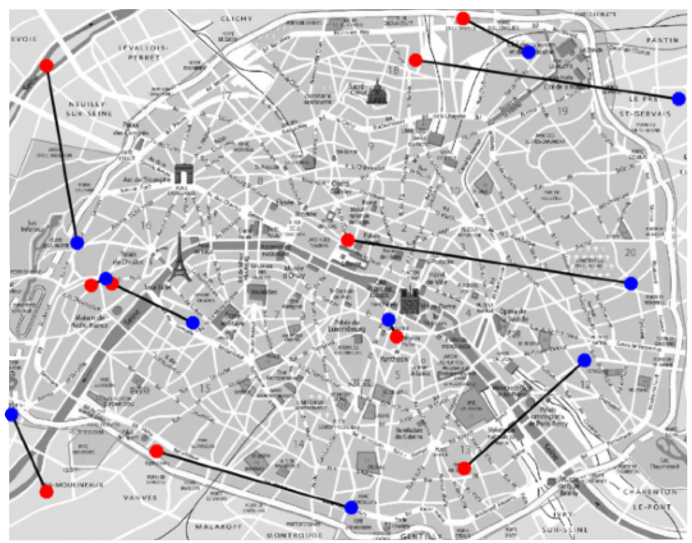


In 2-D

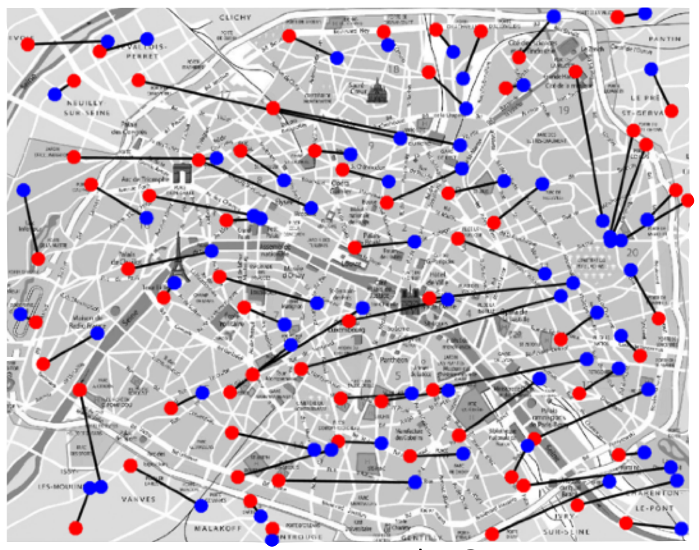
$x_i, y_j \in \mathbb{R}^2$

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

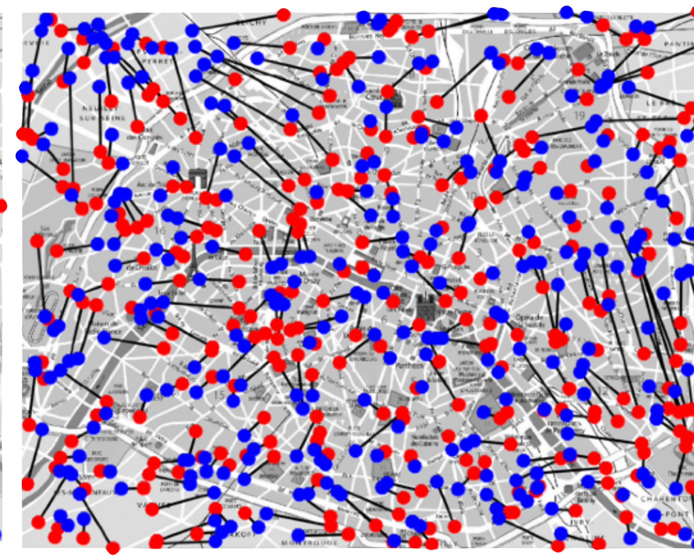




$n = 10$

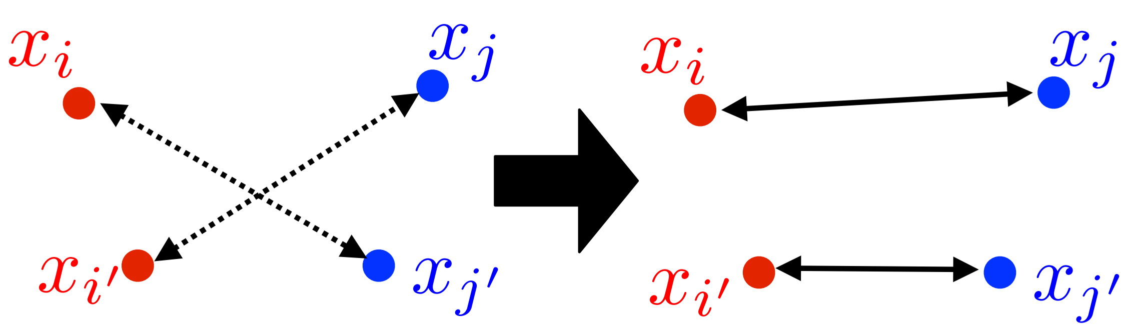


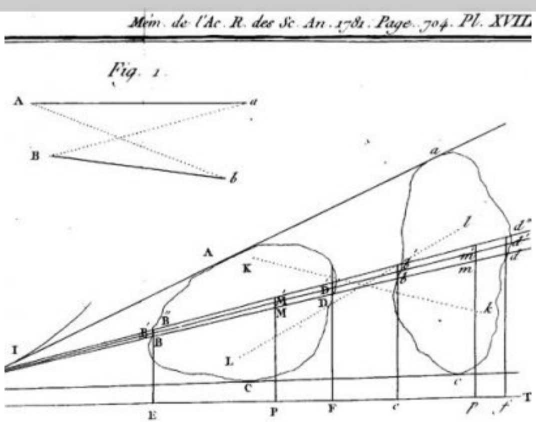
$n = 70$



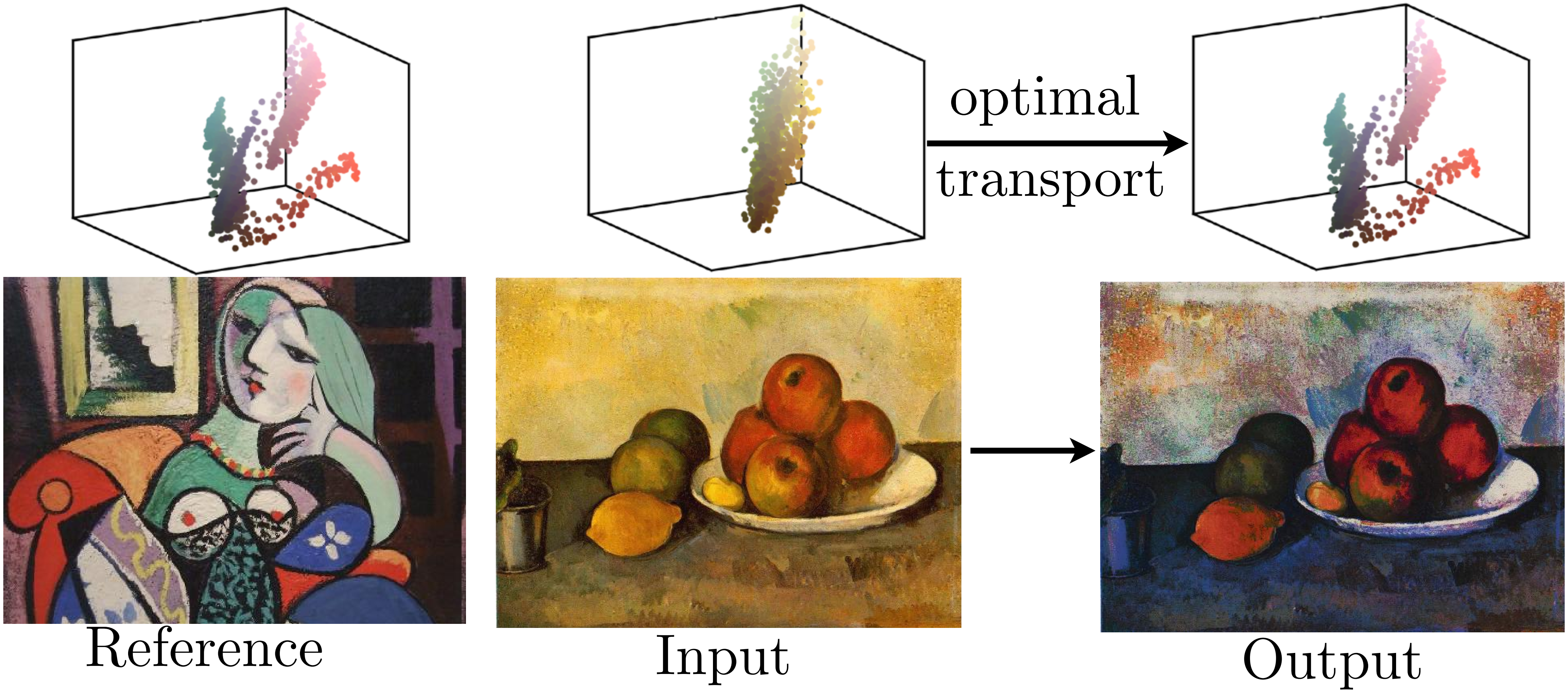
$n = 300$

Proposition: two segments never cross.

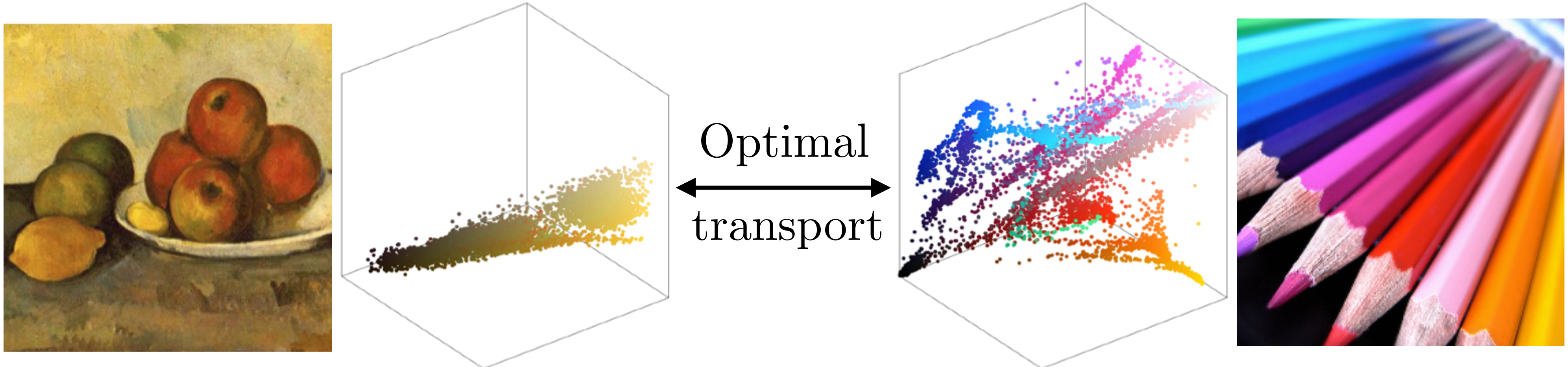




In 3-D: Color Image Palette Equalization



Color Image Palette Equalization



Color Image Palette Equalization

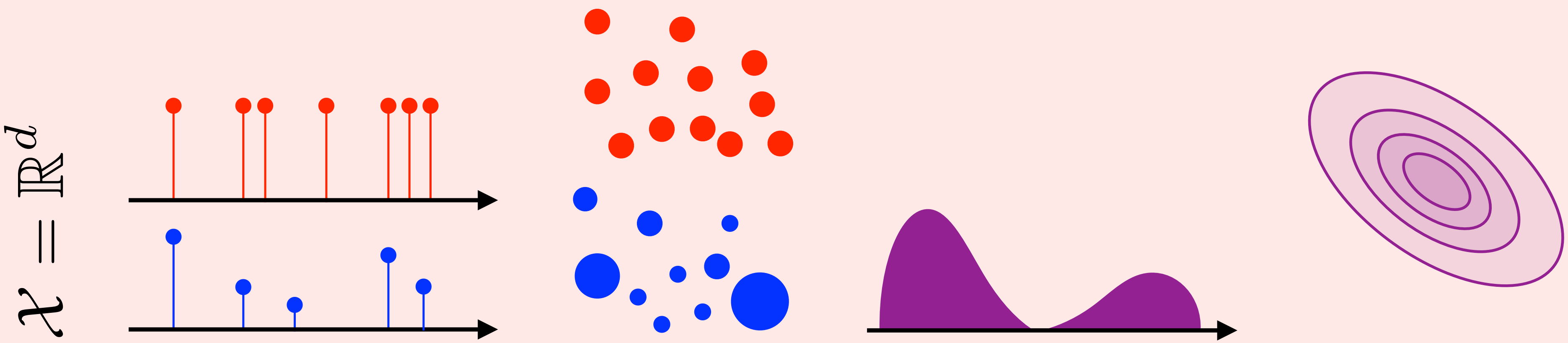
Overview

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

Probability Measures

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

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

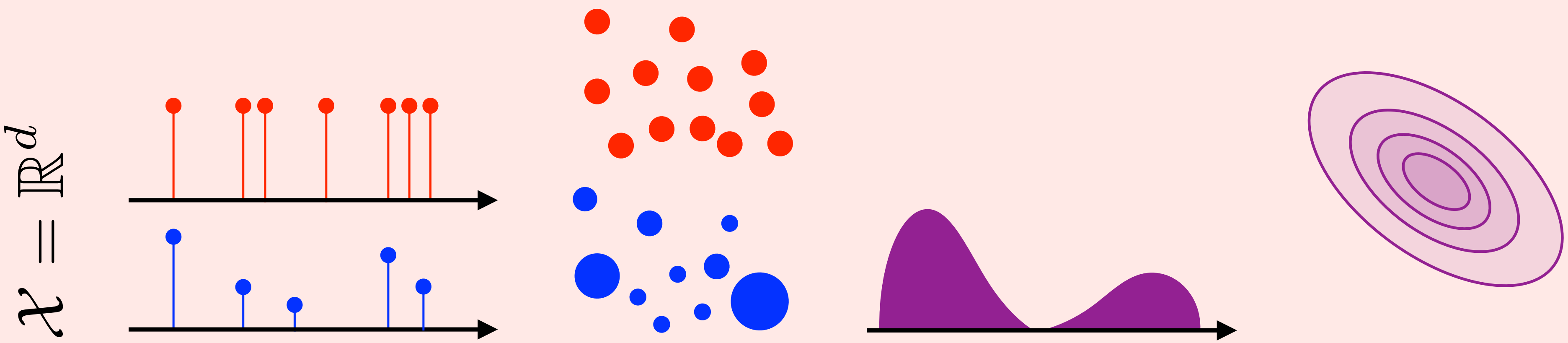


Discrete $d = 1$ Discrete $d = 2$ Density $d = 1$ Density $d = 2$

Probability Measures

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

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



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

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

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

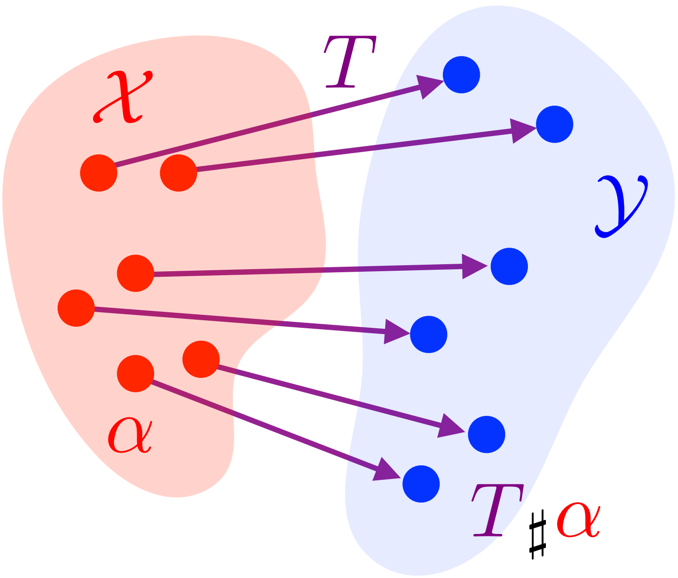
$$\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x \longrightarrow \int_{\mathcal{X}} g\mathrm{d}\alpha = \int_{\mathcal{X}} \rho_\alpha(x)\mathrm{d}x$$

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

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

Push Forward

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

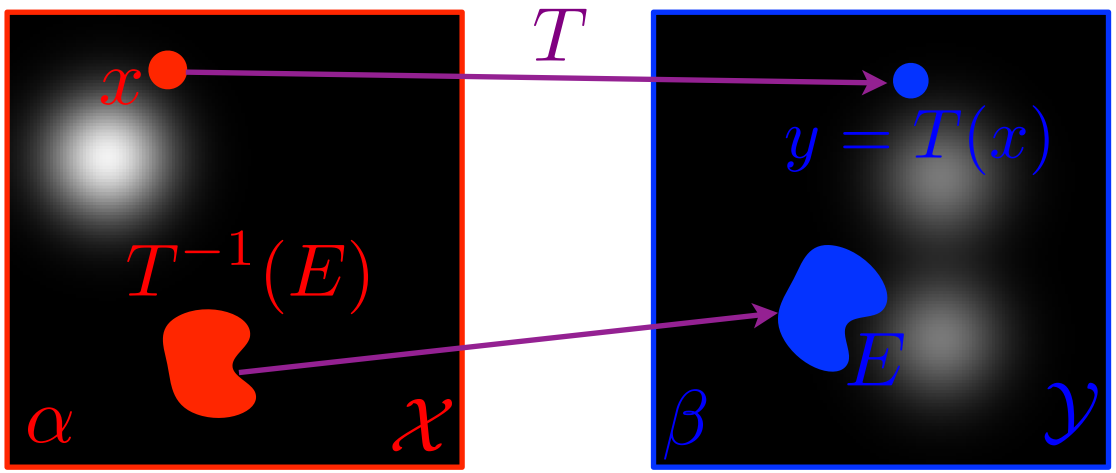


Push-forward:

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

General case:

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



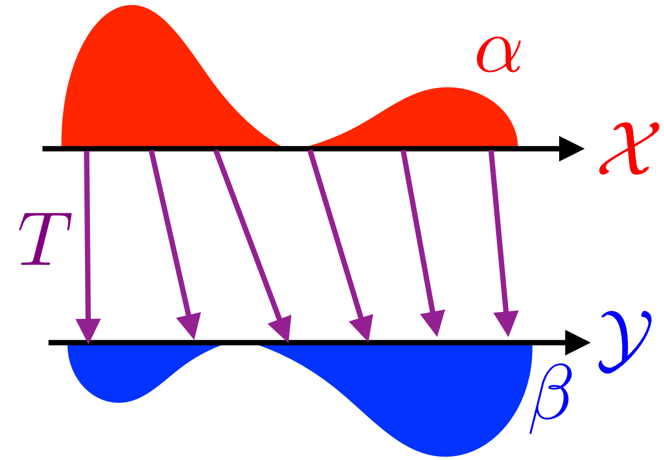
Change of variables:

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

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

Continuous Monge's Problem

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



Discrete case:

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

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

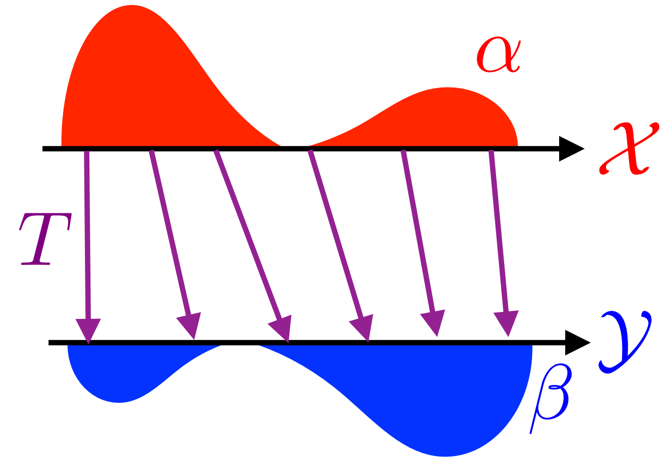
$$\min_{\sigma \in \Sigma_n} \sum_{i=1}^{n} \mathbf{C}_{i,\sigma(i)}$$

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

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

Continuous Monge's Problem

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



Discrete case:

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

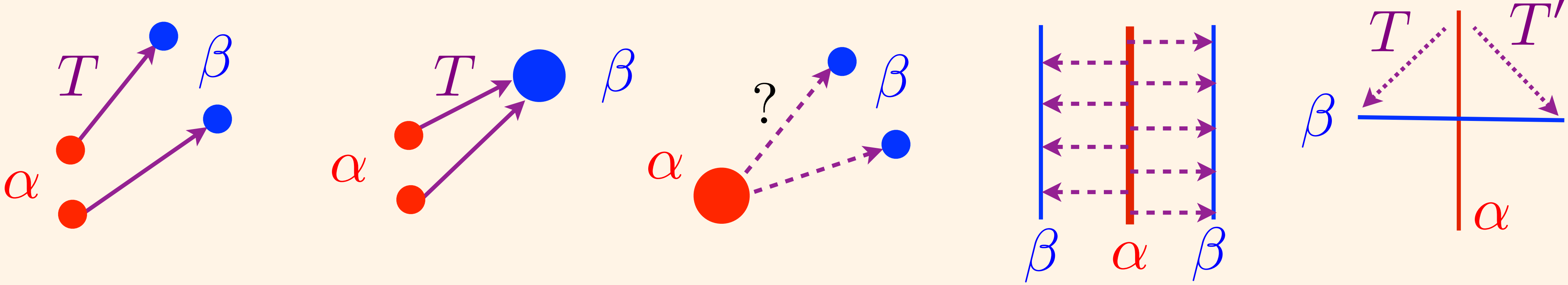
$\beta = \sum_{j=1}^n \delta_{y_j}$

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

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

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

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



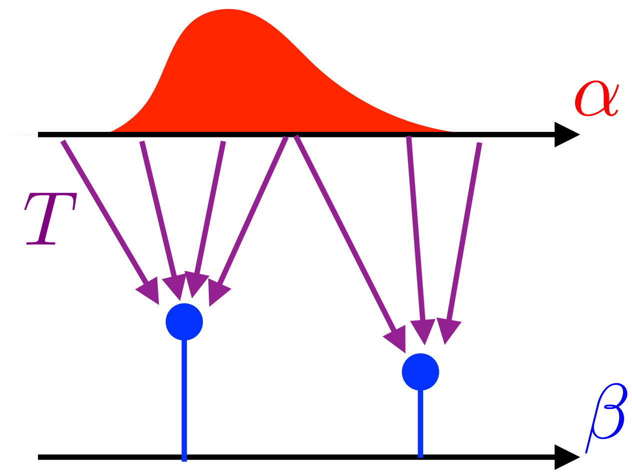
Brenier's Theorem

Hypotheses:

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

$$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \rho_\alpha \text{ density.}$$

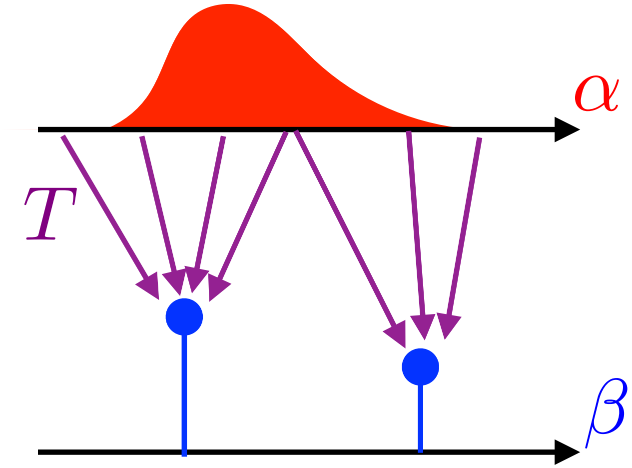
$$\min_{\beta = T_\sharp \alpha} \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x)$$



Brenier's Theorem

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

$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \rho_\alpha$ density.



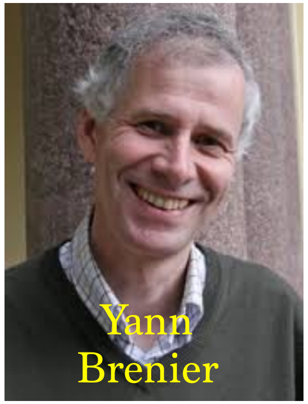
$$\min_{\beta = T_\sharp \alpha} \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x)$$

Theorem: [Brenier, 1991]

There exists a unique Monge map T.

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

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

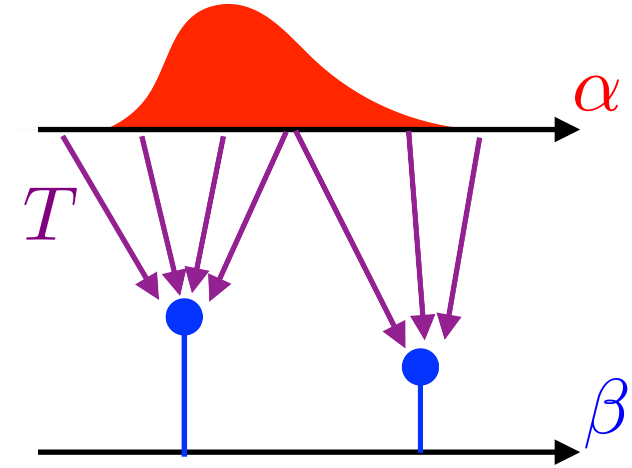


Brenier's Theorem

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

$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \rho_\alpha$ density.

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

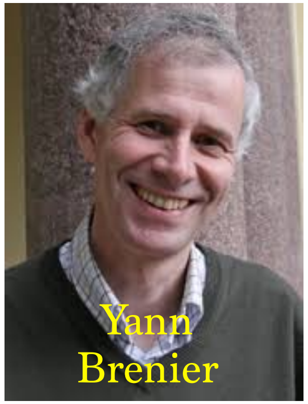


Theorem: [Brenier, 1991]

There exists a unique Monge map T.

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

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

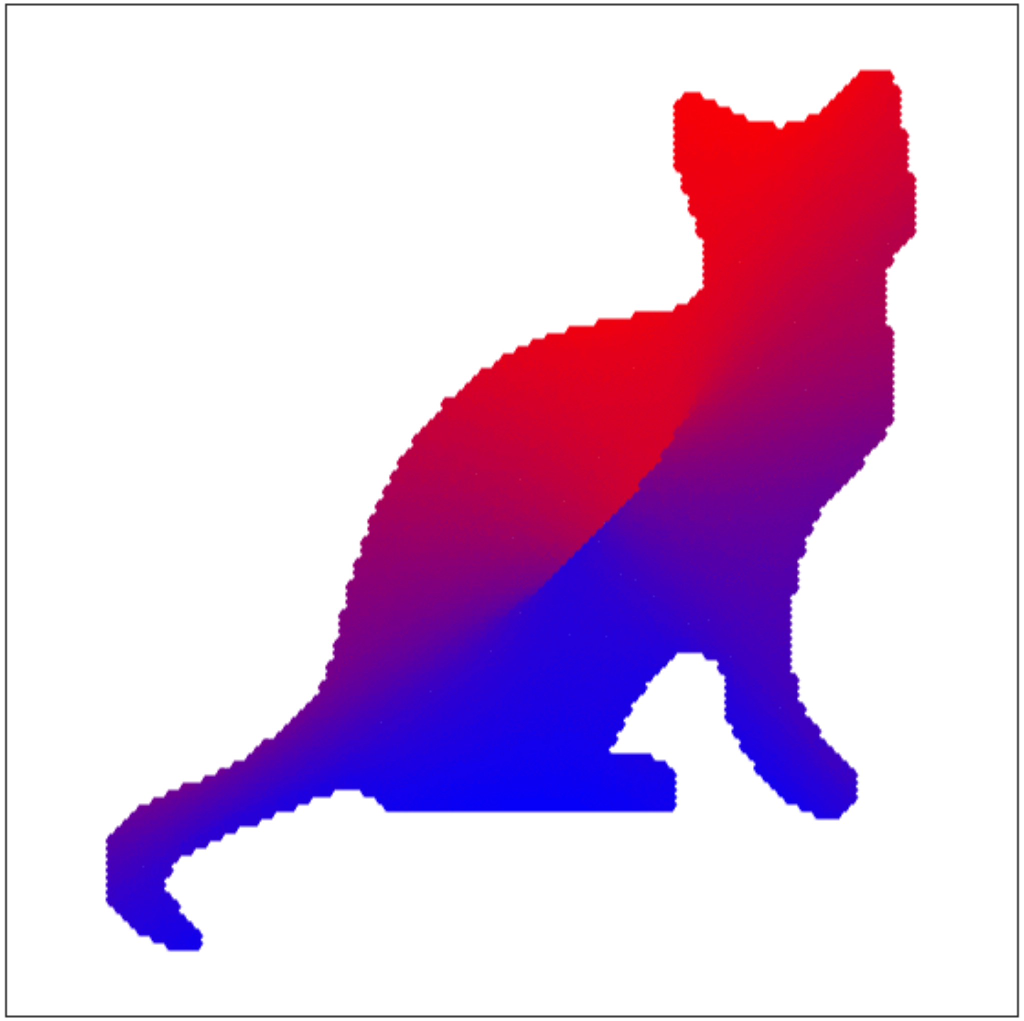


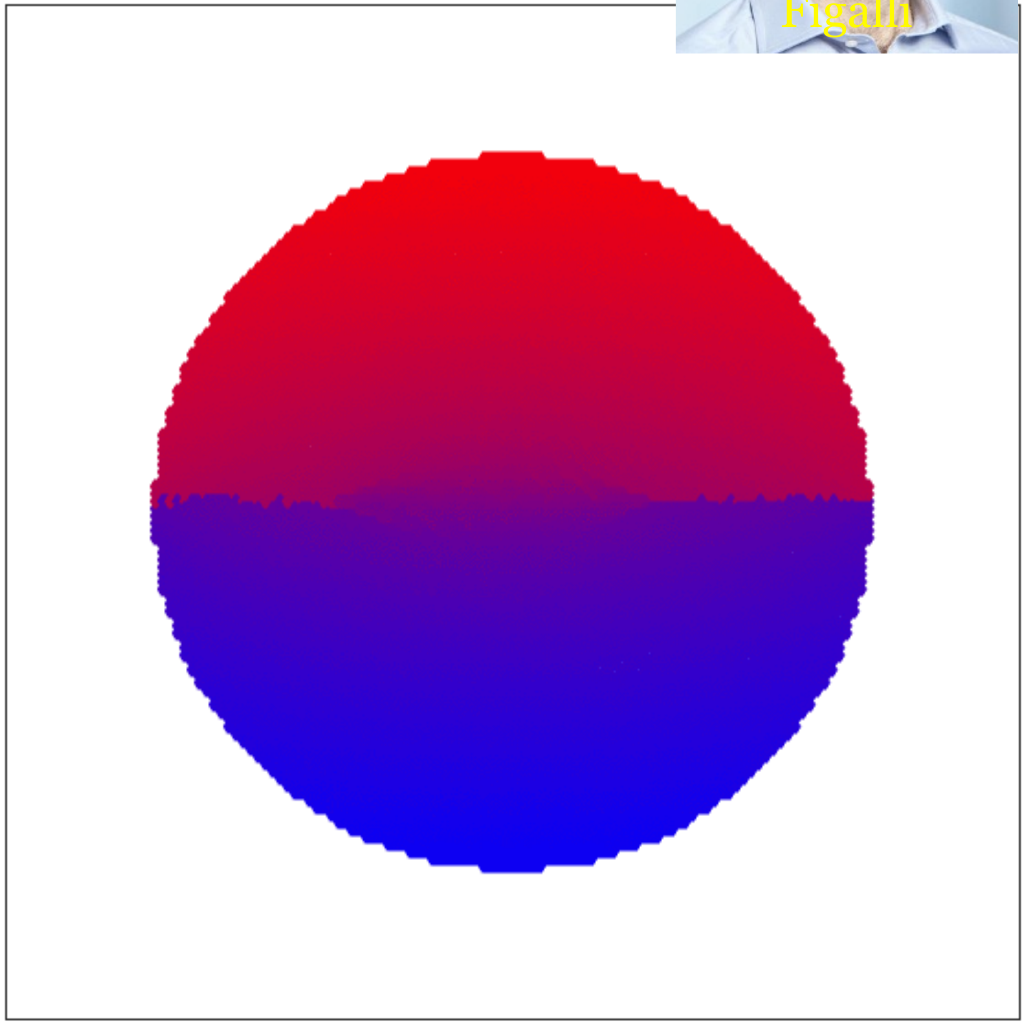
$\rightarrow$ Monge-Ampère equation (non-linear, degenerate elliptic).

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

Regularity Theory

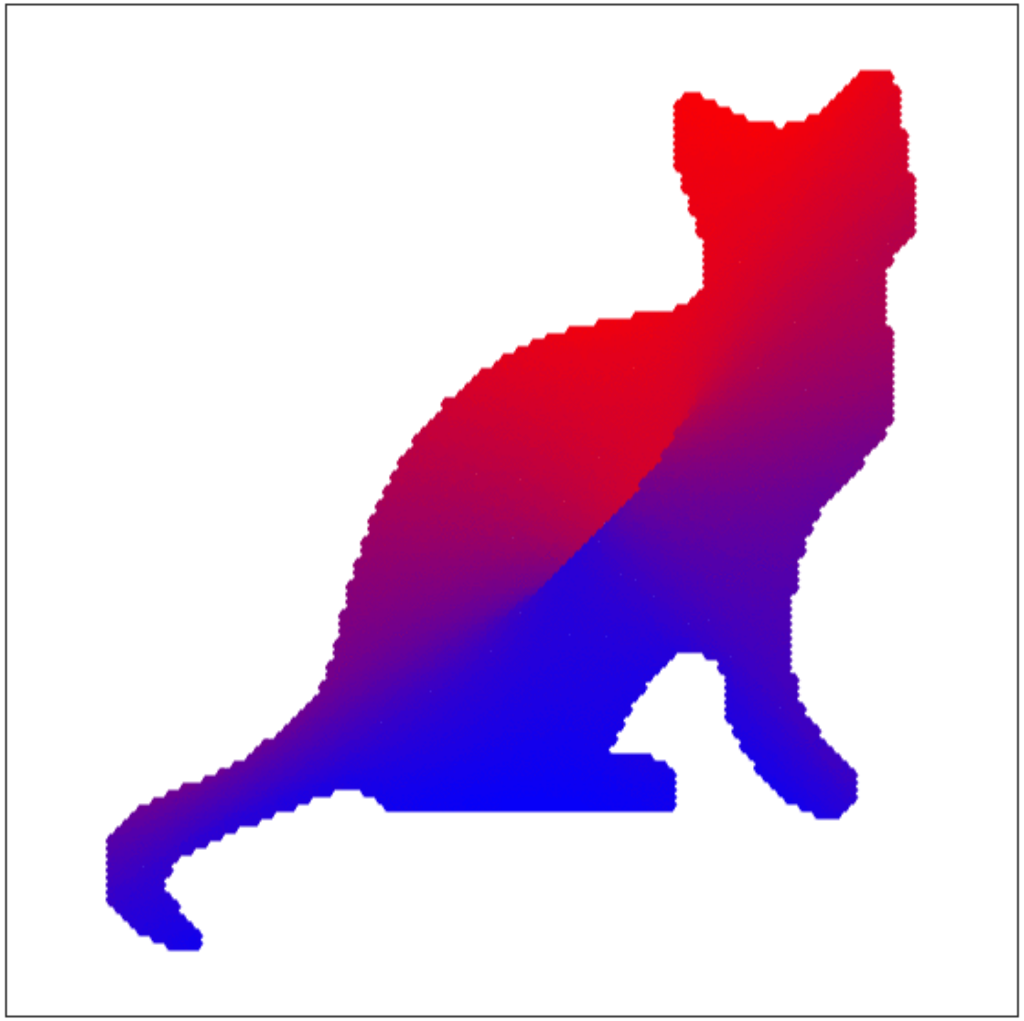
$\rightarrow$ Regularity of T requires convex target.

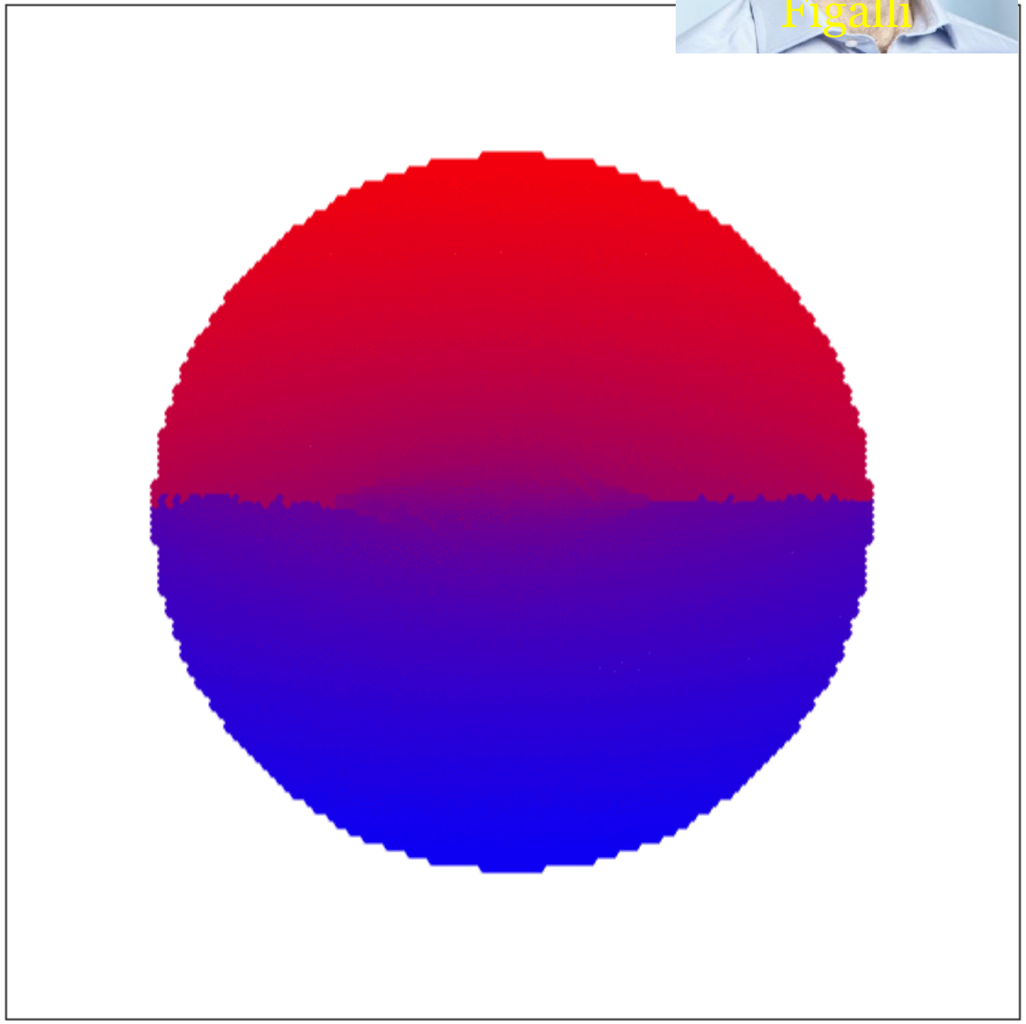




Regularity Theory

→ Regularity of T requires convex target.





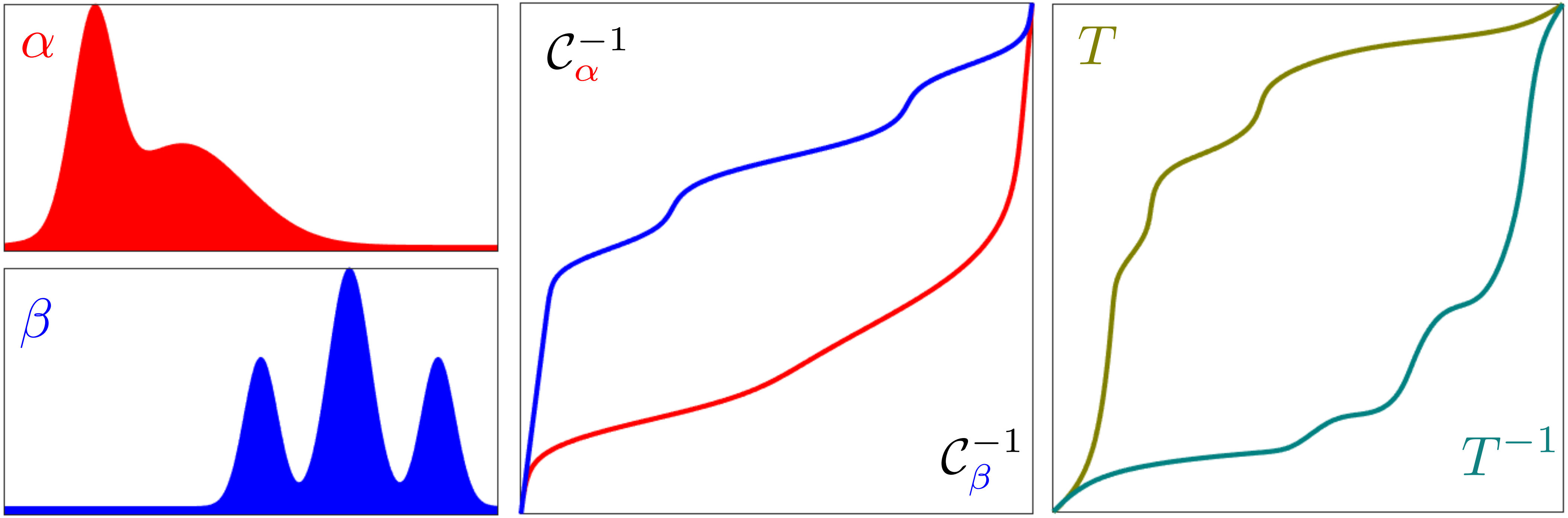
1-D Optimal Transport

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

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

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

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



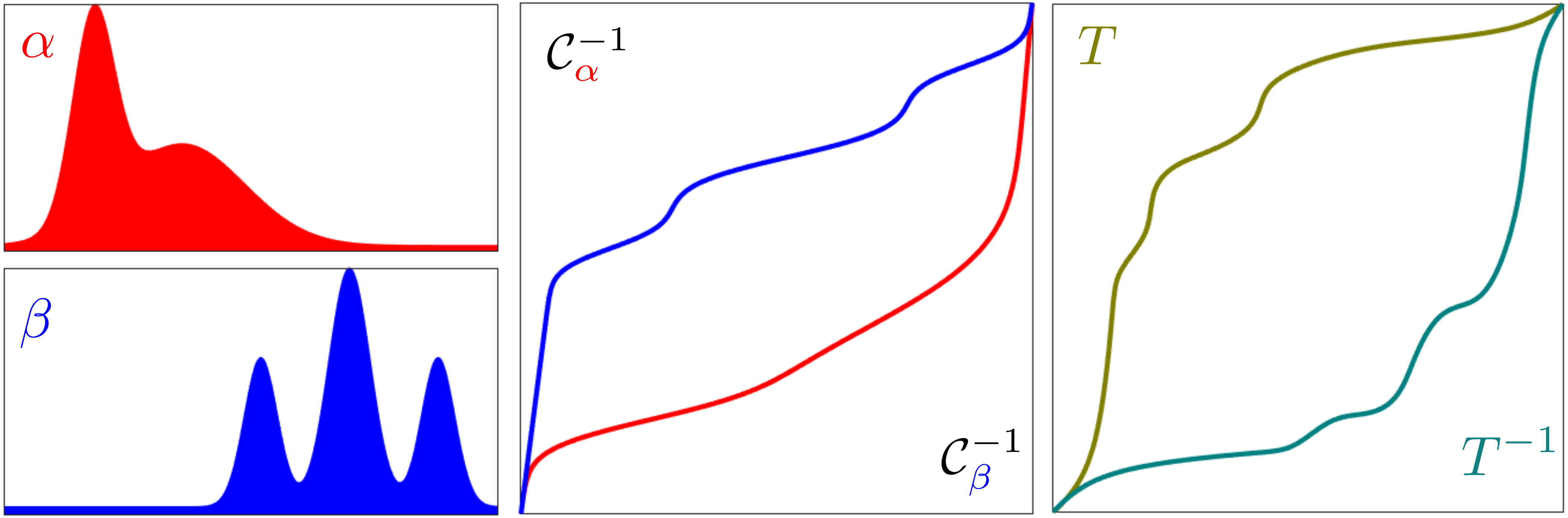
1-D Optimal Transport

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

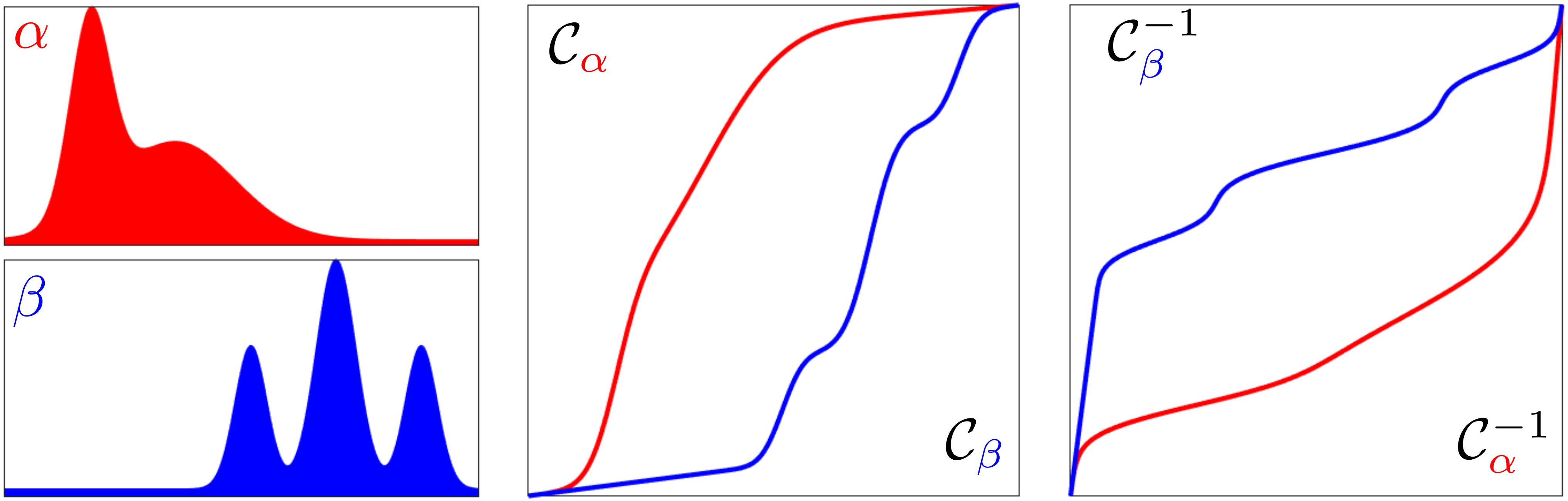
Cumulative function: $C_{\alpha\sharp} : \alpha \longmapsto \mathcal{U}_{[0,1]}$

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

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



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



Wasserstein distance: $W_p(\alpha,\beta)^p = \int_0^1 |\mathcal{C}_\alpha^{-1}(t) - \mathcal{C}_\beta^{-1}(t)|^p \mathrm{d}t$

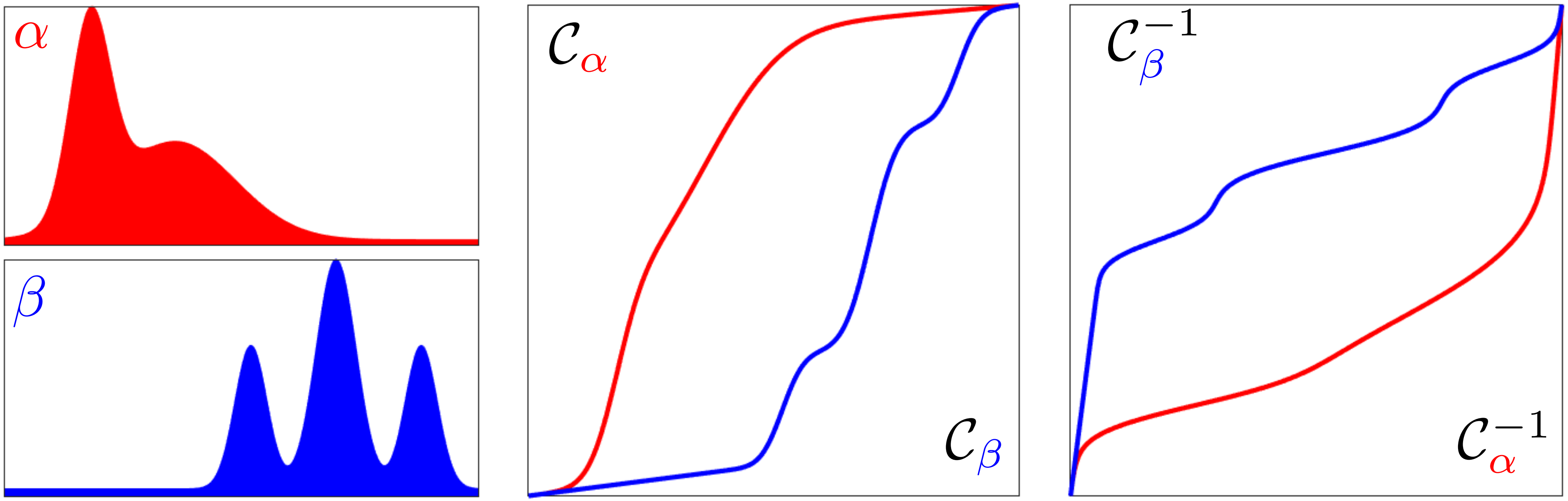
$$W_1(\alpha,\beta) = \|\alpha - \beta\|_{W_1} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| \mathrm{d}x$$

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

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



Wasserstein distance: $W_p(\alpha,\beta)^p = \int_0^1 |\mathcal{C}_\alpha^{-1}(t) - \mathcal{C}_\beta^{-1}(t)|^p \mathrm{d}t$

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

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

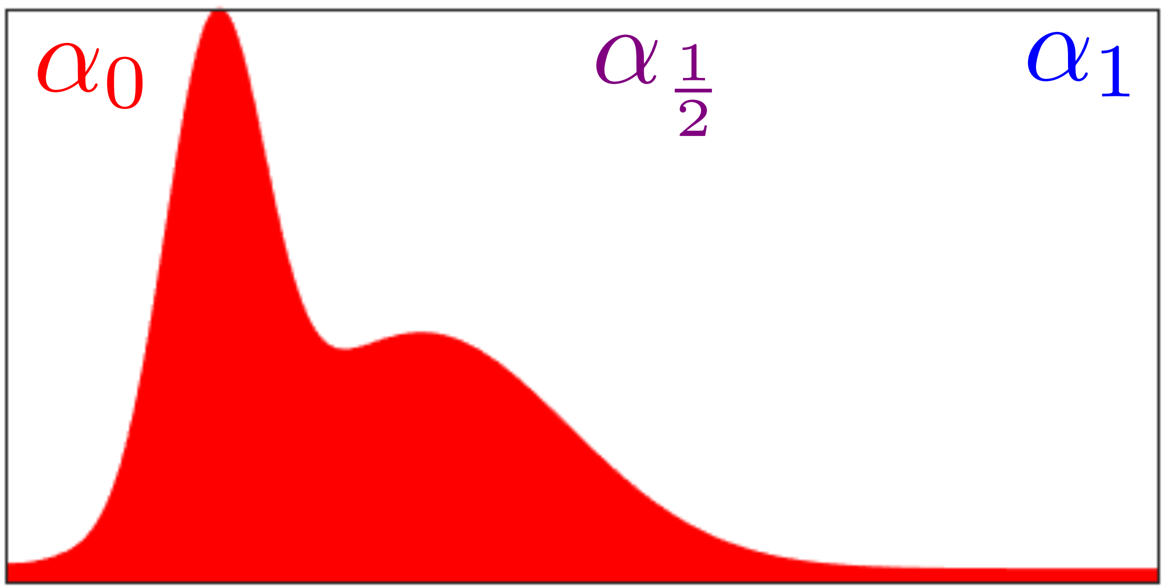
Area under the curve: $\mathrm{AUC}(\alpha,\beta) = 1 - \int_0^1 \mathcal{C}_\alpha \circ \mathcal{C}_\beta{}^{-1}(x)$

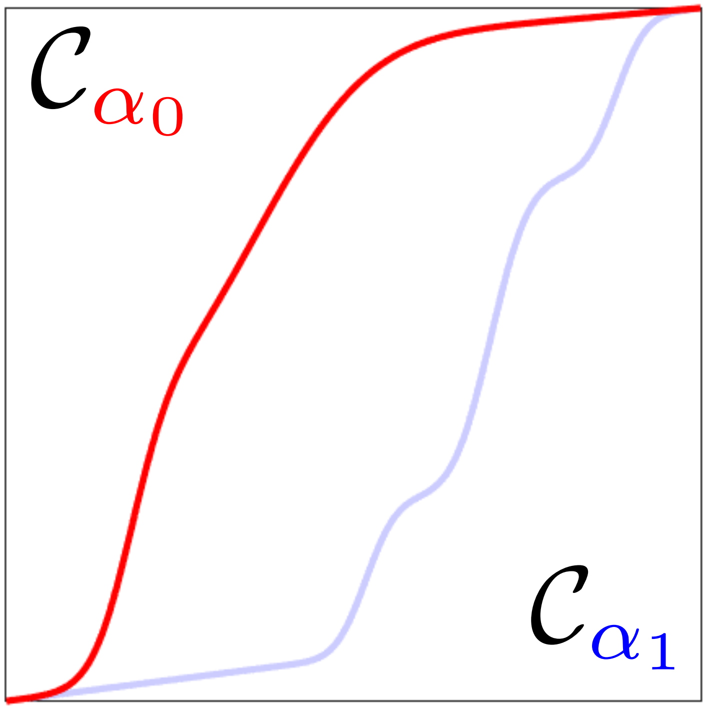
1-D Optimal Transport Interpolation

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

Optimal transport interpolation $\alpha_0 \leftrightarrow \alpha_1$

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



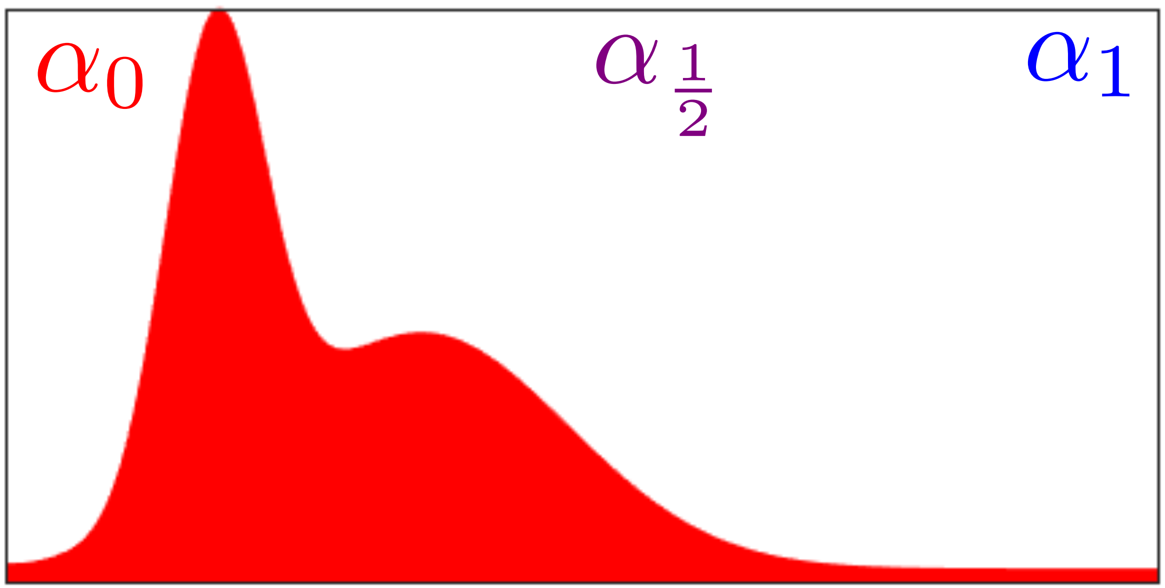


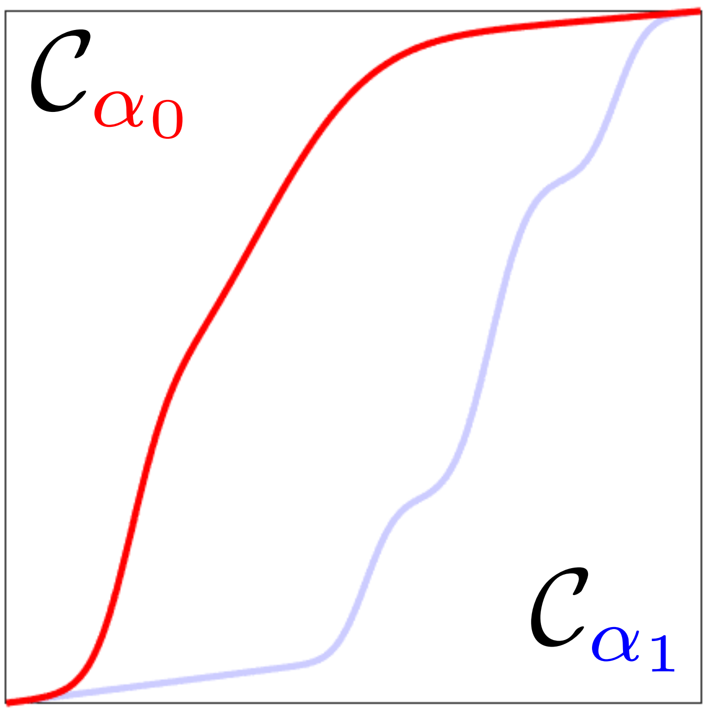
1-D Optimal Transport Interpolation

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

Optimal transport interpolation $\alpha_0 \leftrightarrow \alpha_1$

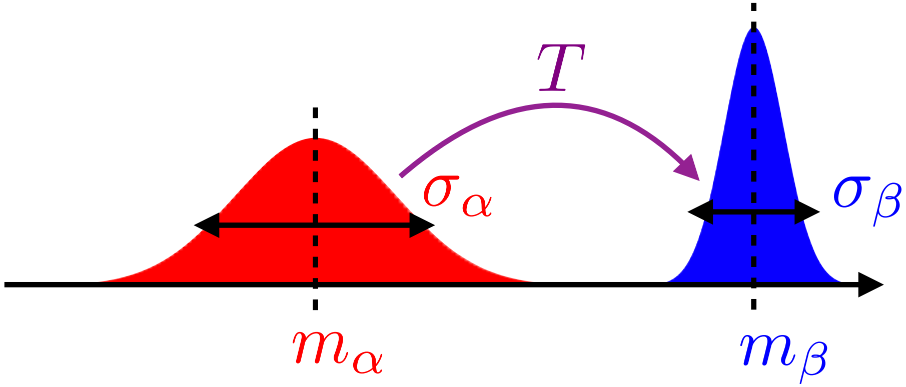
$$\forall\, t \in [0,1], \mathcal{C}_{\alpha_t}^{-1} = (1-t)\mathcal{C}_{\alpha_0}^{-1} + t\mathcal{C}_{\alpha_1}^{-1}$$





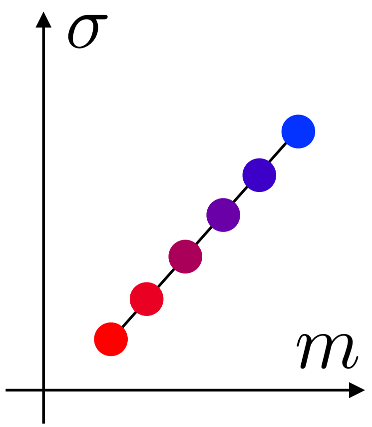
OT Between 1D Gaussians

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



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

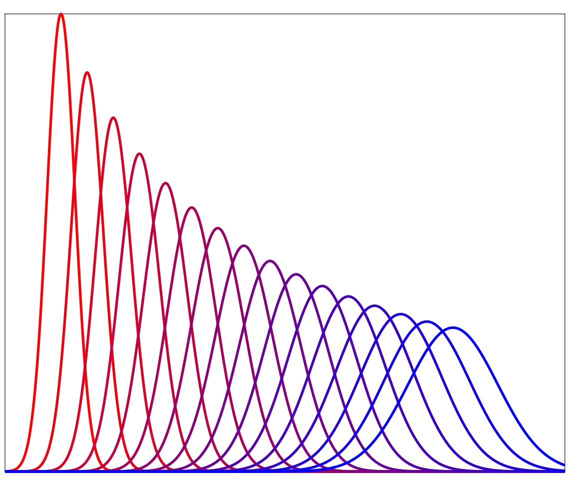
$$\varphi(x) = \frac{\sigma_\beta}{2\sigma_\alpha}(x - m_\alpha)^2 + m_\beta x$$



$T = \nabla\varphi$ $\quad$ φ is convex.

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

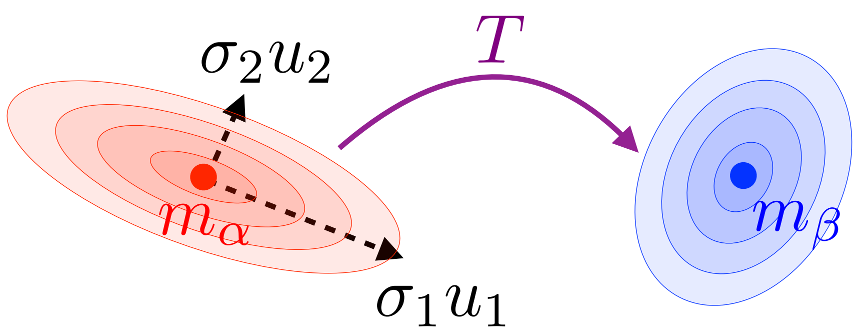
$$\mathrm{W}_2^2(\alpha, \beta) = (m_\alpha - m_\beta)^2 + (\sigma_\alpha - \sigma_\beta)^2$$



OT Between Gaussians

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

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



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

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

Proposition: $A \in \mathcal{S}_+^n$

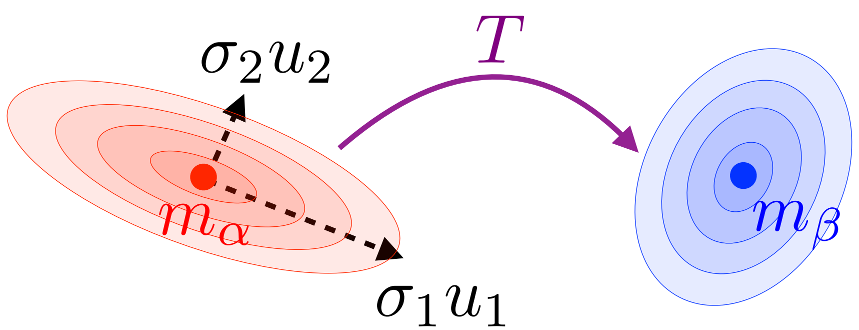
$$A \Sigma_\alpha A = \Sigma_\beta$$

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

OT Between Gaussians

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

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



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

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

Proposition: $A \in \mathcal{S}_+^n$

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

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

$$\overset{\text{Brenier}}{\Longrightarrow}$$

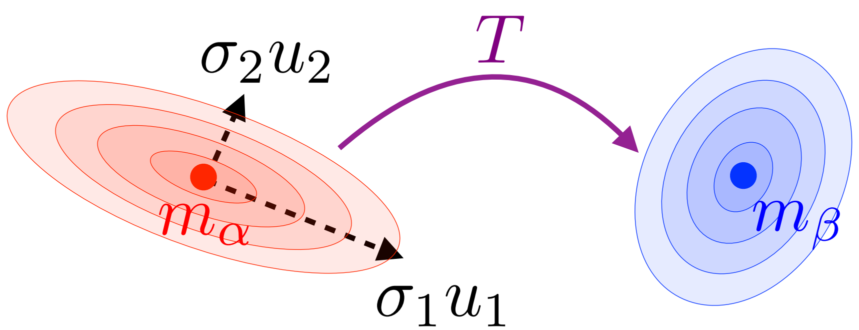
$$T \equiv \mathrm{OT}$$

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

OT Between Gaussians

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

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



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

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

Proposition: $A \in \mathcal{S}_+^n$

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

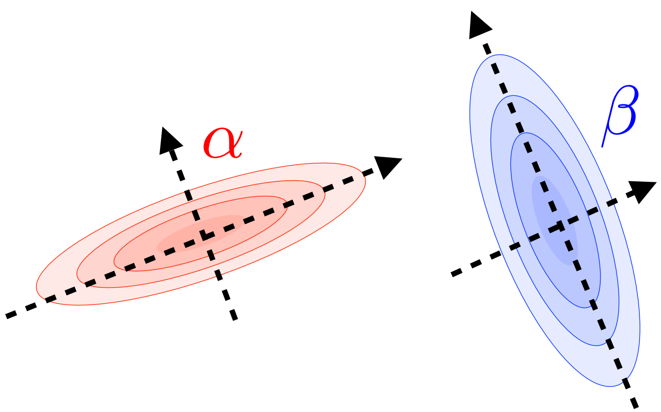
$$\mathrm{W}_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2$$

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

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

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

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

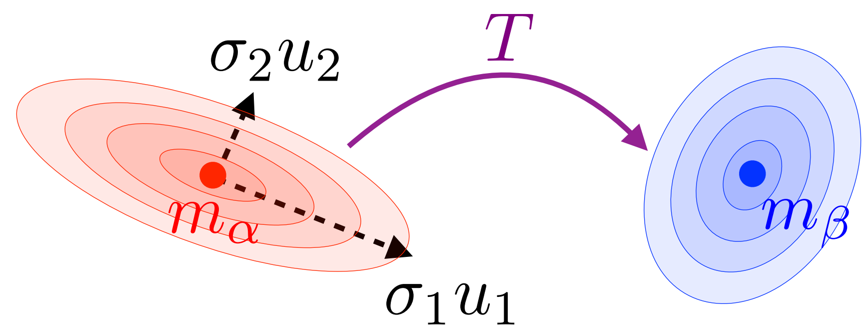


Interpolation Between Gaussians

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

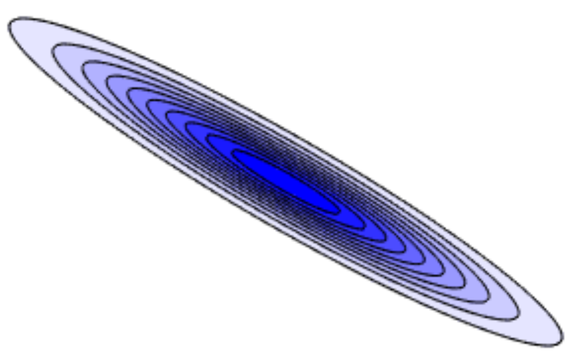
$$T(x) = A(x - m_\alpha) + m_\beta$$

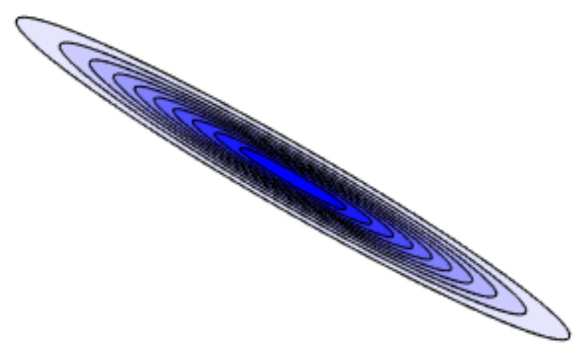
$$A = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}}$$



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

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



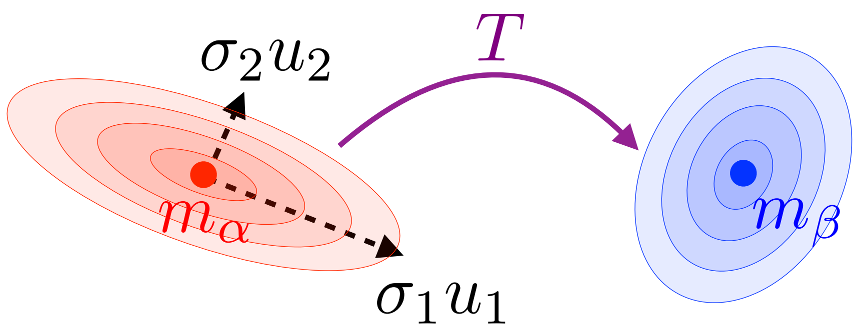


Interpolation Between Gaussians

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

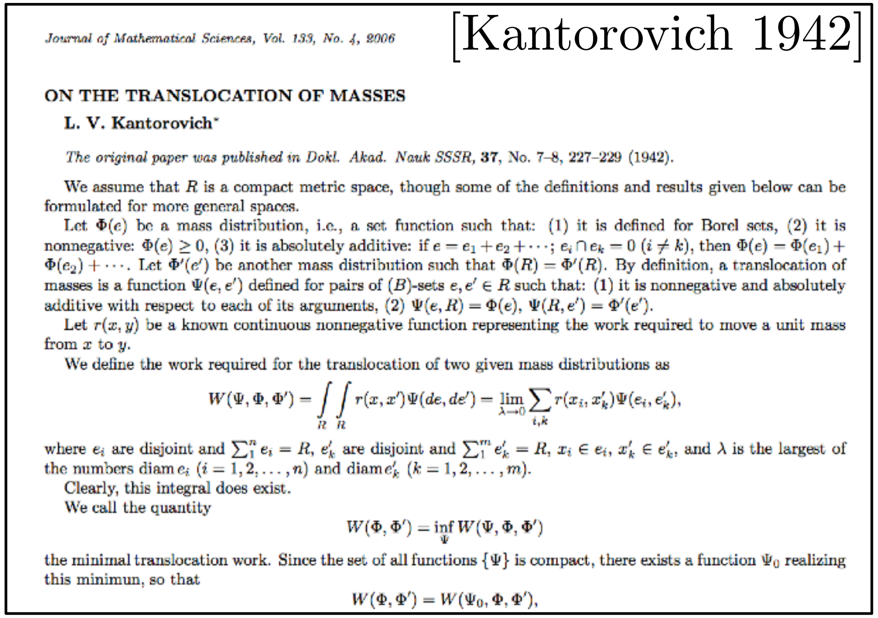
$$m_t = (1-t)m_\alpha + t m_\beta \qquad \Sigma_t = [(1-t)\mathrm{Id} + tA]\Sigma_\alpha[(1-t)\mathrm{Id} + tA]$$

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

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

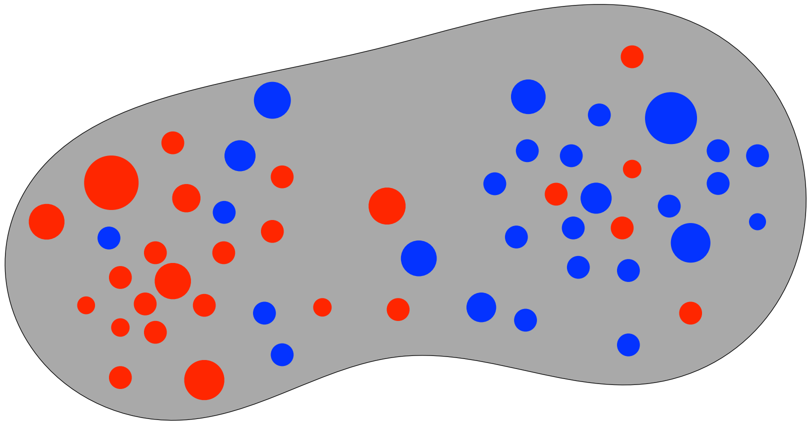
the minimal translocation work. Since the set of all functions $\{\Psi\}$ is compact, there exists a function Ψ_0 realizing this minimun, so that

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

Kantorovitch's Formulation

Input distributions

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



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

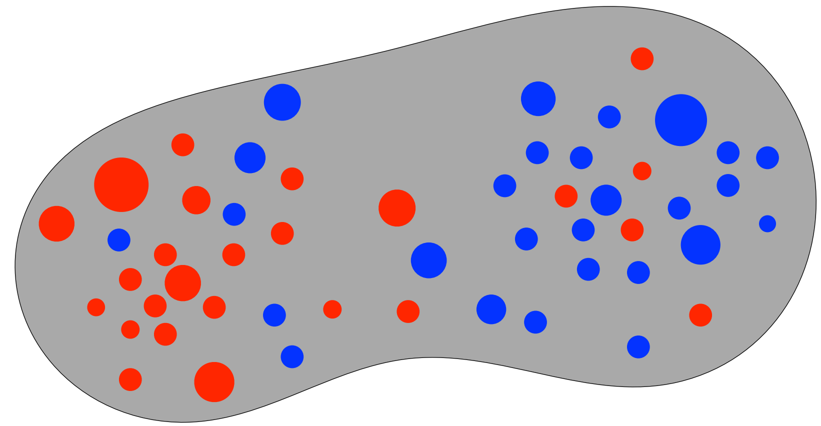
Weights $\mathbf{a}_i \geqslant 0$, $\mathbf{b}_j \geqslant 0$.

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

Kantorovitch's Formulation

Input distributions

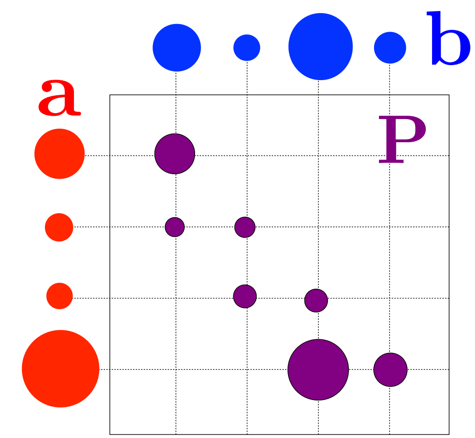
$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$



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

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

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



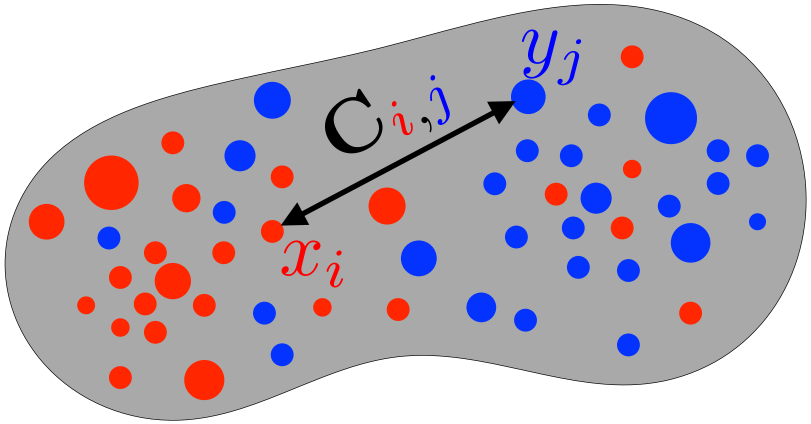
Couplings:

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

Kantorovitch's Formulation

Input distributions

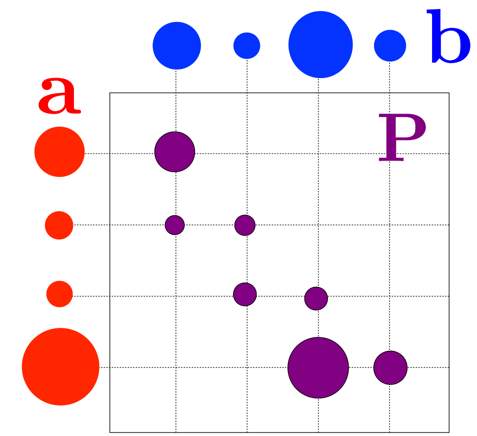
$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i} \quad \beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$



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

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

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



Couplings:

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

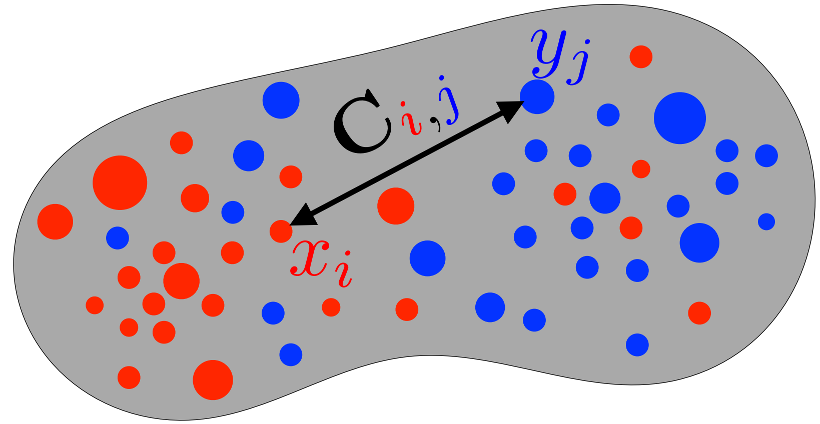
[Kantorovich 1942]

$$\min \left\{ \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j} \; ; \; \mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b}) \right\}$$

Kantorovitch's Formulation

Input distributions

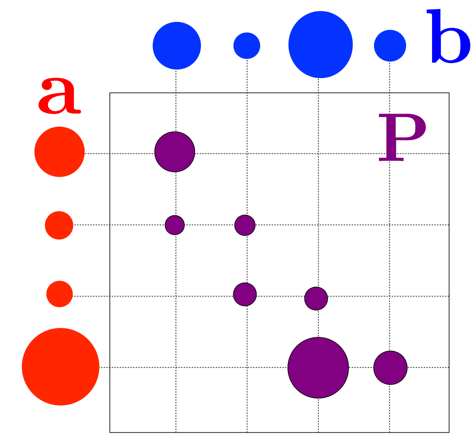
$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$



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

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

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

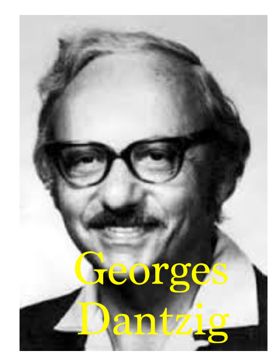


Couplings:

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

[Kantorovich 1942]

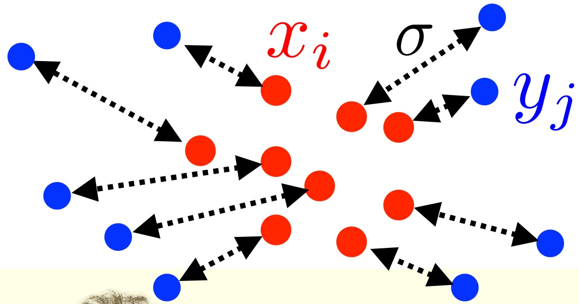
$$\min \left\{ \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j} \; ; \; \mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b}) \right\}$$

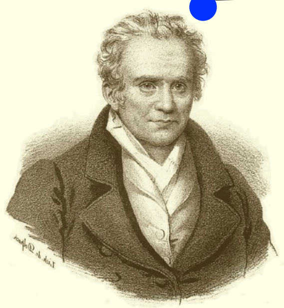


$\rightarrow$ Linear program, simplex $O(n^3 \log(n))$.

Kantorovitch's Exact Relaxation

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



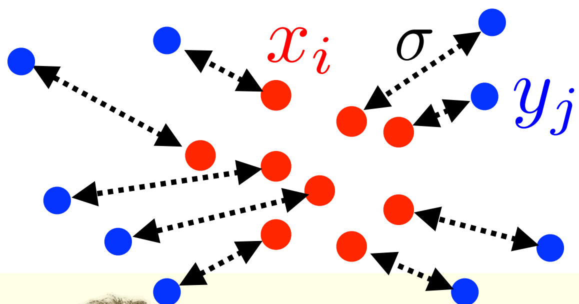


Monge (1784):

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

Kantorovich’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):

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

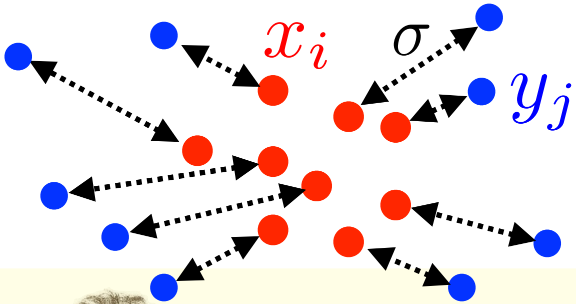
$\geqslant$ (relaxation)

Kantorovich (1942):

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

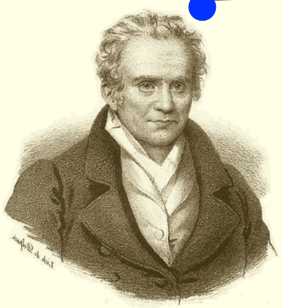
Kantorovitch's Exact Relaxation

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



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

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



Monge (1784):

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

$\geqslant$ (relaxation)

Kantorovitch (1942):

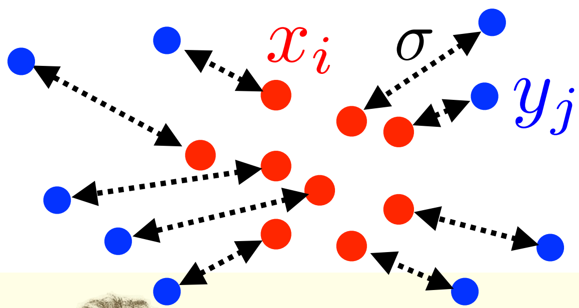
$$\min_{\mathbf{P} \in \text{Bist}_n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{P}_{i,j} \mathbf{C}_{i,j}$$

$n!$ permutations

$O(n^3)$ algorithm

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:

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

Monge (1784):

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

$\geqslant$ (relaxation)

Kantorovitch (1942):

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

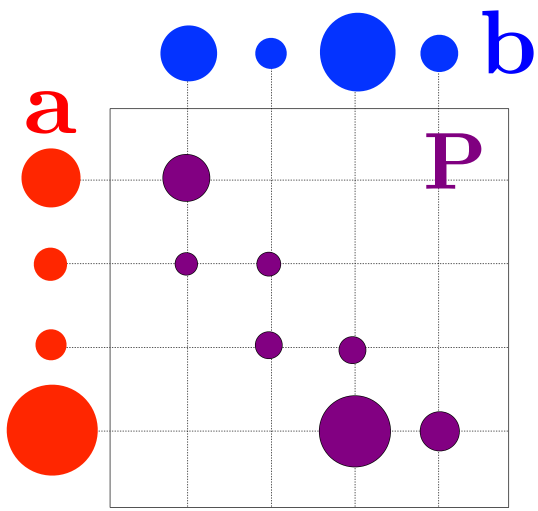
$n!$ permutations

$O(n^3)$ algorithm

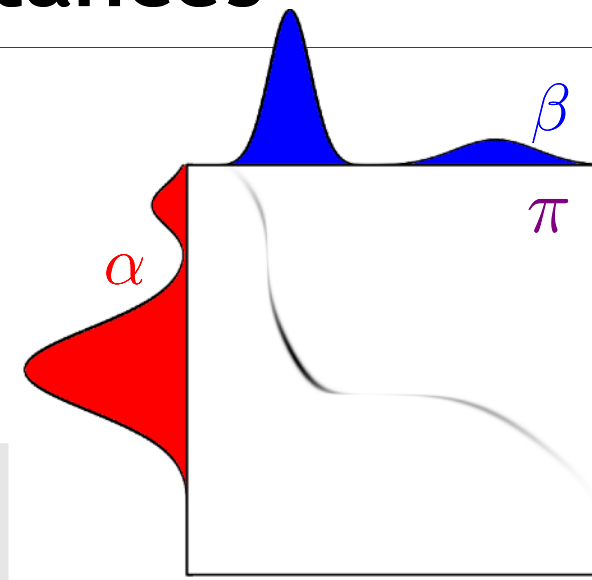
Theorem: [Birkhoff-von Neumann]

"Monge $\Leftrightarrow$ Kantorovitch"

Optimal Transport Distances



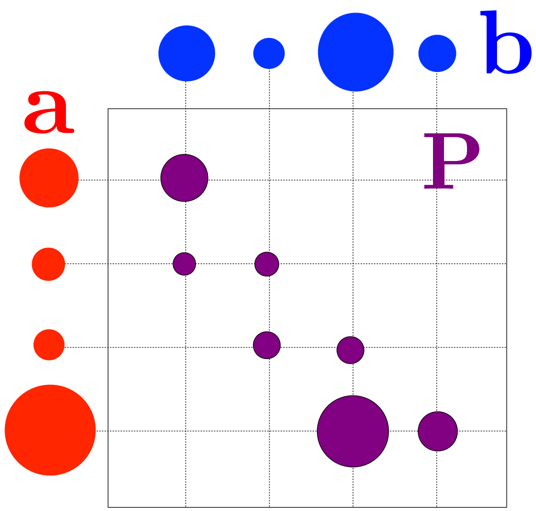
$$\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}$$



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

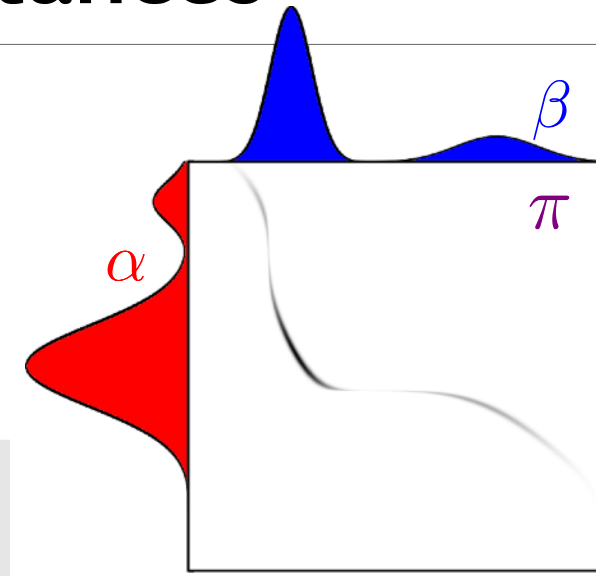
$$\mathrm{W}_p(\alpha, \beta)^p \overset{\text{def.}}{=} \min_{\pi \in \mathcal{M}_+^1(\mathcal{X}^2)} \left\{ \int_{\mathcal{X}^2} d(x, y)^p \mathrm{d}\pi(x, y) \; ; \; \pi_1 = \alpha, \pi_2 = \beta \right\}$$

Optimal Transport Distances



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

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



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

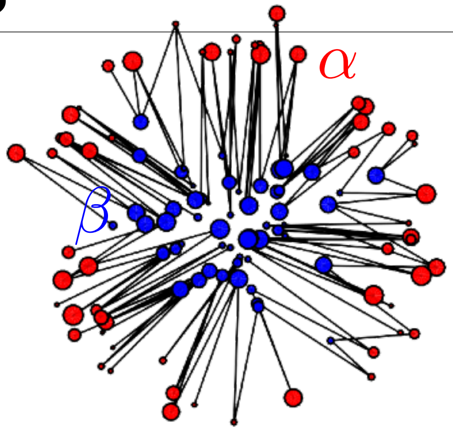
Theorem: W_p is a distance and $\alpha_n \rightharpoonup \beta \Leftrightarrow \mathrm{W}_p(\alpha_n, \beta) \to 0$

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

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

A Glimpse at Algorithms

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

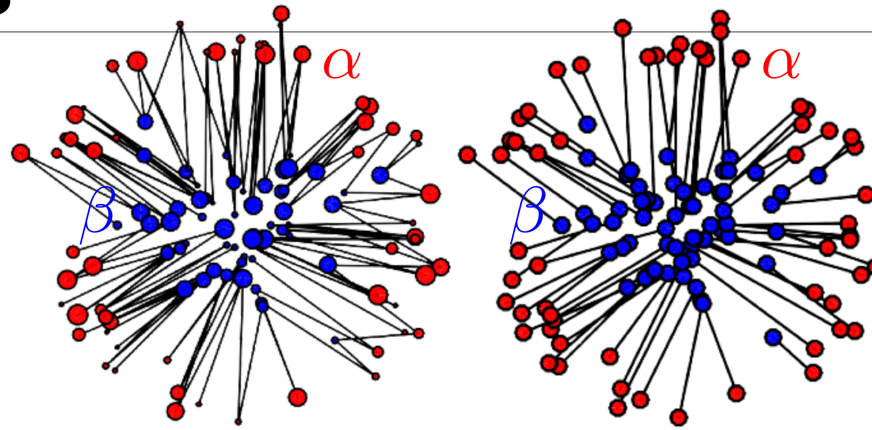


A Glimpse at Algorithms

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

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

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



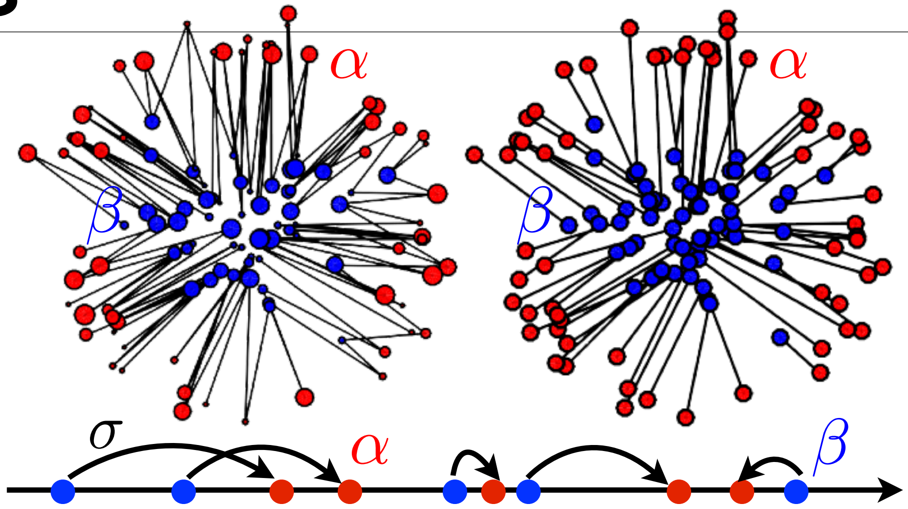
A Glimpse at Algorithms

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

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

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

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

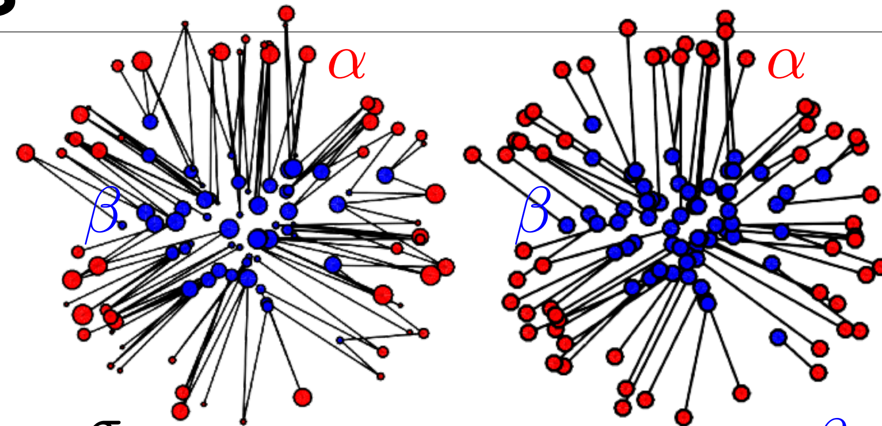


A Glimpse at Algorithms

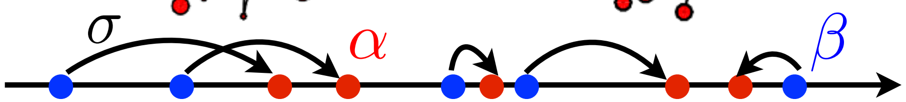
Linear programming: $O(n^3 \log(n)^2)$

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

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



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 $O(n^2 \log(n))$.

A Glimpse at Algorithms

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

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

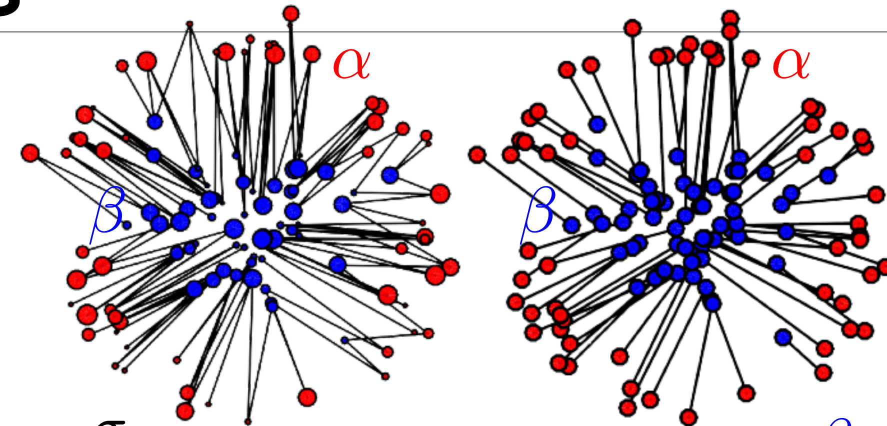
$\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \qquad \beta = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$

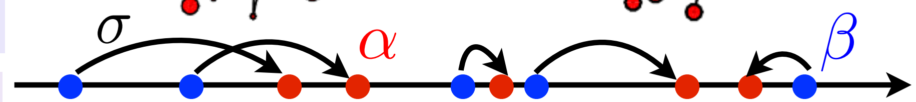
1-D case: sorting $O(n \log(n))$.

$p = 1$
$d = \| \cdot \|$
$$W_1(\alpha, \beta) = \min_{\mathrm{div}(u) = \alpha - \beta} \int \|u(x)\| \mathrm{d}x$$

$\rightarrow$ 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: $O(n^3)$

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

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

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

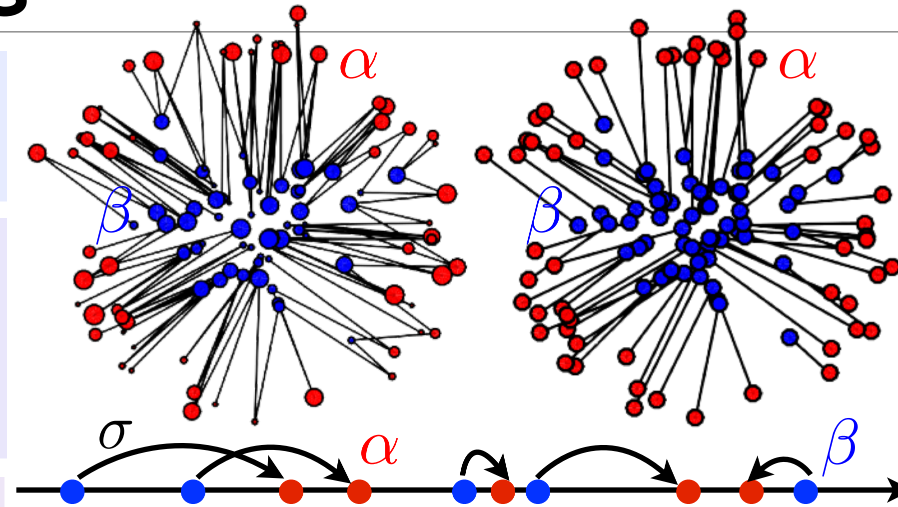
$$W_1(\alpha,\beta) = \min_{\mathrm{div}(u)=\alpha-\beta} \int \|u(x)\| \mathrm{d}x$$

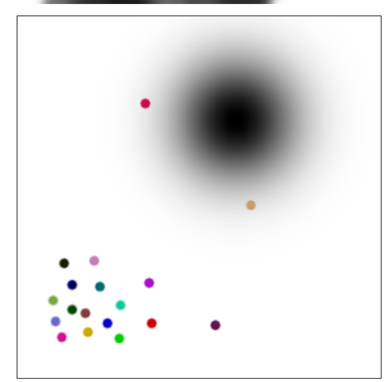
$\rightarrow$ min-cost flow, on graphs $O(n^2\log(n))$.

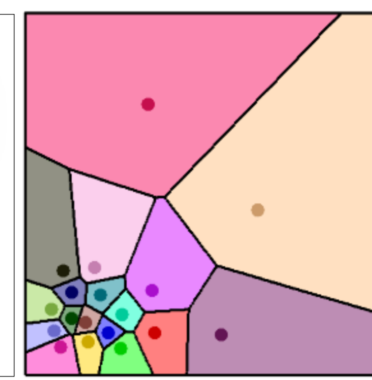
Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.

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

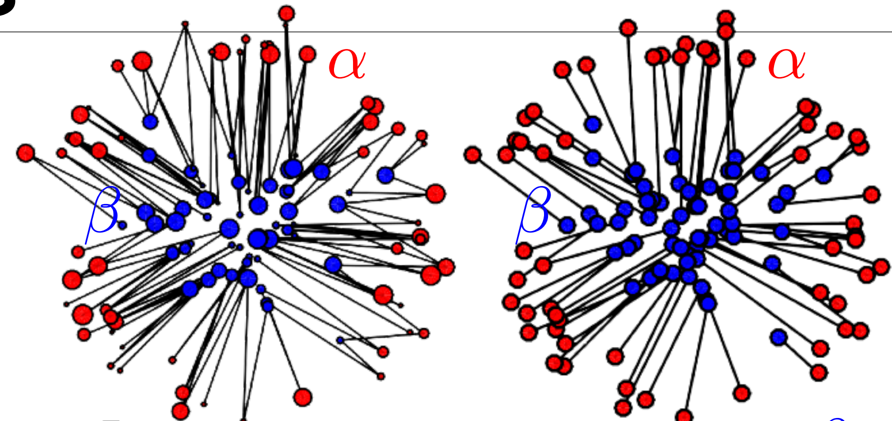
[Merigot 2013]

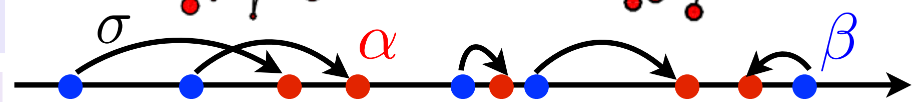






A Glimpse at Algorithms





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

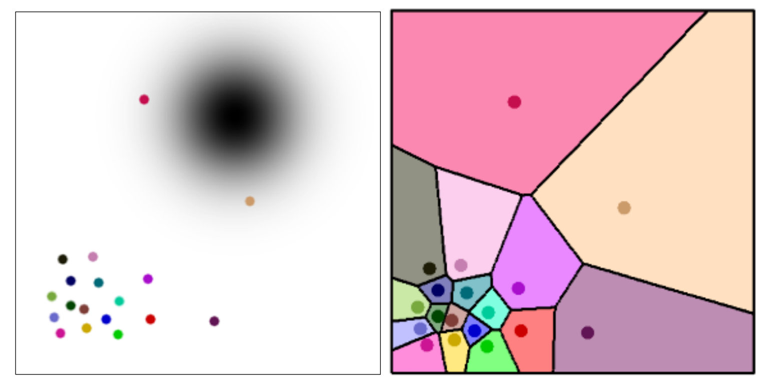
Hungarian/Auction: $O(n^3)$

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

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

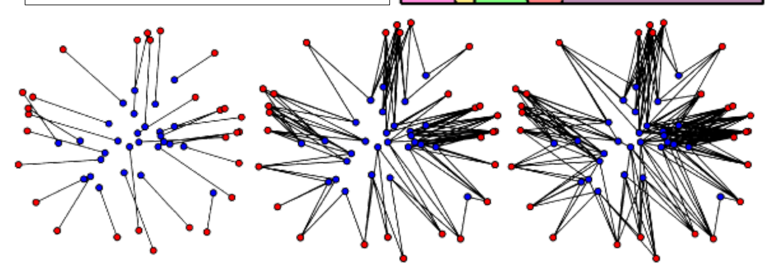
$p = 1$
$d = \| \cdot \|$ $\quad W_1(\alpha, \beta) = \min_{\mathrm{div}(u) = \alpha - \beta} \int \|u(x)\| \mathrm{d}x$

$\to$ 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.

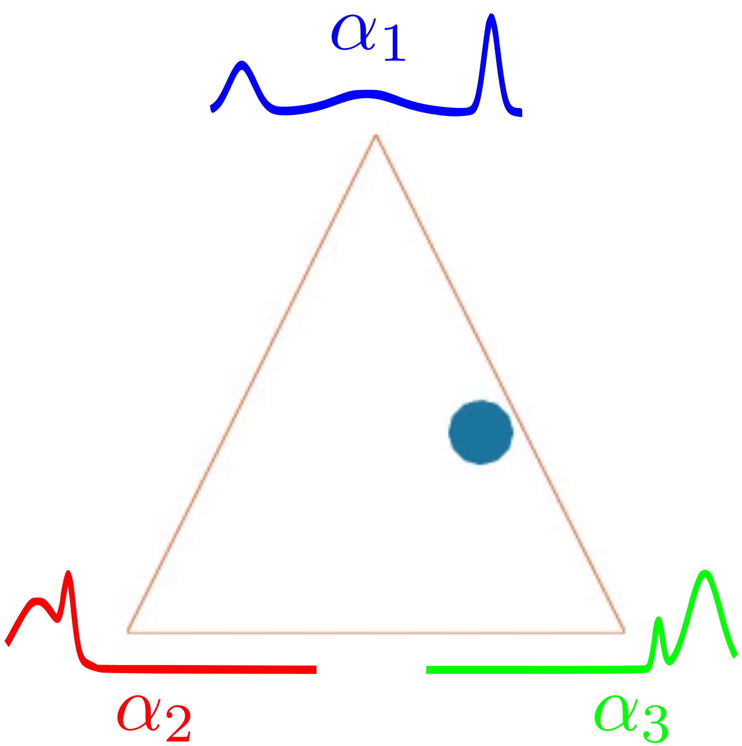
Overview

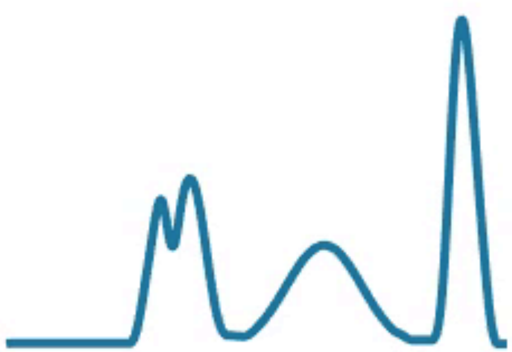
- Monge Formulation
- Continuous Optimal Transport
- Kantorovitch Formulation
- Applications

Wasserstein Barycenters

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

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





$\lambda \in \Sigma_3$

$\min_\alpha \sum_s \lambda_s \mathrm{W}_p^p(\alpha, \alpha_s)$

Wasserstein

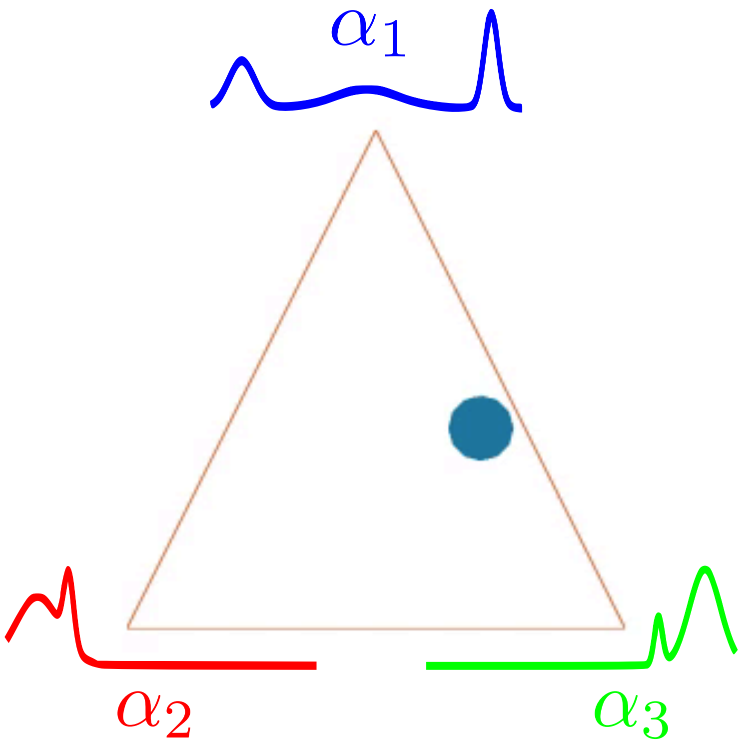
$\sum_s \lambda_s \alpha_s$

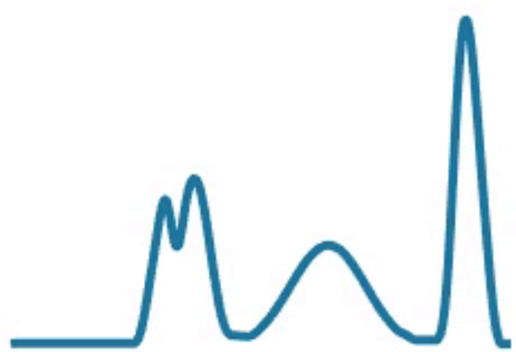
Euclidean

Wasserstein Barycenters

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

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





$\lambda \in \Sigma_3$

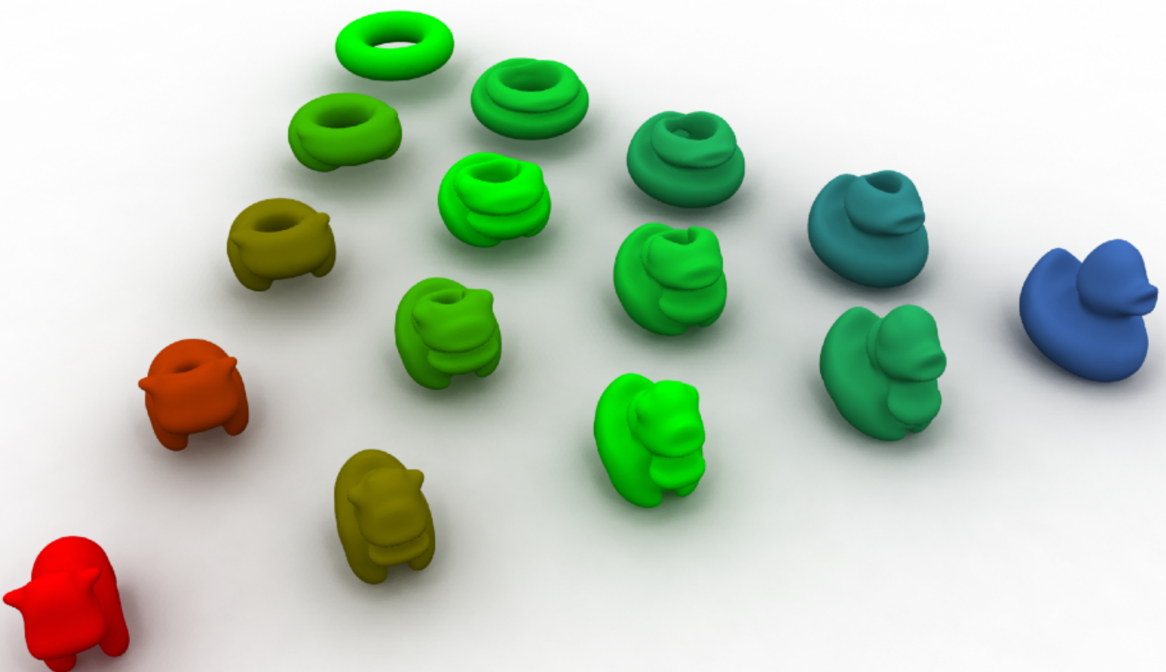
$$\min_{\alpha} \; \sum_s \lambda_s \mathrm{W}_p^p(\alpha, \alpha_s)$$

Wasserstein

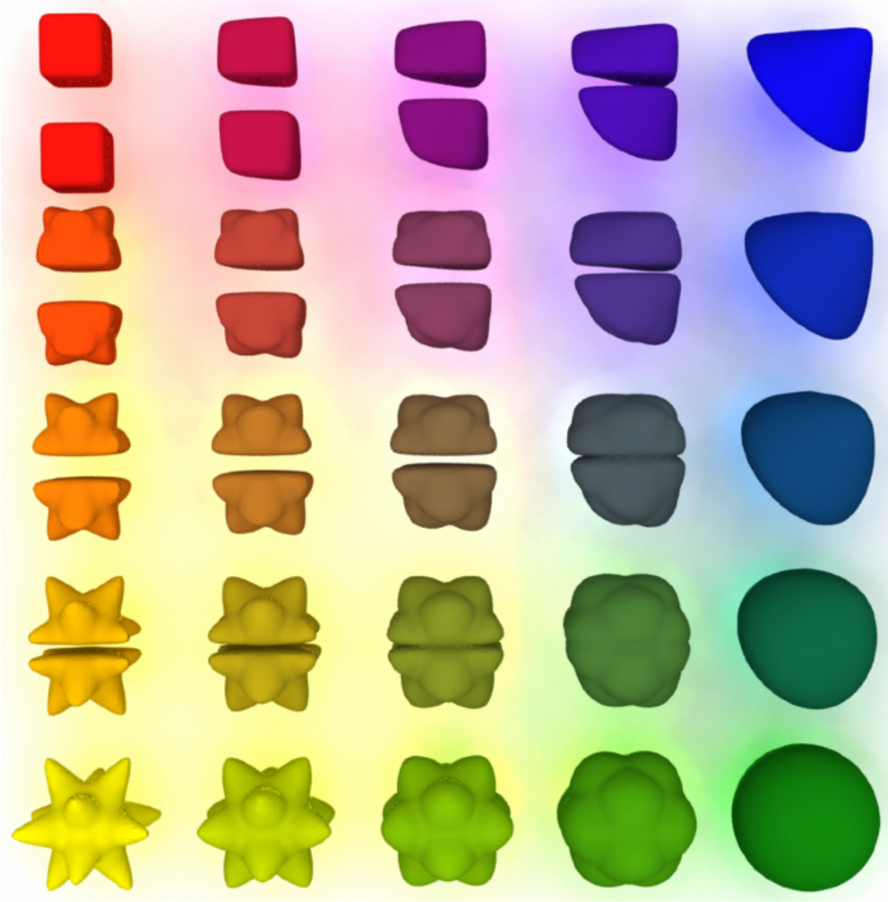
$$\sum_s \lambda_s \alpha_s$$

Euclidean

Examples

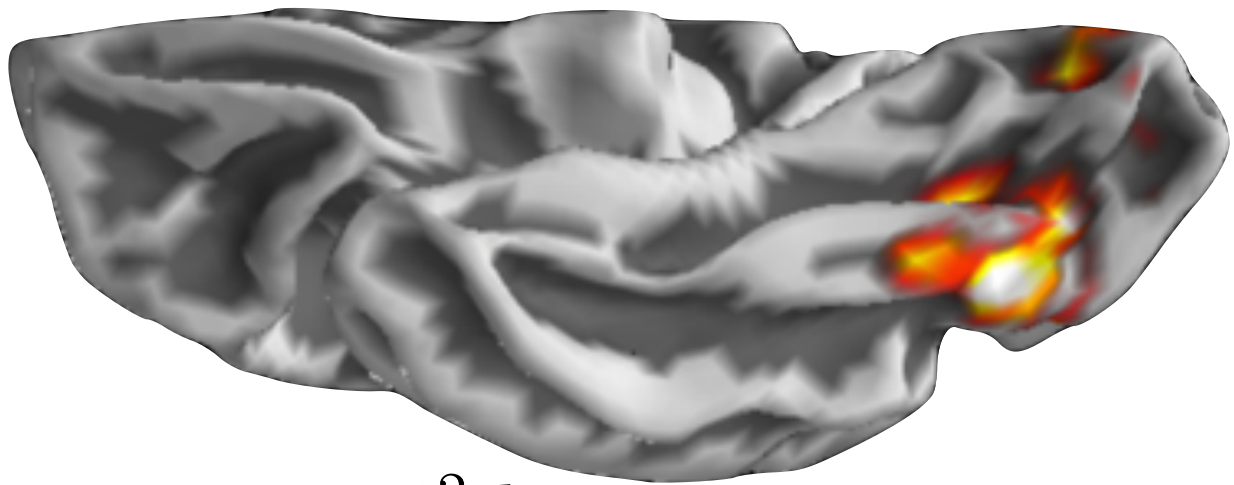


[Solomon et al, SIGGRAPH 2015]

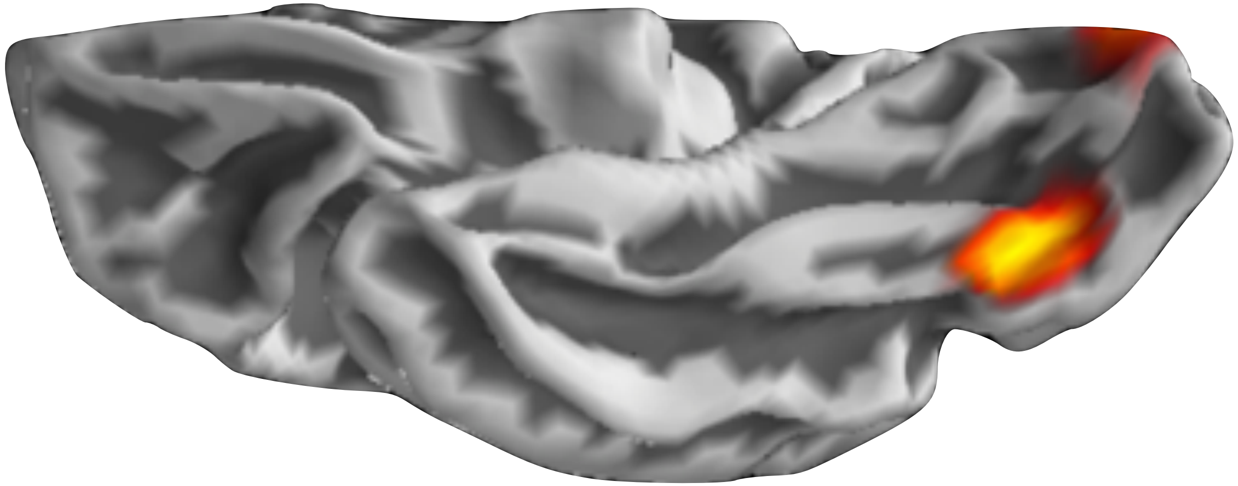


MRI Data Procesing [with A. Gramfort]

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



L^2 barycenter

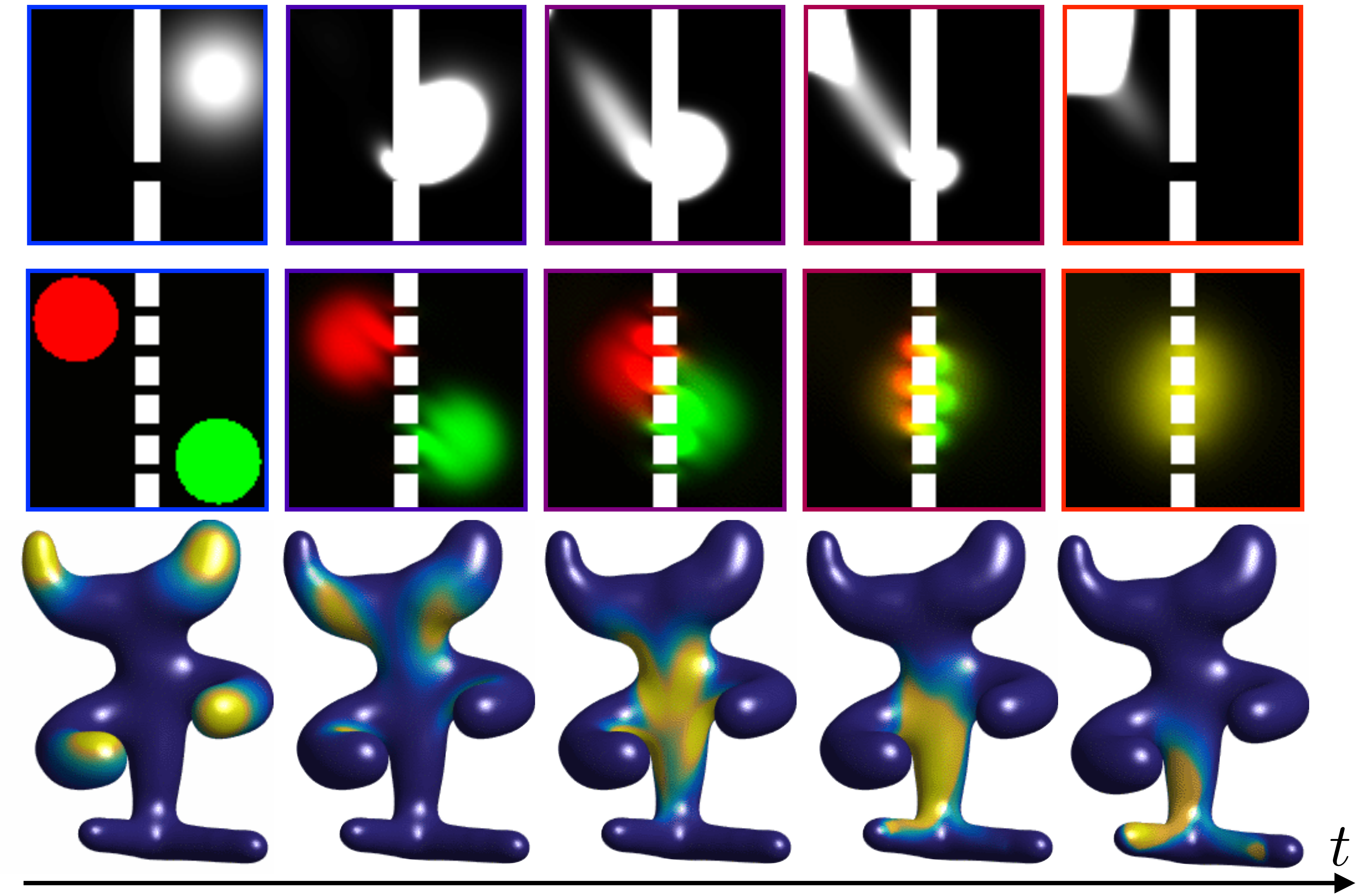


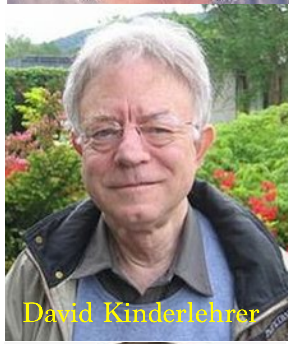
W_2^2 barycenter

Generalizations: Gradient Flows

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

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





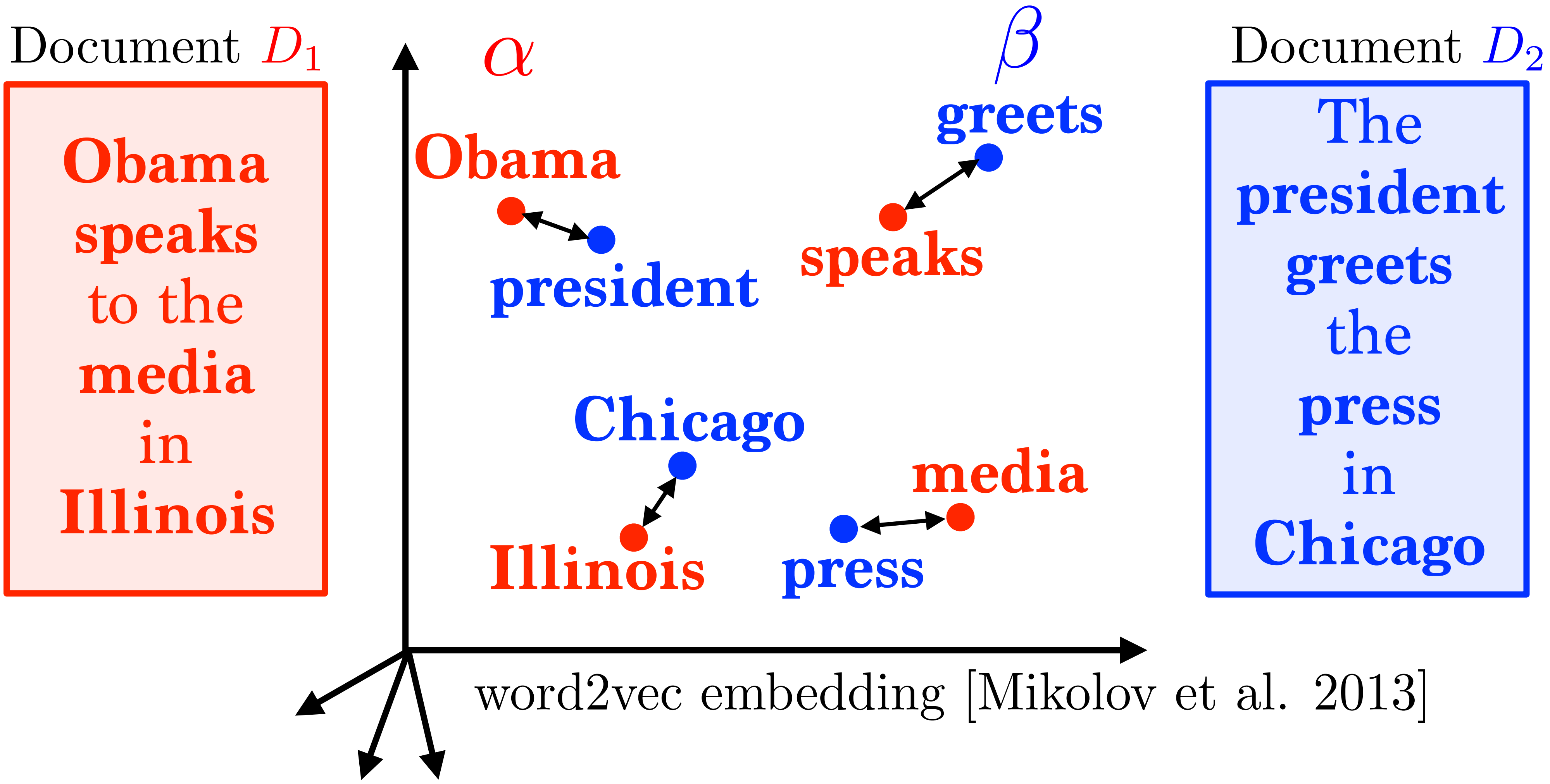
Gradient Flows Simulation

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



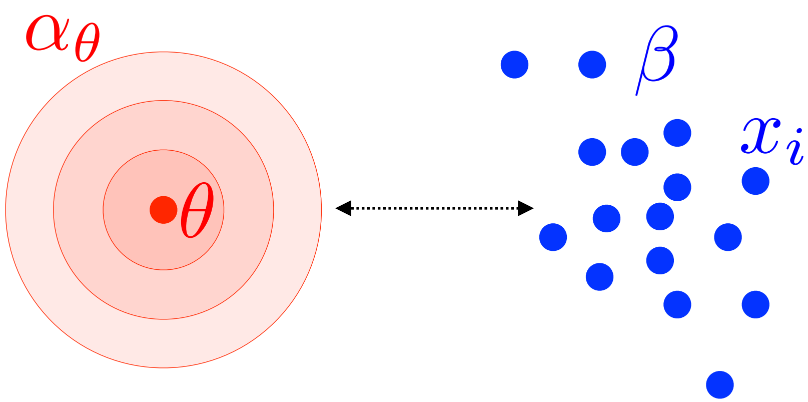
Word mover's distance: [Kusner et al 2015]

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

Density Fitting and Generative Models

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

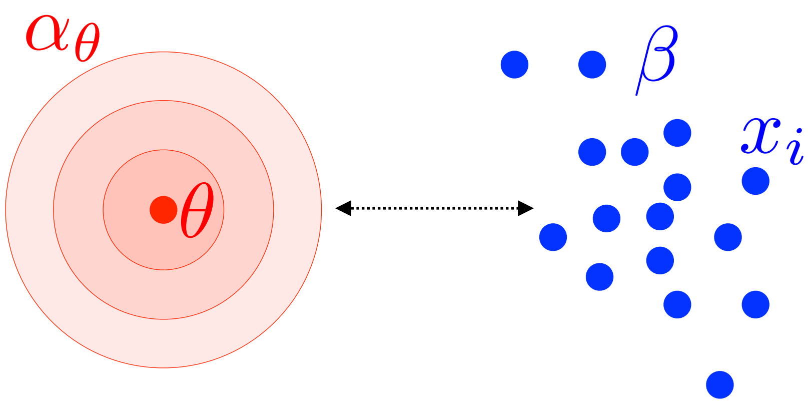
Parametric model: $\theta \mapsto \alpha_\theta$



Density Fitting and Generative Models

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

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



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

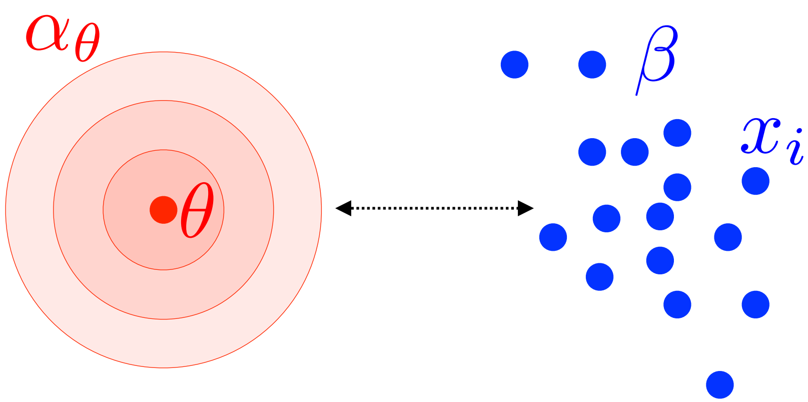
$$\min_\theta \widehat{\mathrm{KL}}(\alpha_\theta|\beta) \overset{\text{def.}}{=} -\sum_i \log(\rho_\theta(x_i))$$

Maximum likelihood (MLE)

Density Fitting and Generative Models

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

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



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

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

Maximum likelihood (MLE)

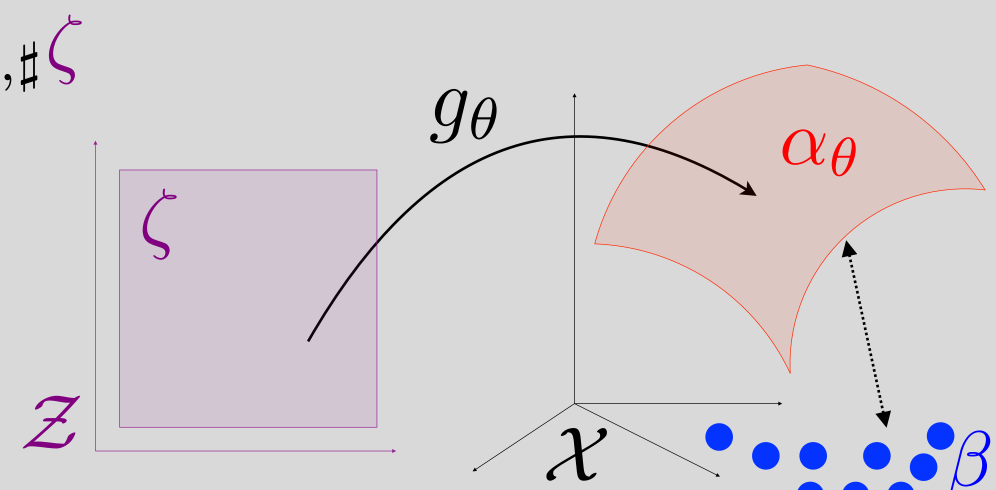
Generative model fit: $\alpha_\theta = g_{\theta,\sharp}\zeta$

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

$\rightarrow$ MLE undefined.

$\rightarrow$ Need a weaker metric.

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

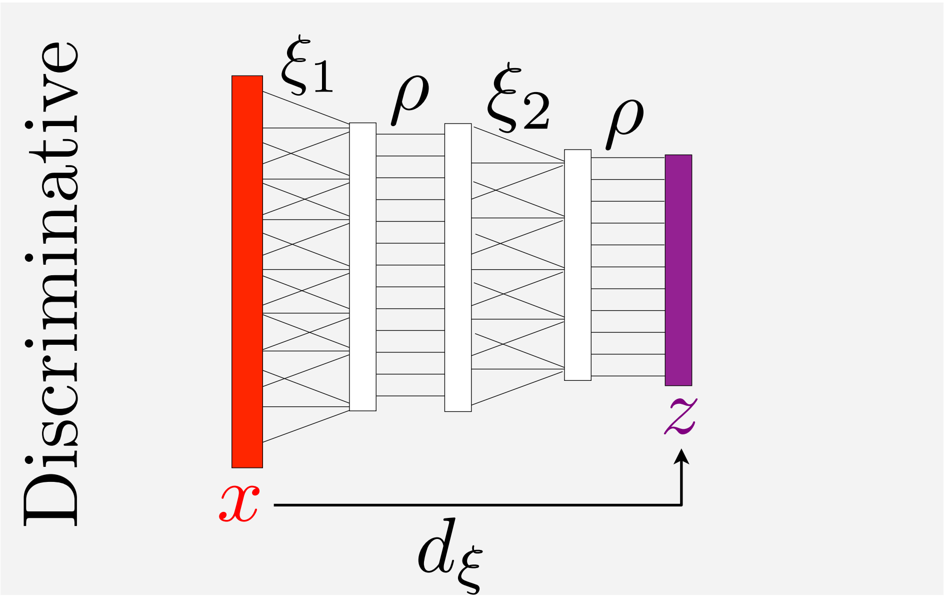


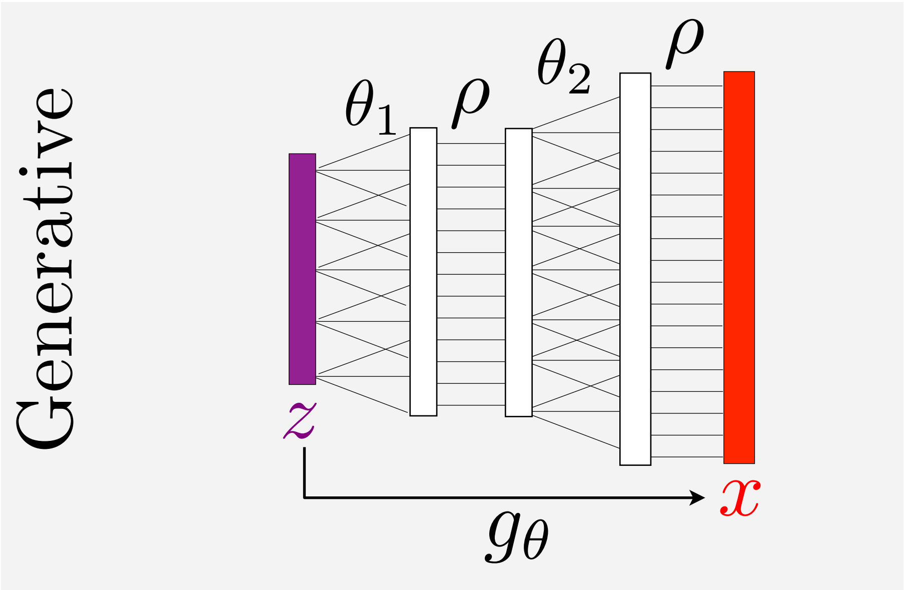
Deep Discriminative vs Generative Models

Deep networks:

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

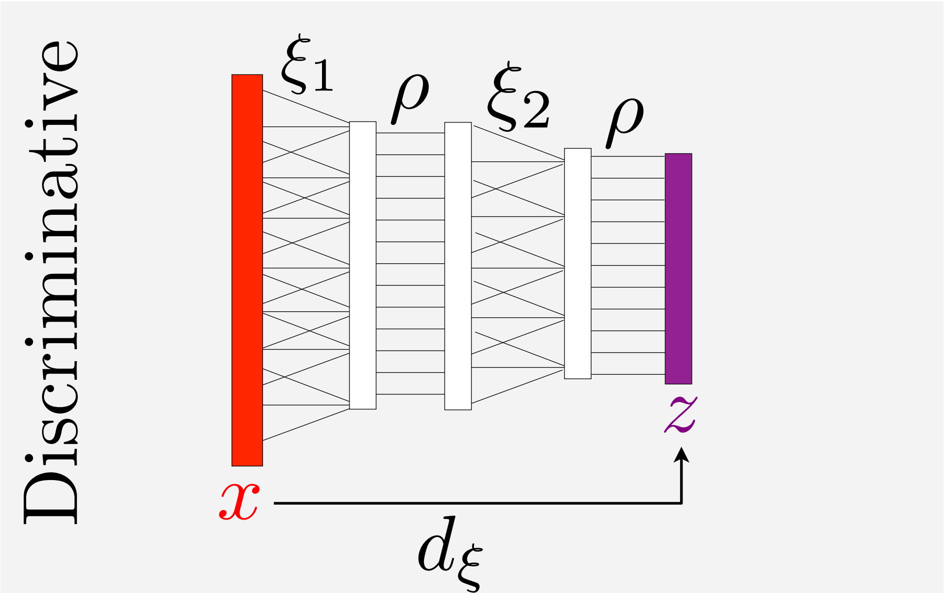
$$g_\theta(z) = \rho(\theta_K(\ldots\rho(\theta_2(\rho(\theta_1(z)\ldots)$$

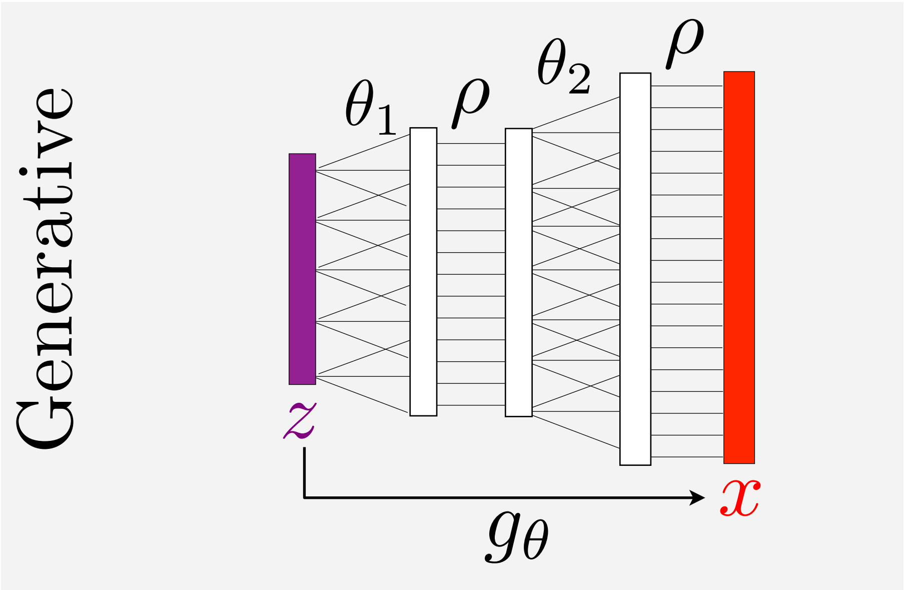


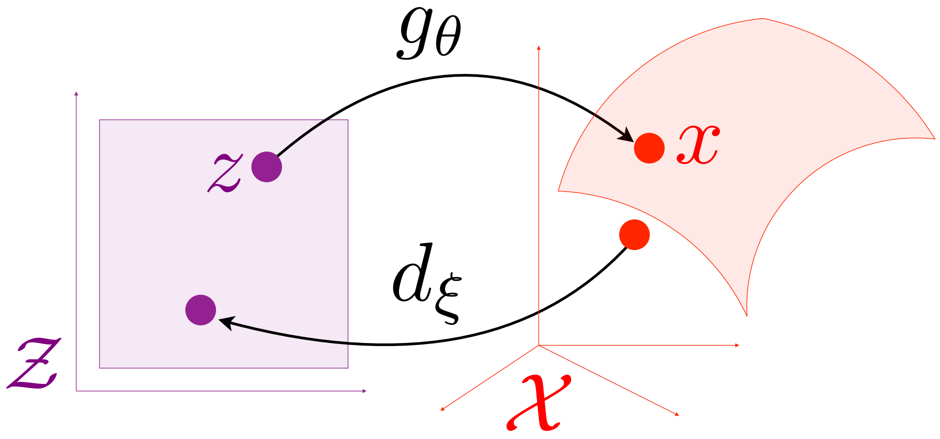


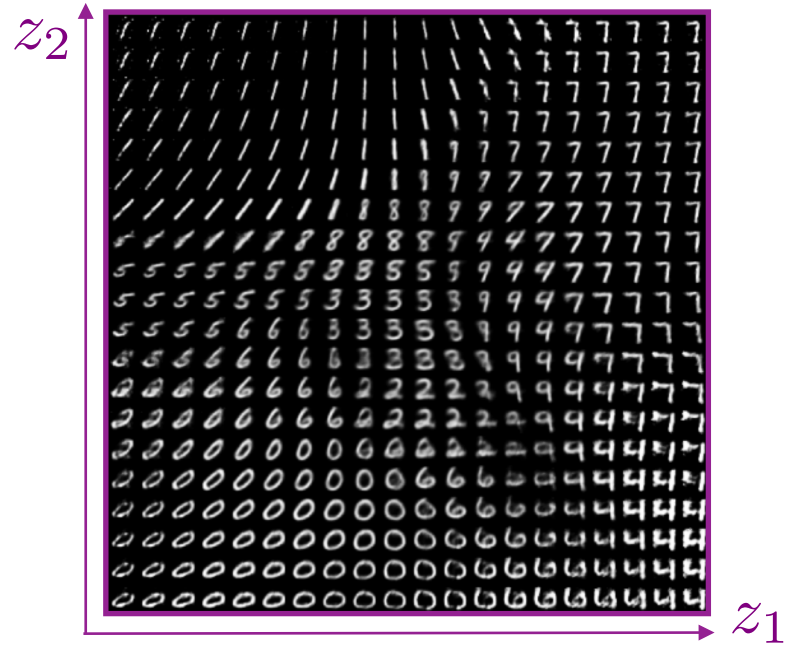
Deep Discriminative vs Generative Models

$$\text{Deep networks:} \quad \begin{cases} d_\xi(x) = \rho(\xi_K(\ldots\rho(\xi_2(\rho(\xi_1(x))\ldots) \\ g_\theta(z) = \rho(\theta_K(\ldots\rho(\theta_2(\rho(\theta_1(z))\ldots) \end{cases}$$





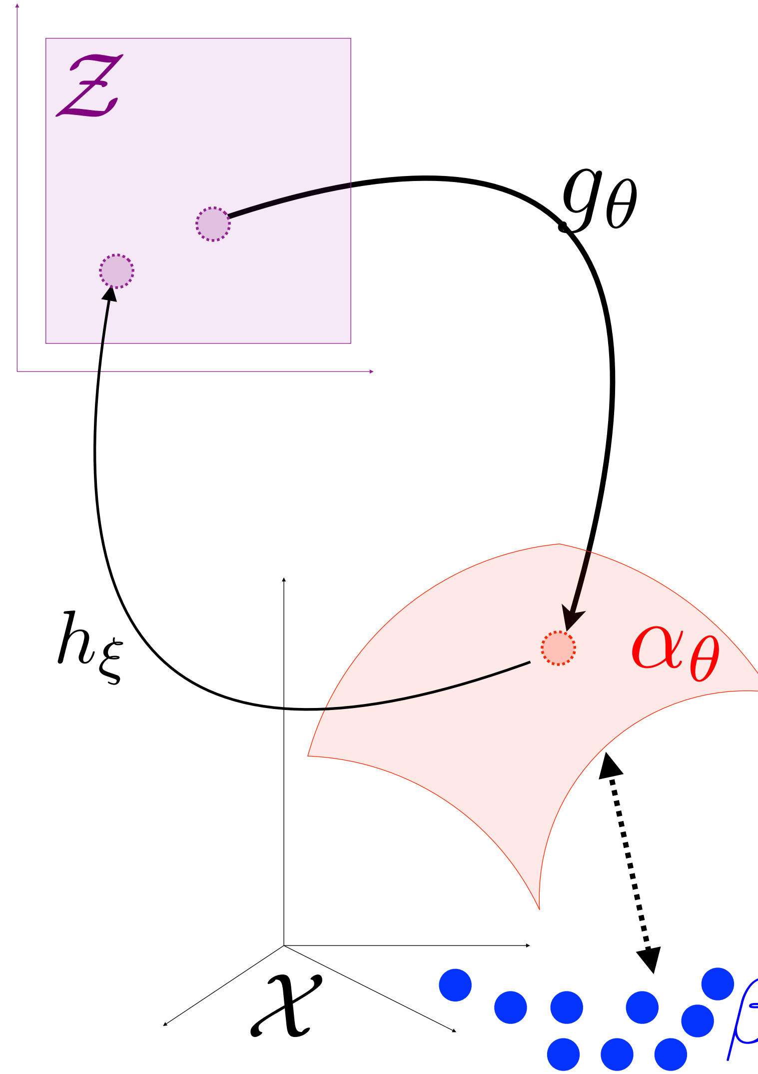




Examples of Images Generation

Inputs β

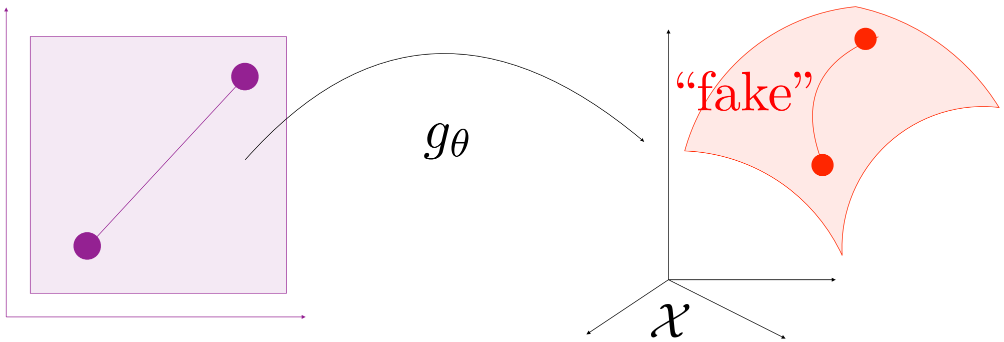
Generated α_θ



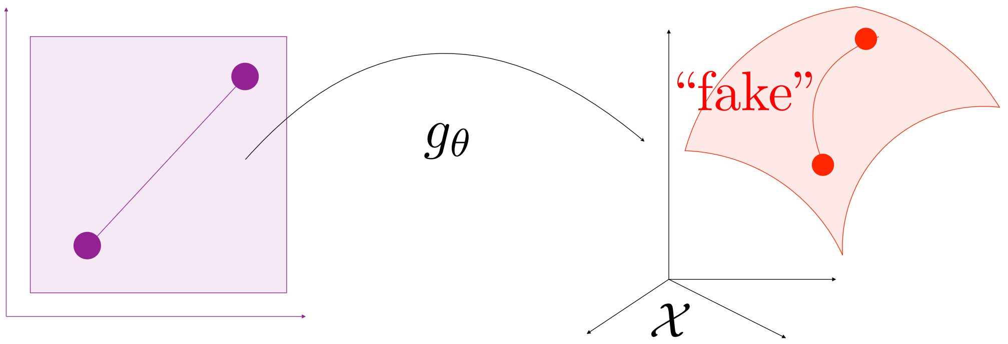
$\rightarrow$ Need to learn the metric $d(x, y) = \|h_\xi(x) - h_\xi(y)\|$ (GANs)

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

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



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



Conclusion: Toward High-dimensional OT