Numerical Optimal Transport

http://optimaltransport.github.io

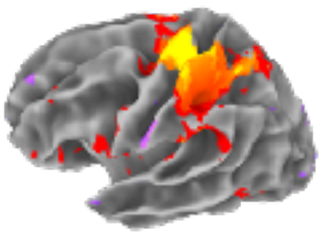
Theoretical Foundations

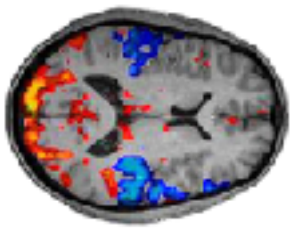
Gabriel Peyré

www.numerical-tours.com

Comparing Measures

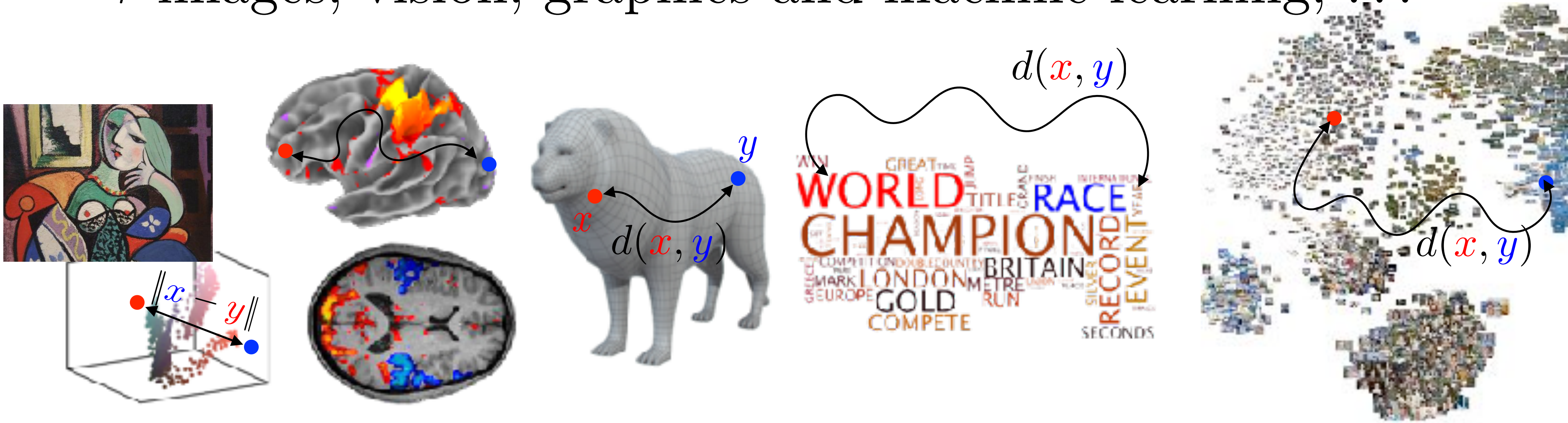
→ images, vision, graphics and machine learning, . . .





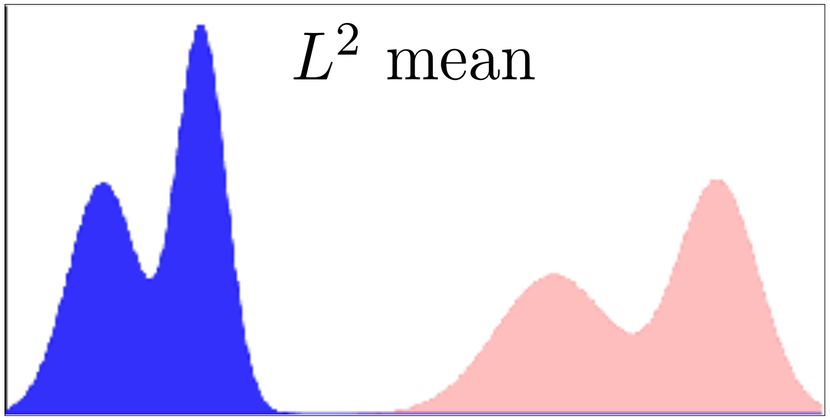
Comparing Measures

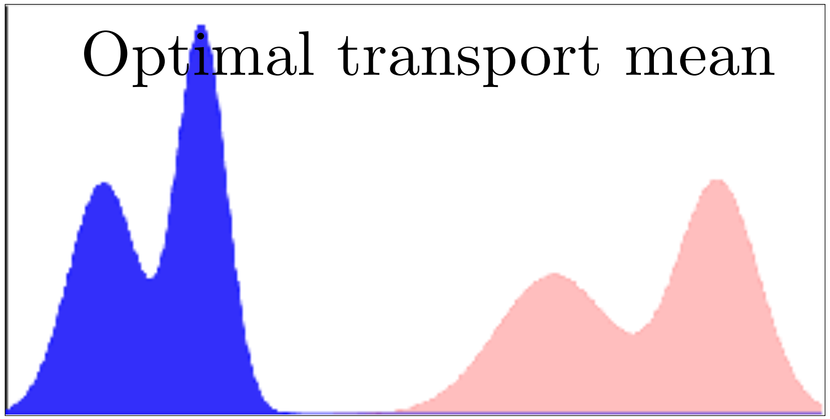
$\rightarrow$ images, vision, graphics and machine learning, ...



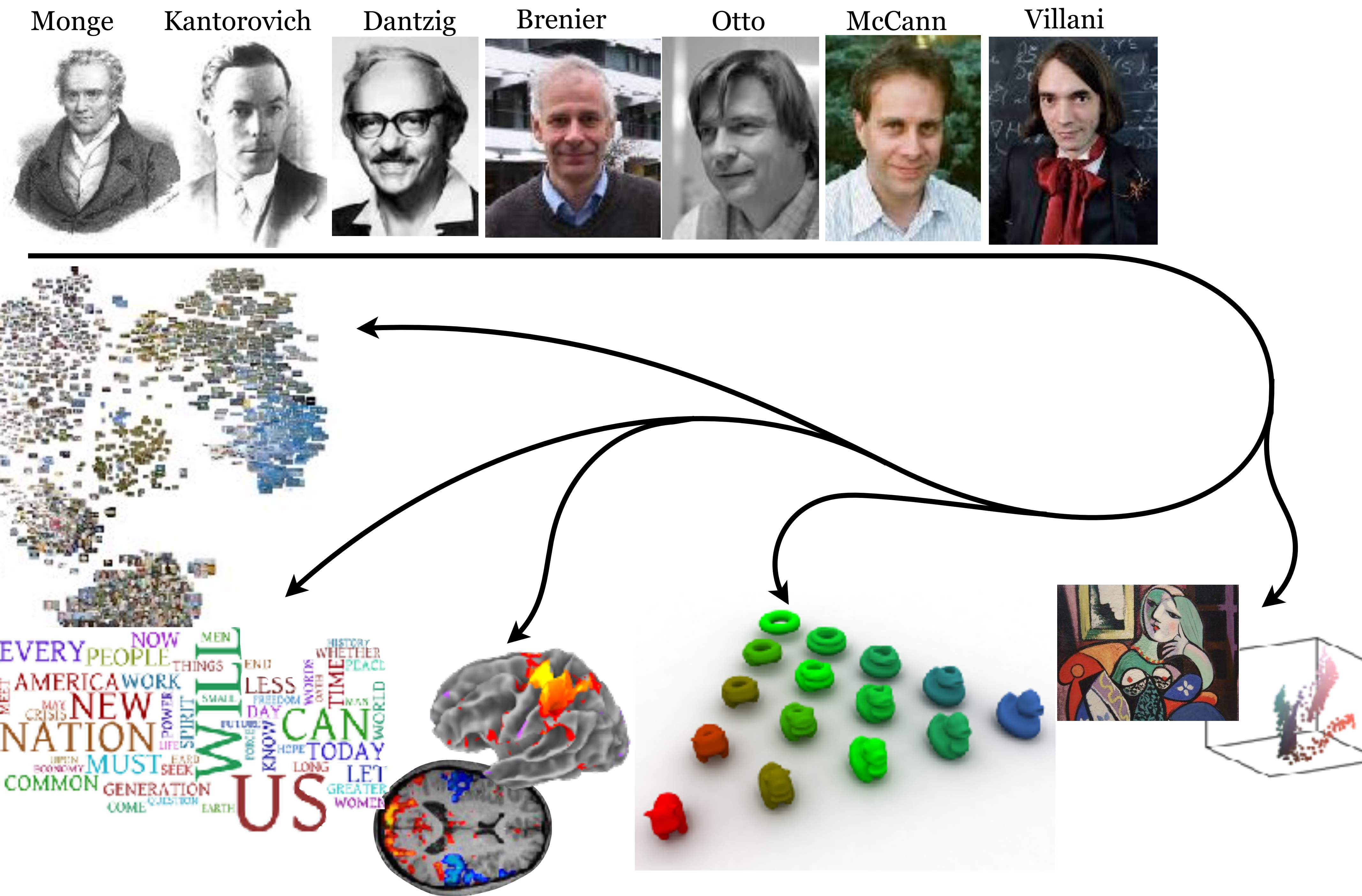
- *Optimal transport*

  $\rightarrow$ takes into account a metric d.





Toward High-dimensional OT



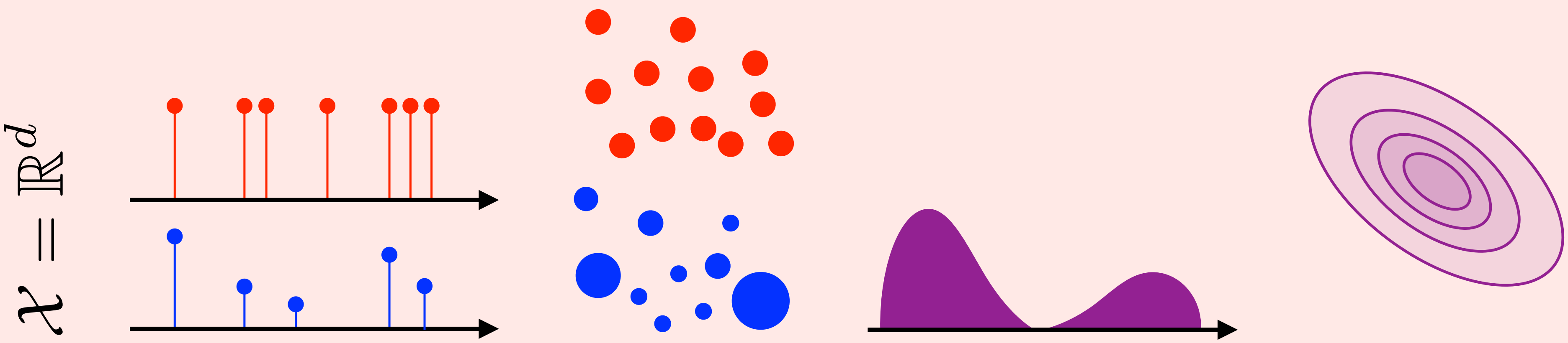
Overview

- **Measures and Histograms**
- From Monge to Kantorovitch Formulations
- Special Cases

Probability Measures

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

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

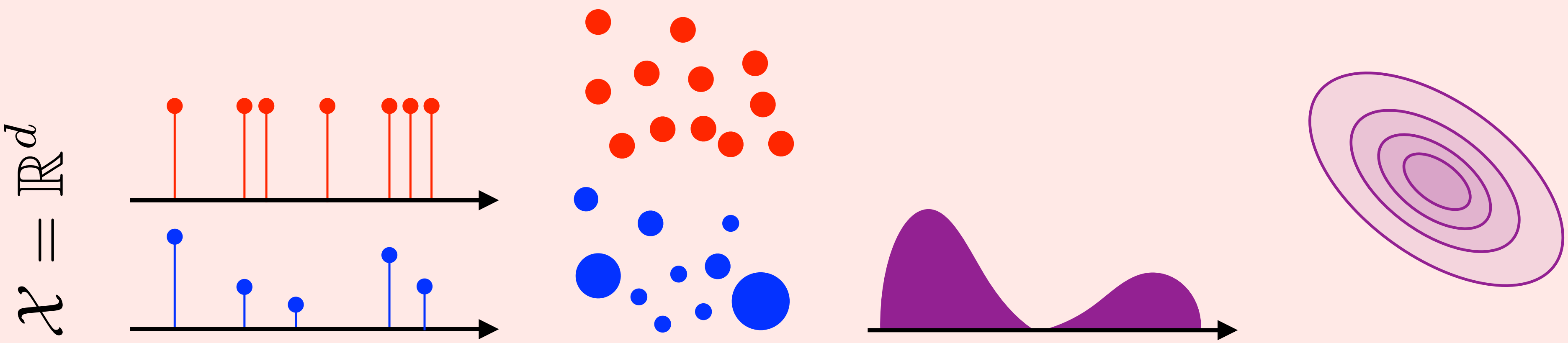


Discrete $d = 1$ Discrete $d = 2$ Density $d = 1$ Density $d = 2$

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

Measures and Random Variables

Random vectors

$$\mathbb{P}(X \in A)$$

Convergence in law:

$\forall$ set A

$$\mathbb{P}(X_n \in A) \xrightarrow{n \to +\infty} \mathbb{P}(X \in A)$$

$=$

Radon measures

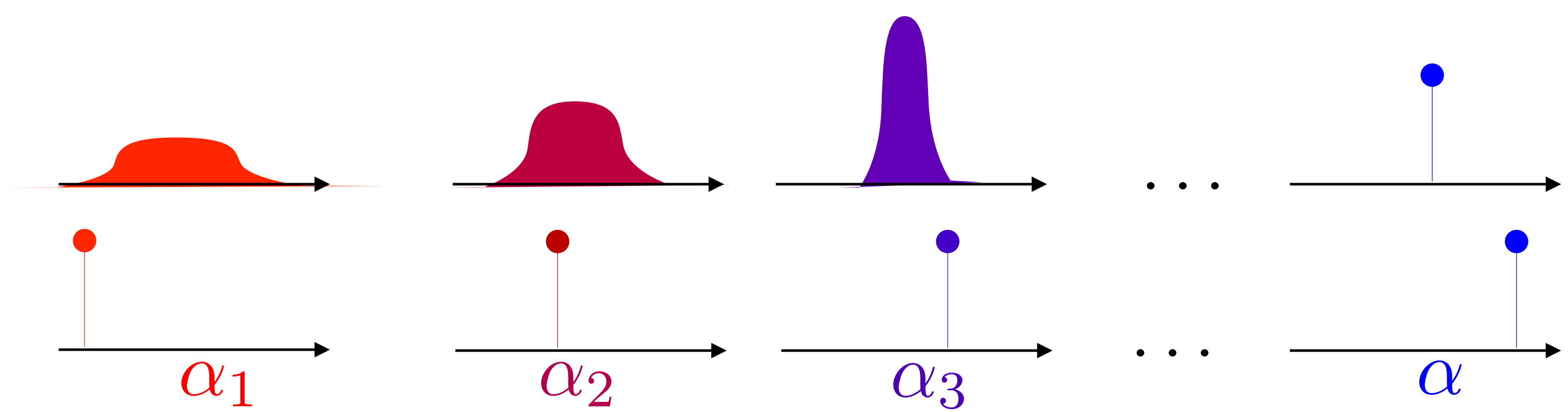
$$\int_A \mathrm{d}\alpha(x)$$

Weak* convergence:

$\forall$ continuous function f

$$\int f \mathrm{d}\alpha_n \xrightarrow{n \to +\infty} \int f \mathrm{d}\alpha$$

Weak convergence:

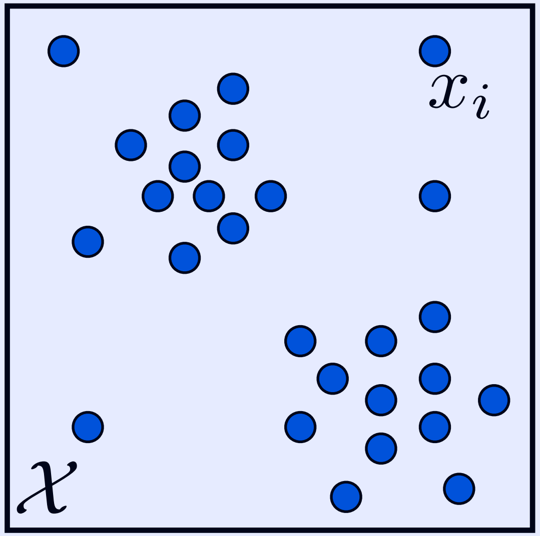


Lagrangian vs. Eulerian Discretization

Discrete measure: $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad x_i \in \mathcal{X}, \quad \sum_i \mathbf{a}_i = 1$

Lagrangian (point clouds)

Constant weights $\mathbf{a}_i = \frac{1}{n}$

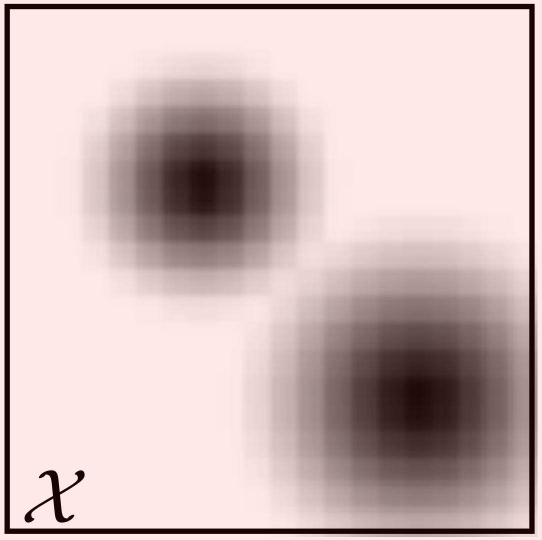


Quotient space:

$\mathcal{X}^n / \mathrm{Perm}(n)$

Eulerian (histograms)

Fixed positions x_i (e.g. grid)



Convex polytope (simplex):

$\{(\mathbf{a}_i)_i \geqslant 0 \ ; \ \sum_i \mathbf{a}_i = 1\}$

Push Forward

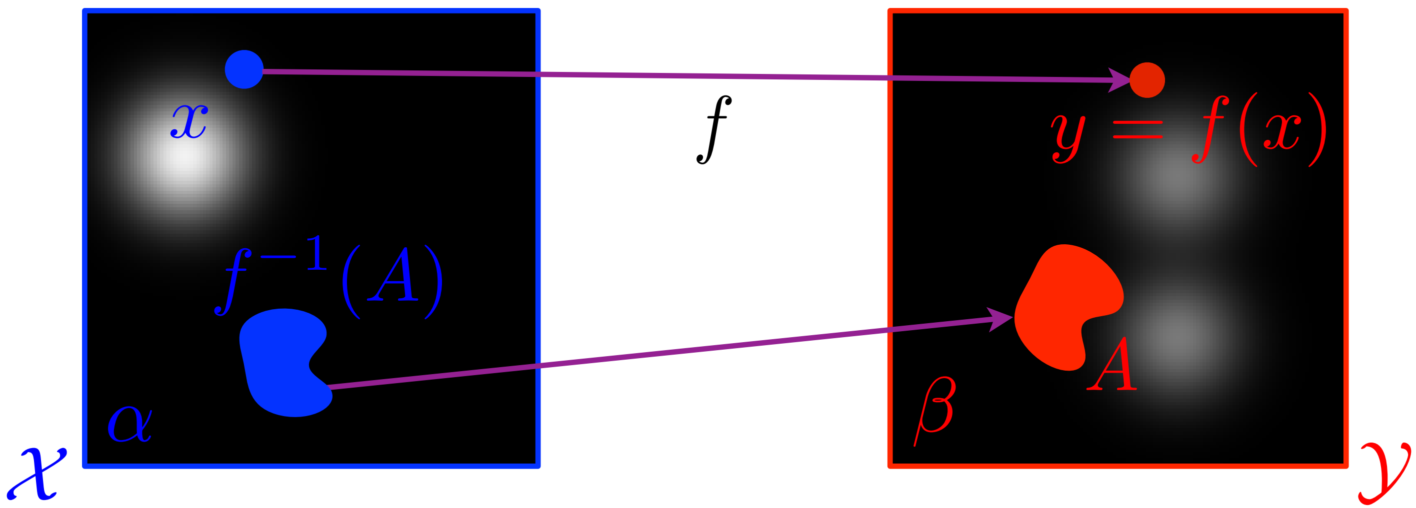
Radon measures (α, β) on $(\mathcal{X}, \mathcal{Y})$.

Transfer of measure by $f : \mathcal{X} \to \mathcal{Y}$: *push forward.*

$\beta = f_\sharp \alpha$ defined by:

$$\beta(A) \overset{\text{def.}}{=} \alpha(f^{-1}(A))$$

$$\iff \int_\mathcal{Y} g(y) \mathrm{d}\beta(y) \overset{\text{def.}}{=} \int_\mathcal{X} g(f(x)) \mathrm{d}\alpha(x)$$



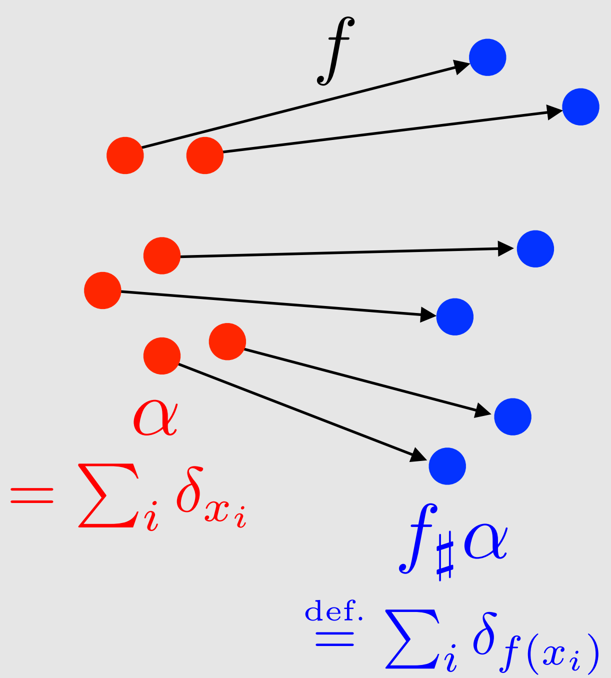
Smooth densities: $\mathrm{d}\alpha = \rho(x)\mathrm{d}x$, $\mathrm{d}\beta = \xi(x)\mathrm{d}x$

$$f_\sharp \alpha = \beta \iff \rho(f(x))|\det(\partial f(x))| = \xi(x)$$

Push-forward vs. Pull-back

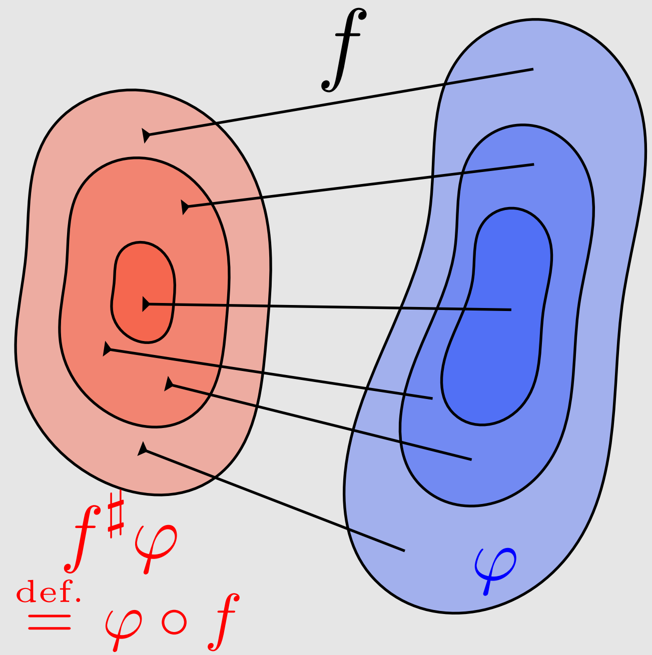
$f : \mathcal{X} \to \mathcal{Y}$

Measures:
push-forward



$f_\sharp : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$

Functions:
pull-back



$f^\sharp : \mathcal{C}(\mathcal{Y}) \to \mathcal{C}(\mathcal{X})$

Remark: $f^\sharp$ and $f_\sharp$ are adjoints

$$\int_{\mathcal{Y}} \varphi \mathrm{d}(f_\sharp \alpha) = \int_{\mathcal{X}} (f^\sharp \varphi) \mathrm{d}\alpha$$

Overview

- Measures and Histograms
- **From Monge to Kantorovitch Formulations**
- Special Cases

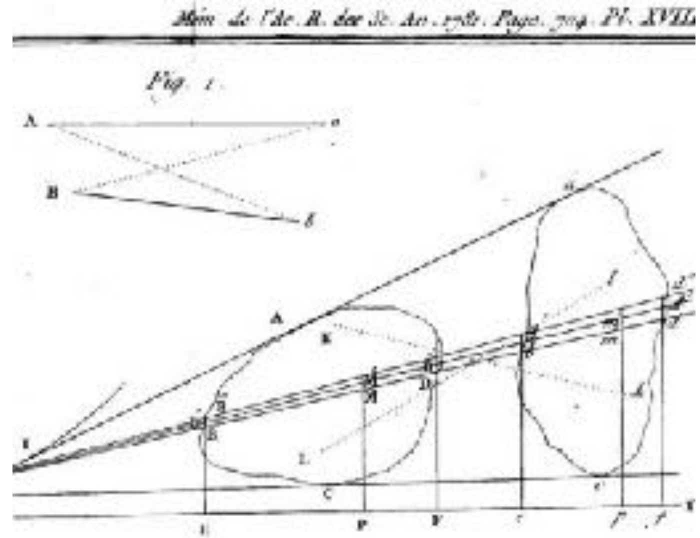
Gaspard Monge (1746-1818)

(1784)

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

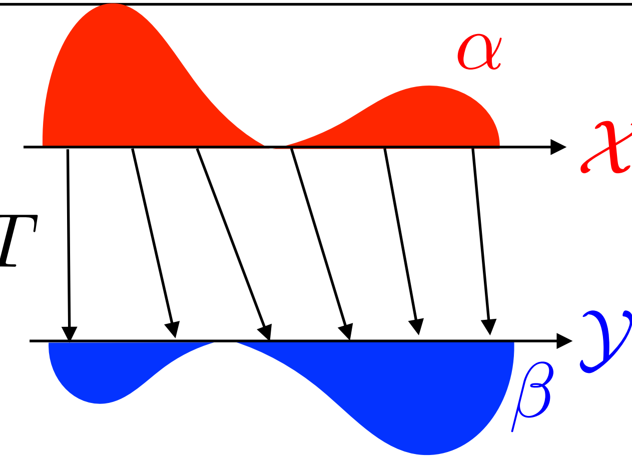
Lorsqu'on doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.

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*.



Monge's Transport

$$\min_{\beta = T_\sharp \alpha} \int_{\mathcal{X}} \|x - T(x)\|^2 \mathrm{d}\alpha(x)$$



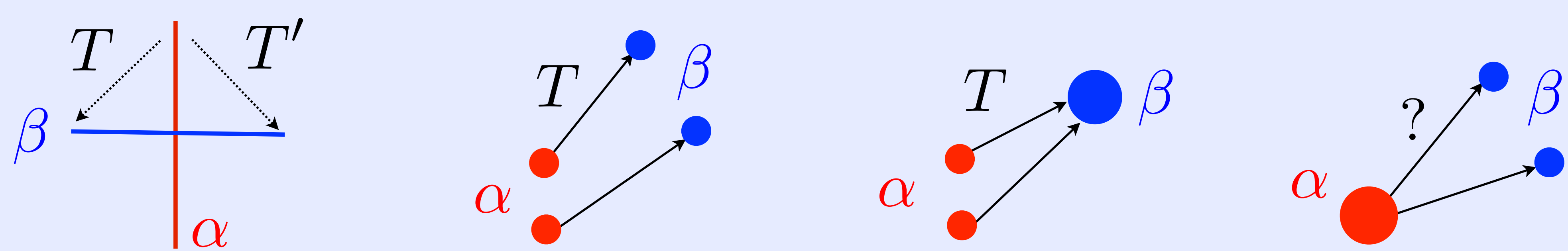
Densities: $\dfrac{\mathrm{d}\alpha}{\mathrm{d}x} = \rho_\alpha, \dfrac{\mathrm{d}\beta}{\mathrm{d}x} = \rho_\beta$

Theorem: [Brenier] Unique $T = \nabla\varphi$ solving

$$\det(\partial^2\varphi(x))\rho_\beta(\nabla\varphi(x)) = \rho_\alpha(x) \qquad \varphi \text{ convex}$$

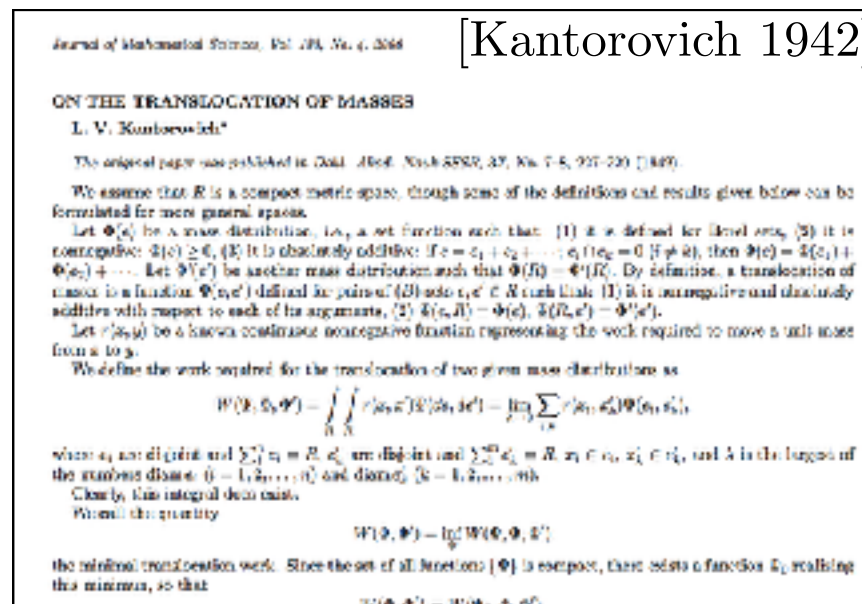
$\rightarrow$ Monge-Ampère equation (non-linear, degenerate elliptic).

Non-uniqueness / non-existence:



Leonid Kantorovich (1912-1986)

Леонид Витальевич Канторович



ON THE TRANSLOCATION OF MASSES

L. V. Kantorovich*

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: ... 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

...

Clearly, this integral does exist.

We call the quantity

...

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

...

[Kantorovich 1942]

Before Kantorovitch

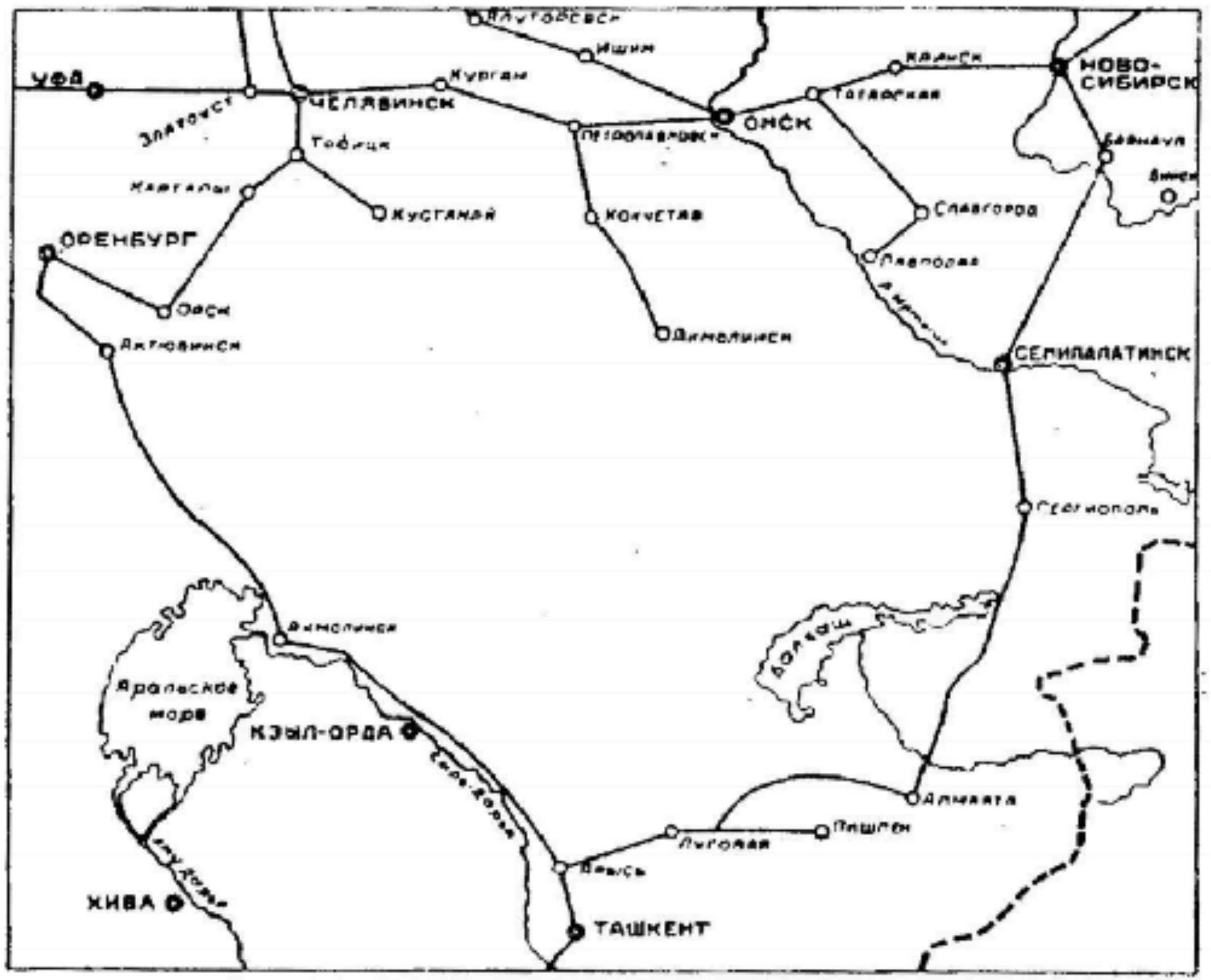


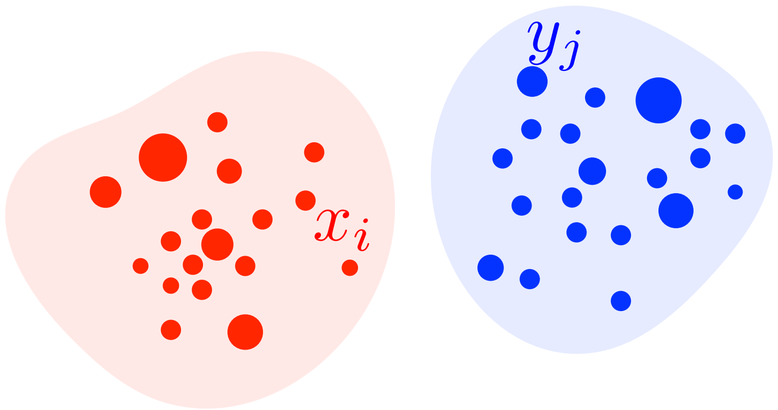
Figure 1: Figure from Tolstoĭ [1930] to illustrate a negative cycle

Optimal Transport was formulated in 1930 by A.N. Tolstoi, 12 years before Kantorovich. He even solved a "large scale" 10×68 instance!

Kantorovitch's Formulation

Input distributions

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



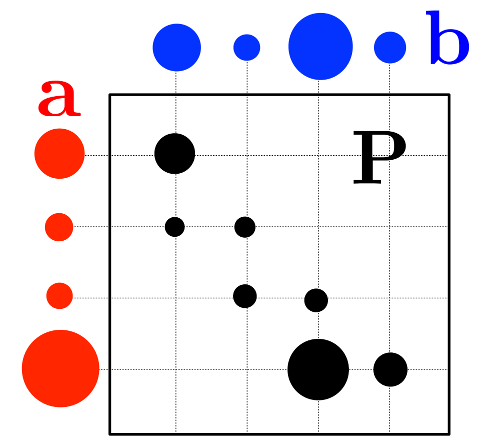
Points $(x_i)_i$, $(y_j)_j$

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

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

Couplings:

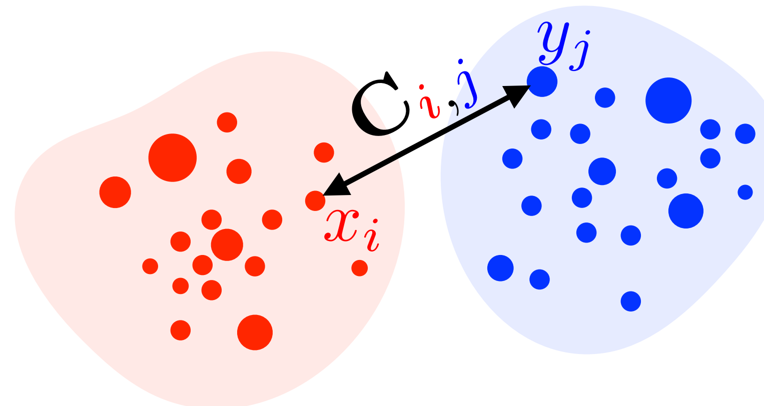
$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \; ; \; \mathbf{P} \mathbb{1}_n = \mathbf{a}, \mathbf{P}^\top \mathbb{1}_m = \mathbf{b} \right\}$$



Kantorovitch's Formulation

Input distributions

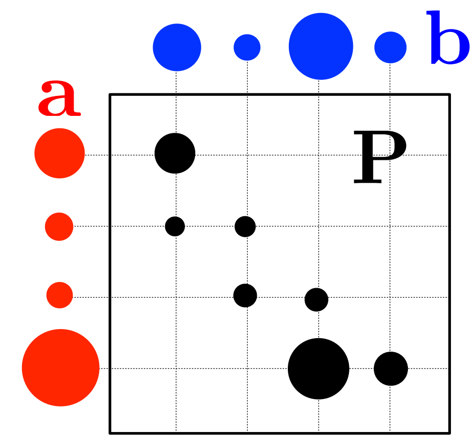
$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i} \quad \beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$



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

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

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



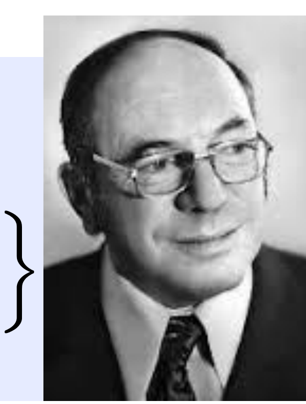
Couplings:

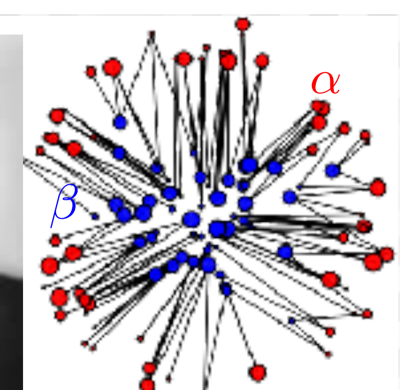
$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \; ; \; \mathbf{P}\mathbb{1}_n = \mathbf{a}, \mathbf{P}^\top \mathbb{1}_m = \mathbf{b} \right\}$$

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

[Kantorovich 1942]

$$L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \min \left\{ \sum_{i,j} \mathbf{P}_{i,j} \mathbf{C}_{i,j} \; ; \; \mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b}) \right\}$$





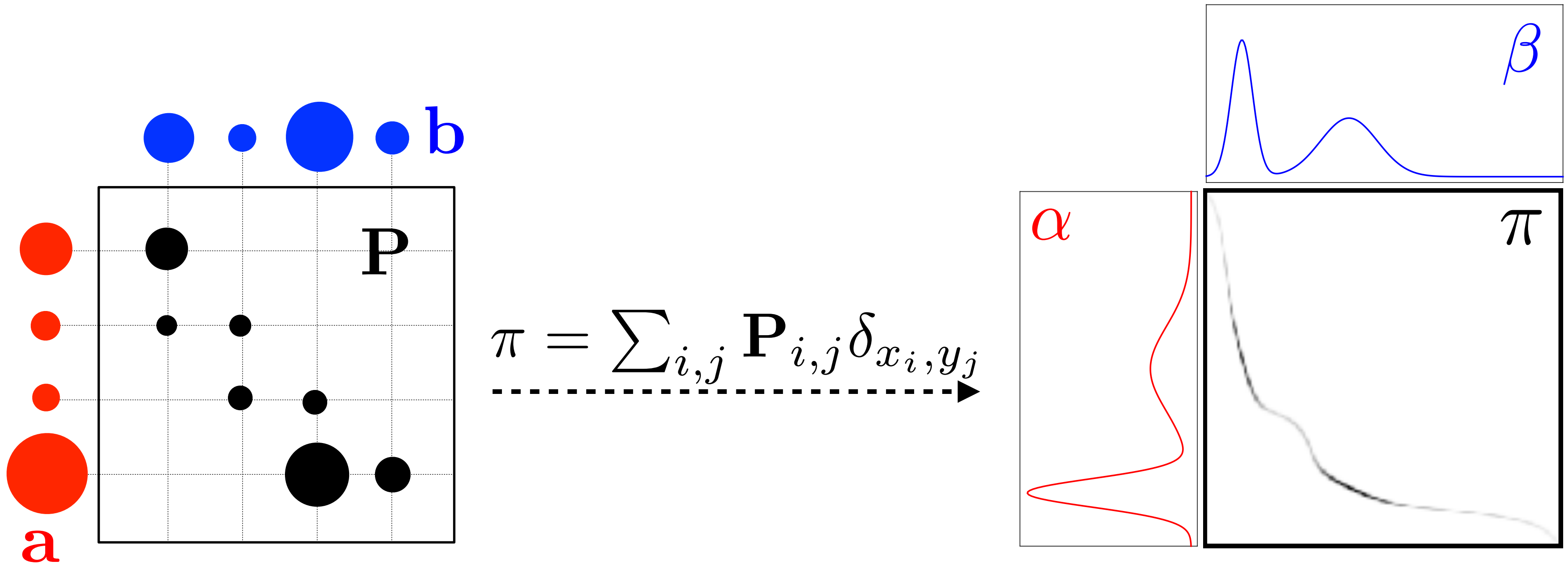
Couplings Between General Measures

Projectors:

$$P_{\mathcal{X}} : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto x \in \mathcal{X}$$
$$P_{\mathcal{Y}} : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto y \in \mathcal{Y}$$

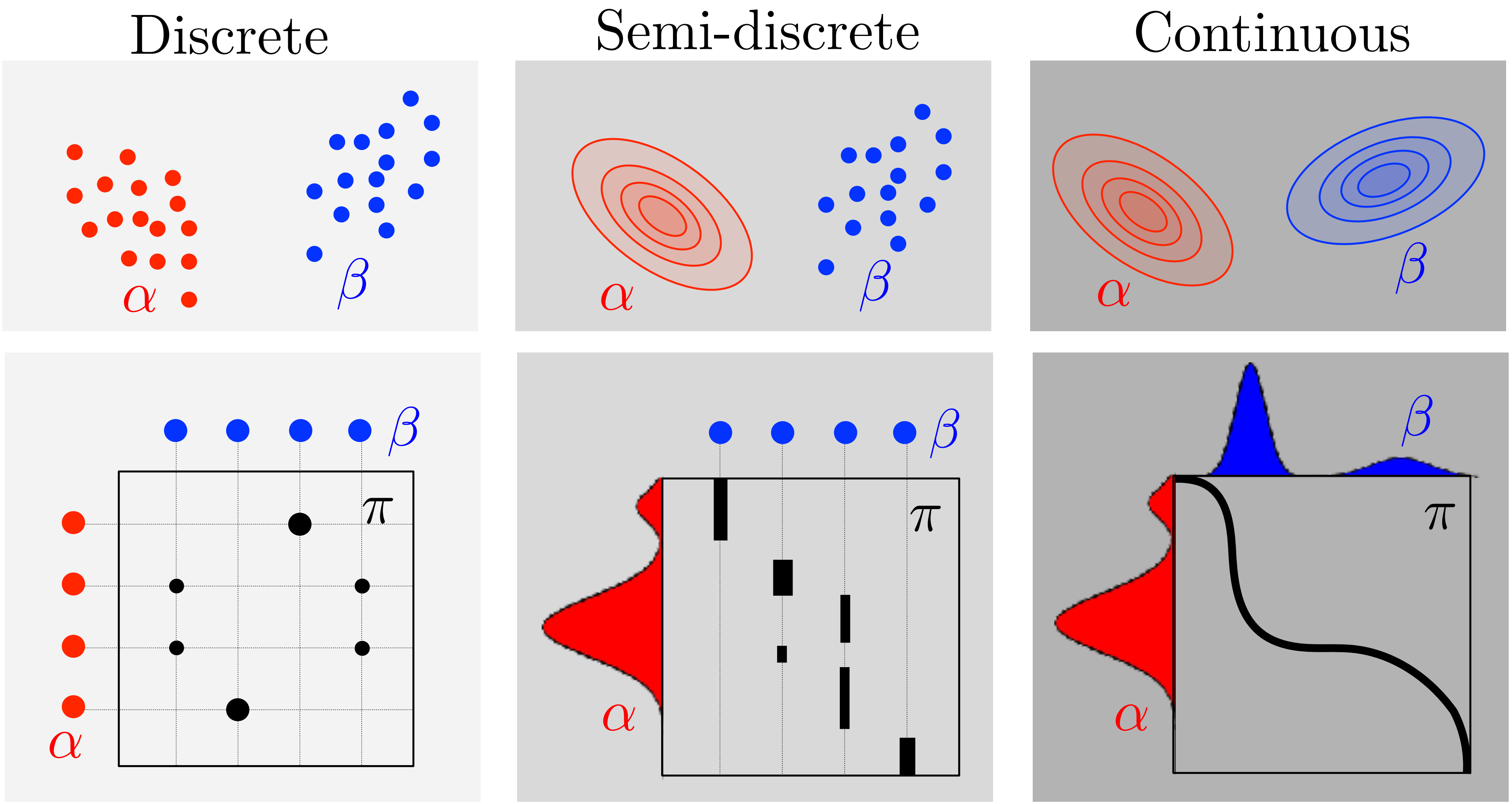
Couplings:

$$\mathcal{U}(\alpha, \beta) \overset{\text{def.}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \;:\; P_{\mathcal{X}\sharp}\pi = \alpha \quad \text{and} \quad P_{\mathcal{Y}\sharp}\pi = \beta \right\}$$

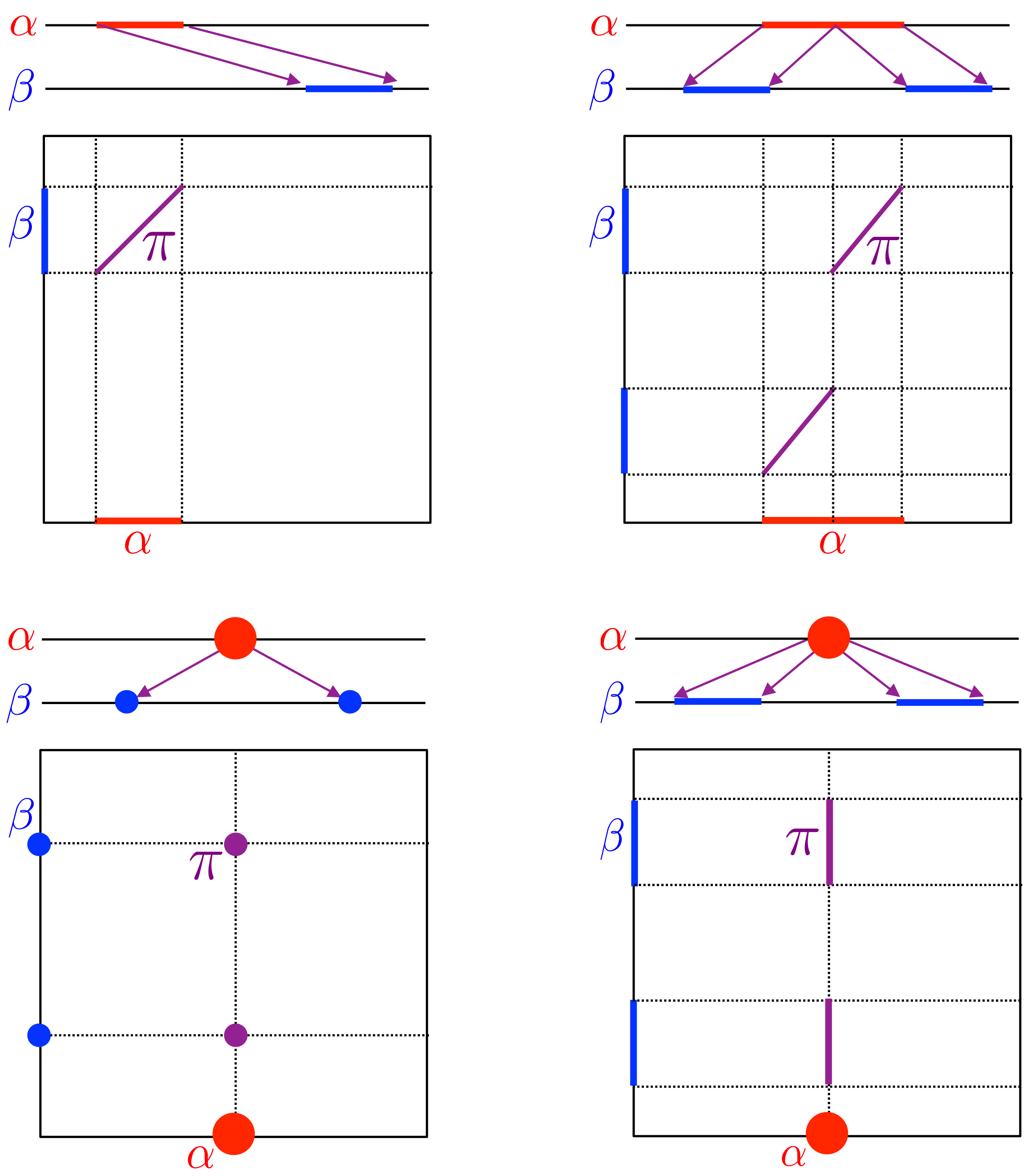


$$P_{\mathcal{X},\sharp}\left(\textstyle\sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)}\right) = \textstyle\sum_i \left(\sum_j \mathbf{P}_{i,j}\right) \delta_{x_i}$$
$$P_{\mathcal{Y},\sharp}\left(\textstyle\sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)}\right) = \textstyle\sum_j \left(\sum_i \mathbf{P}_{i,j}\right) \delta_{y_j}$$

Couplings: the 3 Settings



Examples of Couplings



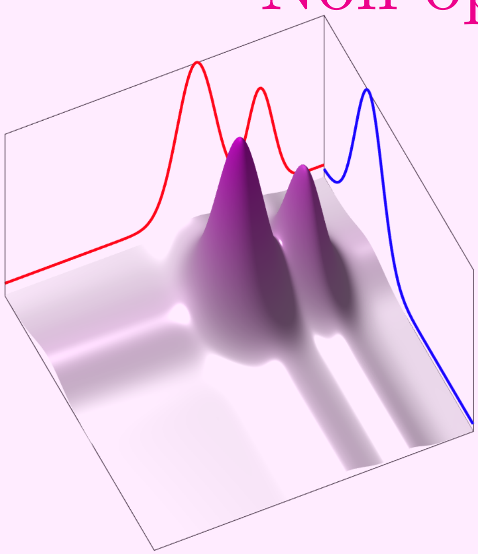
Kantorovitch Problem for General Measures

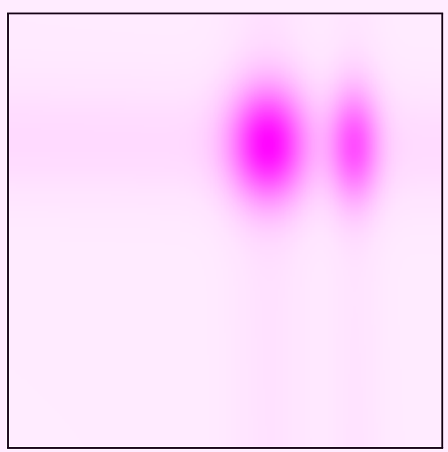
$$\mathcal{L}_c(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y)$$

Probabilistic interpretation:

$$\min_{(X,Y)} \left\{ \mathbb{E}_{(X,Y)}(c(X,Y)) \;:\; X \sim \alpha, Y \sim \beta \right\}$$

Non-optimal

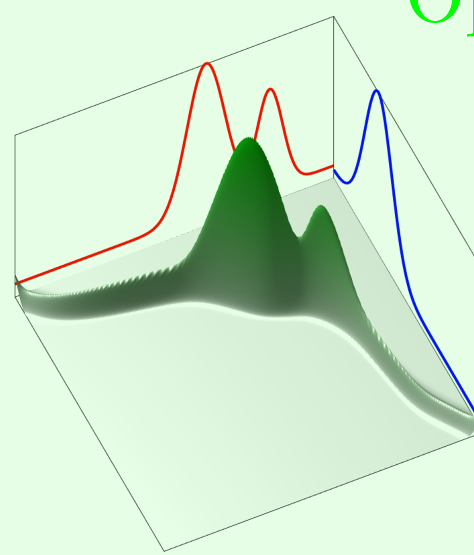


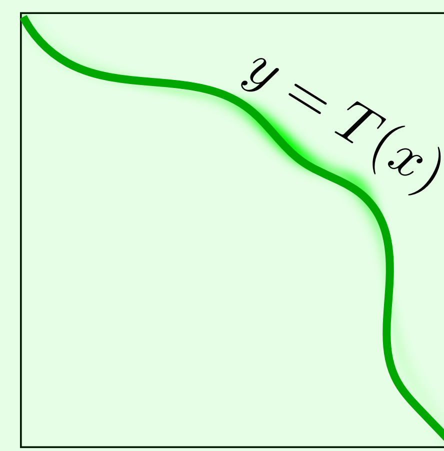


$\pi = \alpha \otimes \beta$

(X,Y) independents

Optimal





$\pi = (\mathrm{Id}, T)_\sharp \alpha$

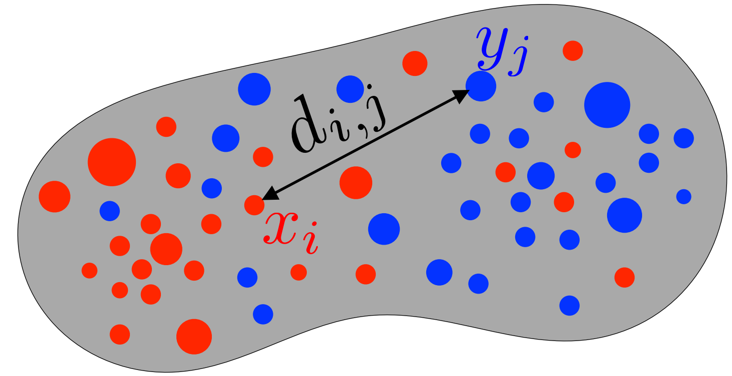
$Y = T(X)$

Wasserstein Distance

Metric spaces $\mathcal{X} = \mathcal{Y}$

Distance $d(x, y)$.

Cost $c(x, y) = d(x, y)^p$, $p \geqslant 1$.



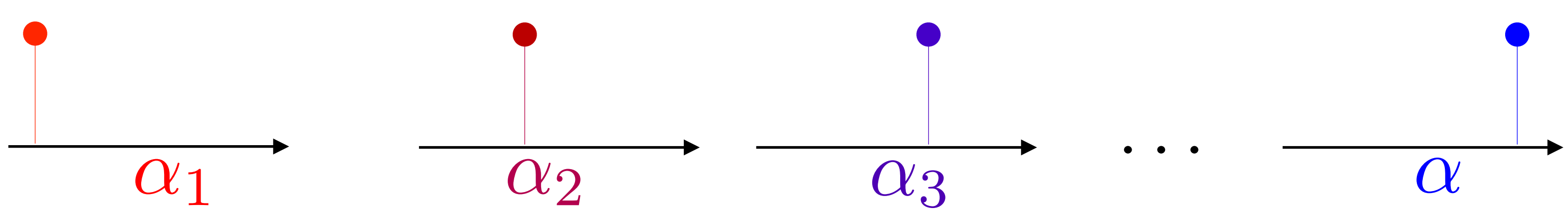
Wasserstein distance:

$$\mathrm{W}_p(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \mathrm{L}_{\mathbf{D}^p}(\mathbf{a}, \mathbf{b})^{1/p} \qquad \mathcal{W}_p(\alpha, \beta) \stackrel{\text{def.}}{=} \mathcal{L}_{d^p}(\alpha, \beta)^{1/p}$$

Theorem: W_p and $\mathcal{W}_p$ are distances.

$$\mathcal{W}_p(\alpha_n, \alpha) \to 0 \iff \alpha_n \stackrel{\text{weak}*}{\to} \alpha$$

Examples: $\mathcal{W}_p(\delta_x, \delta_y) = d(x, y)$.



Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- Special Cases

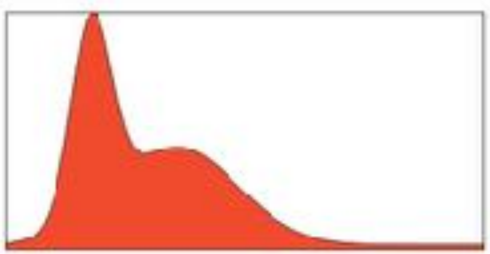
1-D Optimal Transport

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

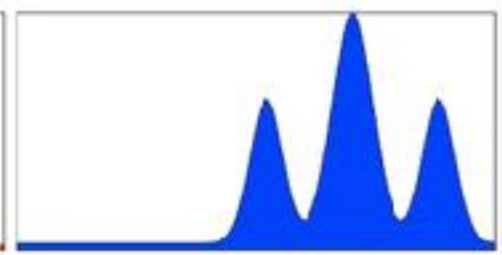
Inverse cumulative: $$\forall r \in [0,1], \quad \mathcal{C}_\alpha^{-1}(r) = \min_x \ \{x \in \mathbb{R} \cup \{-\infty\} \ : \ \mathcal{C}_\alpha(x) \geq r\}$$

Theorem: $$\mathcal{W}_p(\alpha,\beta)^p = \left\| \mathcal{C}_\alpha^{-1} - \mathcal{C}_\beta^{-1} \right\|_{L^p([0,1])}^p = \int_0^1 |\mathcal{C}_\alpha^{-1}(r) - \mathcal{C}_\beta^{-1}(r)|^p \mathrm{d}r$$

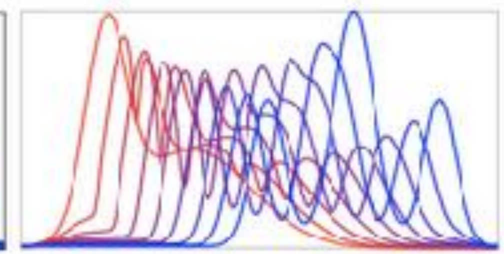
$$\mathcal{W}_1(\alpha,\beta) = \|\mathcal{C}_\alpha - \mathcal{C}_\beta\|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| \mathrm{d}x = \int_{\mathbb{R}} \left| \int_{-\infty}^{x} \mathrm{d}(\alpha - \beta) \right| \mathrm{d}x.$$



α

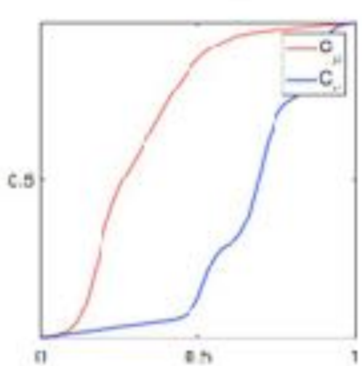


β

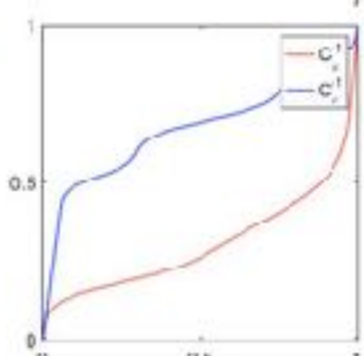


$((t\mathrm{Id} + (1-t)\mathrm{Id})_\sharp \alpha$

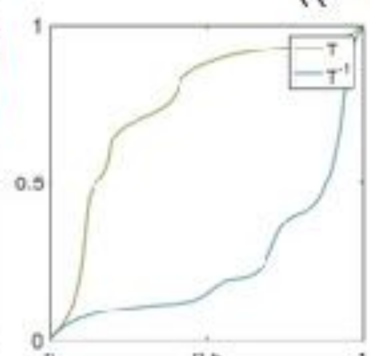
$$T = \mathcal{C}_\beta^{-1} \circ \mathcal{C}_\alpha$$



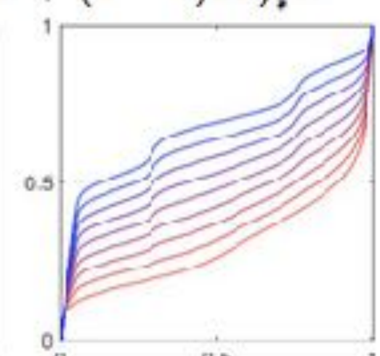
$(\mathcal{C}_\alpha, \mathcal{C}_\beta)$



$(\mathcal{C}_\alpha^{-1}, \mathcal{C}_\beta^{-1})$



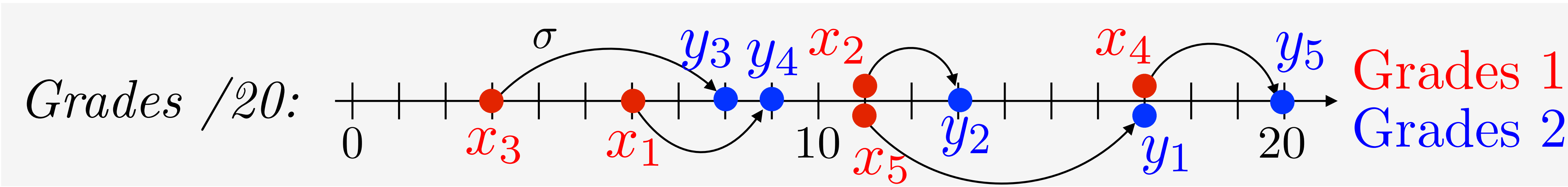
(T, T^{-1})



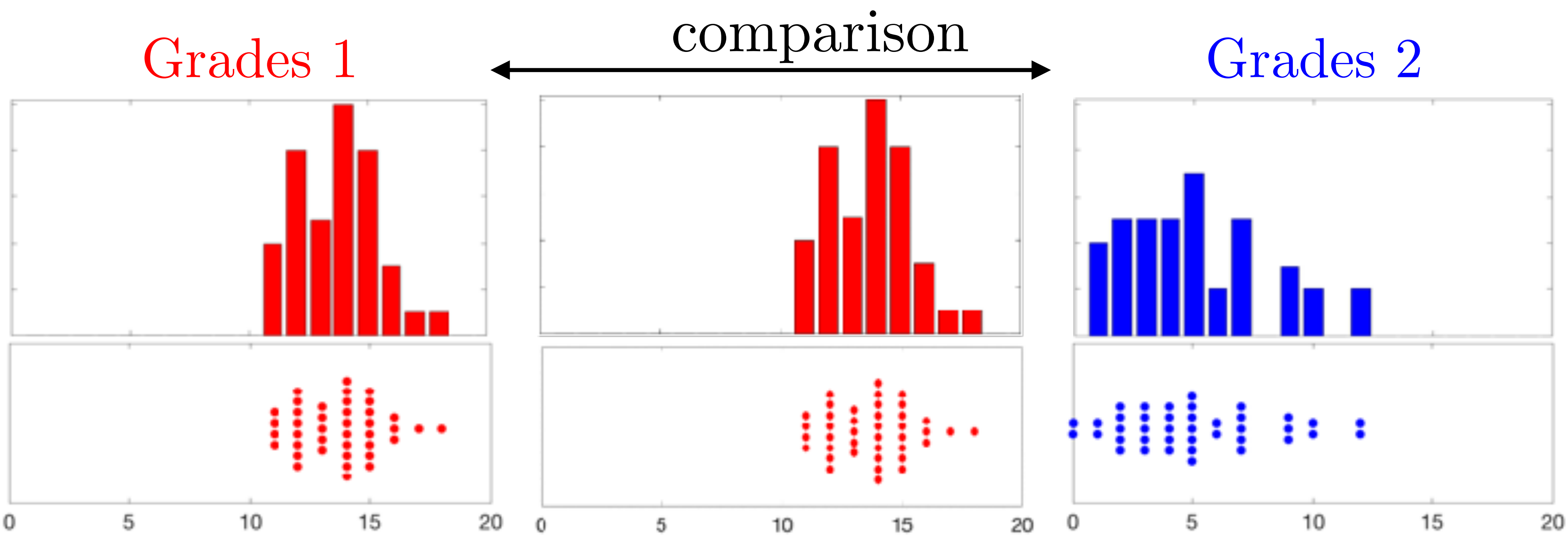
$(1-t)\mathcal{C}_\alpha^{-1} + t\mathcal{C}_\beta^{-1}$

Discrete 1-D Optimal Transport

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

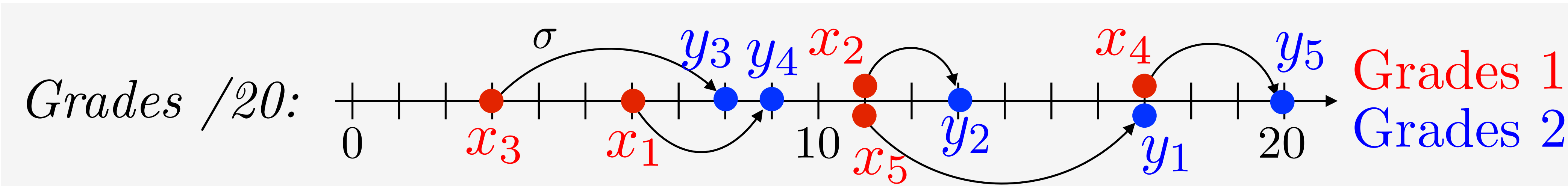


$\rightarrow$ Complexity $O(n \log(n))$.

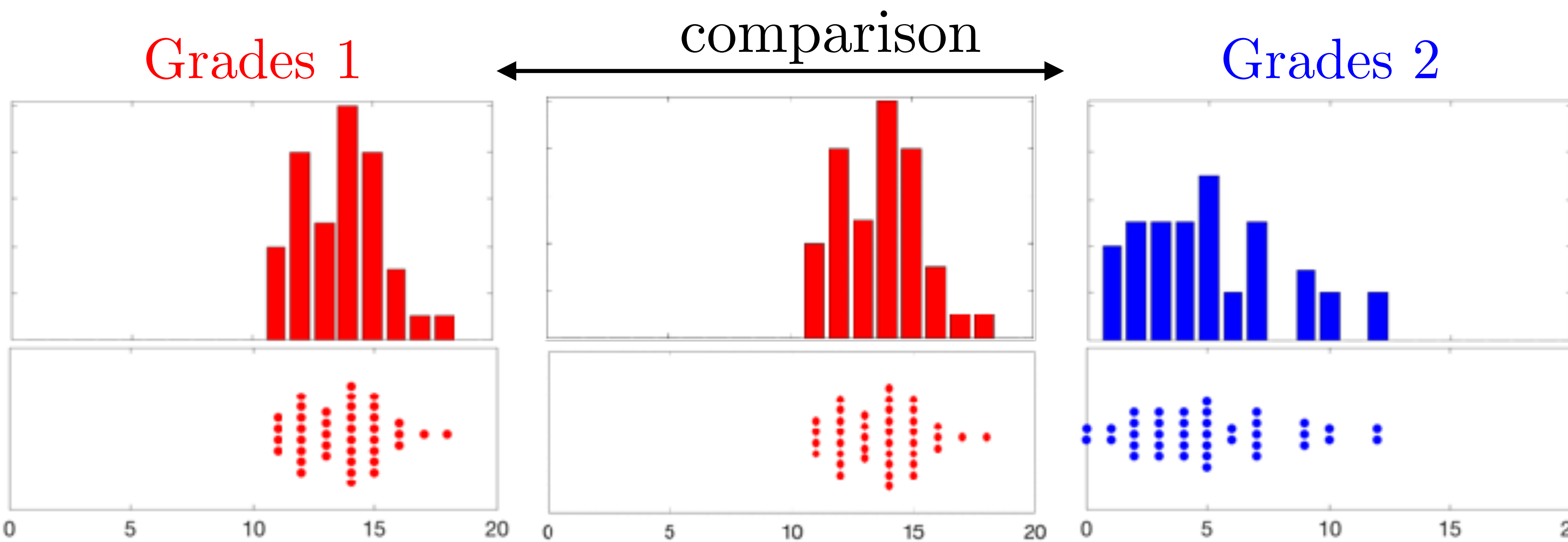


Discrete 1-D Optimal Transport

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



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



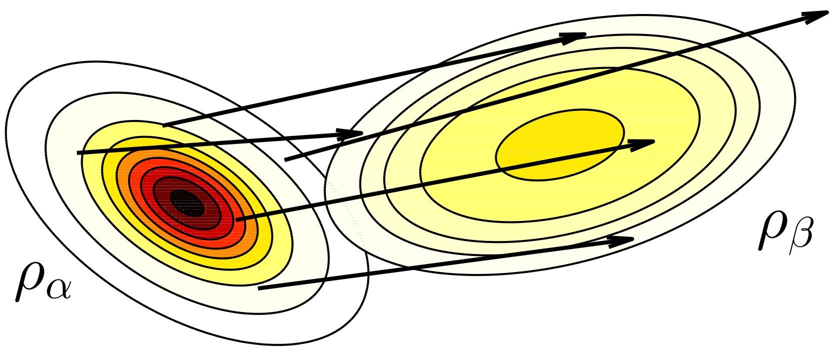
OT Between Gaussians

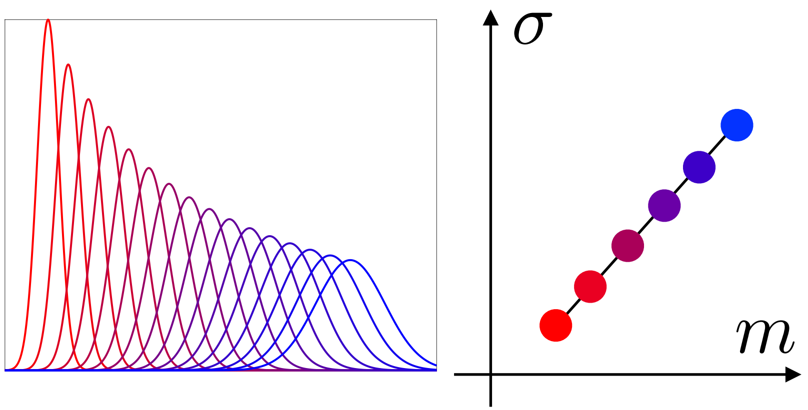
OT between $\begin{matrix}\beta = \mathcal{N}(\mathbf{m}_\beta, \Sigma_\beta) \\ \alpha = \mathcal{N}(\mathbf{m}_\alpha, \Sigma_\alpha)\end{matrix}$ is affine:

$$T : x \mapsto \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha)$$

$$A = \boldsymbol{\Sigma}_\alpha^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}} \boldsymbol{\Sigma}_\beta \boldsymbol{\Sigma}_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \boldsymbol{\Sigma}_\alpha^{-\frac{1}{2}} = A^{\mathrm{T}}$$

$\rightarrow$ Gaussians are OT geodesically convex.





OT Between Gaussians

OT between $\begin{array}{l}\beta = \mathcal{N}(\mathbf{m}_\beta, \Sigma_\beta) \\ \alpha = \mathcal{N}(\mathbf{m}_\alpha, \Sigma_\alpha)\end{array}$ is affine:

$$T : x \mapsto \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha)$$

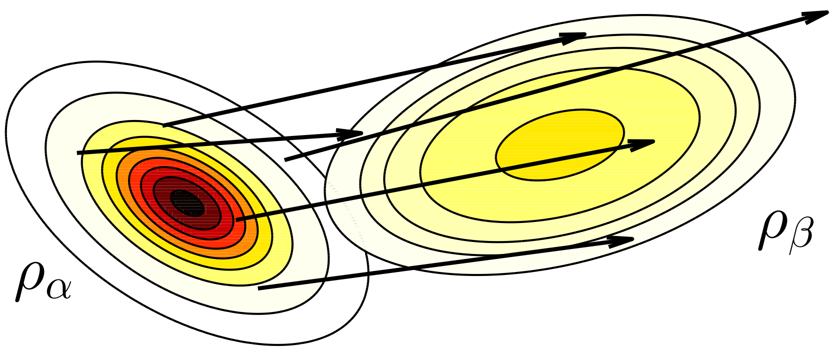
$$A = \mathbf{\Sigma}_\alpha^{-\frac{1}{2}} \left(\mathbf{\Sigma}_\alpha^{\frac{1}{2}} \mathbf{\Sigma}_\beta \mathbf{\Sigma}_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \mathbf{\Sigma}_\alpha^{-\frac{1}{2}} = A^{\mathrm{T}}$$

$\rightarrow$ Gaussians are OT geodesically convex.

$$\mathcal{W}_2^2(\alpha, \beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2$$

Bures distance:

$$\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2 \overset{\text{def.}}{=} \operatorname{tr}\left(\mathbf{\Sigma}_\alpha + \mathbf{\Sigma}_\beta - 2(\mathbf{\Sigma}_\alpha^{1/2} \mathbf{\Sigma}_\beta \mathbf{\Sigma}_\alpha^{1/2})^{1/2} \right)$$



Diagonal covariances:

$$\mathbf{\Sigma}_\alpha = \operatorname{diag}(r_i)_i \qquad \mathbf{\Sigma}_\beta = \operatorname{diag}(s_i)_i$$

$$\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta) = \|\sqrt{r} - \sqrt{s}\|_2$$