Numerical Optimal Transport

http://optimaltransport.github.io

Gabriel Peyré

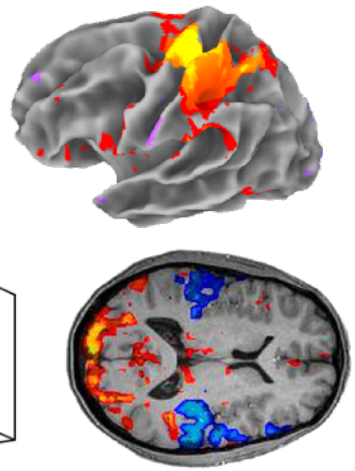
www.numerical-tours.com

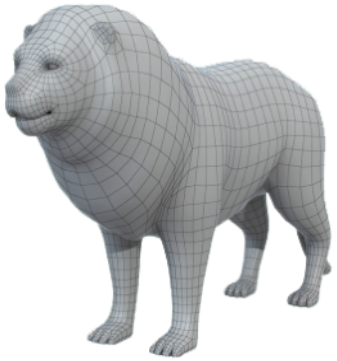
Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein

Comparing Measures and Spaces

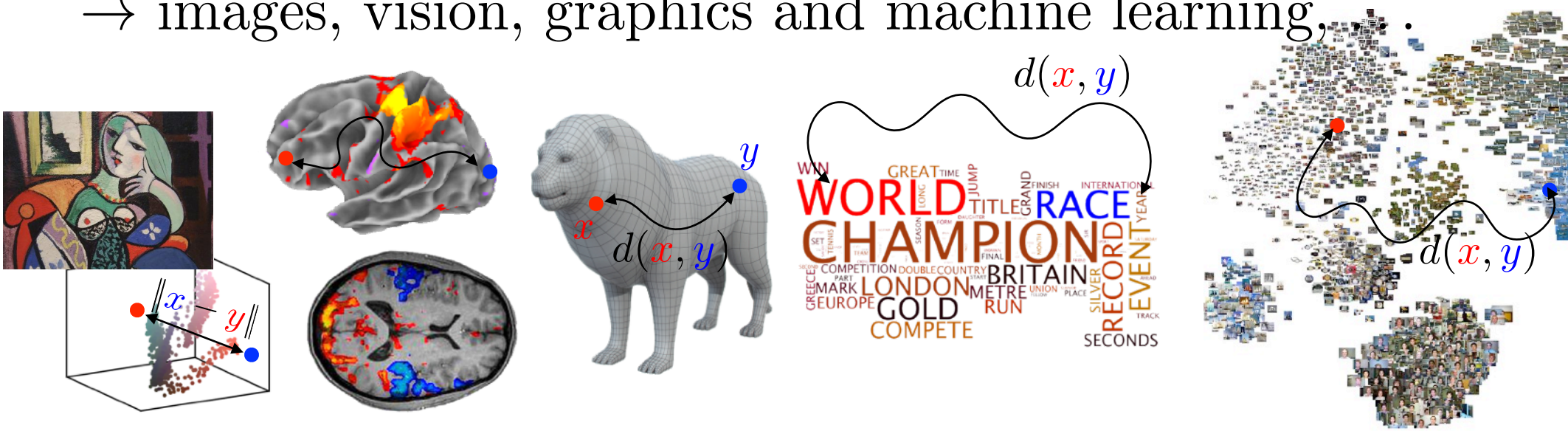
- *Probability distributions and histograms*
  $\rightarrow$ images, vision, graphics and machine learning, . . .



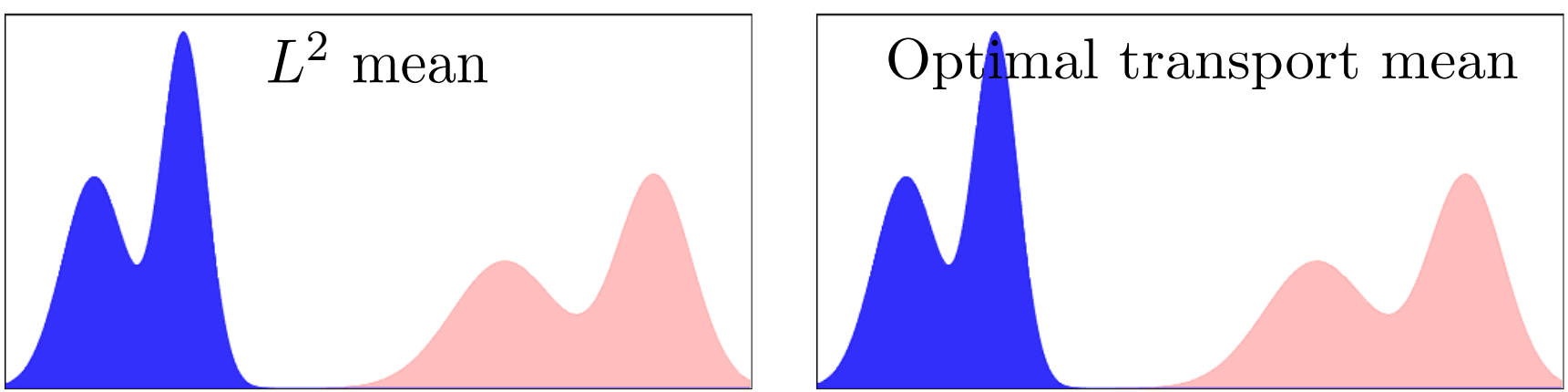


Comparing Measures and Spaces

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



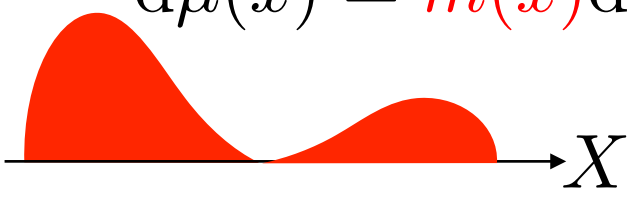
- *Optimal transport*
  $\rightarrow$ takes into account a metric d.



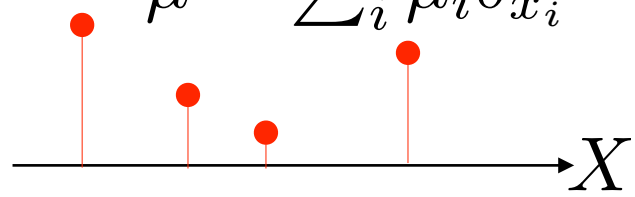
Probability Measures

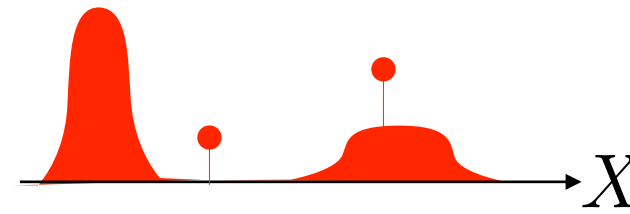
Positive Radon measure μ on a set X.

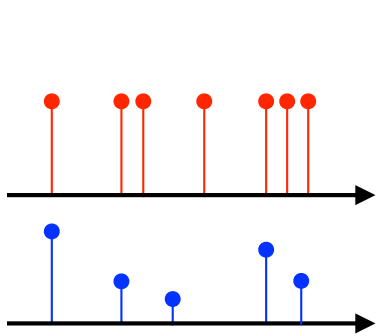
$\mathrm{d}\mu(x) = m(x)\mathrm{d}x$



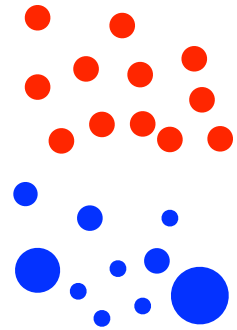
$\mu = \sum_i \mu_i \delta_{x_i}$





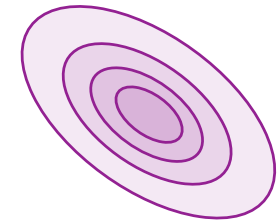


Discrete $d = 1$



Discrete $d = 2$

Density $d = 1$



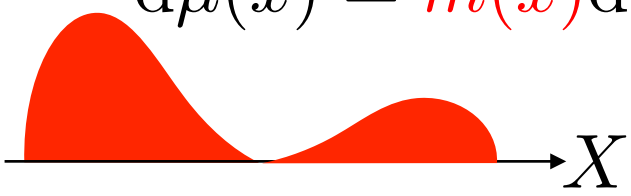
Density $d = 2$

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

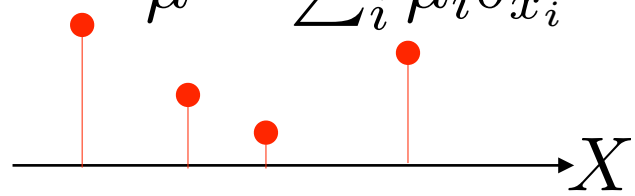
Probability Measures

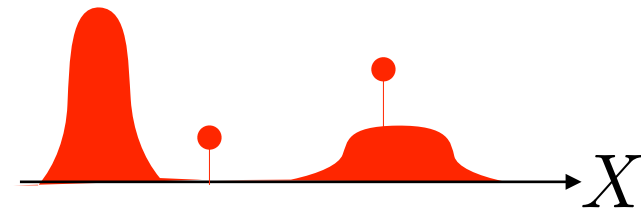
Positive Radon measure μ on a set X.

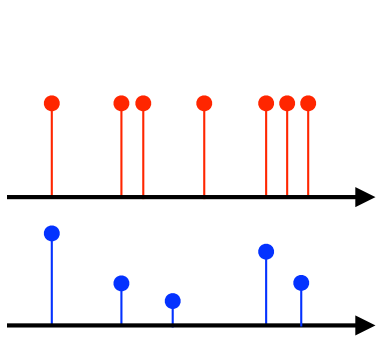
$\mathrm{d}\mu(x) = m(x)\mathrm{d}x$



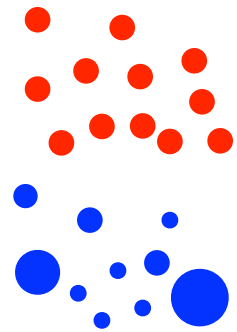
$\mu = \sum_i \mu_i \delta_{x_i}$





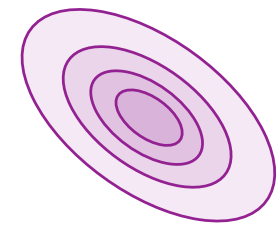


Discrete $d = 1$



Discrete $d = 2$

Density $d = 1$



Density $d = 2$

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

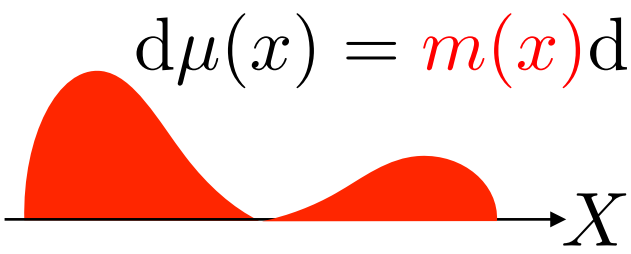
Integration against continuous functions: $\int_X g(x)\mathrm{d}\mu(x) \geqslant 0$

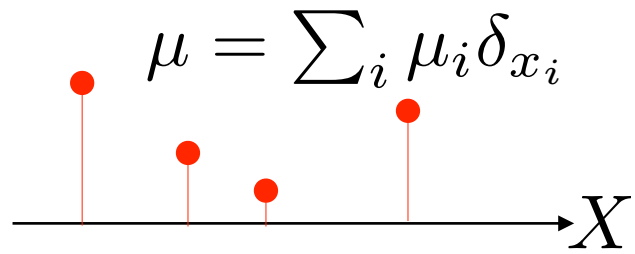
$$\mathrm{d}\mu(x) = m(x)\mathrm{d}x \longrightarrow \int_X g\mathrm{d}\mu = \int_X m(x)\mathrm{d}x$$

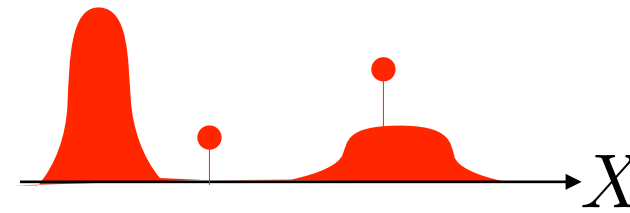
$$\mu = \sum_i \mu_i \delta_{x_i} \longrightarrow \int_X g\mathrm{d}\mu = \sum_i \mu_i g(x_i)$$

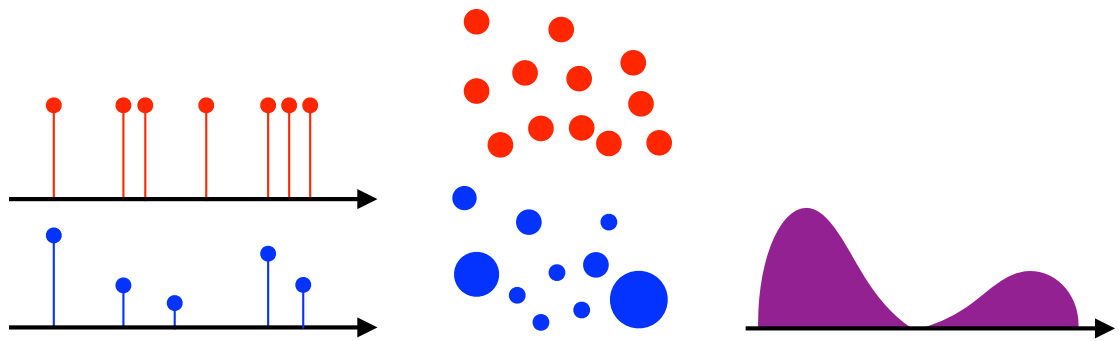
Probability Measures

Positive Radon measure μ on a set X.









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

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

Integration against continuous functions: $\int_X g(x)\mathrm{d}\mu(x) \geqslant 0$

$$\mathrm{d}\mu(x) = m(x)\mathrm{d}x \longrightarrow \int_X g\mathrm{d}\mu = \int_X m(x)\mathrm{d}x$$

$$\mu = \sum_i \mu_i \delta_{x_i} \longrightarrow \int_X g\mathrm{d}\mu = \sum_i \mu_i g(x_i)$$

Probability (normalized) measure: $\mu(X) = \int_X \mathrm{d}\mu(x) = 1$

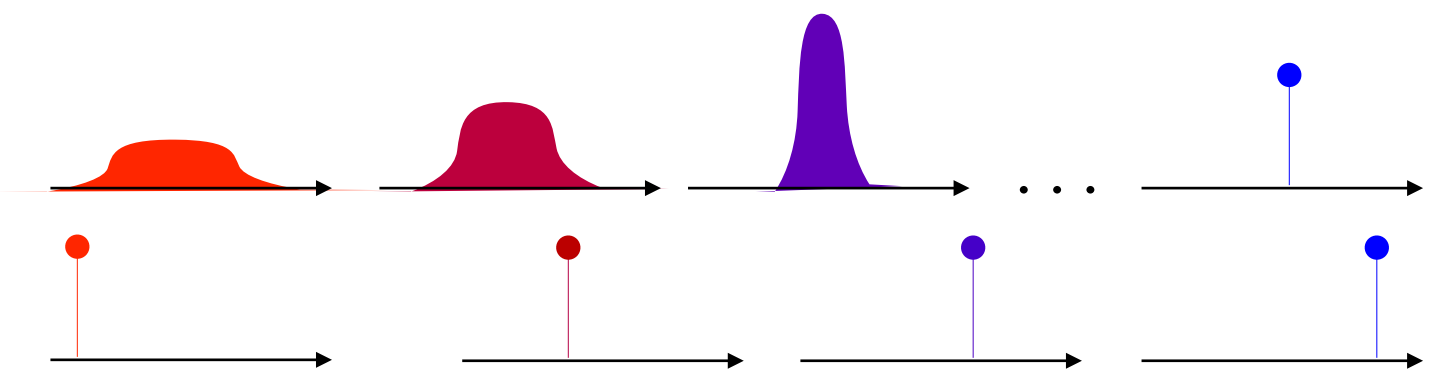
Measures and Random Variables

Random vectors		Radon measures
$\mathbb{P}(X \in A)$	$=$	$\int_A \mathrm{d}\mu(x)$
Weak* convergence:		Convergence in law:
$\forall$ set A $\quad \mathbb{P}(X_n \in A) \overset{n\to+\infty}{\longrightarrow} \mathbb{P}(X \in A)$		$\forall$ continuous function f $\quad \int f \mathrm{d}\mu_n \overset{n\to+\infty}{\longrightarrow} \int f \mathrm{d}\mu$

Measures and Random Variables

Random vectors		Radon measures
$\mathbb{P}(X \in A)$	$=$	$\int_A \mathrm{d}\mu(x)$
Weak* convergence:		Convergence in law:
$\forall$ set A $\quad \mathbb{P}(X_n \in A) \overset{n\to+\infty}{\longrightarrow} \mathbb{P}(X \in A)$		$\forall$ continuous function f $\quad \int f \mathrm{d}\mu_n \overset{n\to+\infty}{\longrightarrow} \int f \mathrm{d}\mu$

Weak convergence:

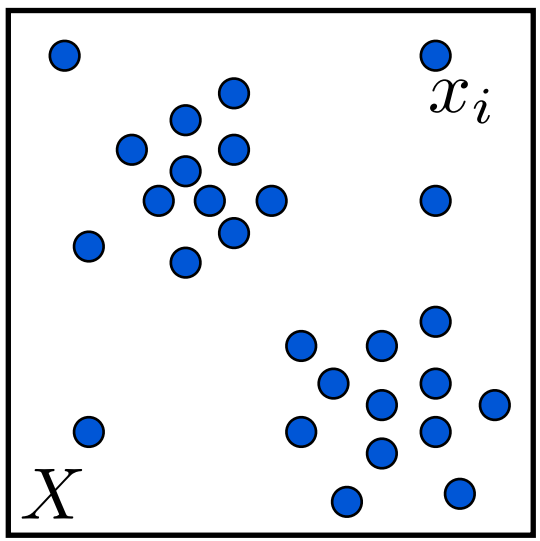


Discretization: Histogram vs. Empirical

Discrete measure: $\mu = \sum_{i=1}^{N} \mu_i \delta_{x_i} \qquad x_i \in X, \quad \sum_i \mu_i = 1$

Lagrangian (point clouds)

Constant weights $\mu_i = \frac{1}{N}$

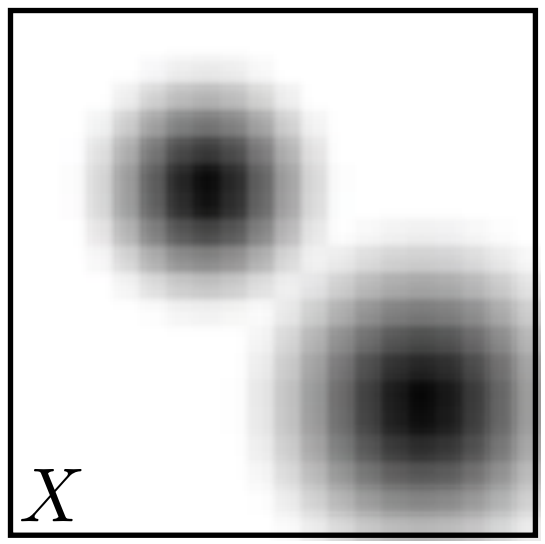


Quotient space:

X^N / Σ_N

Eulerian (histograms)

Fixed positions x_i (e.g. grid)



Convex polytope (simplex):

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

Push Forward

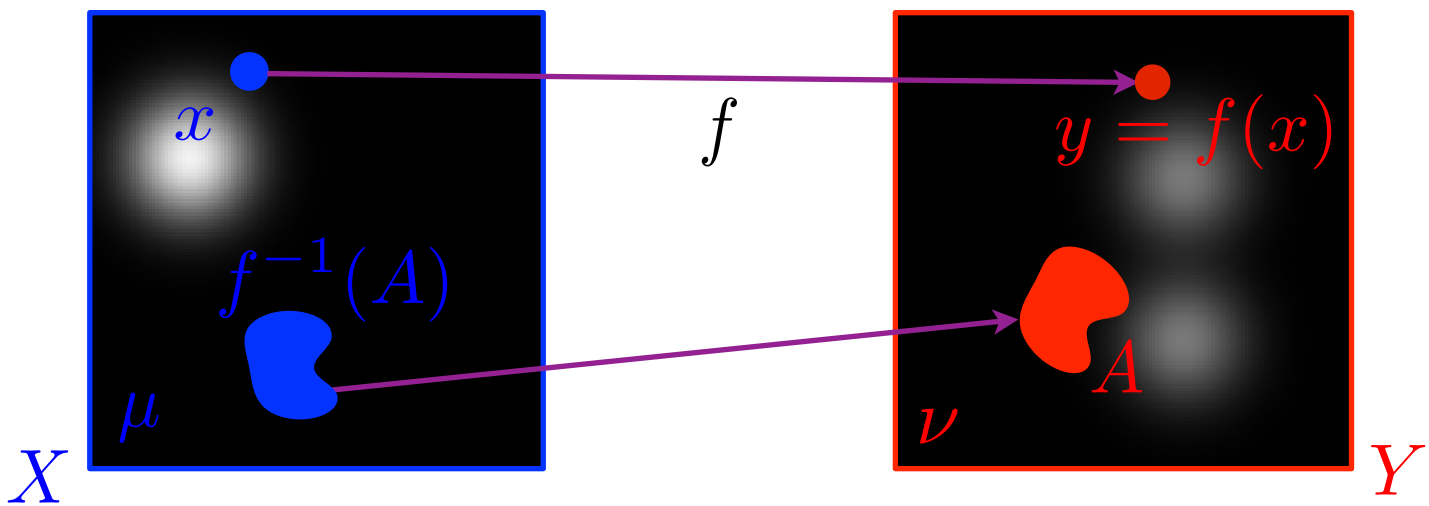
Radon measures (μ, ν) on (X, Y).

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

$\nu = f_\sharp \mu$ defined by:

$$\nu(A) \stackrel{\text{def.}}{=} \mu(f^{-1}(A))$$

$$\iff \int_Y g(y) \mathrm{d}\nu(y) \stackrel{\text{def.}}{=} \int_X g(f(x)) \mathrm{d}\mu(x)$$



Push Forward

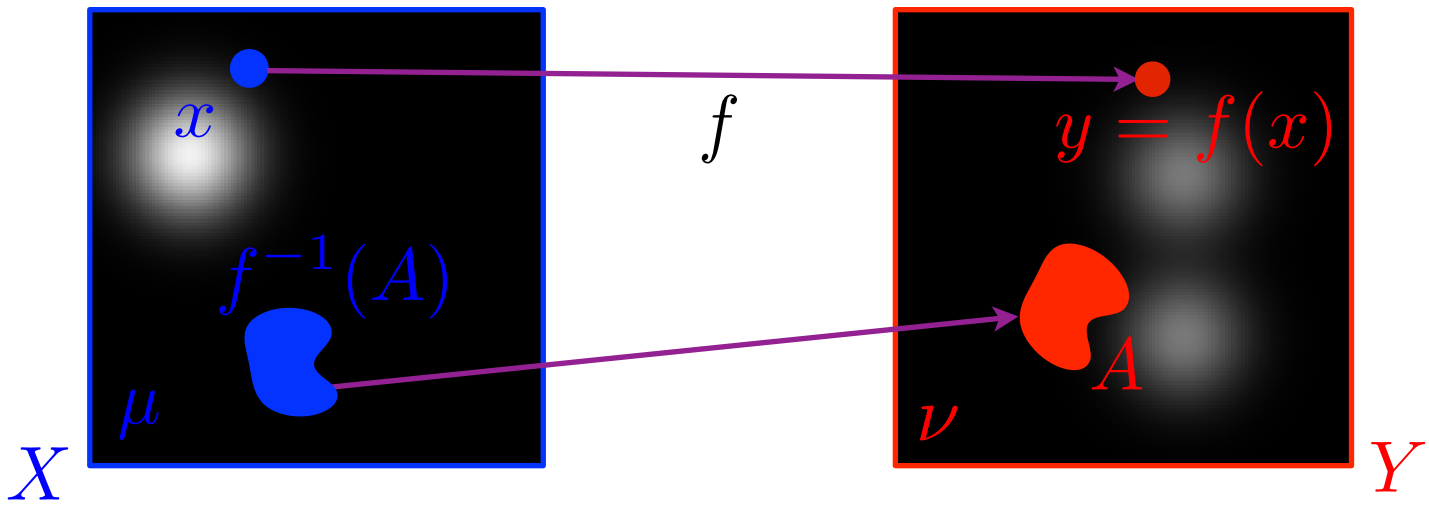
Radon measures (μ, ν) on (X, Y).

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

$\nu = f_\sharp \mu$ defined by:

$$\nu(A) \stackrel{\text{def.}}{=} \mu(f^{-1}(A))$$

$$\Longleftrightarrow \int_Y g(y)\mathrm{d}\nu(y) \stackrel{\text{def.}}{=} \int_X g(f(x))\mathrm{d}\mu(x)$$

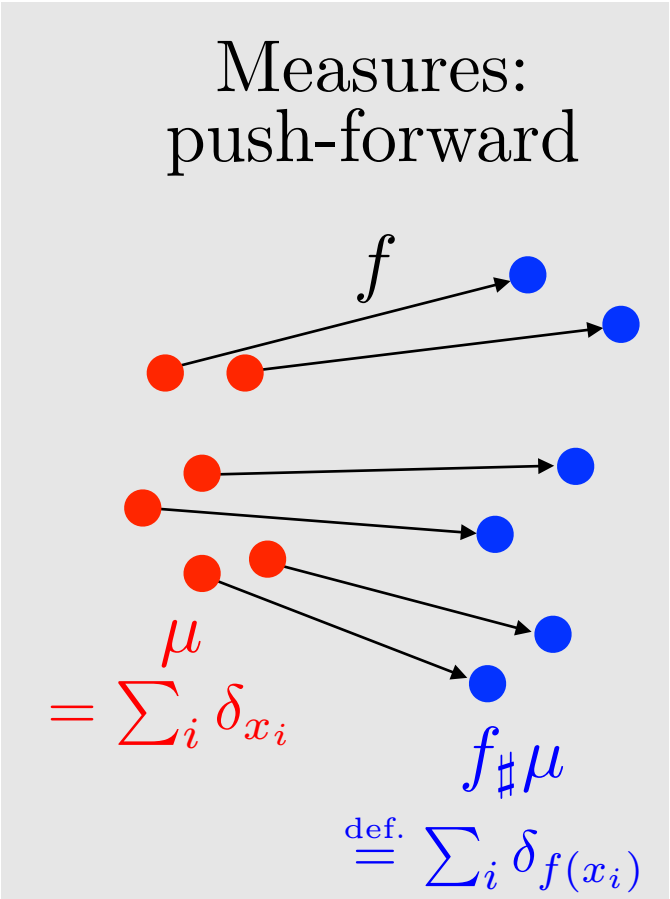


Smooth densities: $\mathrm{d}\mu = \rho(x)\mathrm{d}x$, $\mathrm{d}\nu = \xi(x)\mathrm{d}x$

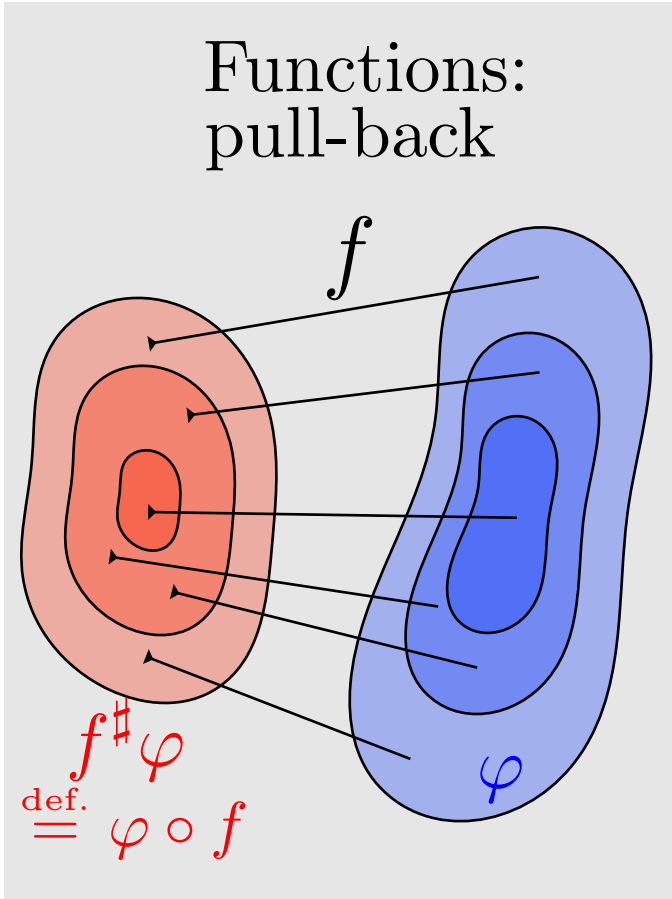
$$f_\sharp \mu = \nu \Longleftrightarrow \rho(f(x))|\det(\partial f(x))| = \xi(x)$$

Push-forward vs. Pull-back

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



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



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

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

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

Convergence of Random Variables

In mean

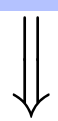
$$\lim_{n \to +\infty} \mathbb{E}(|X_n - X|^p) = 0$$

Almost sure

$$\mathbb{P}(\lim_{n \to +\infty} X_n = X) = 1$$

In probability

$$\forall \varepsilon > 0,\ \mathbb{P}(|X_n - X| > \varepsilon) \overset{n \to +\infty}{\longrightarrow} 0$$



In law

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

(the X_n can be defined on different spaces)

Overview

- Measures and Histograms
- **From Monge to Kantorovitch Formulations**
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein

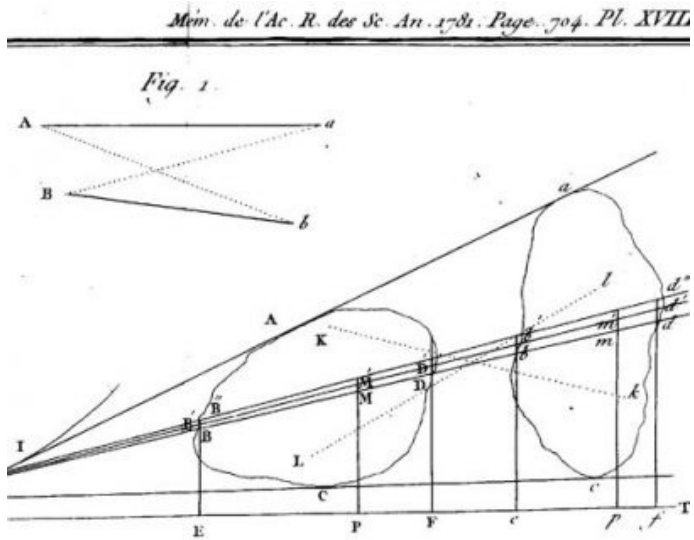
Gaspard Monge (1746-1818)

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.

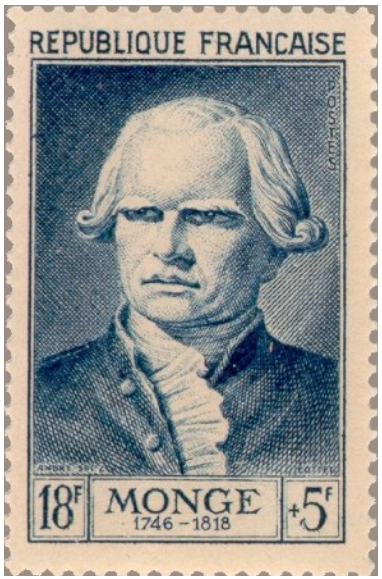
Par M. MONGE.

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*.

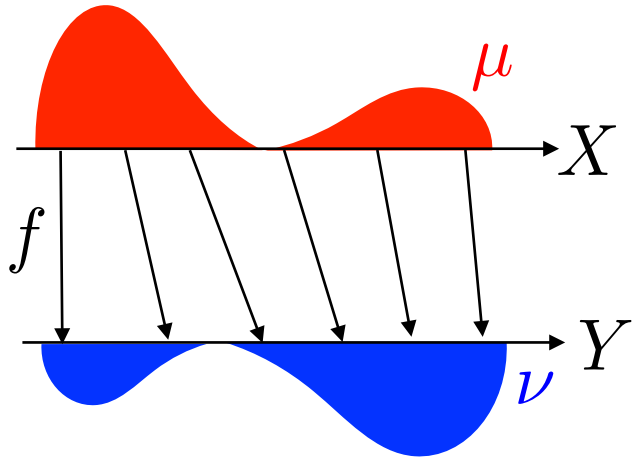


(1784)



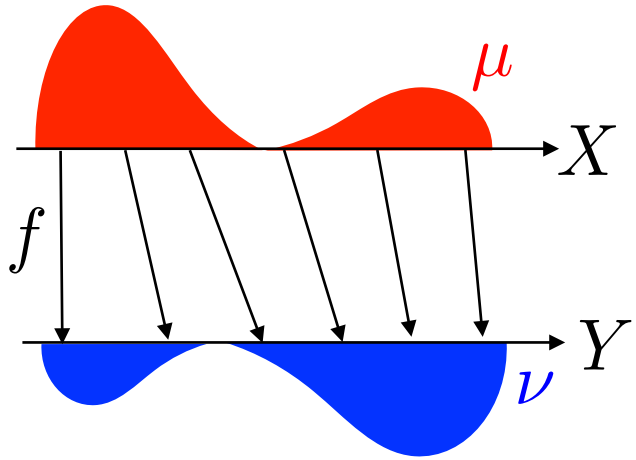
Monge Transport

$$\min_{\nu = f_\sharp \mu} \int_X c(x, f(x)) \mathrm{d}\mu(x)$$



Monge Transport

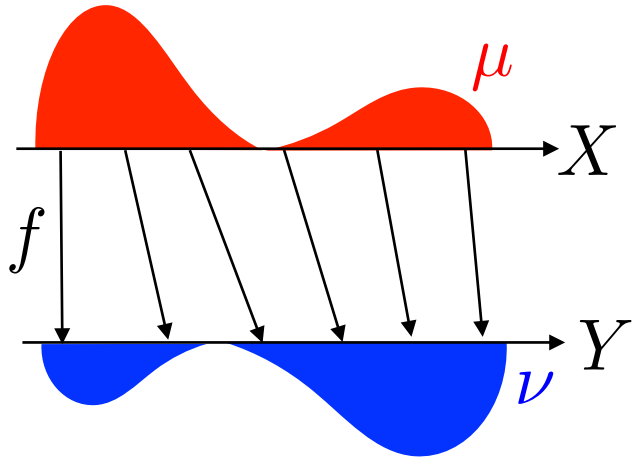
$$\min_{\nu = f_\sharp \mu} \int_X c(x, f(x)) \mathrm{d}\mu(x)$$



Theorem: [Brenier] for $c(x, y) = \|x - y\|^2$, (μ, ν) with density, there exists a unique optimal f. One has $f = \nabla\psi$ where ψ is the unique convex function such that $(\nabla\psi)_\sharp \mu = \nu$

Monge Transport

$$\min_{\nu = f_\sharp \mu} \int_X c(x, f(x)) \mathrm{d}\mu(x)$$

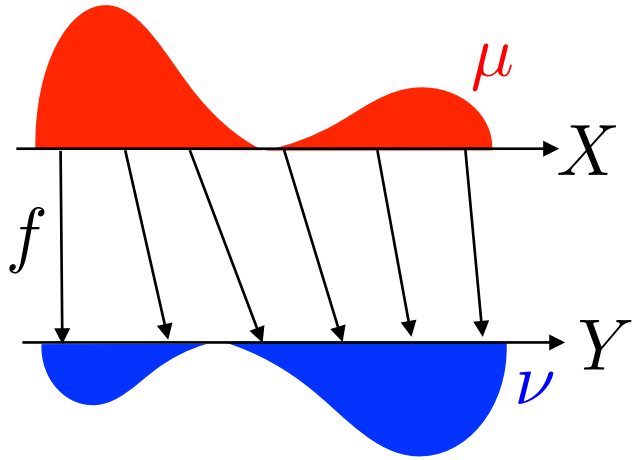


Theorem: [Brenier] for $c(x, y) = \|x - y\|^2$, (μ, ν) with density, there exists a unique optimal f. One has $f = \nabla\psi$ where ψ is the unique convex function such that $(\nabla\psi)_\sharp \mu = \nu$

Monge-Ampère equation: $\rho(\nabla\psi) \det(\partial^2\psi) = \xi$

Monge Transport

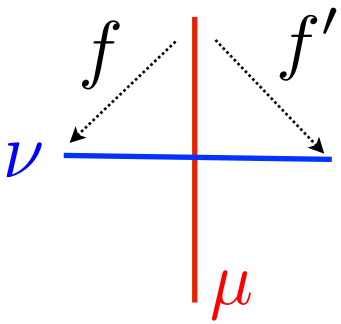
$$\min_{\nu = f_\sharp \mu} \int_X c(x, f(x)) \mathrm{d}\mu(x)$$

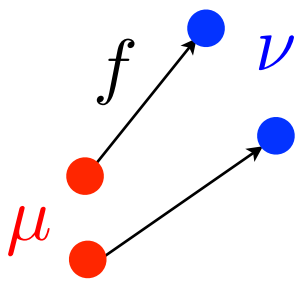


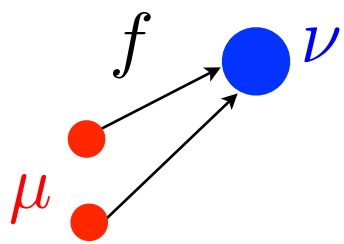
Theorem: [Brenier] for $c(x,y) = \|x-y\|^2$, (μ, ν) with density, there exists a unique optimal f. One has $f = \nabla\psi$ where ψ is the unique convex function such that $(\nabla\psi)_\sharp \mu = \nu$

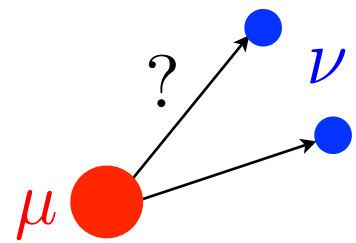
Monge-Ampère equation: $\rho(\nabla\psi)\det(\partial^2\psi) = \xi$

Non-uniqueness / non-existence:









Leonid Kantorovich (1912-1986)

Леонид Витальевич Канторович

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: $\Phi(e) \ge 0$, (3) it is absolutely additive: if $e = e_1 + e_2 + \cdots$; $e_i \cap e_k = 0$ $(i \ne k)$, 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

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

Before Kantorovitch

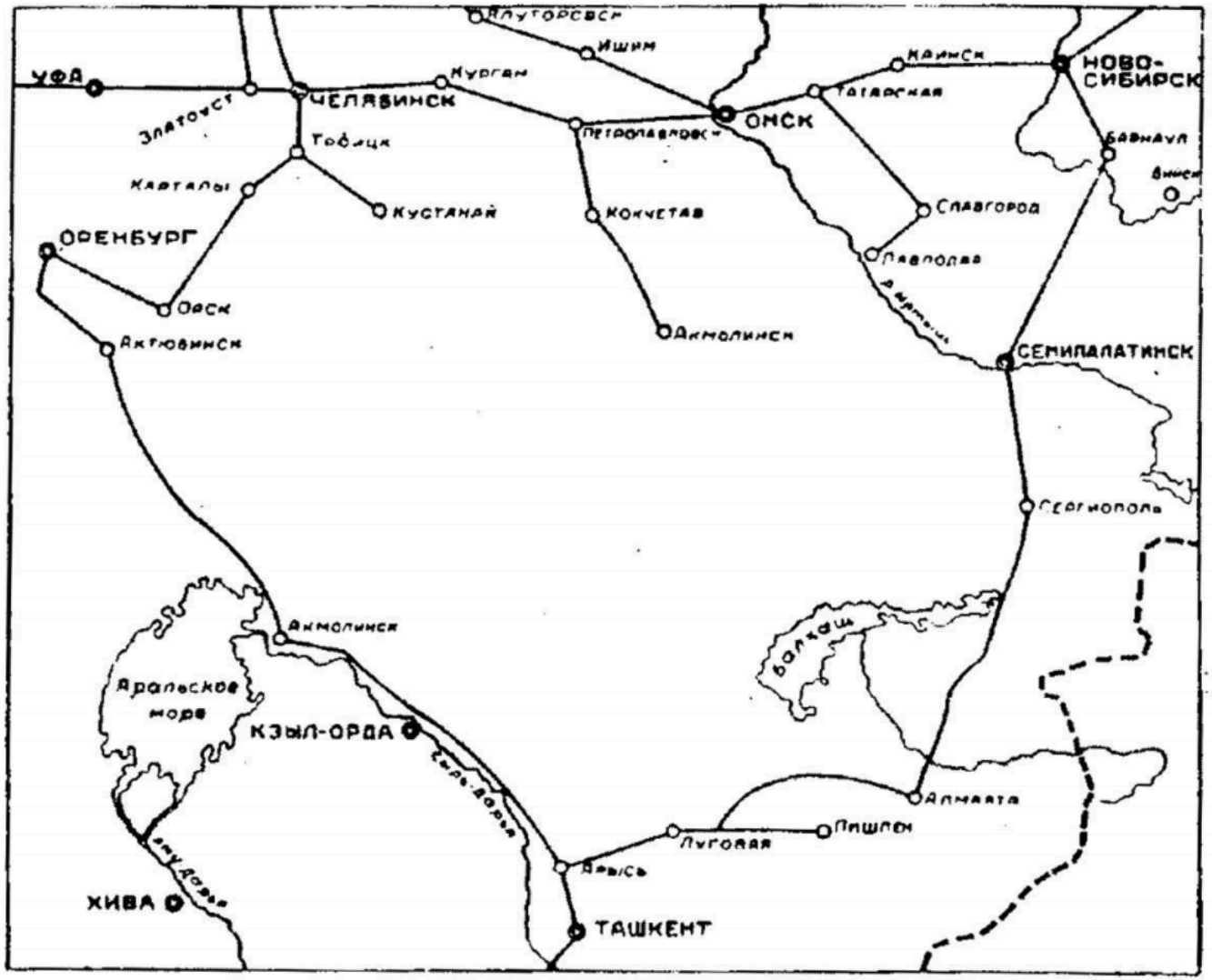


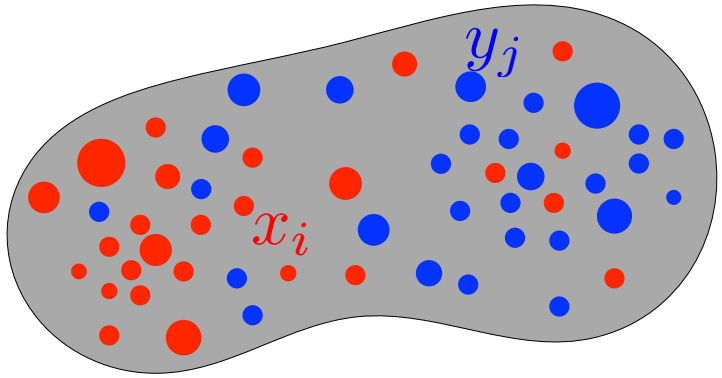
Figure 1: Figure from Tolstoĭ [1930] to illustrate a negative cycle

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

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

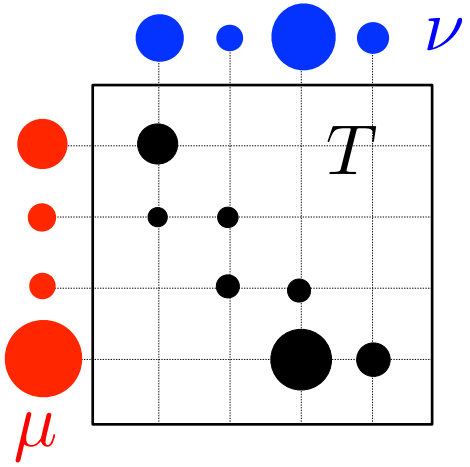


Points $(x_i)_i$, $(y_j)_j$

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

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

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



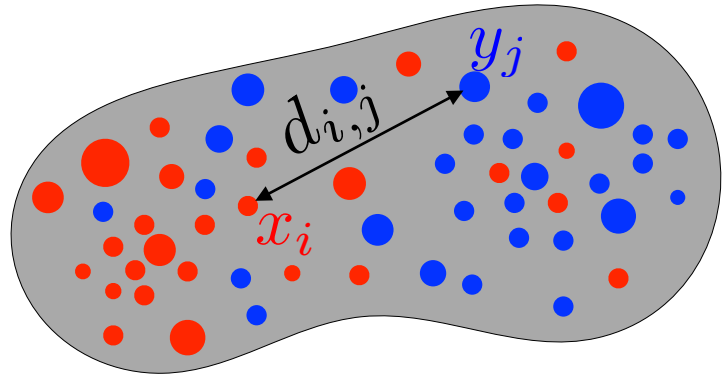
Def. *Couplings*

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

Kantorovitch's Formulation

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

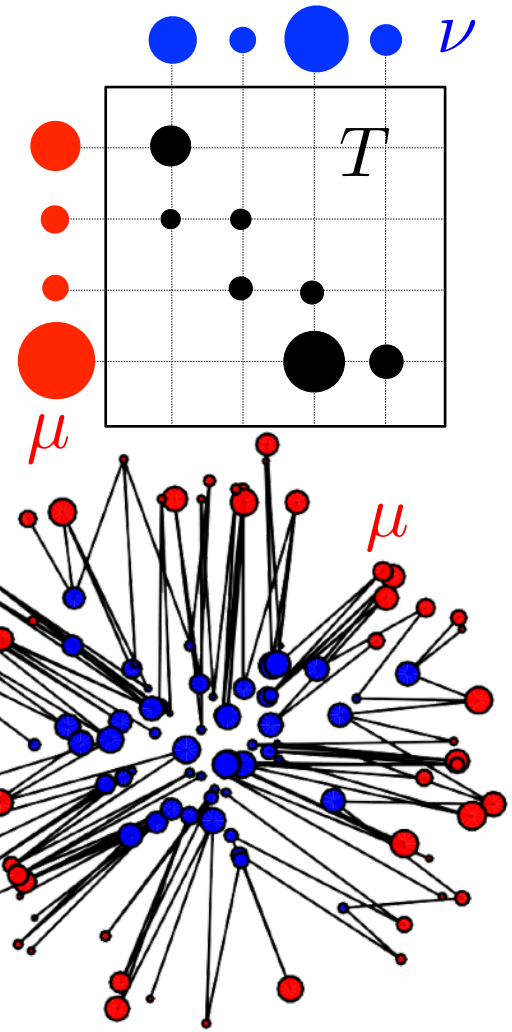
$\nu = \sum_j \nu_j \delta_{y_j}$



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

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

$\sum_{i=1}^{N_1} \mu_i = \sum_{j=1}^{N_2} \nu_j = 1$ $\quad d_{i,j} = d(x_i, y_j)$



Def. *Couplings*

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

Def. *Wasserstein Distance / EMD*

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

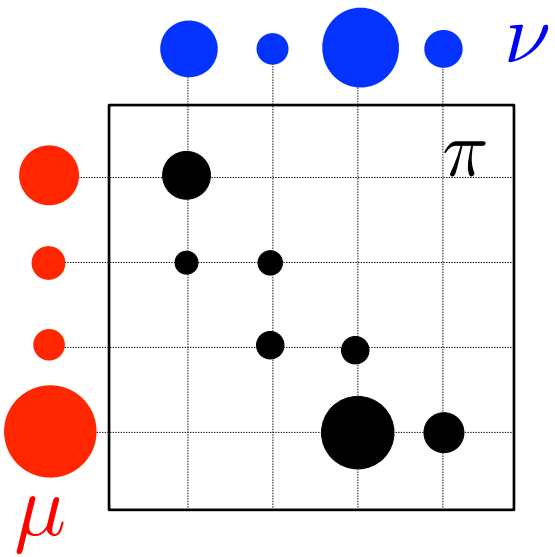
[Kantorovich 1942]

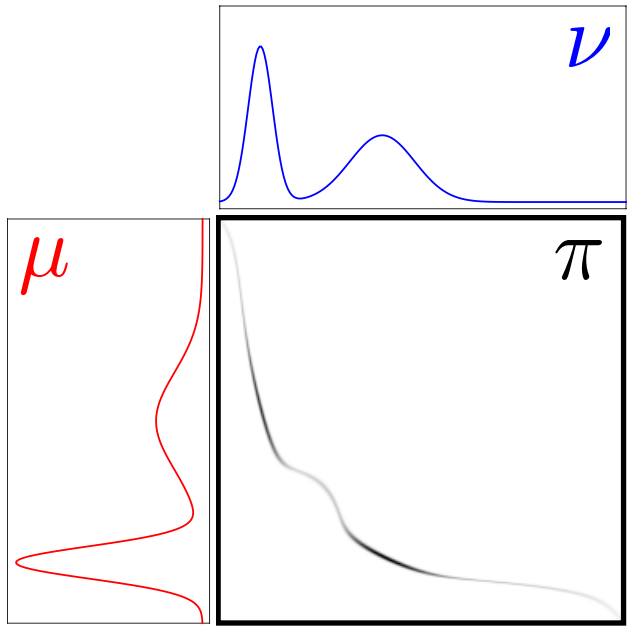
$\rightarrow W_p$ is a distance over Radon probability measures.

OT Between General Measures

Couplings: $\Pi(\mu,\nu) \stackrel{\text{def.}}{=} \{\pi \in \mathcal{M}_+(X \times X)\ ;\ P_{1\sharp}\pi = \mu, P_{2\sharp}\pi = \nu\}$

Marginals: $P_{1\sharp}\pi(S) \stackrel{\text{def.}}{=} \pi(S, X)$ $\quad$ $P_{2\sharp}\pi(S) \stackrel{\text{def.}}{=} \pi(X, S)$

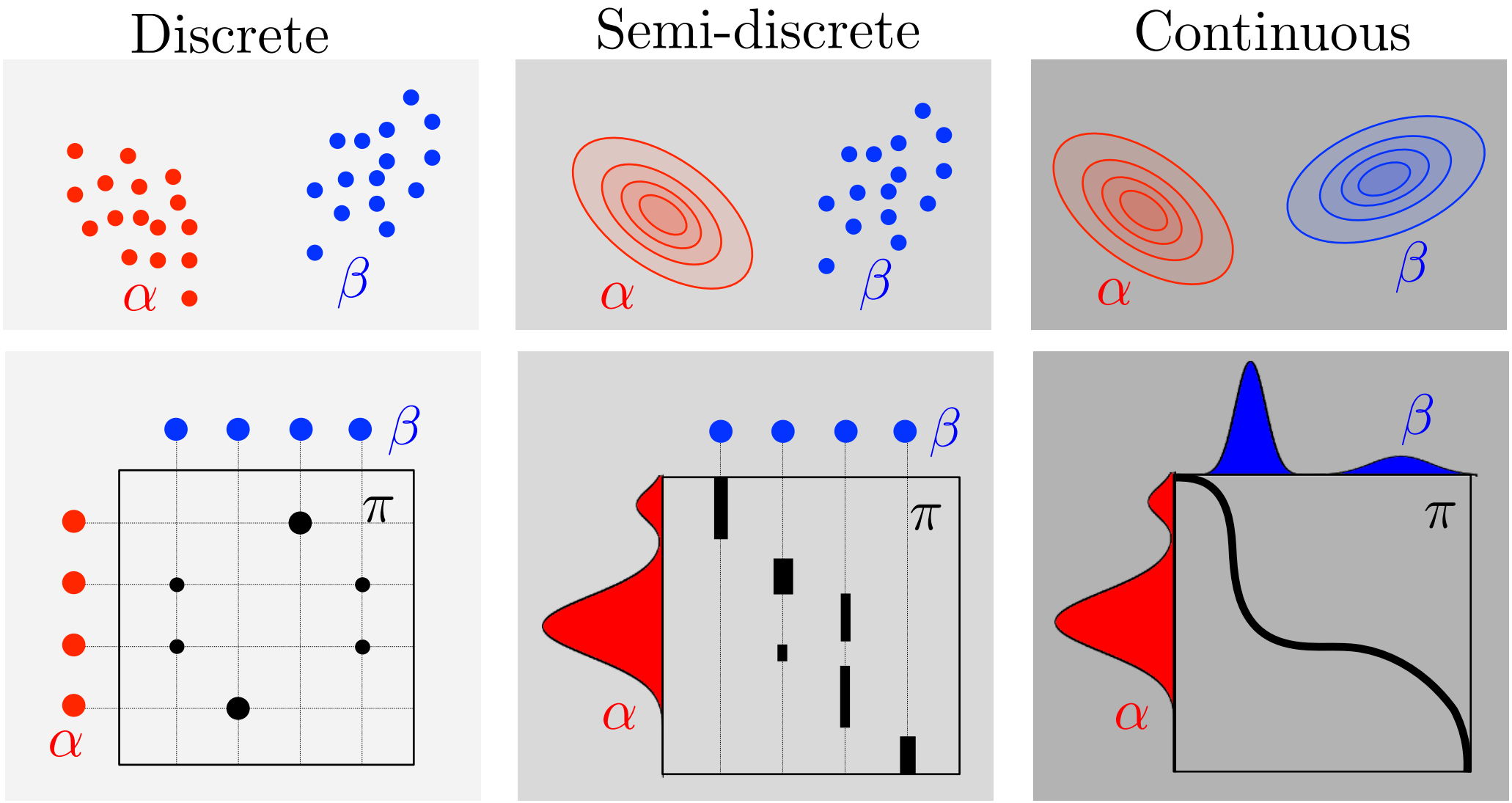




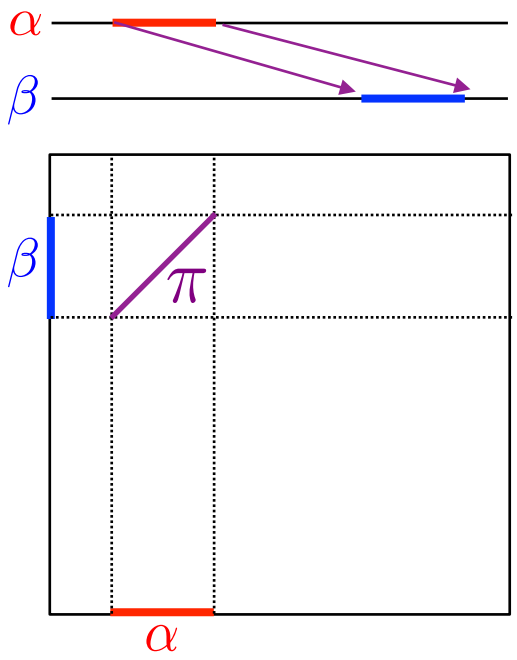
Optimal transport: [Kantorovitch 1942]

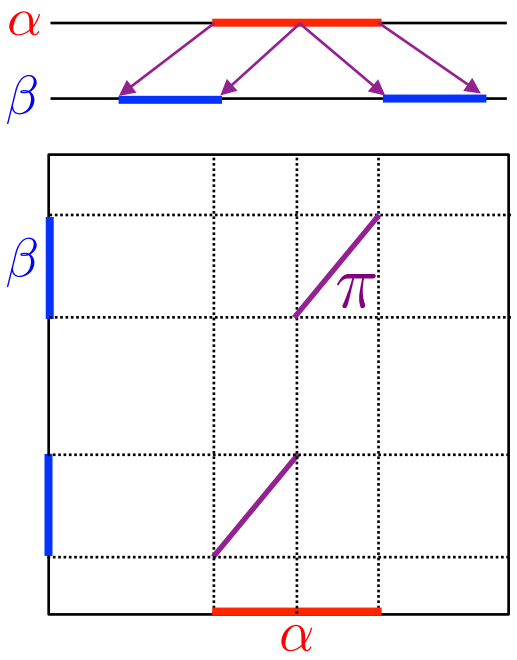
$$W_p^p(\mu,\nu) \stackrel{\text{def.}}{=} \min_{\pi} \left\{ \langle d^p, \pi \rangle = \int_{X\times X} d(x,y)^p \mathrm{d}\pi(x,y)\ ;\ \pi \in \Pi(\mu,\nu) \right\}$$

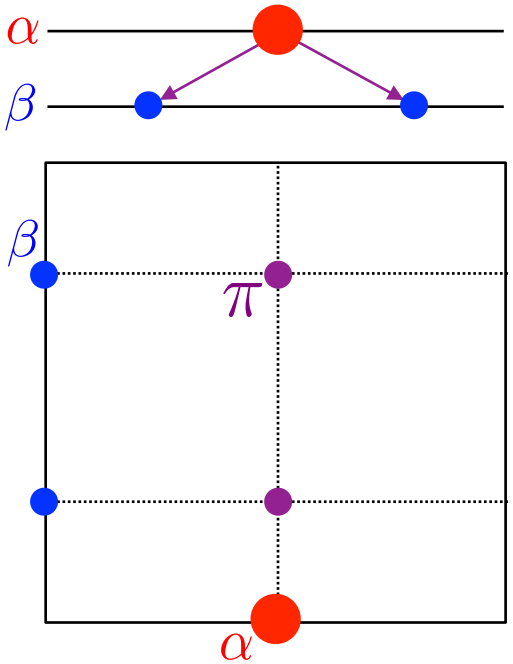
Couplings: the 3 Settings

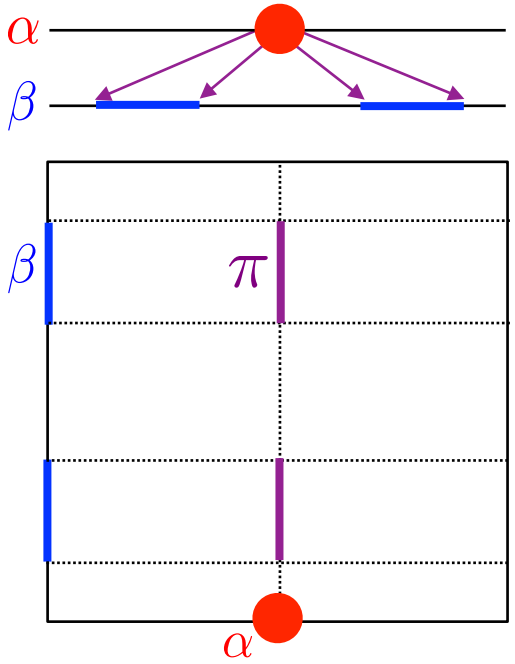


Couplings





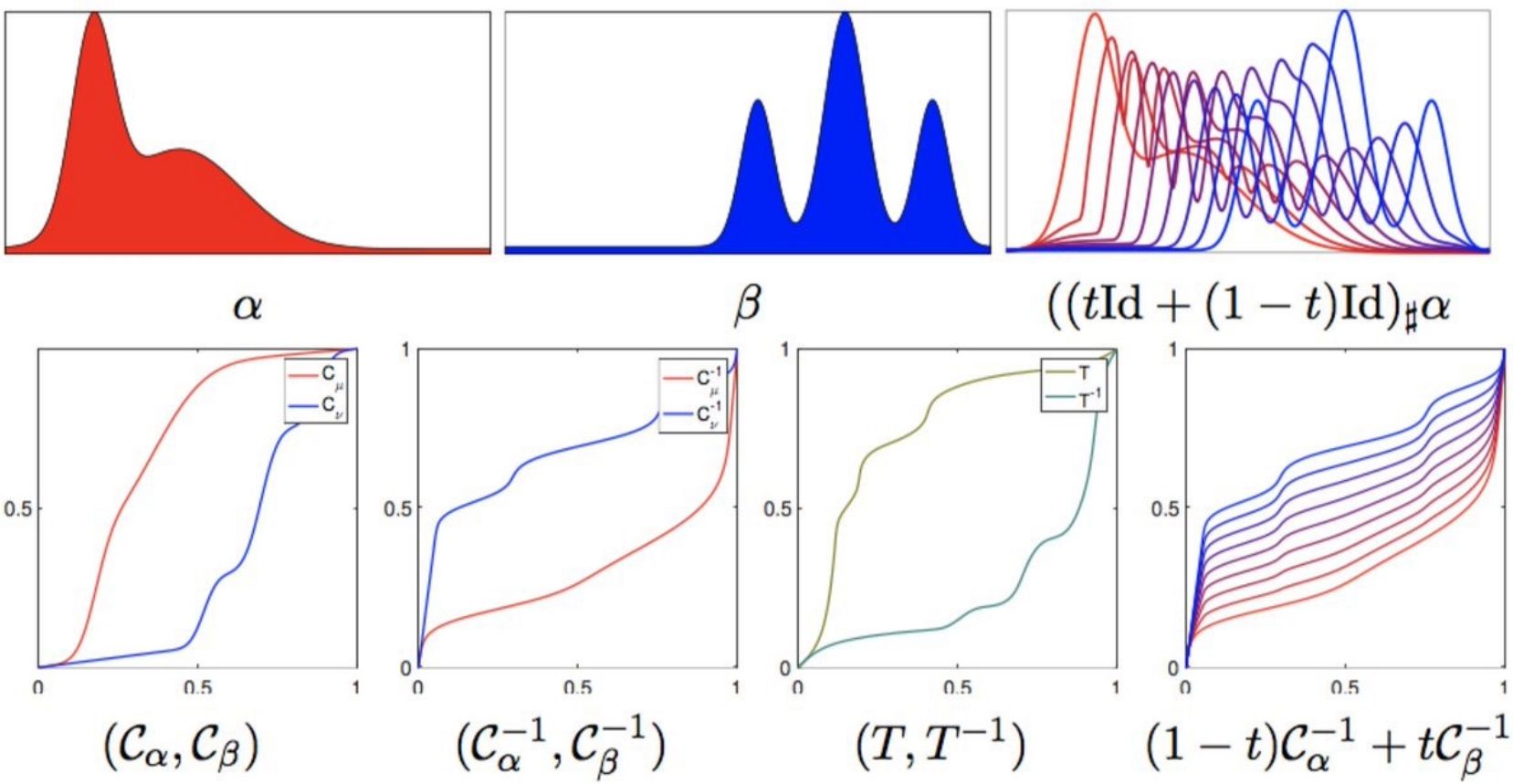




1-D Optimal Transport

Remark. If $\Omega = \mathbb{R}$, $\boldsymbol{c}(x, y) = \boldsymbol{c}(|x - y|)$, $\boldsymbol{c}$ convex, $F_{\boldsymbol{\mu}}^{-1}, F_{\boldsymbol{\nu}}^{-1}$ quantile functions,

$$W(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int_0^1 \boldsymbol{c}(|F_{\boldsymbol{\mu}}^{-1}(x) - F_{\boldsymbol{\nu}}^{-1}(x)|)dx$$



OT Between Gaussians

Remark. If $\Omega = \mathbb{R}^d$, $\boldsymbol{c}(x,y) = \|x-y\|^2$, and $\boldsymbol{\mu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}), \boldsymbol{\nu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\nu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})$ then

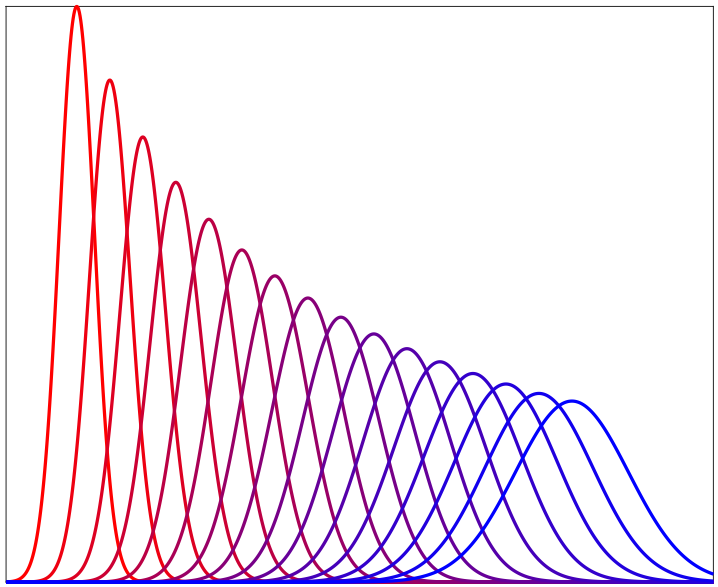
$$W_2^2(\boldsymbol{\mu},\boldsymbol{\nu}) = \|\mathbf{m}_{\boldsymbol{\mu}} - \mathbf{m}_{\boldsymbol{\nu}}\|^2 + B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2$$

where B is the Bures metric

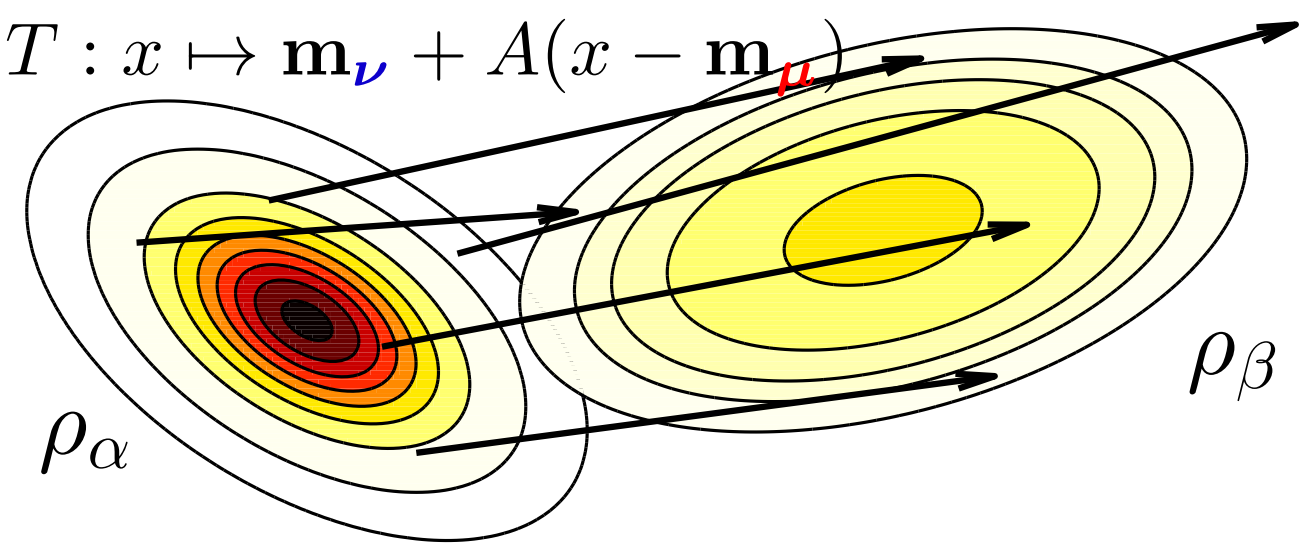
$$B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2 = \mathrm{trace}(\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_{\boldsymbol{\nu}} - 2(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\nu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2})^{1/2}).$$

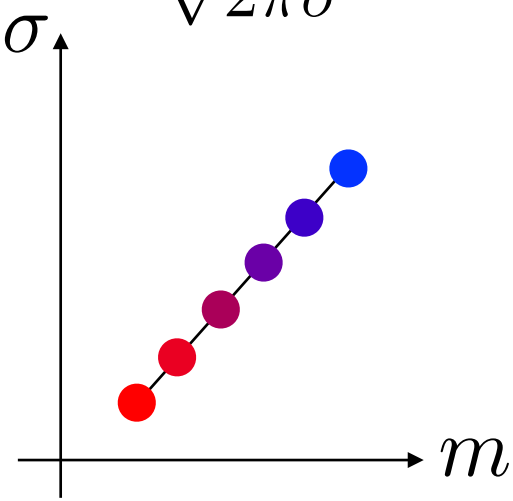
The map $T : x \mapsto \mathbf{m}_{\boldsymbol{\nu}} + A(x - \mathbf{m}_{\boldsymbol{\mu}})$ is **optimal**,

where $A = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\frac{1}{2}} \right)^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-\frac{1}{2}}$.



$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$





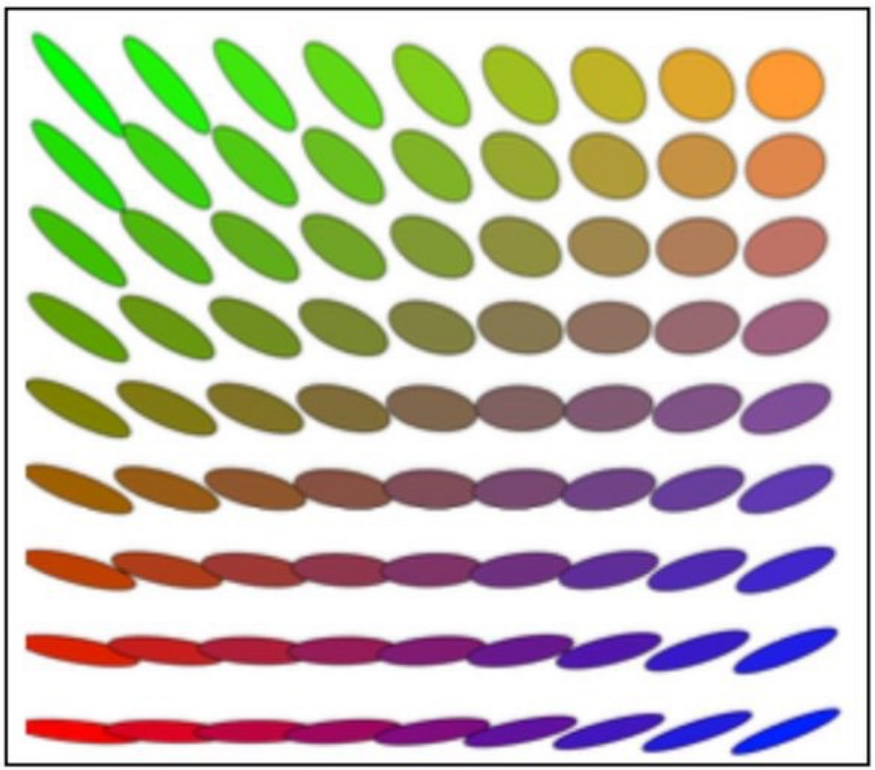
OT on Gaussians and Bures' Distance

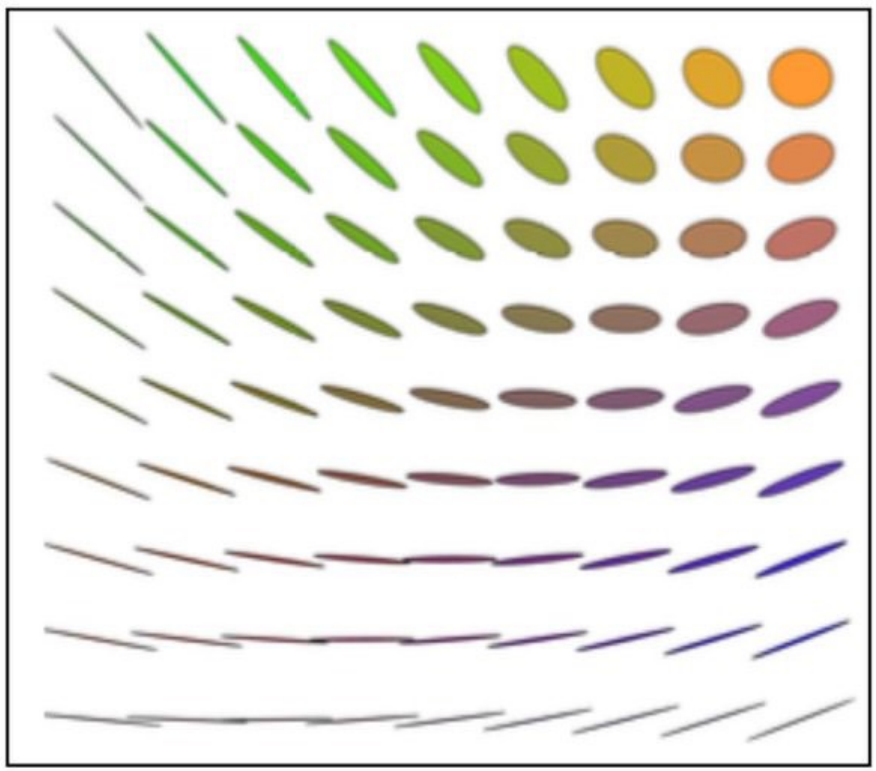
Remark 2.11 (Distance between Gaussians). If $\alpha = \mathcal{N}(m_\alpha, C_\alpha)$ and $\beta = \mathcal{N}(m_\beta, C_\beta)$, then one can show that

$$\mathcal{W}_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}(C_\alpha, C_\beta)^2 \tag{2.19}$$

where $\mathcal{B}$ is the so-called Bures metric

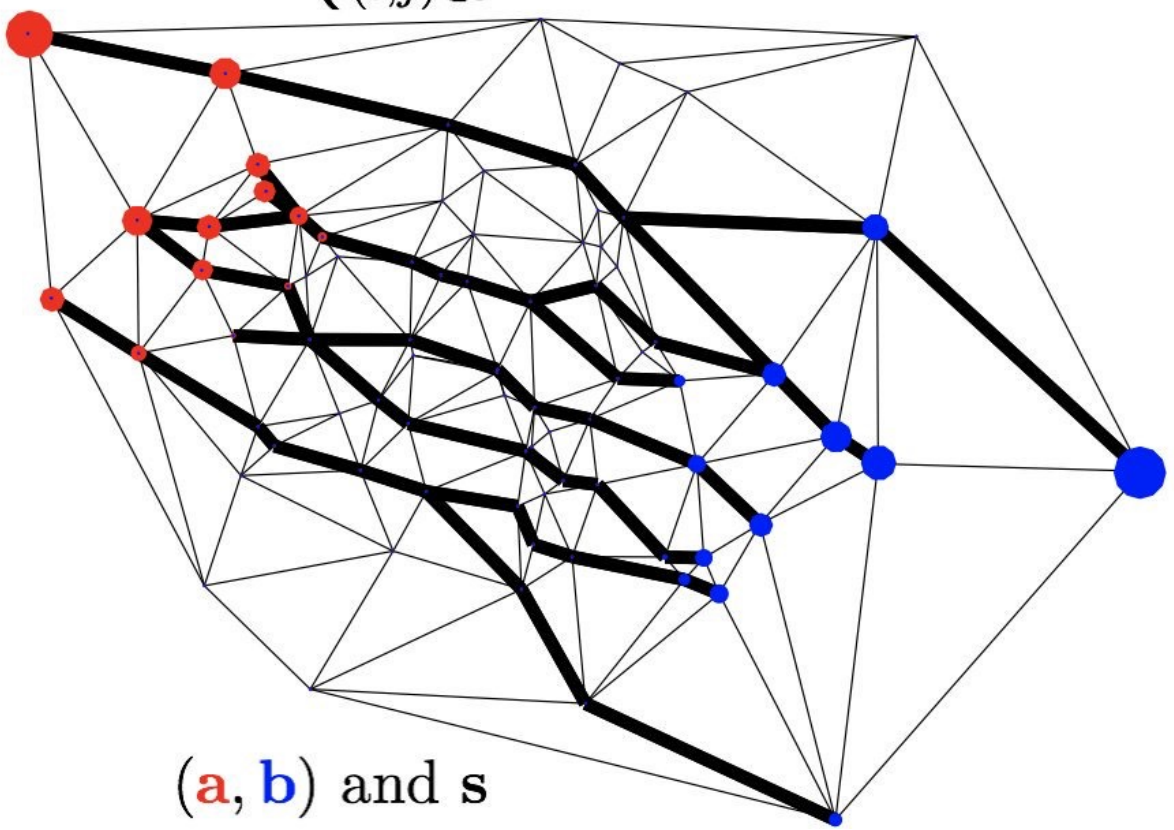
$$\mathcal{B}(C_\alpha, C_\beta)^2 \stackrel{\text{def.}}{=} \operatorname{tr}\left(C_\alpha + C_\alpha - 2(C_\alpha^{1/2} C_\beta C_\alpha^{1/2})^{1/2}\right) \tag{2.20}$$





W1 OT and Min-cost Flows

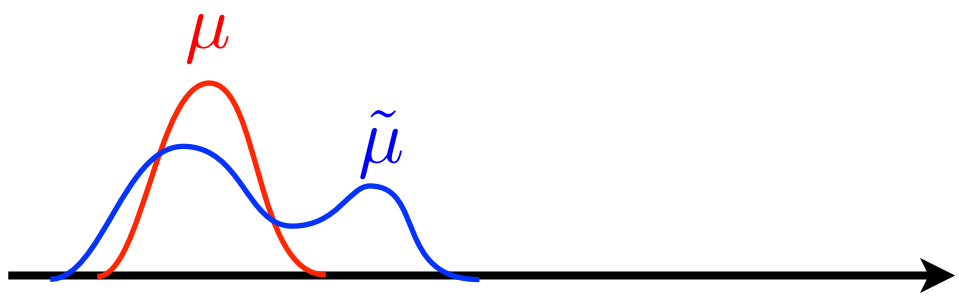
$$\mathrm{W}_1(\mathbf{a},\mathbf{b}) = \min_{\mathbf{s}\in\mathbb{R}_+^{\mathcal{E}}} \left\{ \sum_{(i,j)\in\mathcal{E}} \mathbf{w}_{i,j}\mathbf{s}_{i,j} \;:\; \mathrm{div}(\mathbf{s}) = \mathbf{a}-\mathbf{b} \right\}$$



$(\mathbf{a}, \mathbf{b})$ and $\mathbf{s}$

Metrics on the Space of Measures

$$\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$$
$$\mathrm{d}\tilde{\mu}(x) = \tilde{\rho}(x)\mathrm{d}x$$



Bins-to-bins metrics:

Kullback-Leibler divergence:

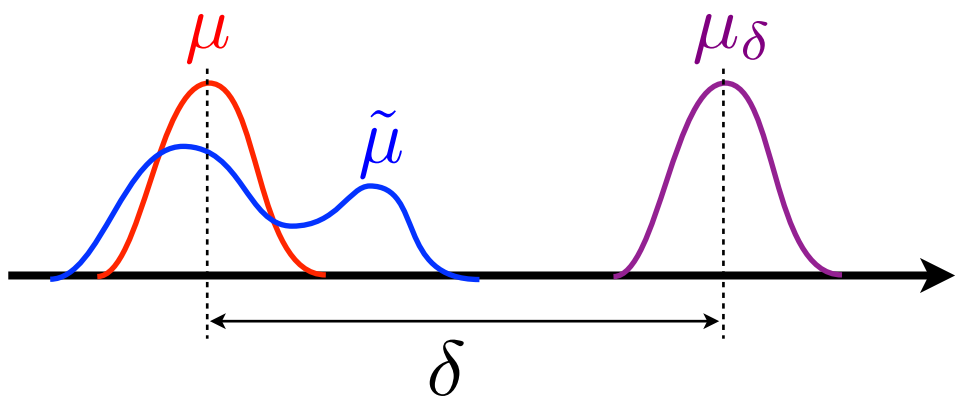
$$D_{\mathrm{KL}}(\mu, \tilde{\mu}) = \int \rho(x) \log \frac{\rho(x)}{\tilde{\rho}(x)} \mathrm{d}x$$

Hellinger distance:

$$D_{\mathrm{H}}(\mu, \tilde{\mu})^2 = \int \left(\sqrt{\rho(x)} - \sqrt{\tilde{\rho}(x)} \right)^2 \mathrm{d}x$$

Metrics on the Space of Measures

$$\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$$
$$\mathrm{d}\tilde{\mu}(x) = \tilde{\rho}(x)\mathrm{d}x$$
$$\mathrm{d}\mu_\delta(x) = \rho(x-\delta)\mathrm{d}x$$



Bins-to-bins metrics:

Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(\mu, \tilde{\mu}) = \int \rho(x) \log \frac{\rho(x)}{\tilde{\rho}(x)} \mathrm{d}x$$

Hellinger distance:

$$D_{\mathrm{H}}(\mu, \tilde{\mu})^2 = \int \left(\sqrt{\rho(x)} - \sqrt{\tilde{\rho}(x)} \right)^2 \mathrm{d}x$$

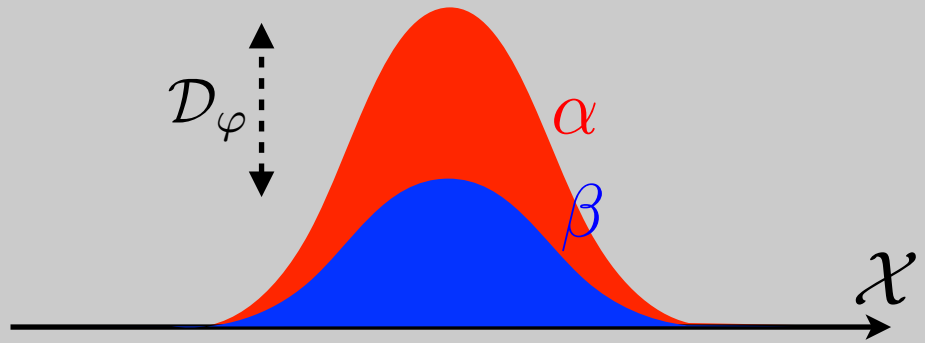
Effect of translation:

$$D(\mu, \mu_\delta) \approx \mathrm{cst} \qquad W_2(\mu, \mu_\delta) = \delta$$

Csiszar Divergence vs Dual Norms

Csiszár divergences:

$$\mathcal{D}_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta$$

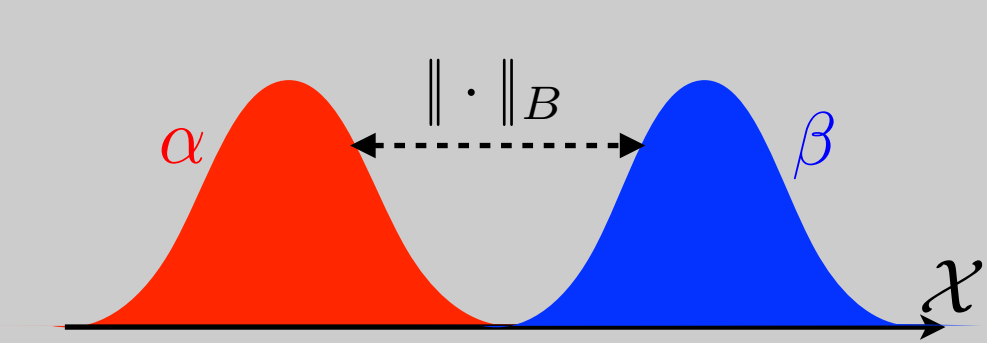


Strong topology

$\longrightarrow$ KL, TV, χ^2, Hellinger ...

Dual norms:

$$\|\alpha - \beta\|_B \overset{\text{def.}}{=} \max_{f \in B} \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x))$$



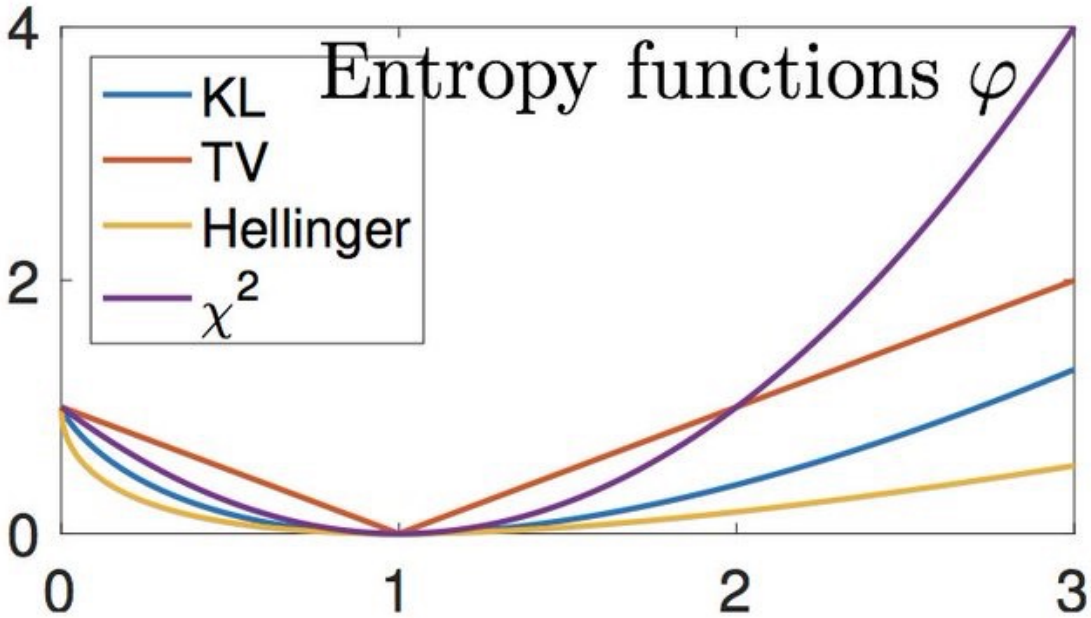
Weak topology

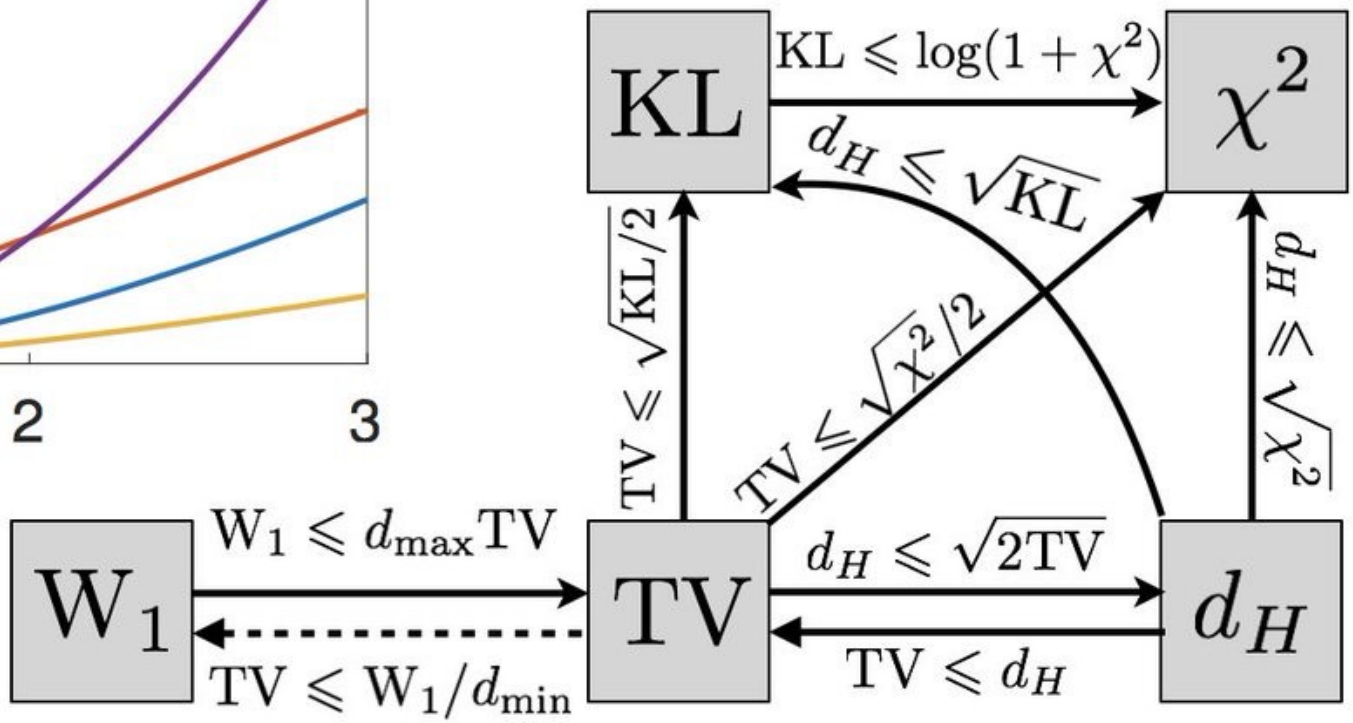
$\longrightarrow$ W_1, flat, RKHS*, energy dist, ...

Csiszar Divergence

$$\mathcal{D}_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta + \varphi'_\infty \alpha^\perp(\mathcal{X})$$

$$\varphi'_\infty = \lim_{x \uparrow +\infty} \varphi(x)/x \in \mathbb{R} \cup \{\infty\}$$





Csiszár divergences, a unifying way to define losses between arbitrary positive measures (discrete & densities). https://en.wikipedia.org/wiki/F-divergence …

Dual Norms

Dual norms: (aka Integral Probability Metrics)

$$\|\alpha - \beta\|_B \stackrel{\text{def.}}{=} \max \left\{ \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x)) \; ; \; f \in B \right\}$$

Wasserstein 1: $B = \{f \; ; \; \|\nabla f\|_\infty \leqslant 1\}$.

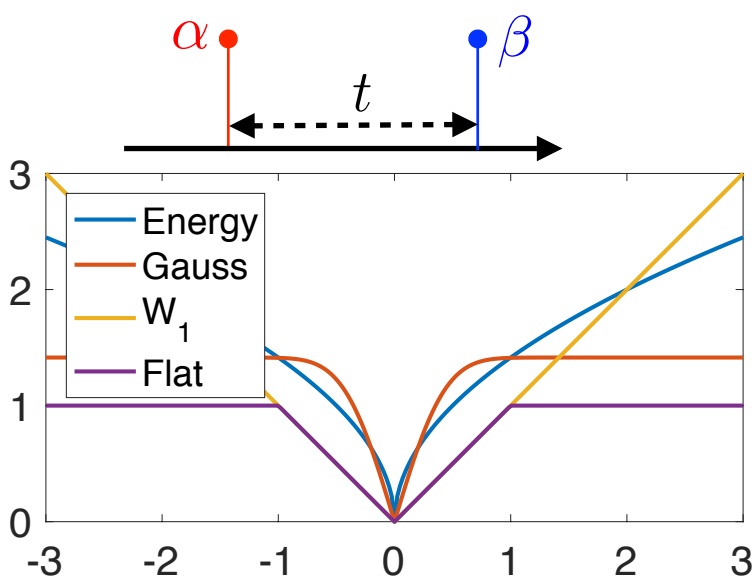
Flat norm: $B = \{f \; ; \; \|f\|_\infty \leqslant 1, \|\nabla f\|_\infty \leqslant 1\}$.

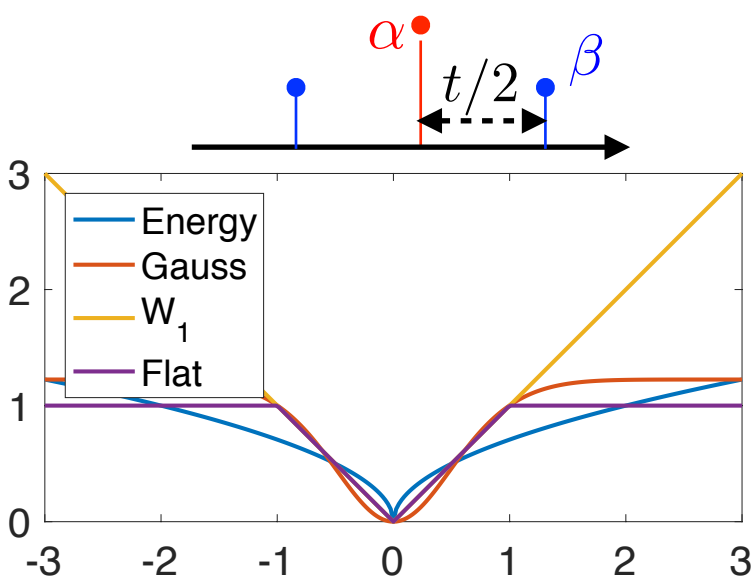
RKHS: $B = \left\{f \; ; \; \|f\|_k^2 \leqslant 1\right\}$.

$$\|\alpha - \beta\|_B^2 = \int k(x,x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int k(x,x')\mathrm{d}\beta(y)\mathrm{d}\beta(y') - 2\int k(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$$

Energy distance: $k(x,y) = -\|x-y\|$

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





RKHS Norms aka Maximum Mean Discrepency

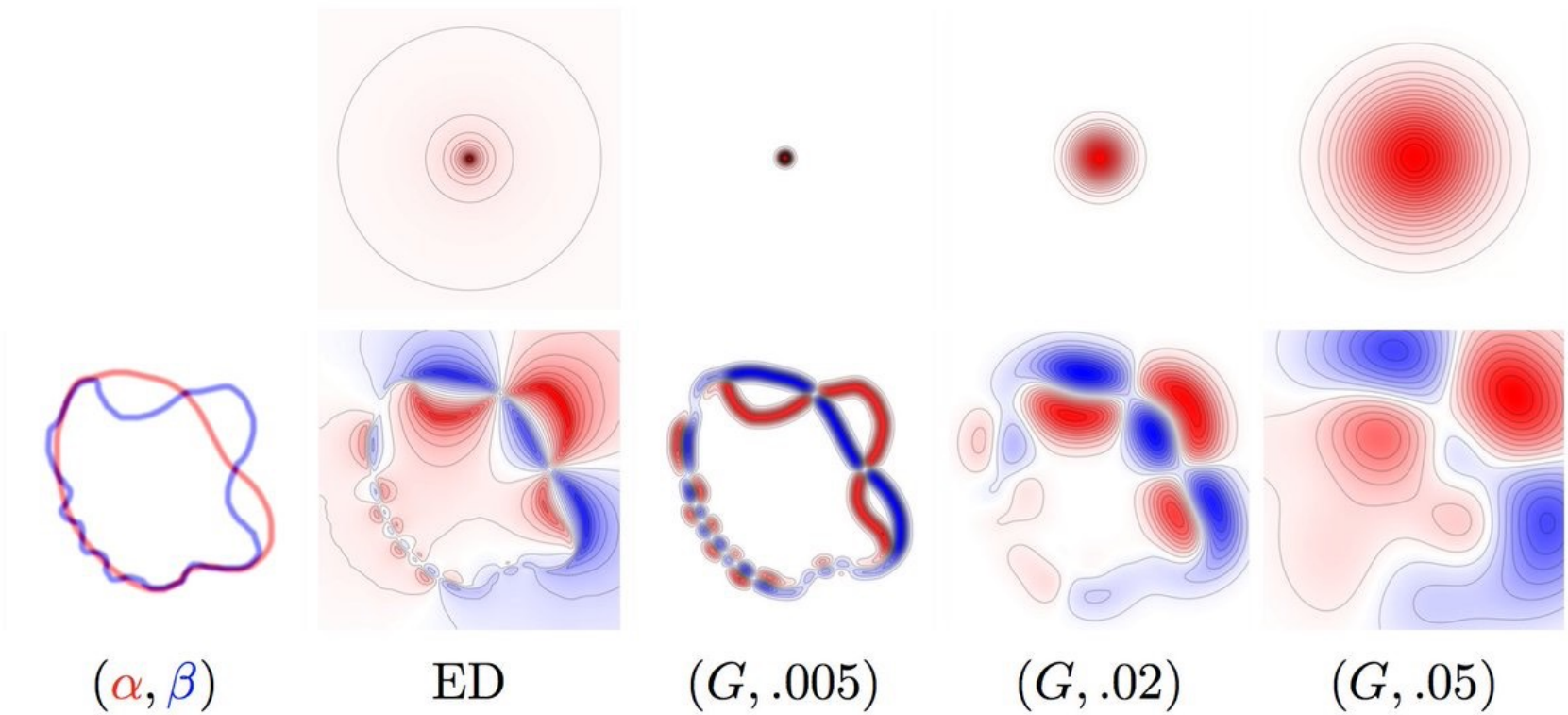
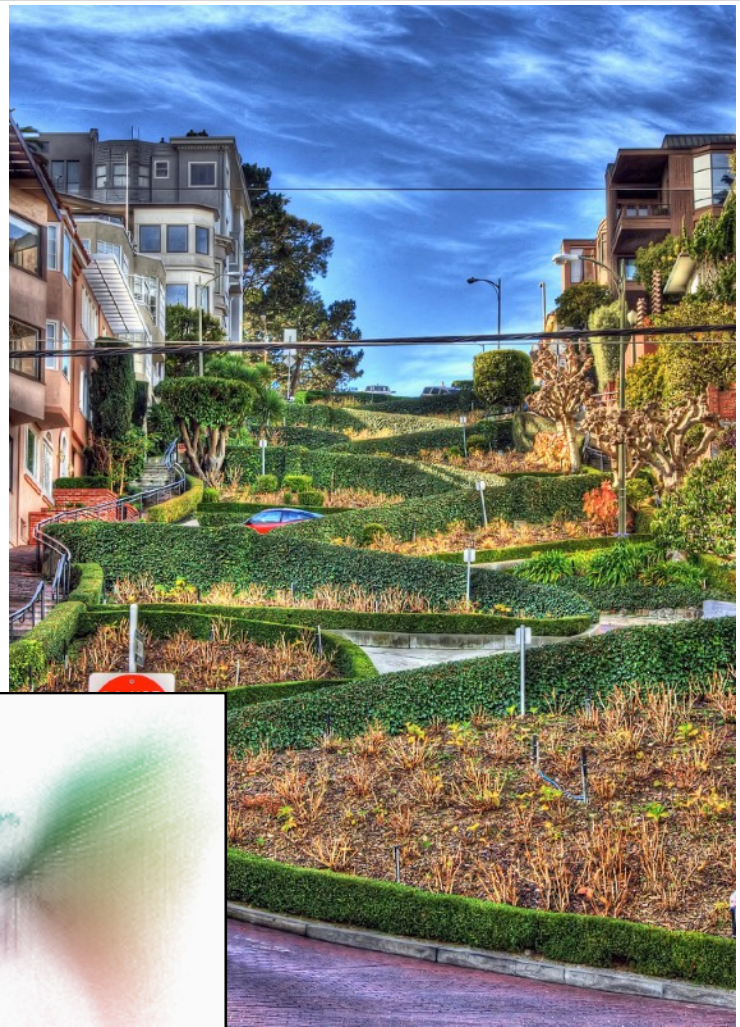
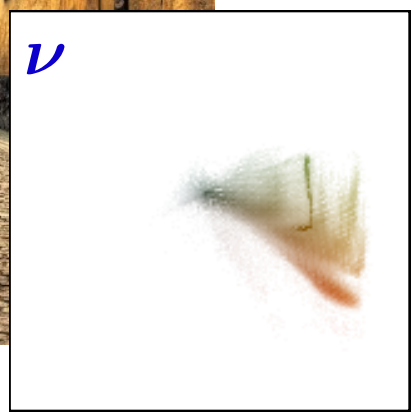


Figure 8.4: Top row: display of ψ such that $\|\alpha - \beta\|_k = \|\psi \star (\alpha - \beta)\|_{L^2(\mathbb{R}^2)}$, formally defined over Fourier as $\hat{\psi}(\omega) = \sqrt{\hat{\varphi}(\omega)}$ where $k^*(x, x') = \varphi(x - x')$. Bottom row: display of $\psi \star (\alpha - \beta)$. (G,$\sigma$) stands for Gaussian kernel of variance σ^2 and ED for Energy Distance kernel (in which case $\psi(x) = 1/\sqrt{\|x\|}$).

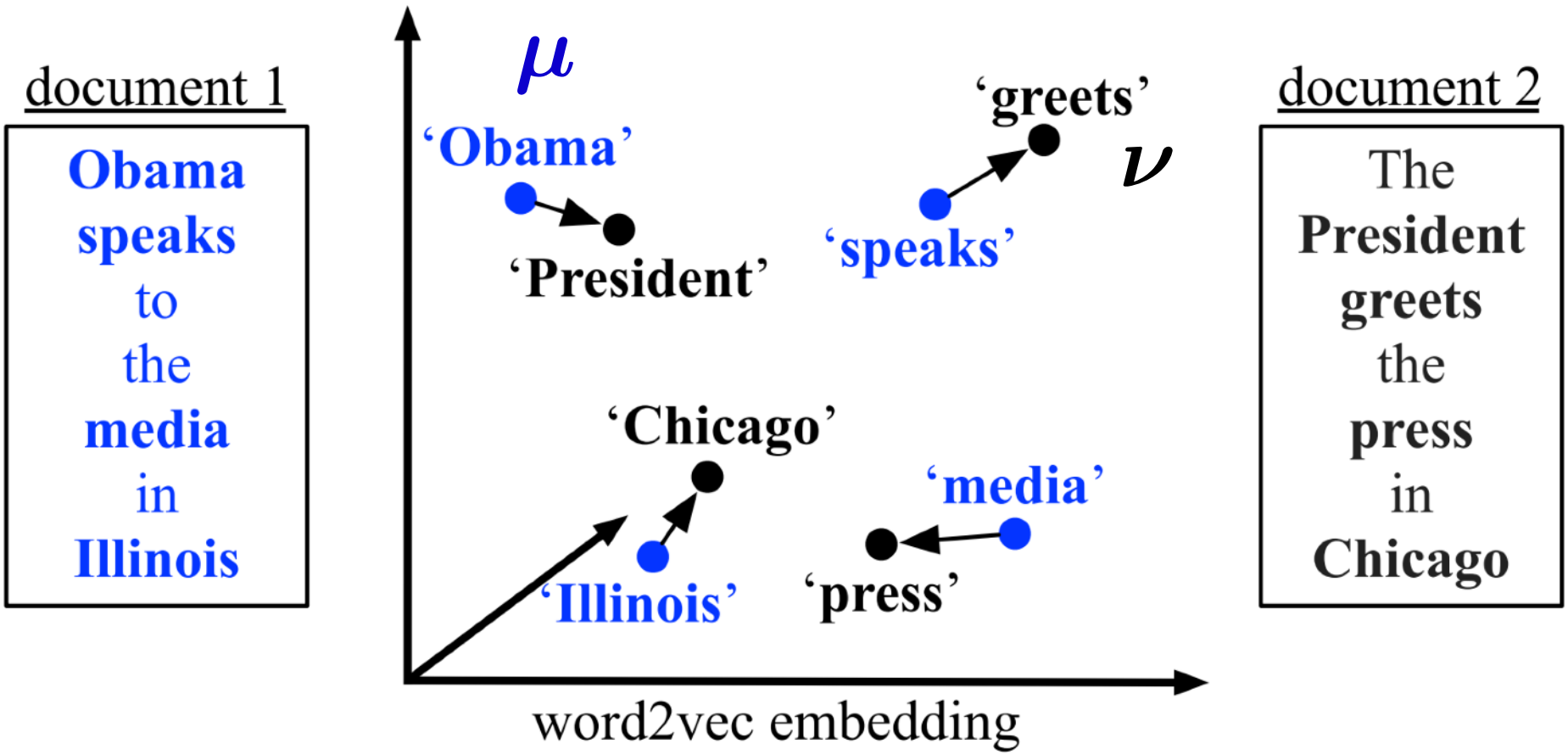
The Earth Mover's Distance





[Rubner'98] $\mathrm{dist}(I_1, I_2) = W_1(\mu, \nu)$

The Word Mover's Distance



[Kusner'15] $\text{dist}(D_1, D_2) = W_2(\mu, \nu)$

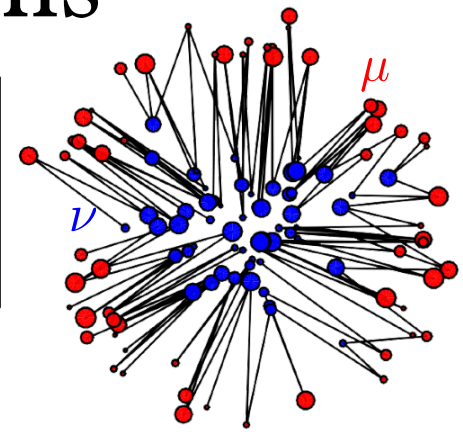
Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- **Linear Programming and Semi-discrete**
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein

Algorithms

Linear programming:

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



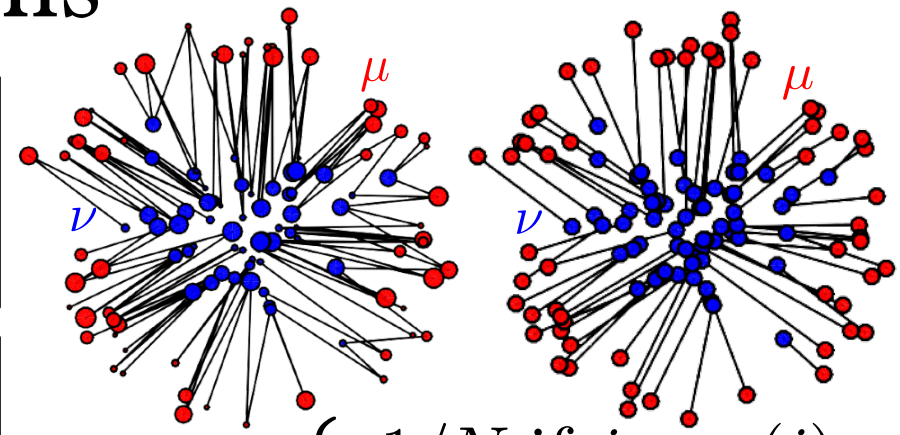
Algorithms

Linear programming:

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

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

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



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

Algorithms

Linear programming:

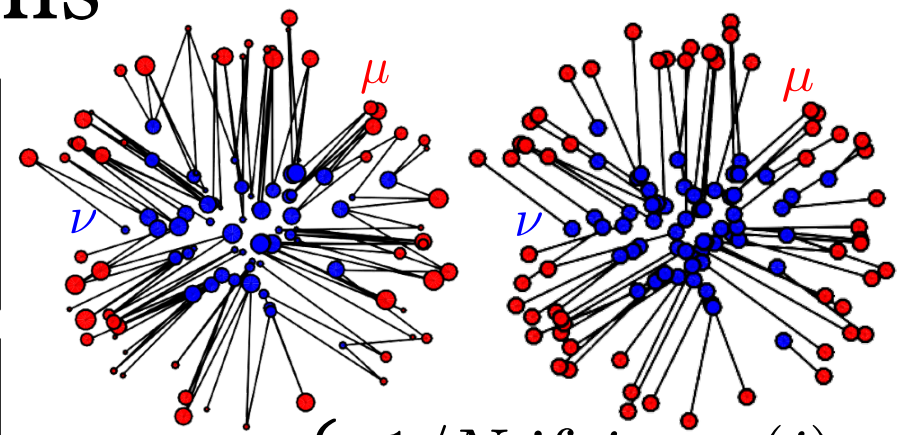
$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$

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

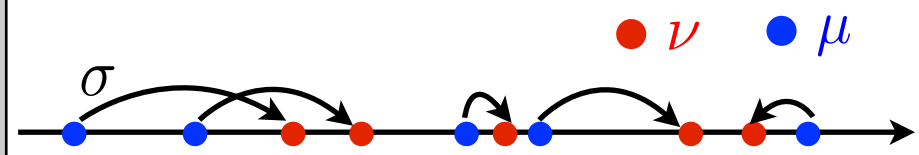
$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$

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

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



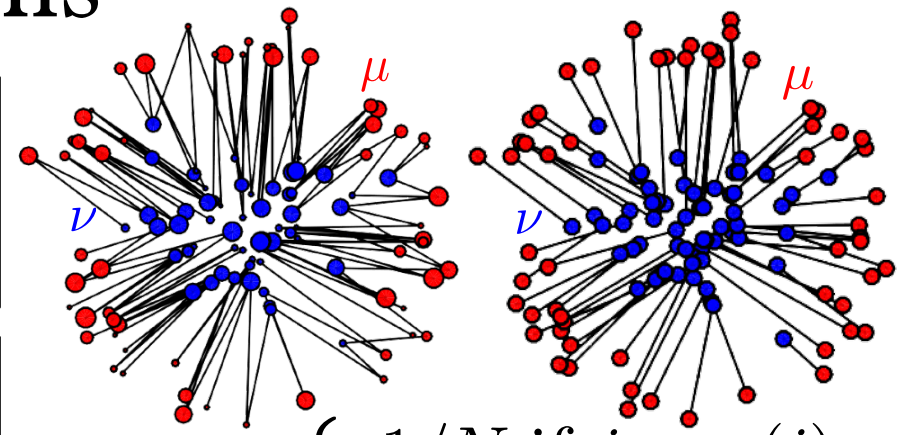
$$T_{i,j} = \begin{cases} 1/N \text{ if } j = \sigma(i), \\ 0 \text{ otherwise.} \end{cases}$$



Algorithms

Linear programming:

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



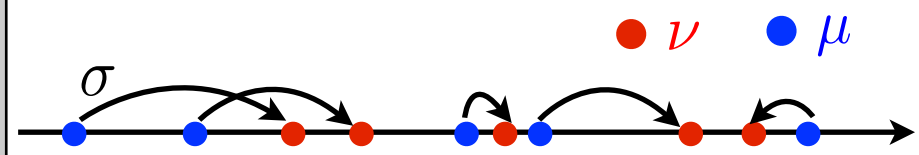
Hungarian/Auction: $\sim O(N^3)$

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

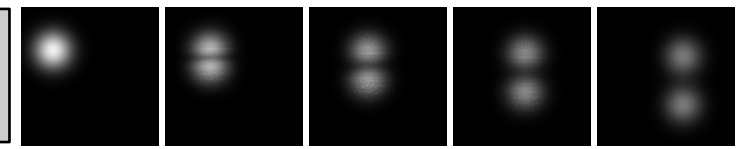
$$T_{i,j} = \begin{cases} 1/N \text{ if } j = \sigma(i), \\ 0 \text{ otherwise.} \end{cases}$$

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

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

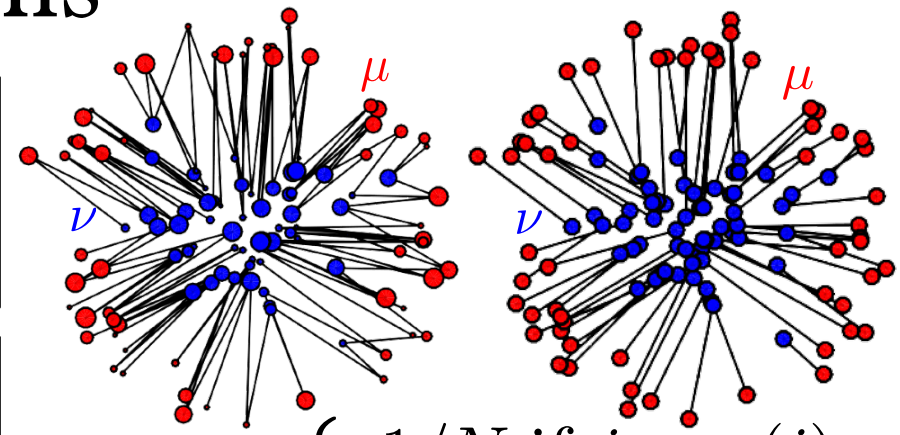


Monge-Ampère/Benamou-Brenier, $d = \|\cdot\|_2^2$.



Algorithms

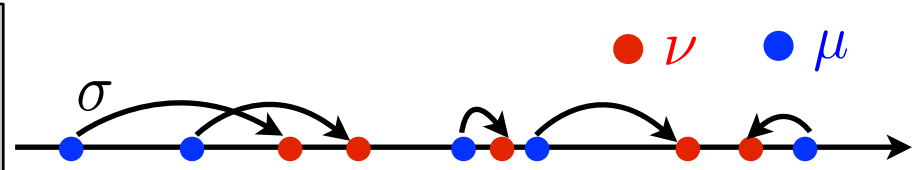
Linear programming:
$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$



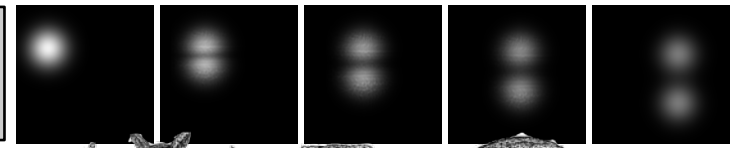
Hungarian/Auction: $\sim O(N^3)$
$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$

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

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



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

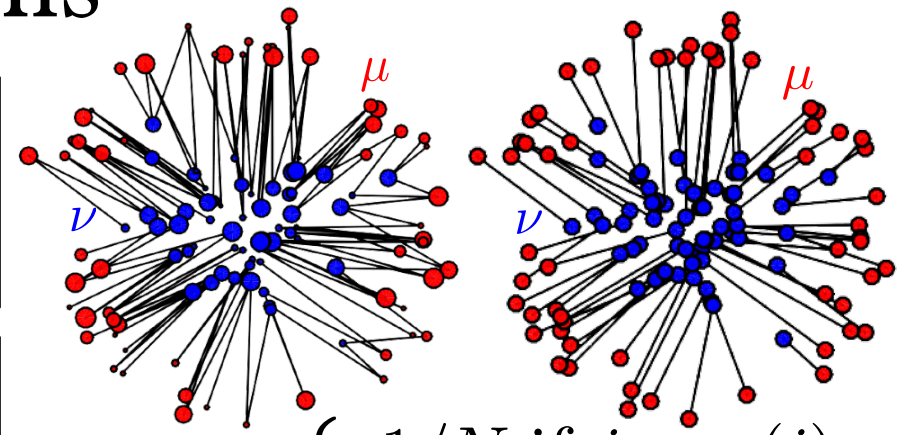


Semi-discrete: Laguerre cells, $d = \|\cdot\|_2^2$.
[Merigot 2013]

[Levy,'15]

Algorithms

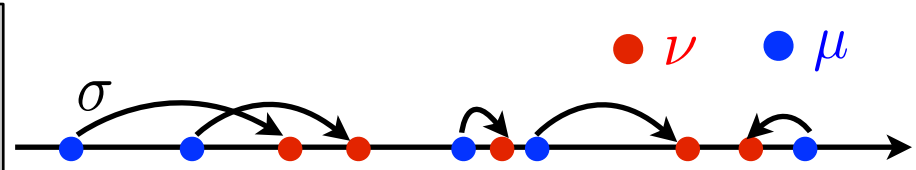
Linear programming:
$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$



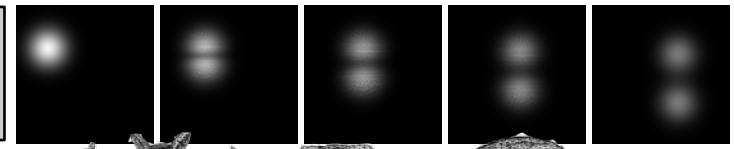
Hungarian/Auction: $\sim O(N^3)$
$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$

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

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



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



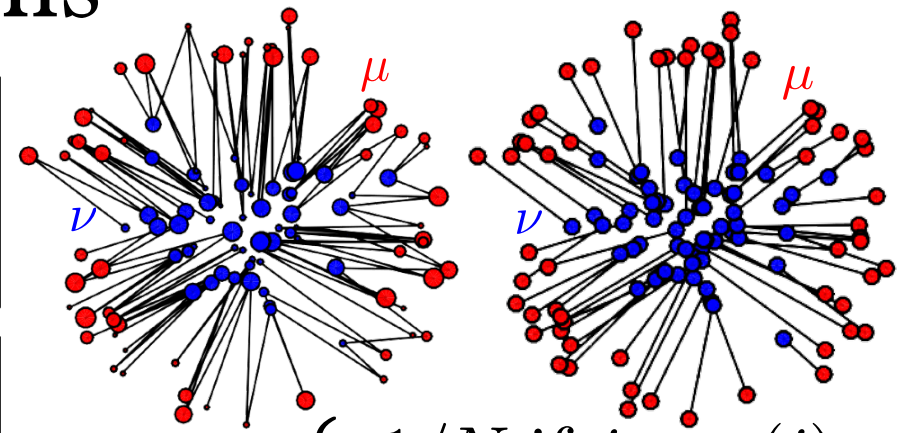
Semi-discrete: Laguerre cells, $d = \|\cdot\|_2^2$.
[Merigot 2013]


[Levy,'15]

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

Algorithms

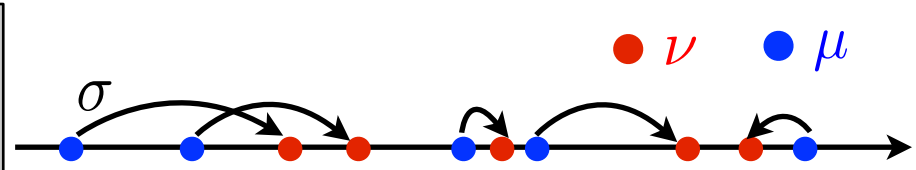
Linear programming:
$$\mu = \sum_{i=1}^{N_1} p_i \delta_{x_i}, \nu = \sum_{j=1}^{N_2} p_j \delta_{y_i}$$



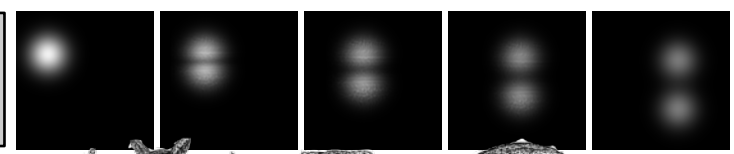
Hungarian/Auction: $\sim O(N^3)$
$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}$$

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

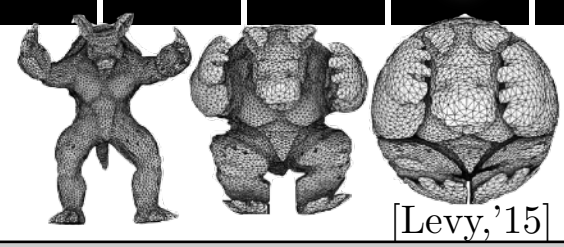
1-D case, $d = |\cdot|^p, p \geqslant 1$.
$\rightarrow$ sorting, $O(N \log(N))$ operations.



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



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


[Levy,'15]

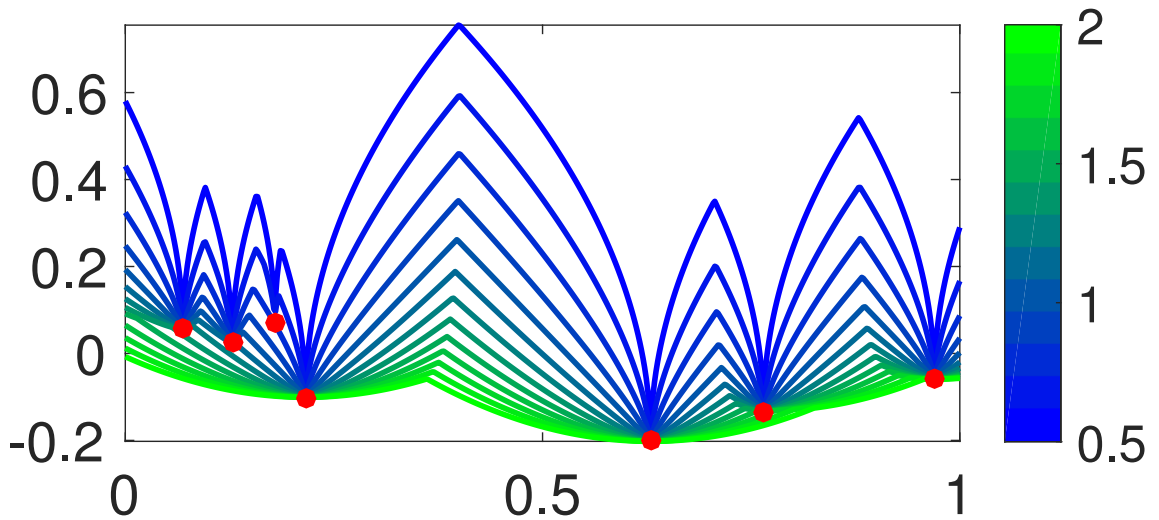
$$d = \|\cdot\|, p = 1 : W_1(\mu, \nu) = \min_{\text{div}(v)=\mu-\nu} \int \|u(x)\| \mathrm{d}x \rightarrow \text{max-flow.}$$

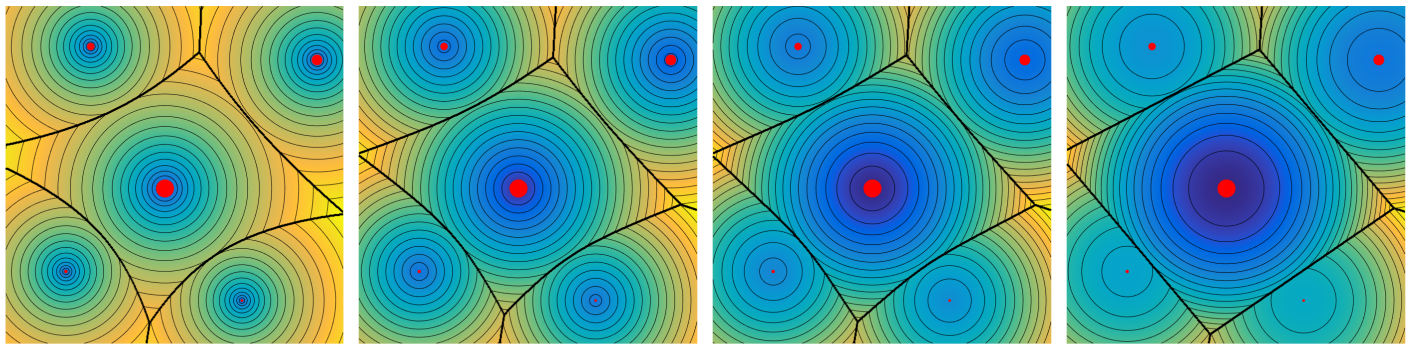
Need for fast approximate algorithms for generic c.

C-Transform

$$\forall\, y \in \mathcal{Y}, \quad f^c(y) \overset{\text{def.}}{=} \inf_{x \in \mathcal{X}} c(x,y) - f(x),$$
$$\forall\, x \in \mathcal{X}, \quad g^{\bar{c}}(x) \overset{\text{def.}}{=} \inf_{y \in \mathcal{Y}} c(x,y) - g(y),$$

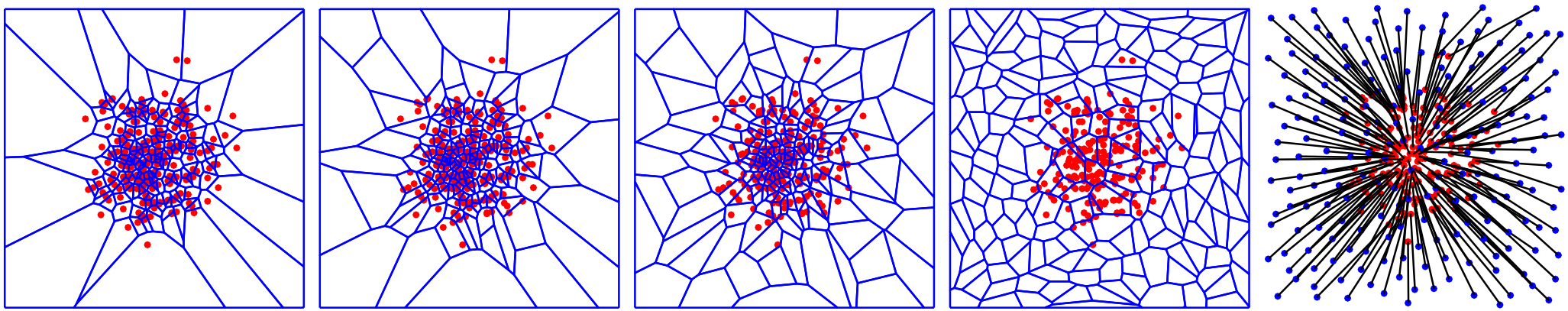
$$\begin{aligned}\mathcal{L}_c(\alpha,\beta) &= \max_{f \in \mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} f^c(y)\mathrm{d}\beta(y), \\ &= \max_{g \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} g^{\bar{c}}(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y).\end{aligned}$$





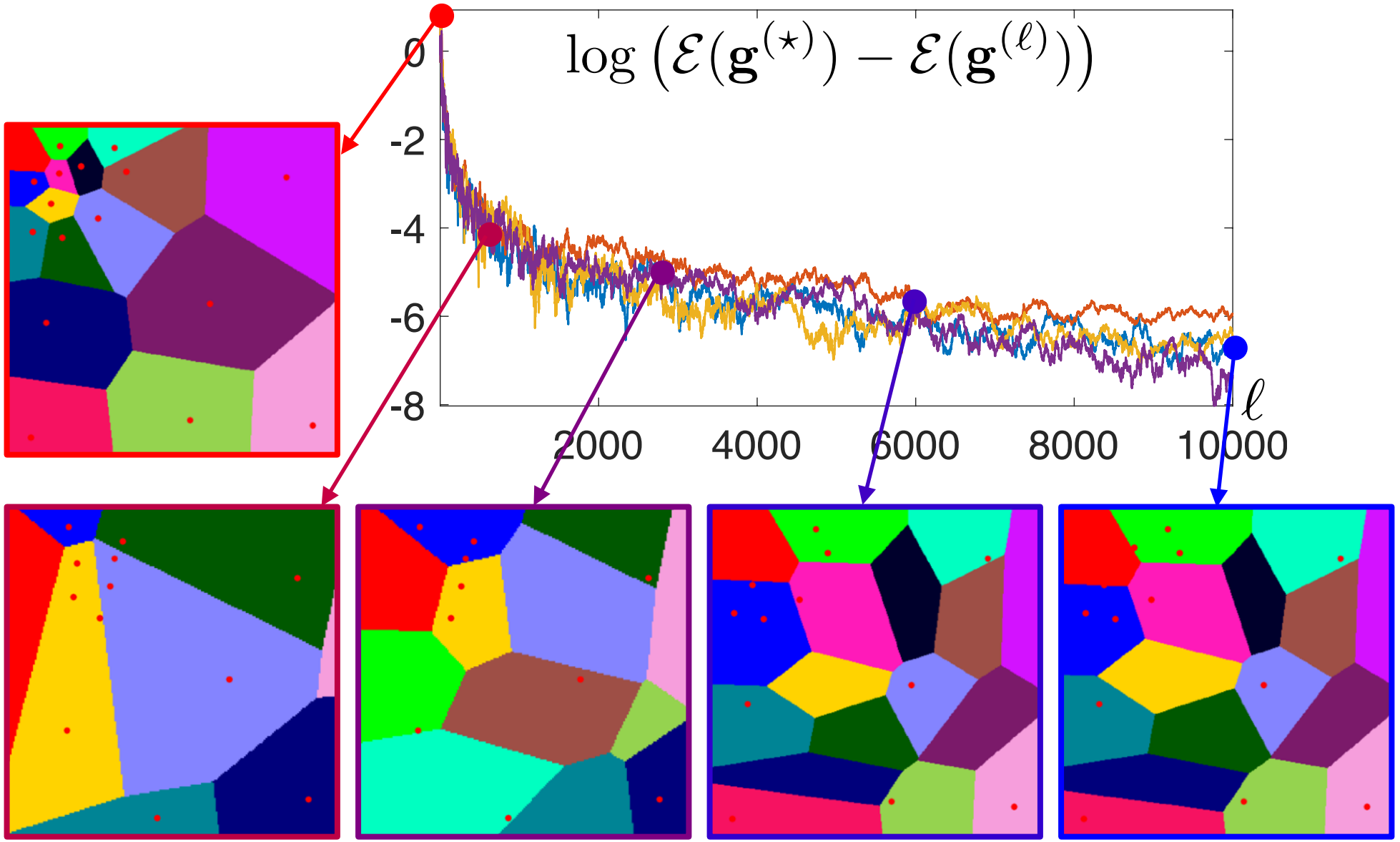
$p = 1/2$ $p = 1$ $p = 3/2$ $p = 2$

Semi-discrete Descent Algorithm



$\ell = 1$ $\ell = 3$ $\ell = 50$ $\ell = 100$ Matching

Semi-discrete Stochastic Descent



Stochastic gradient descent for the semi-discrete Optimal Transport,
illustration of convergence and corresponding Laguerre cells.
https://arxiv.org/abs/1605.08527

Entropic Regularization

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

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

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

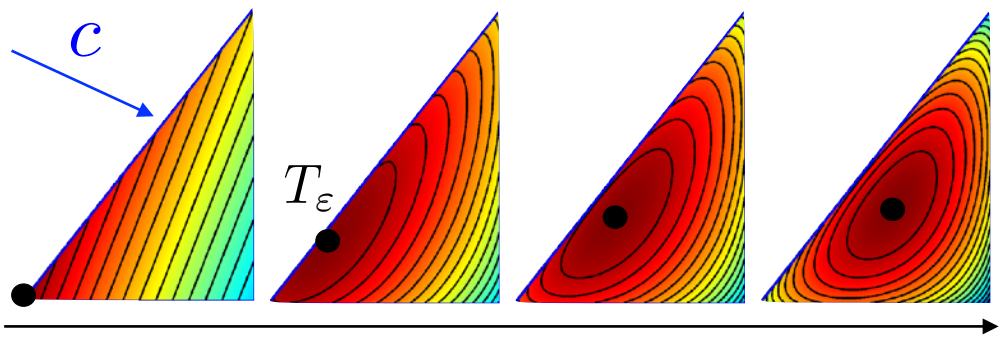
Entropic Regularization

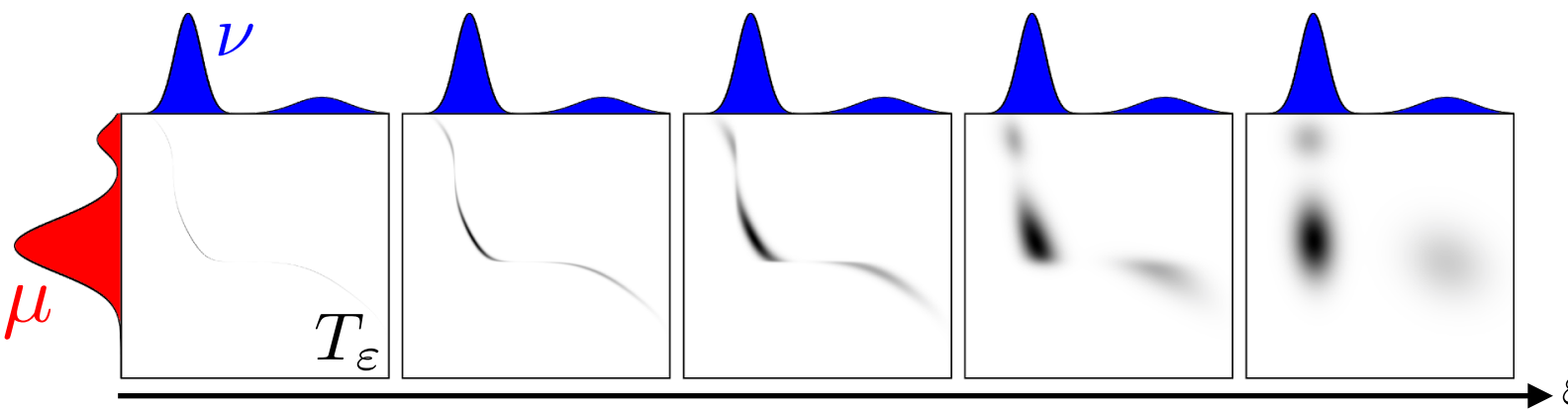
Entropy: $H(T) \stackrel{\text{def.}}{=} -\sum_{i,j=1}^{N} T_{i,j}(\log(T_{i,j}) - 1)$

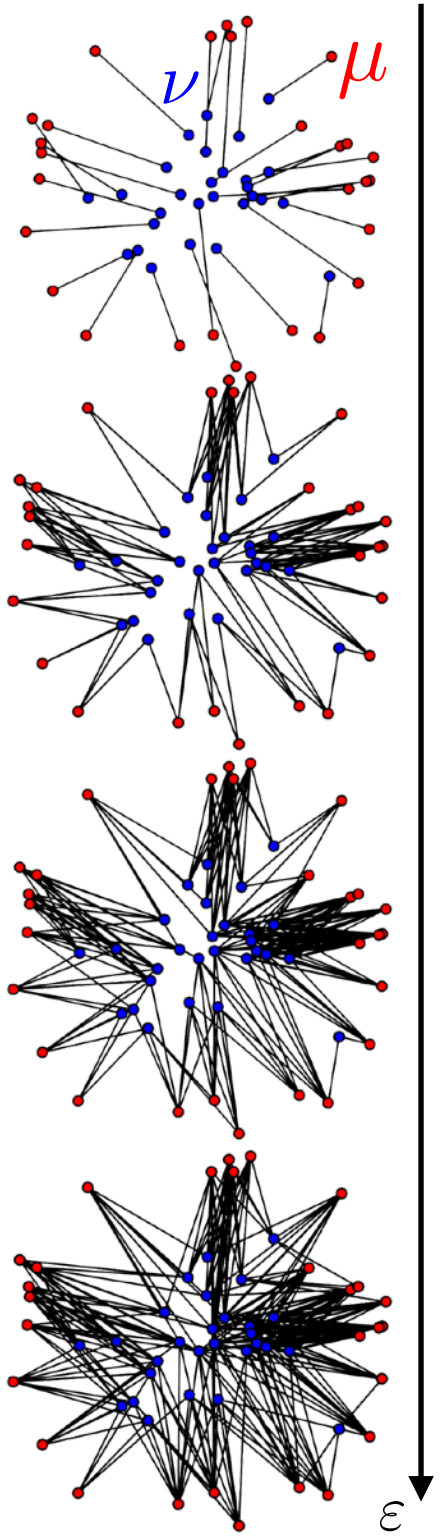
Def. *Regularized OT:* [Cuturi NIPS'13]

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

Regularization impact on solution:







Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- **Entropic Regularization and Sinkhorn**
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein

Sinkhorn's Algorithm

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

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

Sinkhorn's Algorithm

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

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

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

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

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

Sinkhorn's Algorithm

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

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

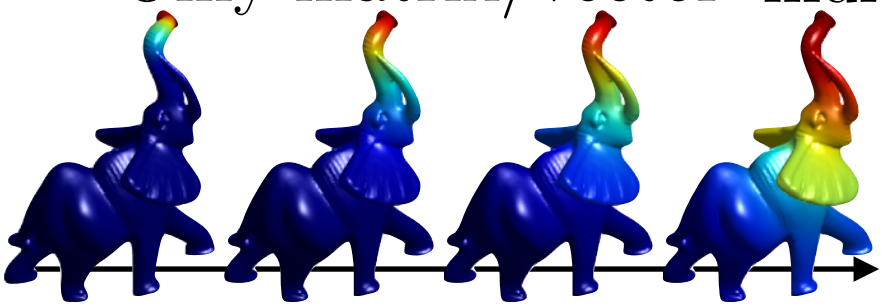
Row constraint: $T \mathbb{1}_{N_2} = \mu \iff a \odot (Kb) = \mu$

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

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

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

$\rightarrow$ Streams well on GPU.



$\rightarrow$ convolutive/heat structure for K

Sinkhorn's Algorithm

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

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

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

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

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

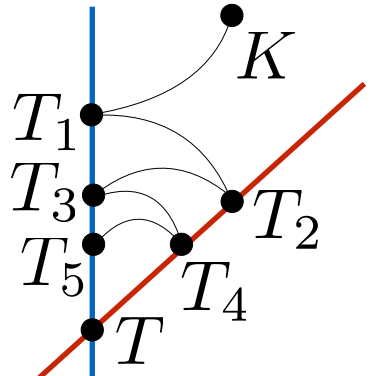
Only matrix/vector multiplications. $\rightarrow$ Parallelizable.

$\rightarrow$ Streams well on GPU.

$\rightarrow$ convolutive/heat structure for K

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

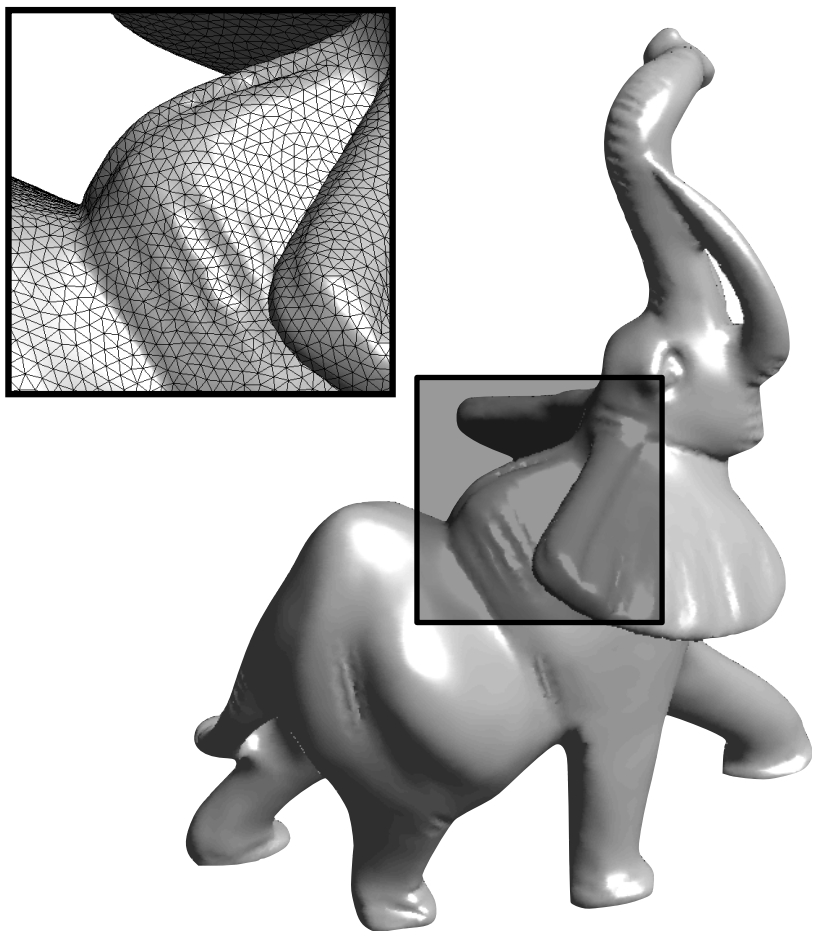
Sinkhorn $\iff$ iterative projections.



Optimal Transport on Surfaces

Triangulated mesh M.

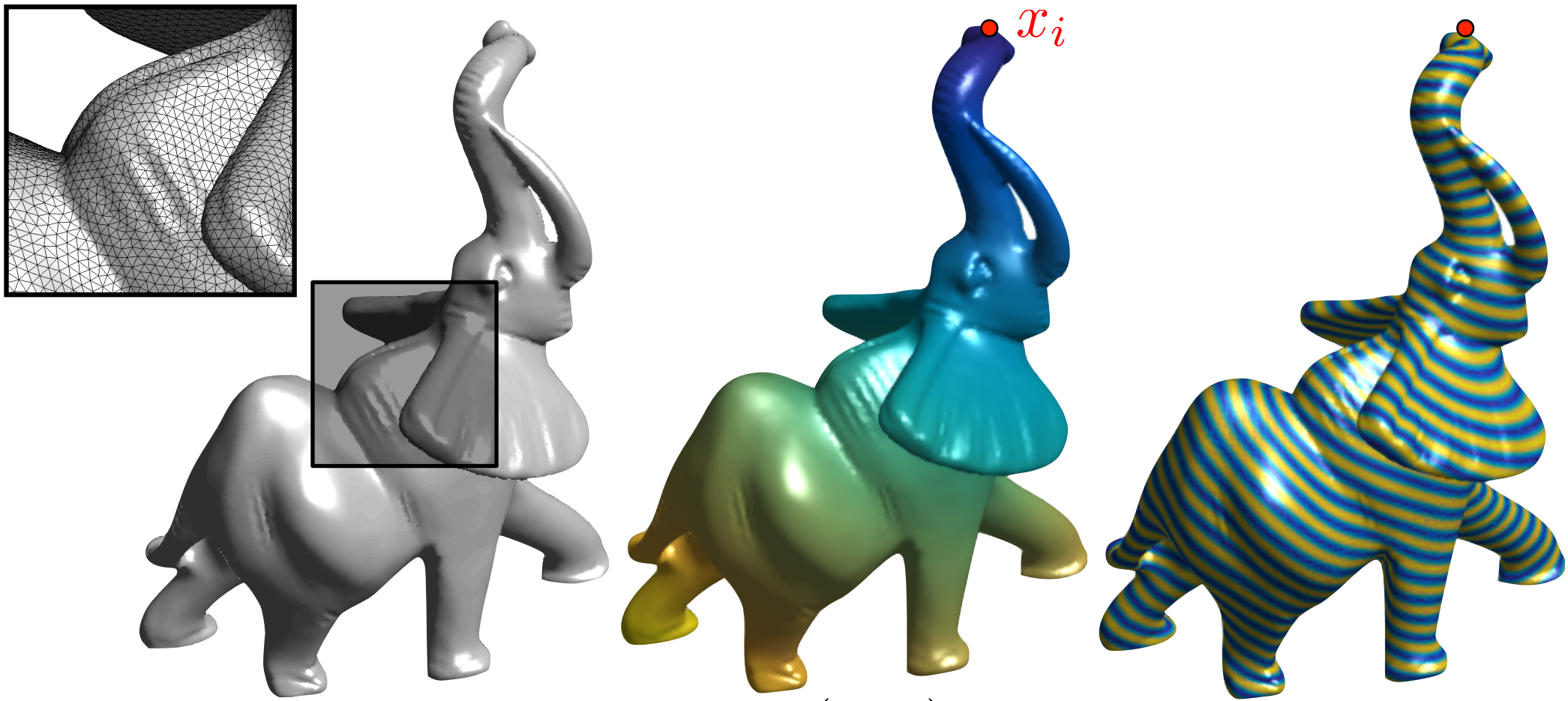
Geodesic distance d_M.



Optimal Transport on Surfaces

Triangulated mesh M. Geodesic distance d_M.

Ground cost: $c(x, y) = d_M(x, y)^\alpha$.

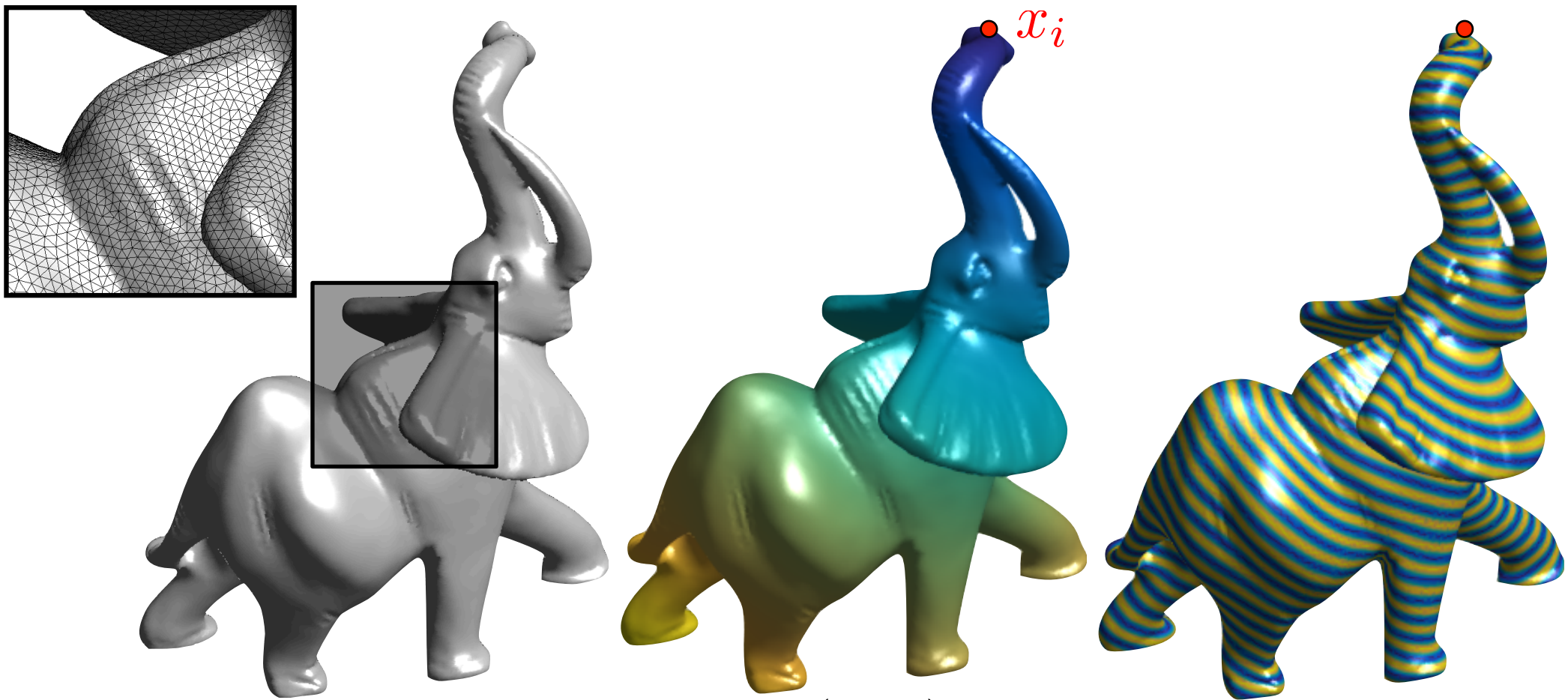


$d(x_i, \cdot)$ Level sets

Optimal Transport on Surfaces

Triangulated mesh M. Geodesic distance d_M.

Ground cost: $c(x, y) = d_M(x, y)^\alpha$.

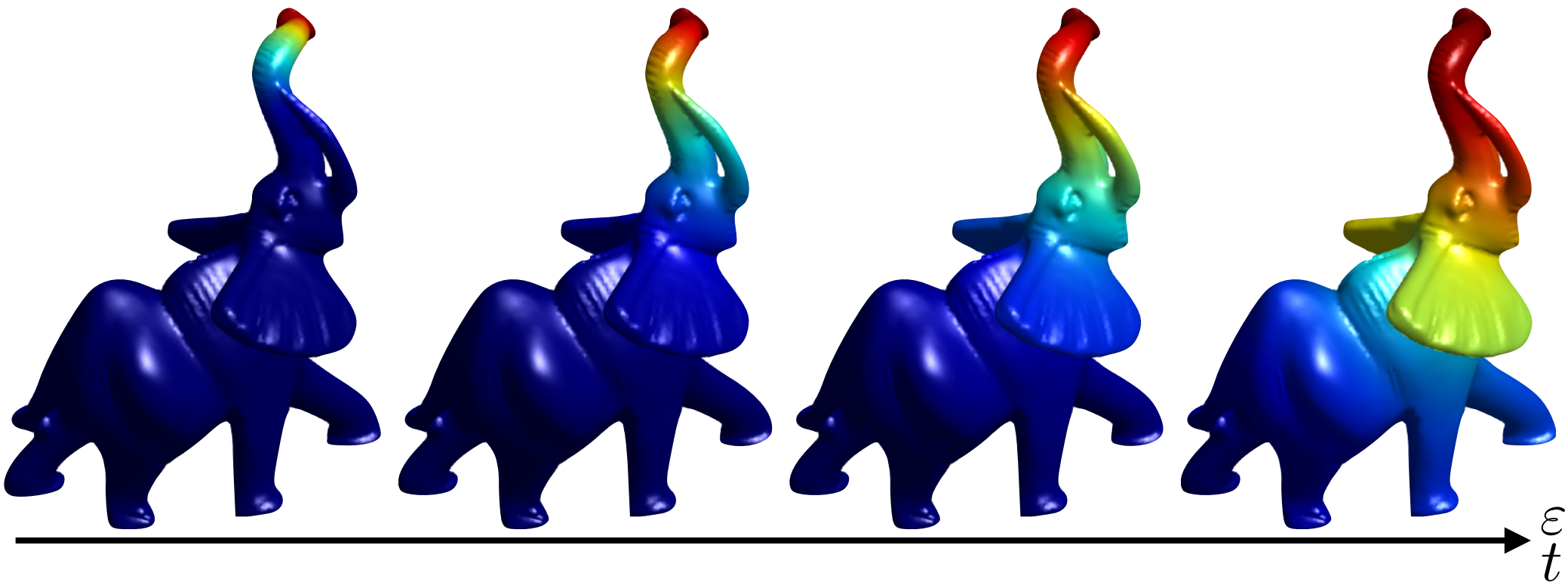


$d(x_i, \cdot)$ Level sets

Computing c (Fast-Marching): $N^2 \log(N) \rightarrow$ too costly.

Entropic Transport on Surfaces

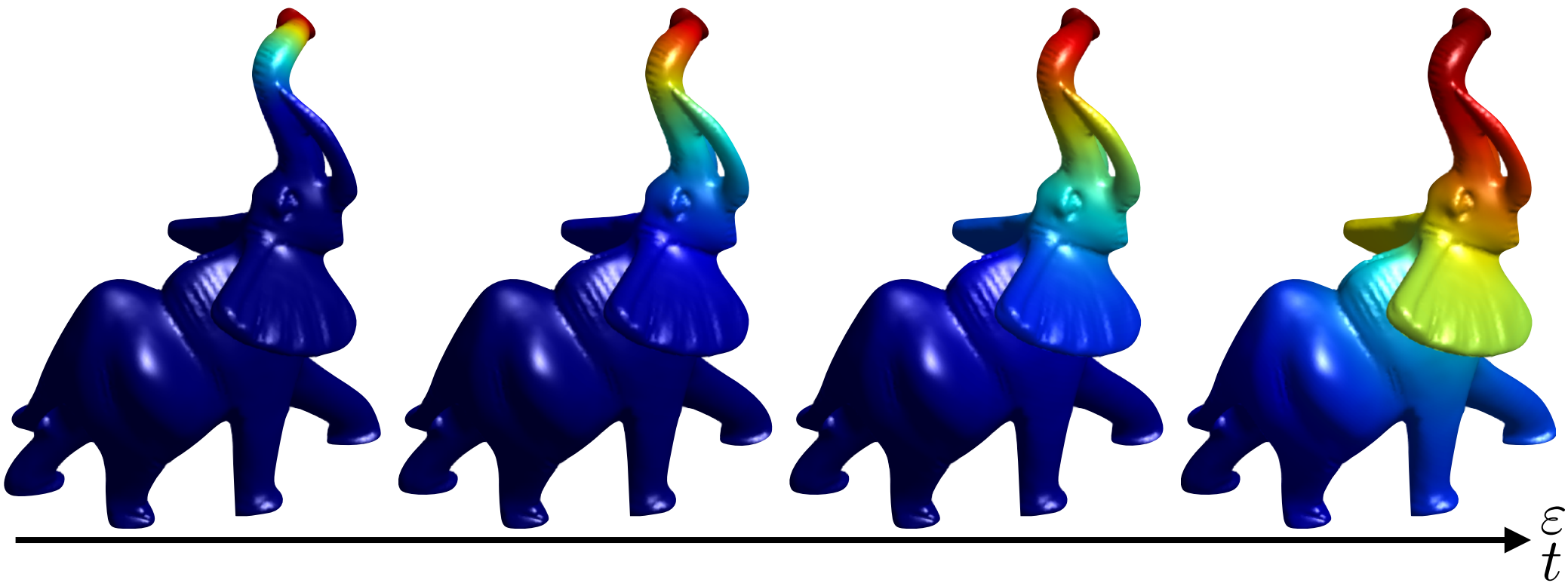
Heat equation on M: $\partial_t u_t(x,\cdot) = \Delta_M u_t(x,\cdot),\ u_{t=0}(x,\cdot) = \delta_x$



Entropic Transport on Surfaces

Heat equation on M: $\partial_t u_t(x,\cdot) = \Delta_M u_t(x,\cdot),\ u_{t=0}(x,\cdot) = \delta_x$

Theorem: [Varadhan] $\quad -\varepsilon \log(u_\varepsilon) \overset{\varepsilon\to 0}{\longrightarrow} d_M^2$

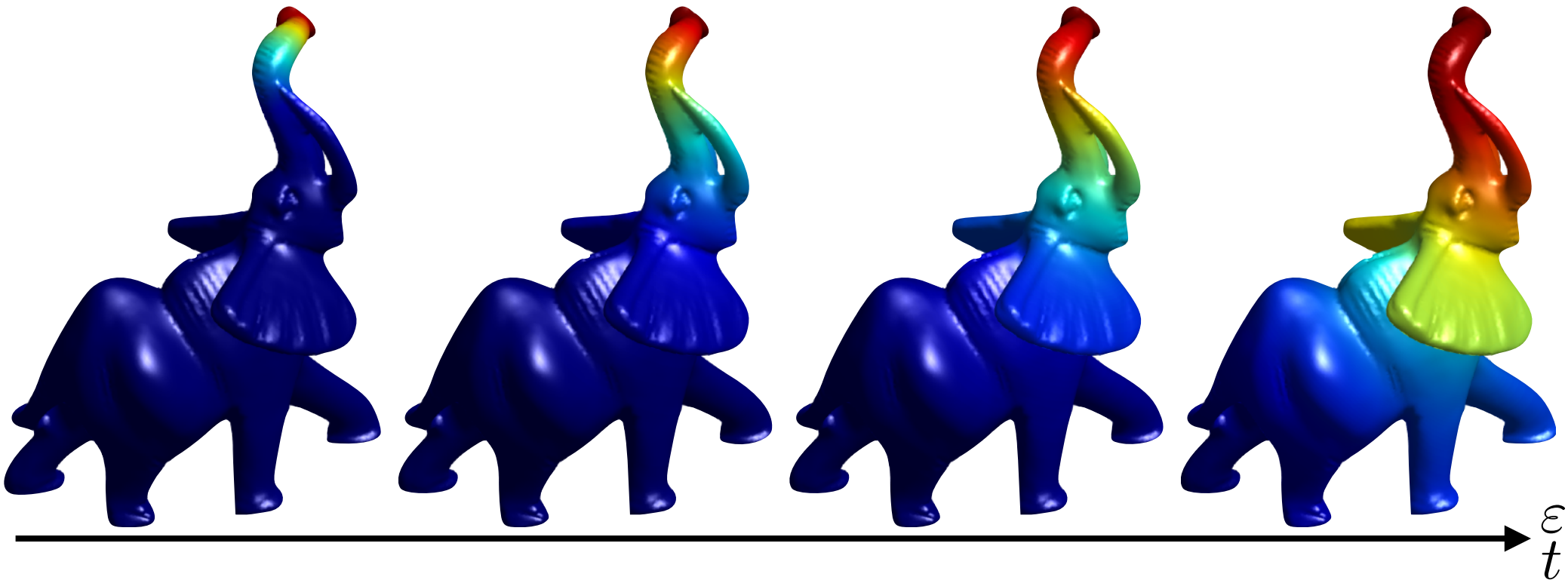


Entropic Transport on Surfaces

Heat equation on M: $\partial_t u_t(x,\cdot) = \Delta_M u_t(x,\cdot),\ u_{t=0}(x,\cdot) = \delta_x$

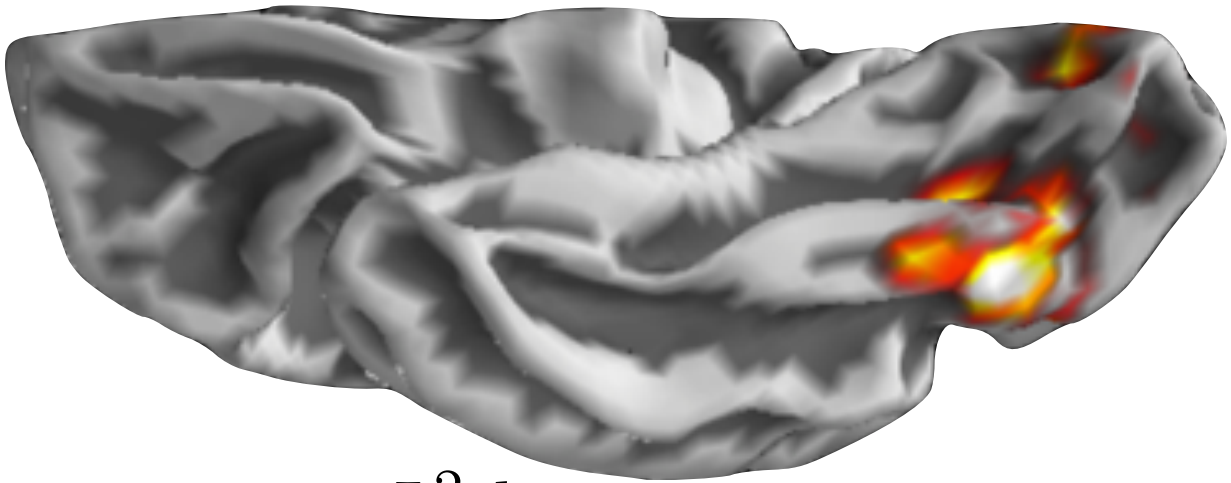
Theorem: [Varadhan] $\quad -\varepsilon \log(u_\varepsilon) \overset{\varepsilon\to 0}{\longrightarrow} d_M^2$

Sinkhorn kernel: $\quad K \overset{\text{def.}}{=} e^{-\frac{d_M^2}{\varepsilon}} \approx u_\varepsilon \approx \left(\mathrm{Id} - \frac{\varepsilon}{\ell}\Delta_M\right)^{-\ell}$

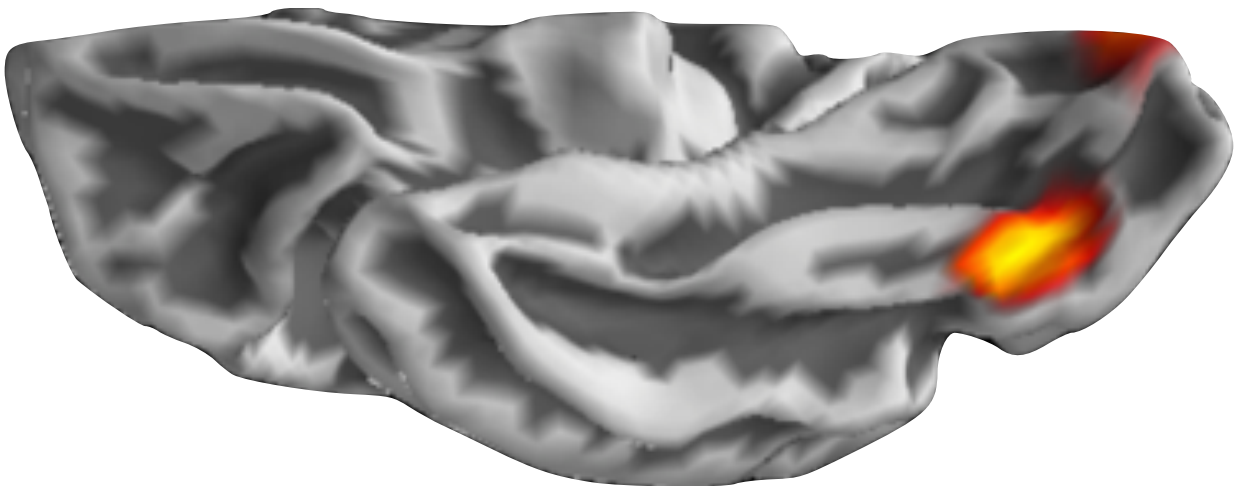


MRI Data Procesing [with A. Gramfort]

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



L^2 barycenter



W_2^2 barycenter

Regularization for General Measures

$$\pi_\varepsilon \overset{\text{def.}}{=} \operatorname*{argmin}_{\pi} \ \{\langle d^p, \pi\rangle + \varepsilon \mathrm{KL}(\pi|\pi_0) \ ; \ \pi \in \Pi(\mu, \nu)\}$$

Schrödinger's problem: $\pi_\varepsilon = \operatorname*{argmin}_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi|K)$

$$K(x,y) \overset{\text{def.}}{=} e^{-\frac{d^p(x,y)}{\varepsilon}} \pi_0(x,y)$$

Landmark computational paper: [Cuturi 2013].

Regularization for General Measures

$$\pi_\varepsilon \overset{\text{def.}}{=} \operatorname*{argmin}_{\pi} \{\langle d^p, \pi\rangle + \varepsilon \mathrm{KL}(\pi|\pi_0) \;;\; \pi \in \Pi(\mu,\nu)\}$$

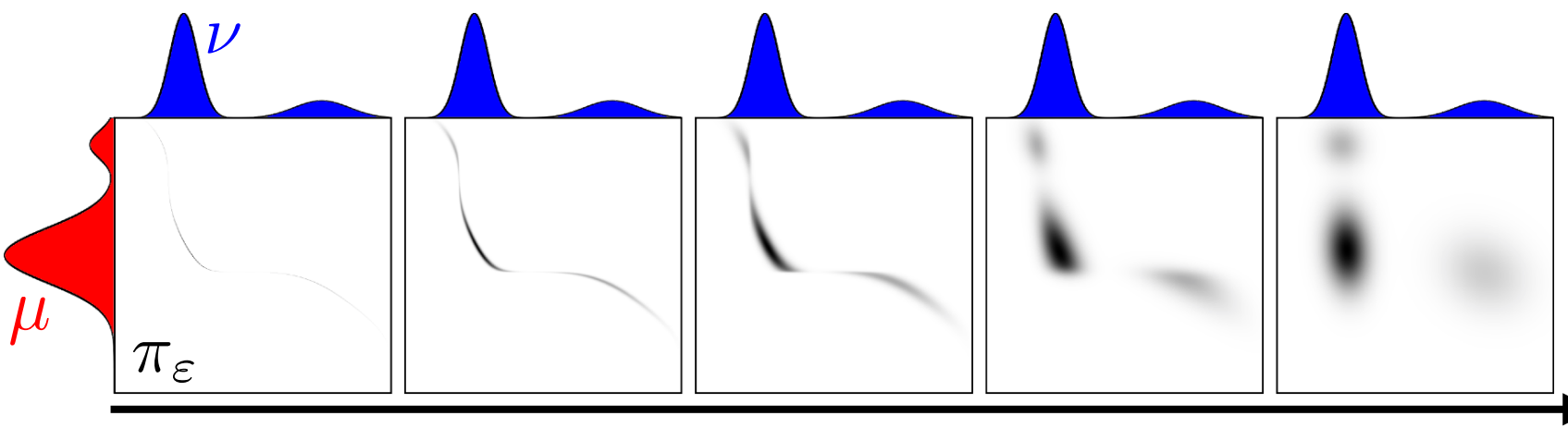
Schrödinger's problem: $\pi_\varepsilon = \operatorname*{argmin}_{\pi\in\Pi(\mu,\nu)} \mathrm{KL}(\pi|K)$

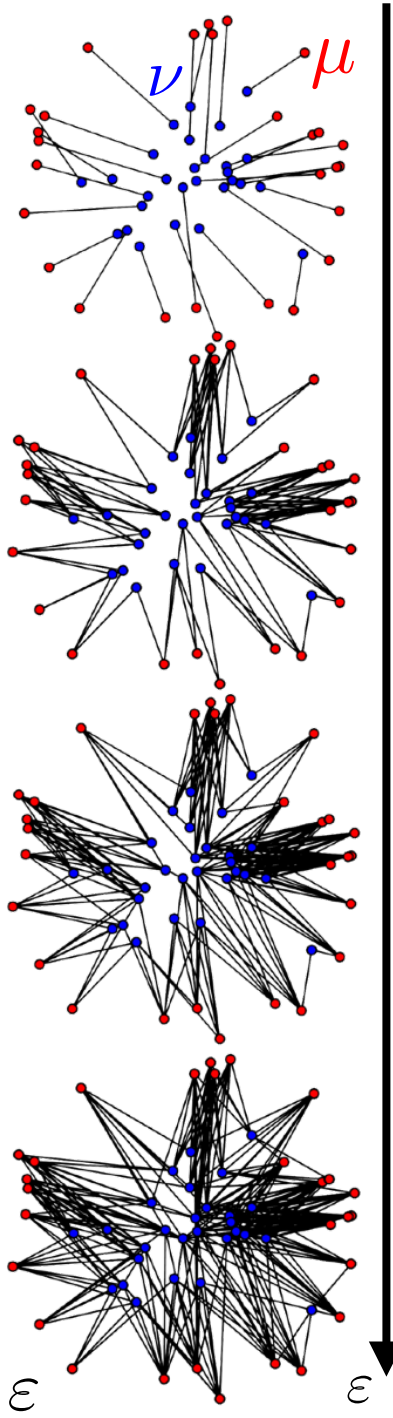
$$K(x,y) \overset{\text{def.}}{=} e^{-\frac{d^p(x,y)}{\varepsilon}} \pi_0(x,y)$$

Landmark computational paper: [Cuturi 2013].

Proposition: [Carlier, Duval, Peyré, Schmitzer 2015]

$$\pi_\varepsilon \overset{\varepsilon\to 0}{\longrightarrow} \operatorname*{argmin}_{\pi\in\Pi(\mu,\nu)} \langle d^p, \pi\rangle \qquad \pi_\varepsilon \overset{\varepsilon\to+\infty}{\longrightarrow} \mu(x)\nu(y)$$





Back to Sinkhorn's Algorithm

Optimal transport problem:

$$f_1 = \iota_\mu \longrightarrow \mathrm{Prox}^{\mathrm{KL}}_{f_1/\varepsilon}(\tilde{\mu}) = \mu$$
$$f_2 = \iota_\nu \longrightarrow \mathrm{Prox}^{\mathrm{KL}}_{f_2/\varepsilon}(\tilde{\nu}) = \nu$$

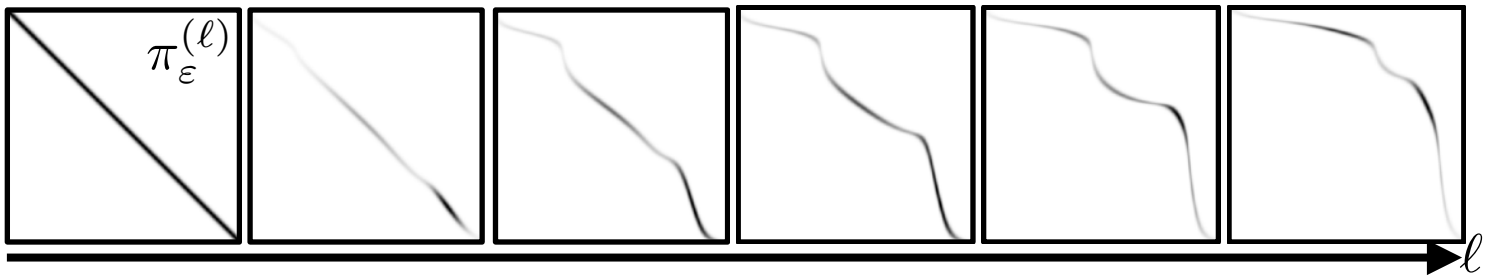
Back to Sinkhorn's Algorithm

Optimal transport problem:

$$f_1 = \iota_\mu \longrightarrow \mathrm{Prox}^{\mathrm{KL}}_{f_1/\varepsilon}(\tilde{\mu}) = \mu$$
$$f_2 = \iota_\nu \longrightarrow \mathrm{Prox}^{\mathrm{KL}}_{f_2/\varepsilon}(\tilde{\nu}) = \nu$$

Sinkhorn/IPFP algorithm: [Sinkhorn 1967][Deming,Stephan 1940]

$$a^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mu}{K b^{(\ell)}} \quad \text{and} \quad b^{(\ell+1)} \overset{\text{def.}}{=} \frac{\nu}{K^* a^{(\ell+1)}}$$



Back to Sinkhorn's Algorithm

Optimal transport problem:

$$f_1 = \iota_\mu \longrightarrow \mathrm{Prox}^{\mathrm{KL}}_{f_1/\varepsilon}(\tilde{\mu}) = \mu$$
$$f_2 = \iota_\nu \longrightarrow \mathrm{Prox}^{\mathrm{KL}}_{f_2/\varepsilon}(\tilde{\nu}) = \nu$$

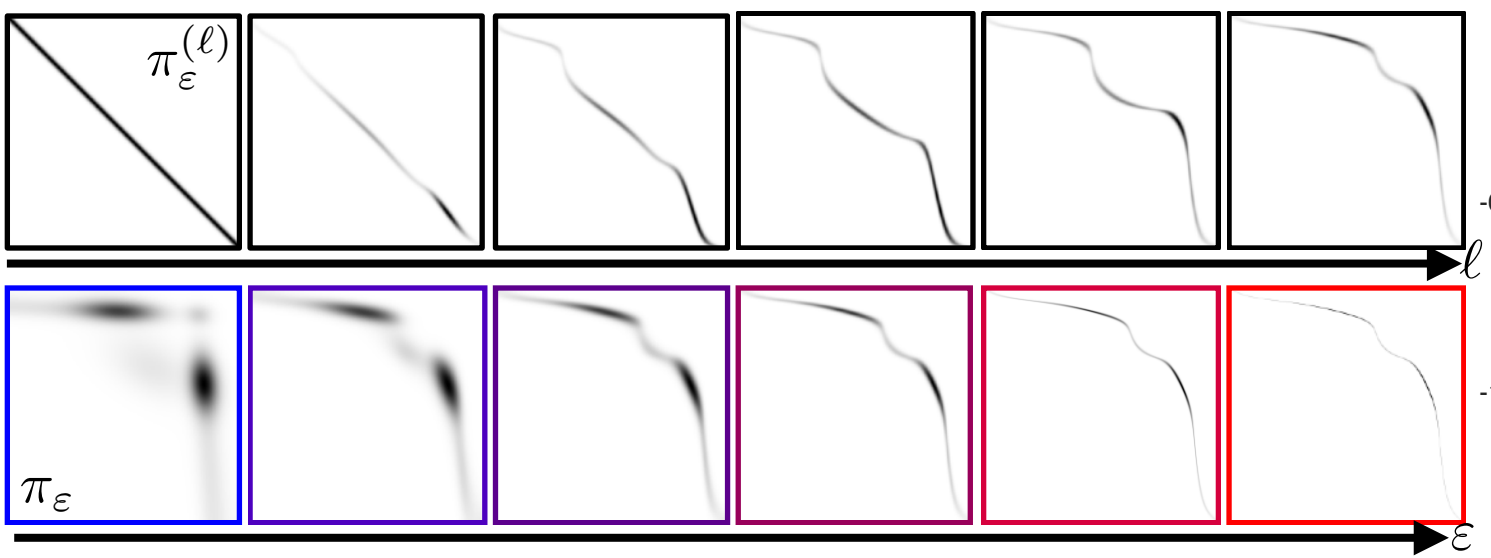
Sinkhorn/IPFP algorithm: [Sinkhorn 1967][Deming,Stephan 1940]

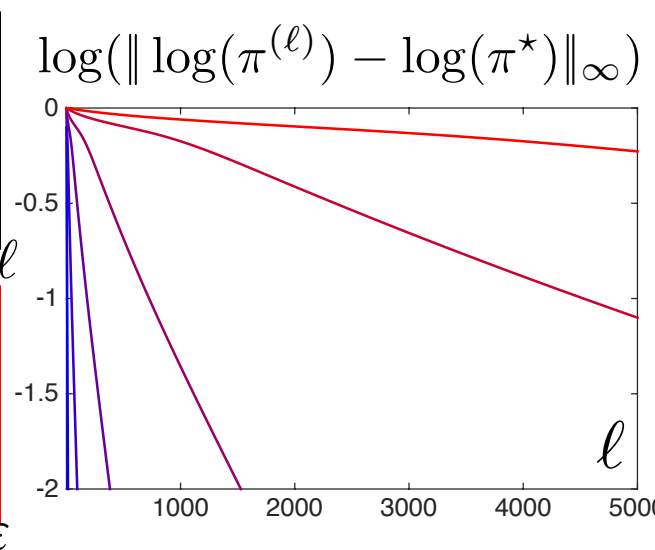
$$a^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mu}{K b^{(\ell)}} \quad \text{and} \quad b^{(\ell+1)} \overset{\text{def.}}{=} \frac{\nu}{K^* a^{(\ell+1)}}$$

Proposition: $\|\log(\pi^{(\ell)}) - \log(\pi^\star)\|_\infty = O(1-\delta)^\ell, \quad \delta \sim \kappa_c^{-1/\varepsilon}$

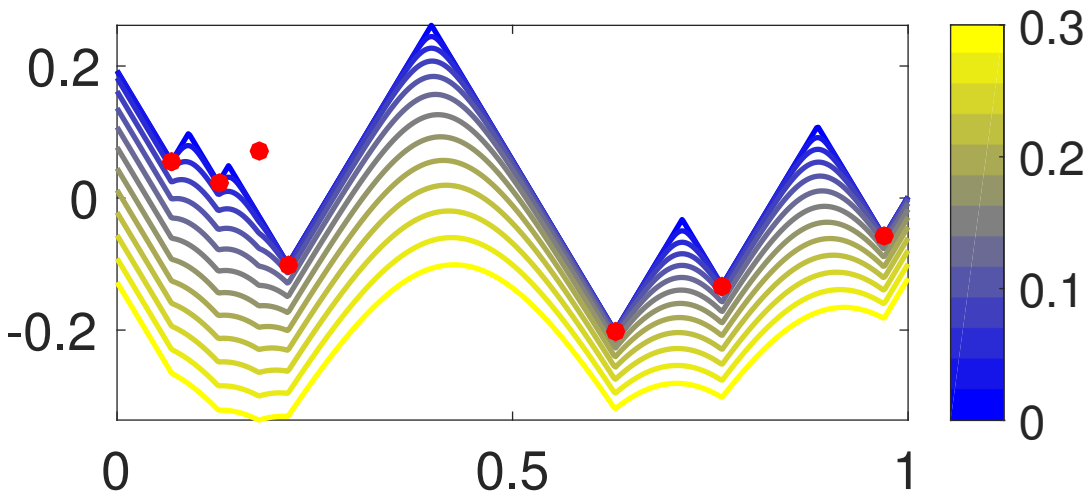
$$\pi^{(\ell)} \overset{\text{def.}}{=} \mathrm{diag}(a^{(\ell)}) K \, \mathrm{diag}(b^{(\ell)})$$

[Franklin,Lorenz 1989]
Local rate: [Knight 2008]



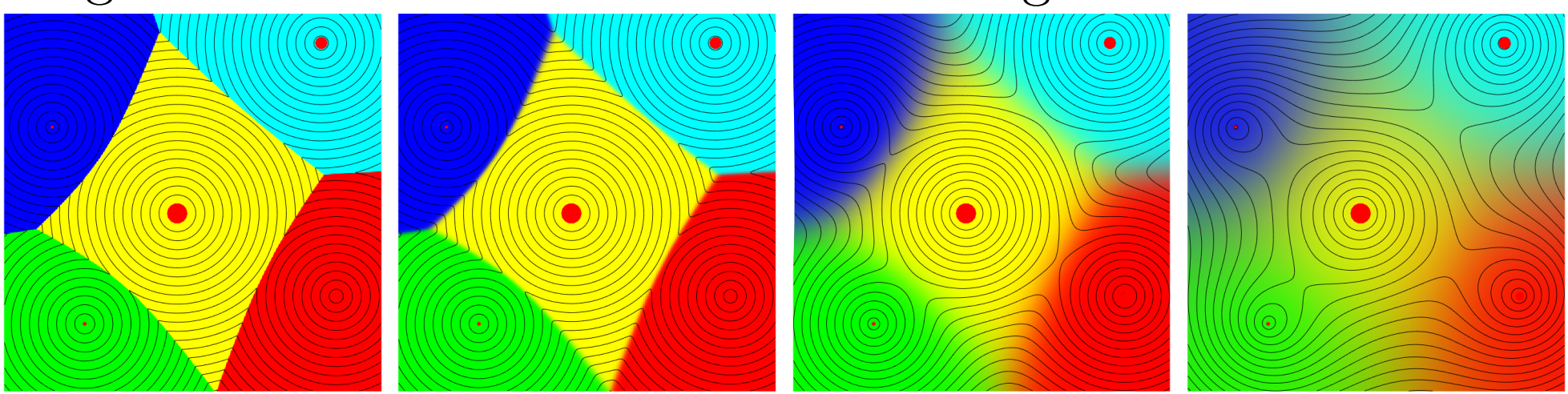


Semi-discrete OT and Entropy



Laguerre cells

"Sinkhorn" Laguerre cells



$\varepsilon = 0$ $\varepsilon = 0.01$ $\varepsilon = 0.1$ $\varepsilon = 0.3$

Semi-discrete Optimal Transport is a multiclass SVM. Sinkhorn version is logistic regression. Decomposes space in entropic Laguerre diagrams

Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- Entropic Regularization and Sinkhorn
- **Barycenters**
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein

Wasserstein Barycenters

Barycenters of measures $(\mu_k)_k$: $\quad \sum_k \lambda_k = 1$

$$\mu^\star \in \operatorname*{argmin}_{\mu} \sum_k \lambda_k W_2^2(\mu_k, \mu)$$

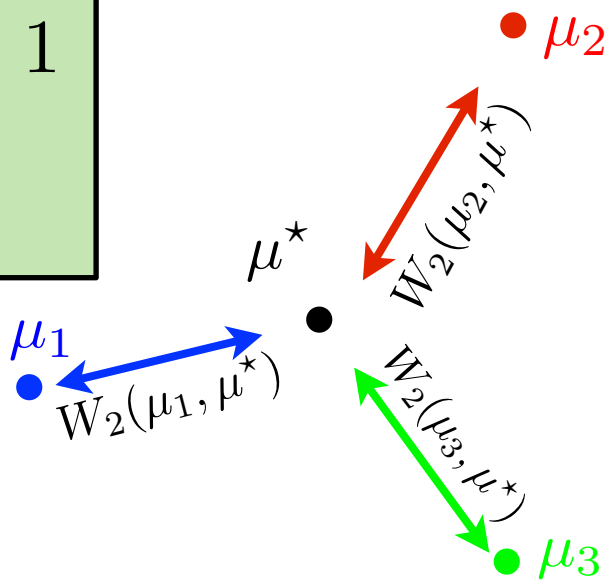
Wasserstein Barycenters

Barycenters of measures $(\mu_k)_k$: $\sum_k \lambda_k = 1$

$$\mu^\star \in \operatorname*{argmin}_{\mu} \sum_k \lambda_k W_2^2(\mu_k, \mu)$$

Generalizes Euclidean barycenter:

If $\mu_k = \delta_{x_k}$ then $\mu^\star = \delta_{\sum_k \lambda_k x_k}$



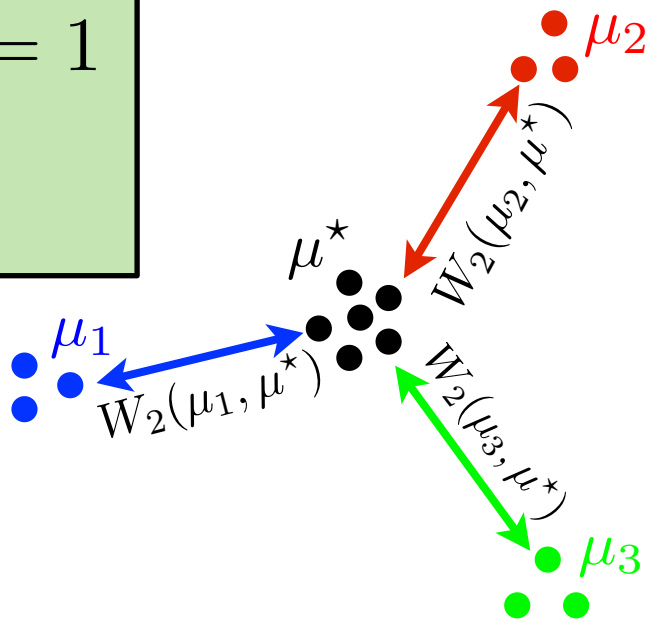
Wasserstein Barycenters

Barycenters of measures $(\mu_k)_k$: $\sum_k \lambda_k = 1$

$$\mu^\star \in \operatorname*{argmin}_{\mu} \sum_k \lambda_k W_2^2(\mu_k, \mu)$$

Generalizes Euclidean barycenter:

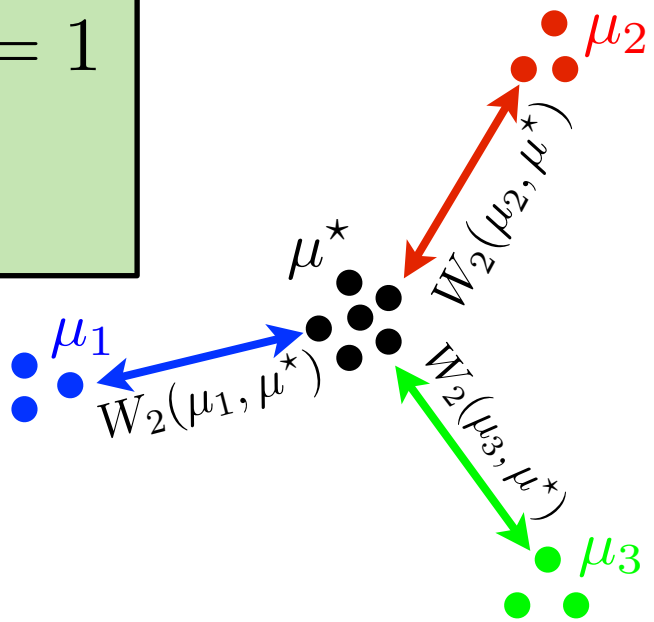
If $\mu_k = \delta_{x_k}$ then $\mu^\star = \delta_{\sum_k \lambda_k x_k}$



Wasserstein Barycenters

Barycenters of measures $(\mu_k)_k$: $\sum_k \lambda_k = 1$

$$\mu^\star \in \operatorname*{argmin}_\mu \sum_k \lambda_k W_2^2(\mu_k, \mu)$$



Generalizes Euclidean barycenter:

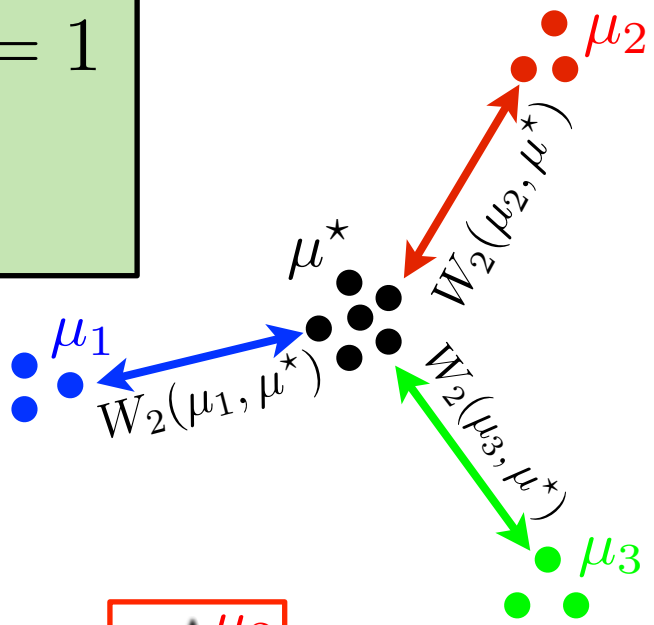
If $\mu_k = \delta_{x_k}$ then $\mu^\star = \delta_{\sum_k \lambda_k x_k}$

Mc Cann's displacement interpolation.

Wasserstein Barycenters

Barycenters of measures $(\mu_k)_k$: $\sum_k \lambda_k = 1$

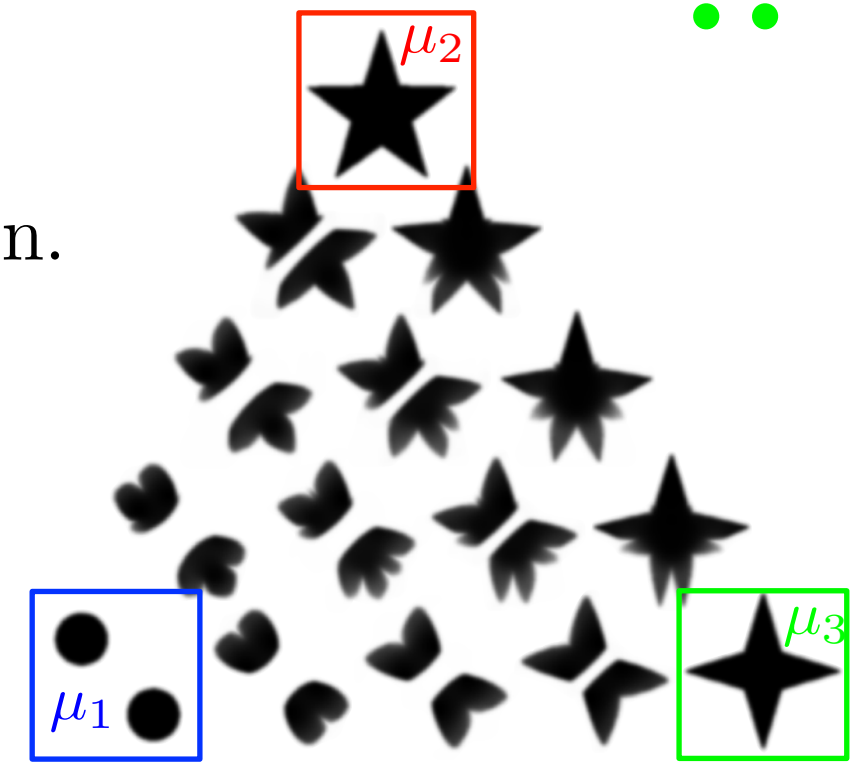
$$\mu^\star \in \operatorname*{argmin}_\mu \sum_k \lambda_k W_2^2(\mu_k, \mu)$$



Generalizes Euclidean barycenter:

If $\mu_k = \delta_{x_k}$ then $\mu^\star = \delta_{\sum_k \lambda_k x_k}$

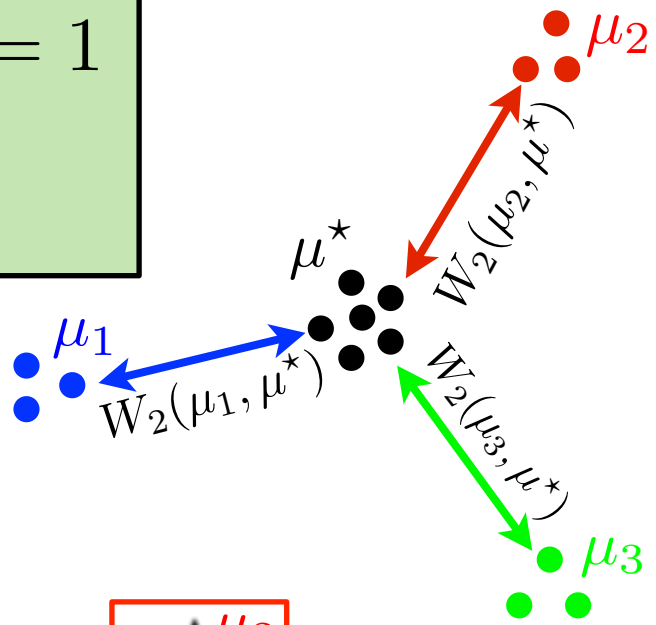
Mc Cann's displacement interpolation.



Wasserstein Barycenters

Barycenters of measures $(\mu_k)_k$: $\sum_k \lambda_k = 1$

$$\mu^\star \in \operatorname*{argmin}_\mu \sum_k \lambda_k W_2^2(\mu_k, \mu)$$



Generalizes Euclidean barycenter:

If $\mu_k = \delta_{x_k}$ then $\mu^\star = \delta_{\sum_k \lambda_k x_k}$

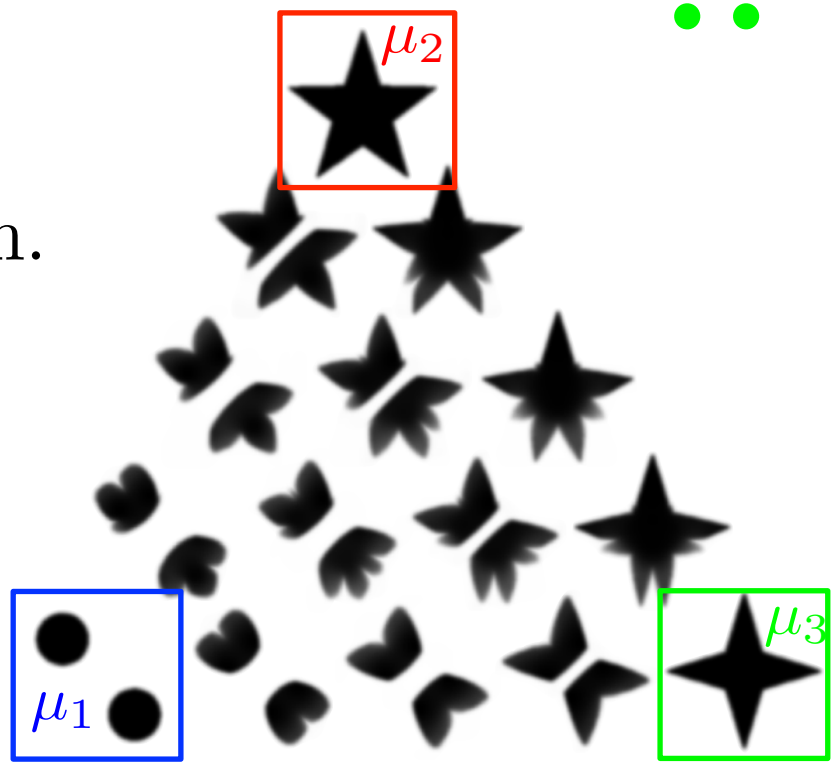
Mc Cann's displacement interpolation.

Theorem: [Agueh, Carlier, 2010]

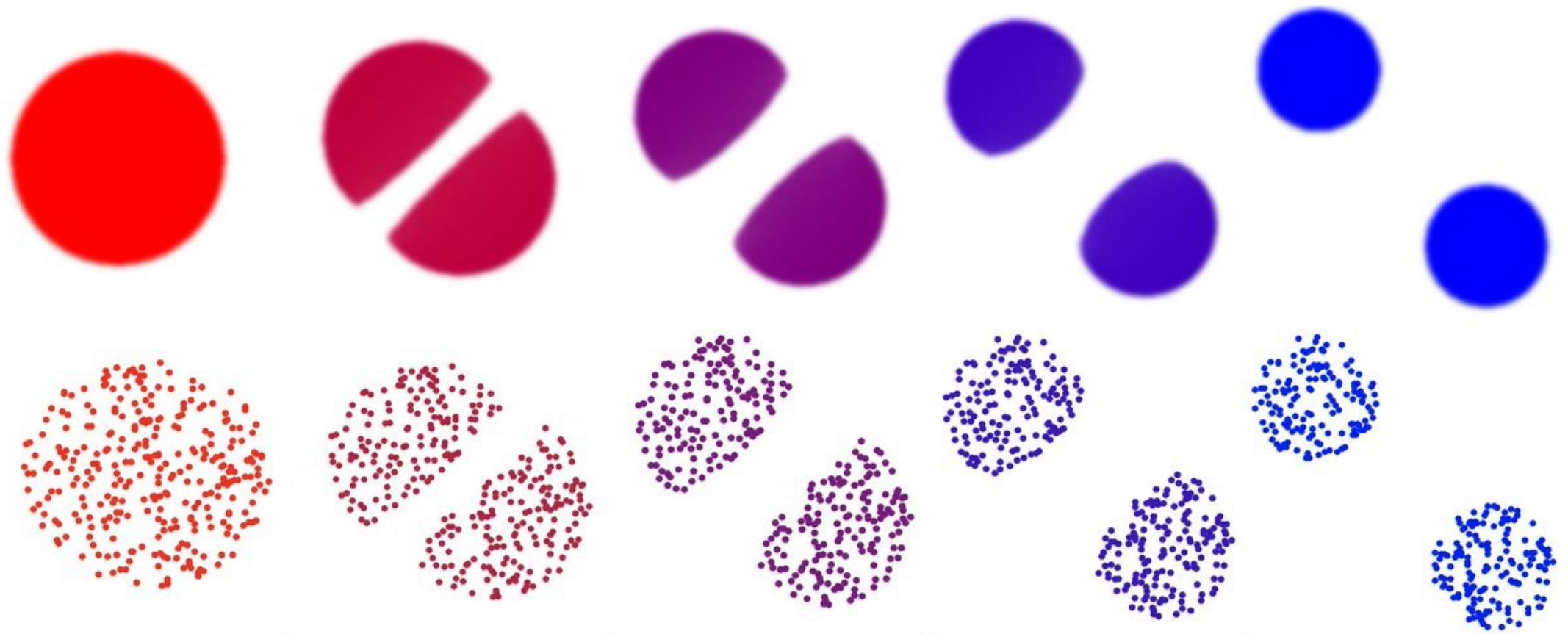
(for $c(x,y) = \|x-y\|^2$)

if μ_1 does not vanish on small sets,

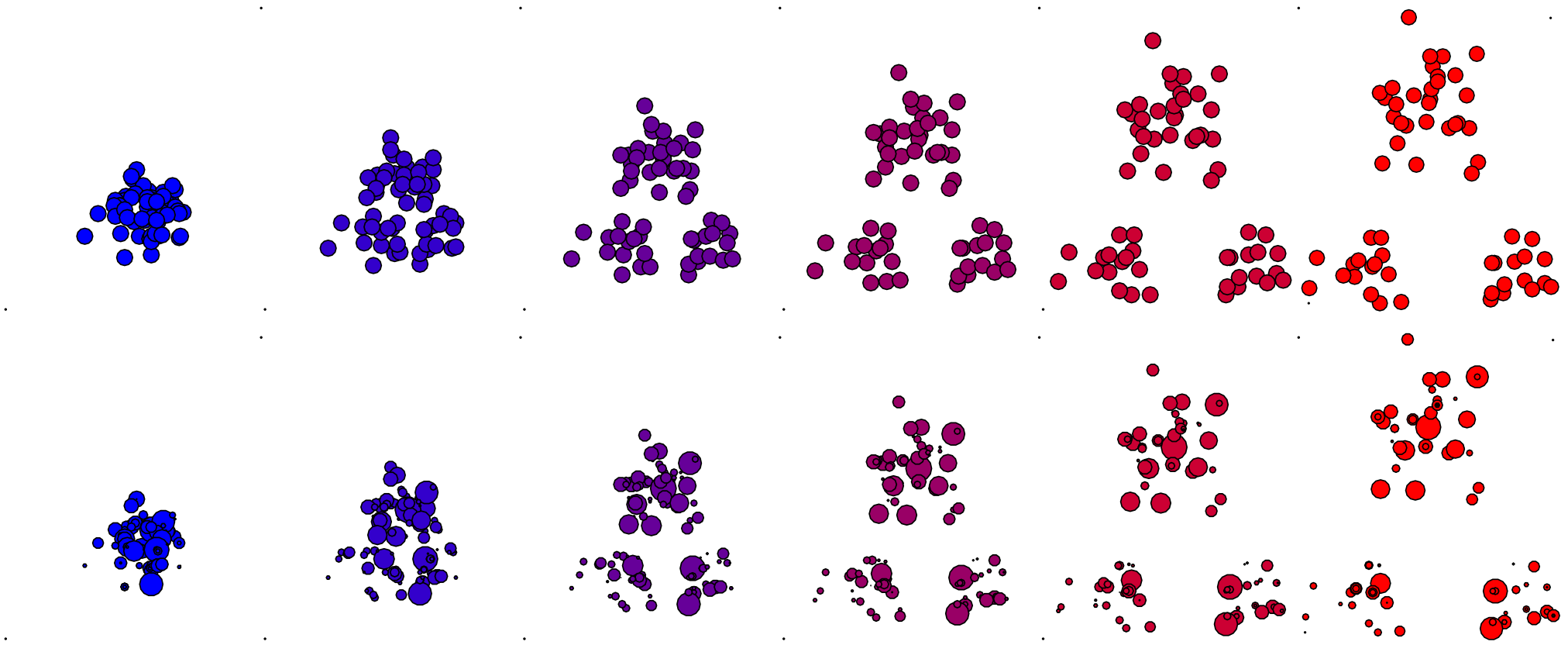
$\mu^\star$ exists and is unique.



Displacement Interpolation

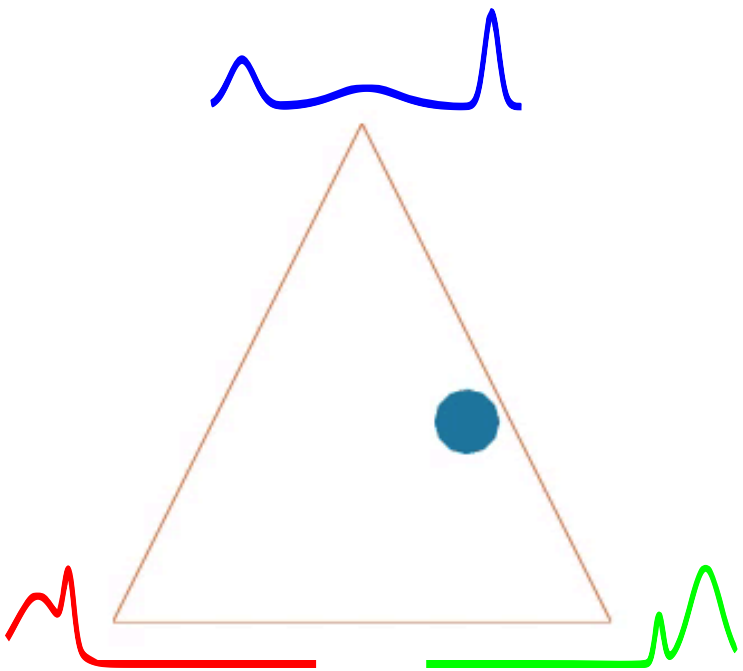


Displacement Interpolation

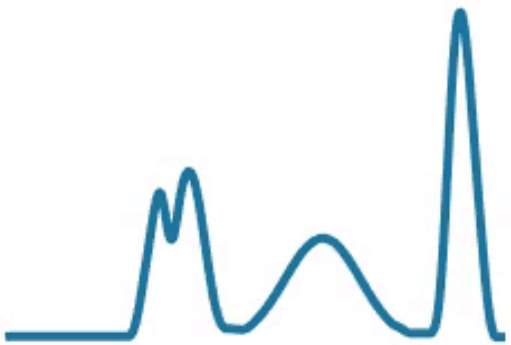


α_0 $\alpha_{1/5}$ $\alpha_{2/5}$ $\alpha_{3/5}$ $\alpha_{4/5}$ α_1

Wasserstein Barycenters



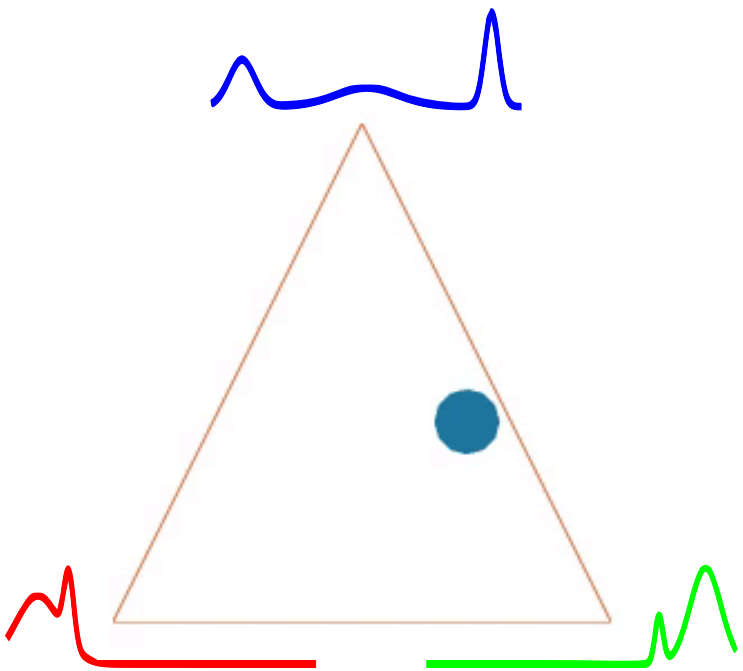
$\lambda \in \Sigma_3$



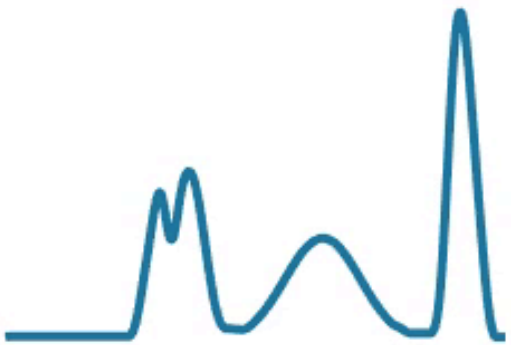
Wasserstein mean

L_2 mean

Wasserstein Barycenters



$\lambda \in \Sigma_3$



Wasserstein mean

L_2 mean

Regularized Barycenters

$$\min_{(\pi_k)_k, \mu} \left\{ \sum_k \lambda_k \left(\langle c,\, \pi_k \rangle + \varepsilon \mathrm{KL}(\pi_k | \pi_{0,k}) \right) \;;\; \forall k, \pi_k \in \Pi(\mu_k, \mu) \right\}$$

Regularized Barycenters

$$\min_{(\pi_k)_k, \mu} \left\{ \sum_k \lambda_k \left(\langle c, \pi_k \rangle + \varepsilon \mathrm{KL}(\pi_k | \pi_{0,k}) \right) \ ; \ \forall k, \pi_k \in \Pi(\mu_k, \mu) \right\}$$

$\longrightarrow$ Need to fix a discretization grid for μ, i.e. choose $(\pi_{0,k})_k$

Regularized Barycenters

$$\min_{(\pi_k)_k, \mu} \left\{ \sum_k \lambda_k \left(\langle c, \pi_k \rangle + \varepsilon \mathrm{KL}(\pi_k | \pi_{0,k}) \right) \ ; \ \forall k, \pi_k \in \Pi(\mu_k, \mu) \right\}$$

$\longrightarrow$ Need to fix a discretization grid for μ, i.e. choose $(\pi_{0,k})_k$

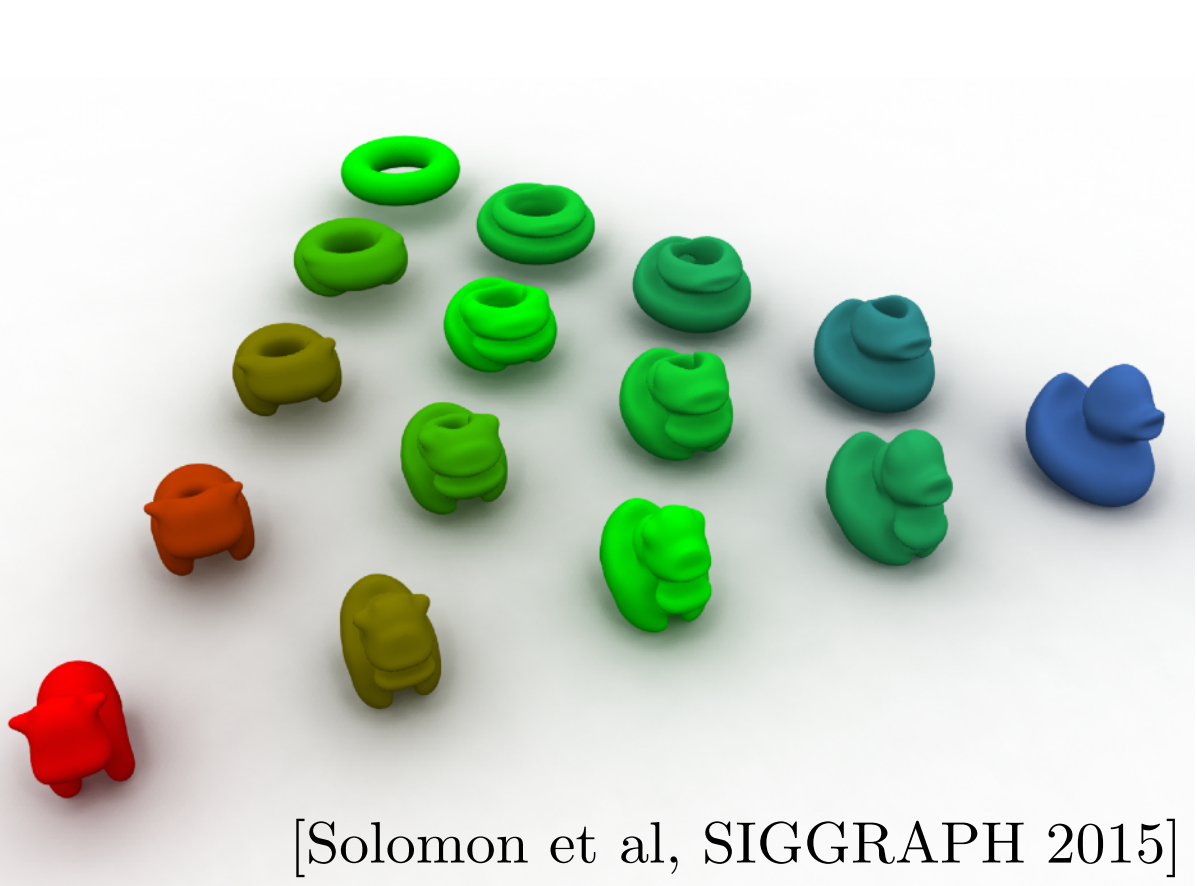
$\longrightarrow$ Sinkhorn-like algorithm [Benamou, Carlier, Cuturi, Nenna, Peyré, 2015].

Regularized Barycenters

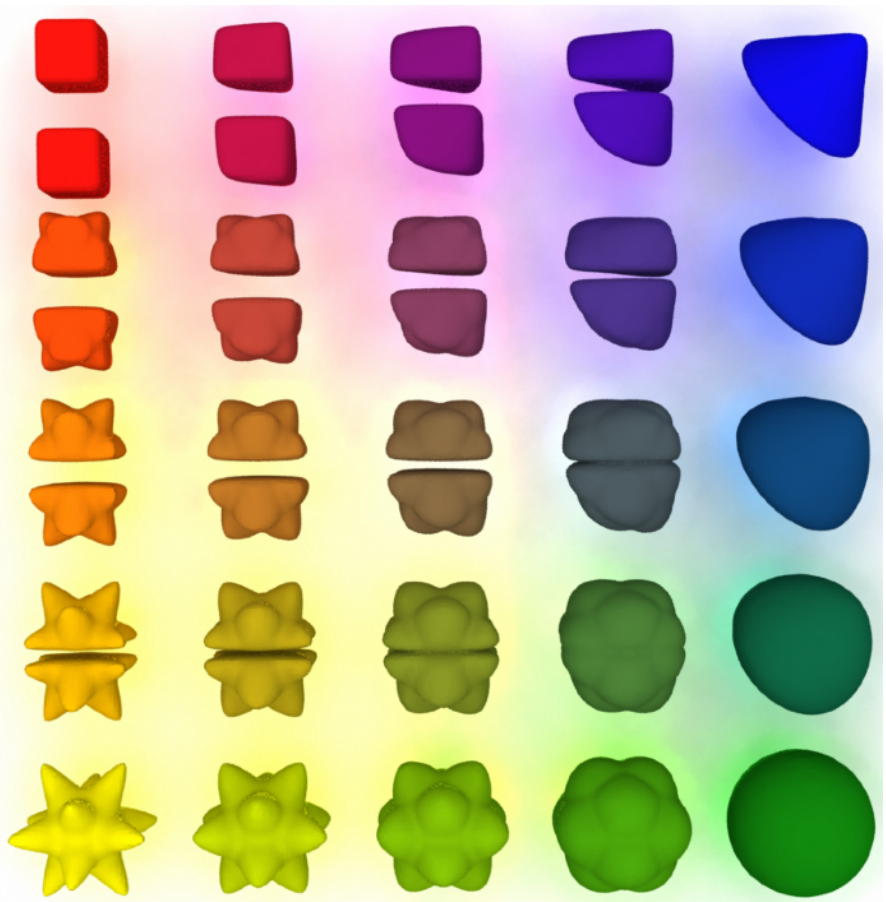
$$\min_{(\pi_k)_k,\mu} \left\{ \sum_k \lambda_k \left(\langle c,\, \pi_k \rangle + \varepsilon \mathrm{KL}(\pi_k | \pi_{0,k}) \right) \;;\; \forall k, \pi_k \in \Pi(\mu_k, \mu) \right\}$$

$\longrightarrow$ Need to fix a discretization grid for μ, i.e. choose $(\pi_{0,k})_k$

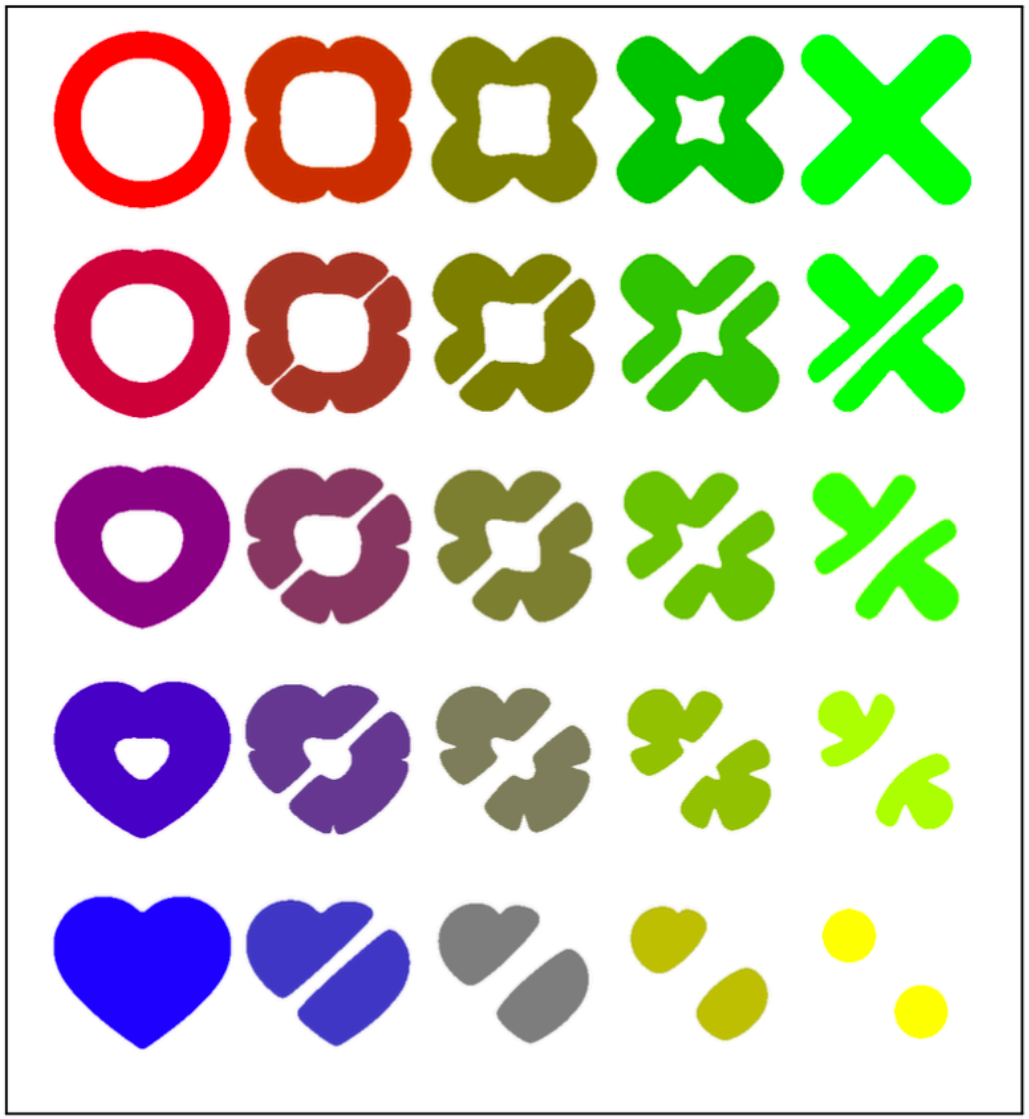
$\longrightarrow$ Sinkhorn-like algorithm [Benamou, Carlier, Cuturi, Nenna, Peyré, 2015].

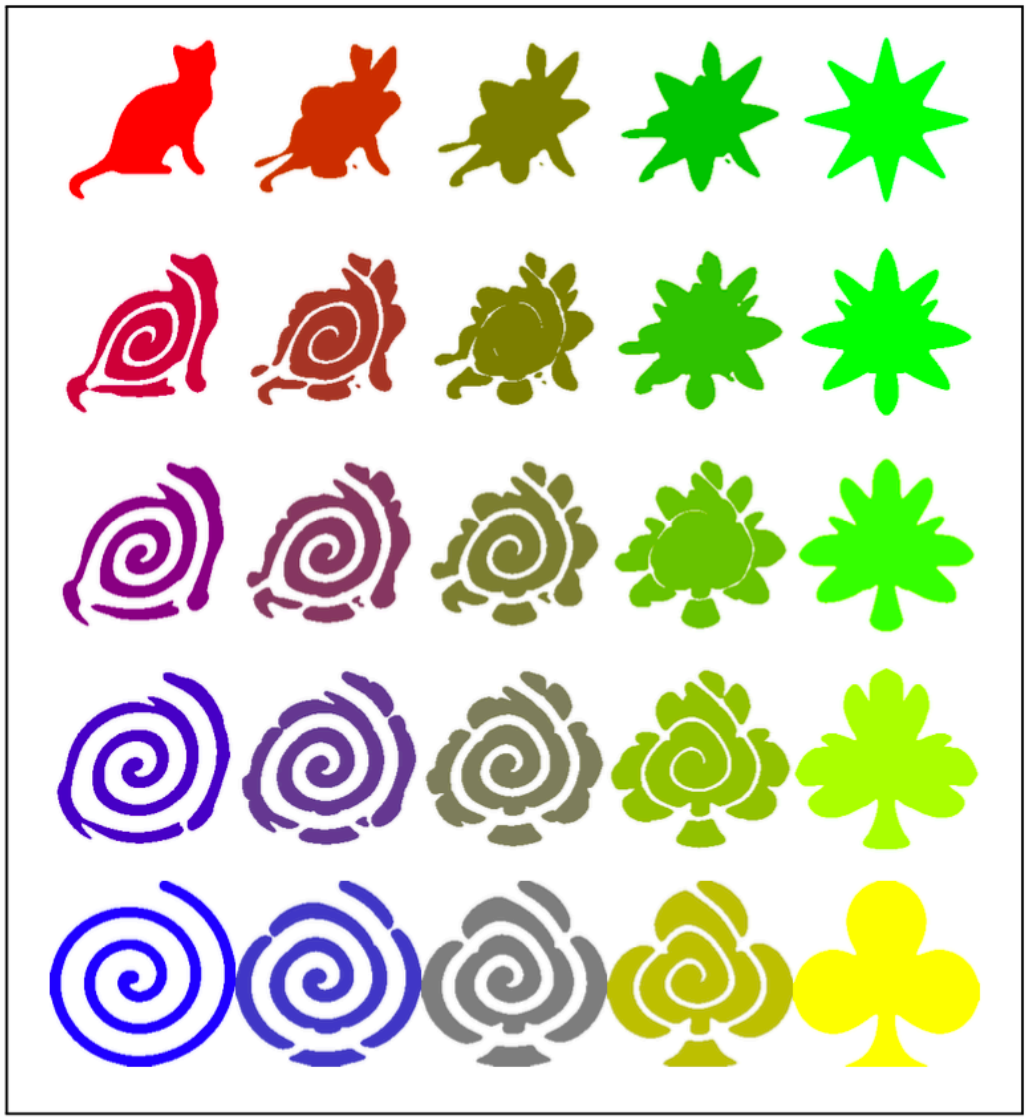


[Solomon et al, SIGGRAPH 2015]

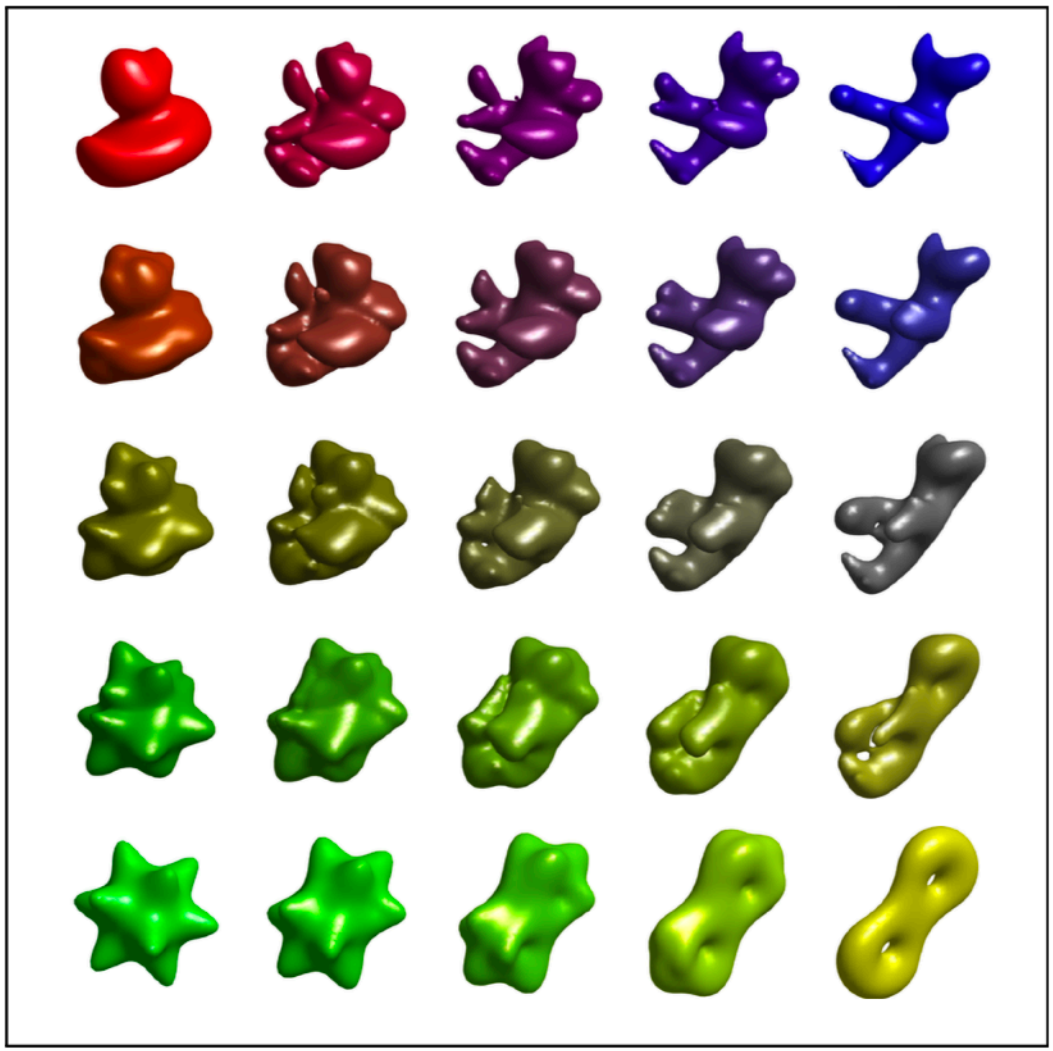


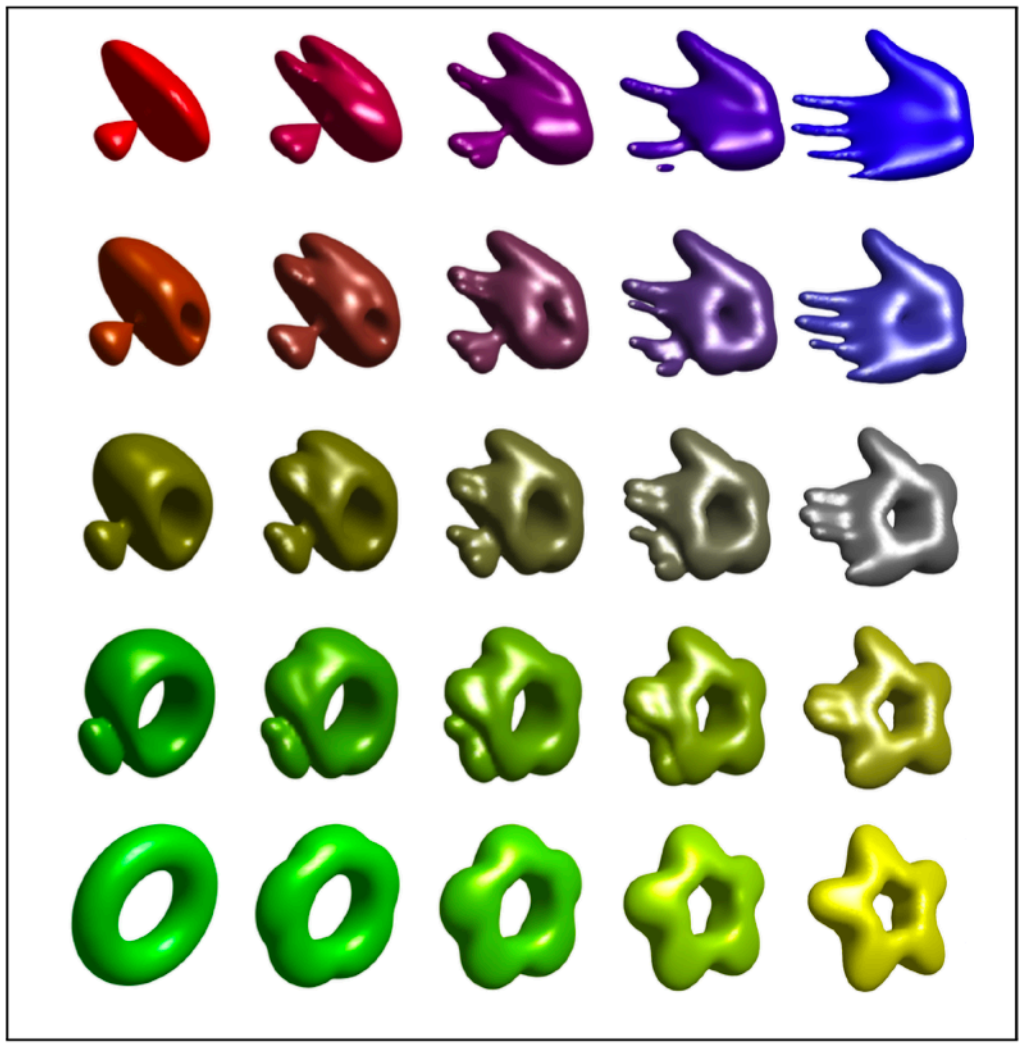
Barycenters of 2D Shapes



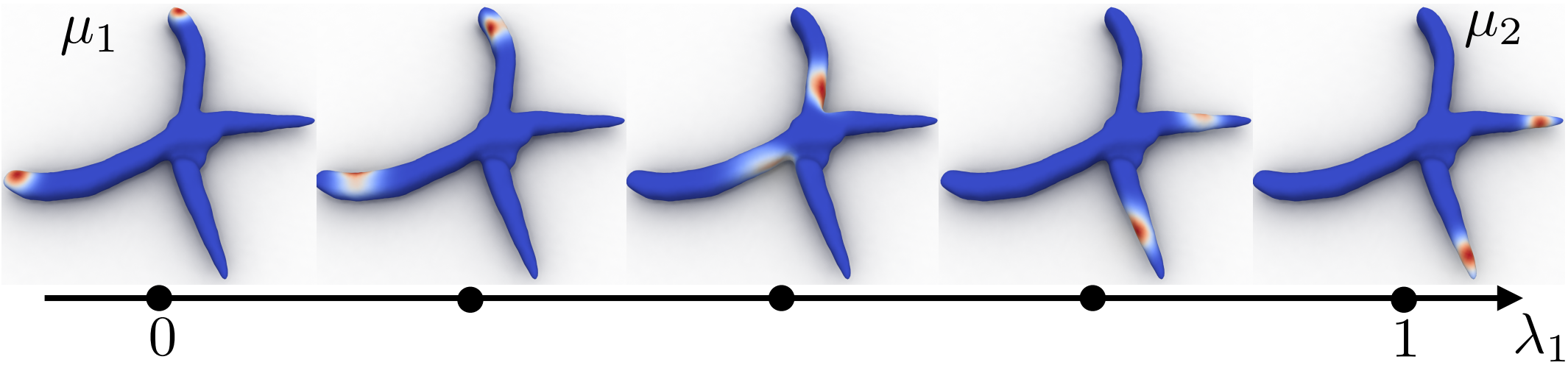


Barycenters of 3D Shapes

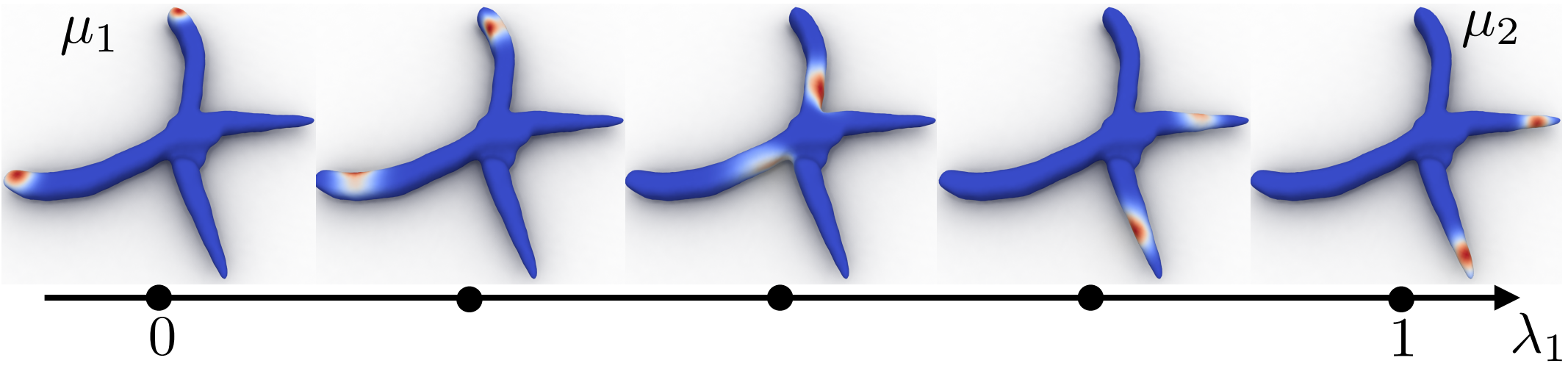


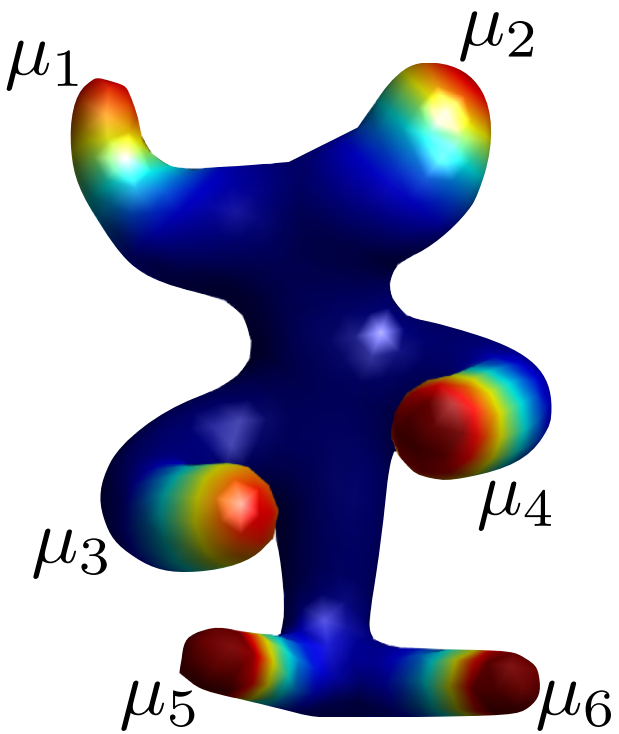


Barycenter on a Surface

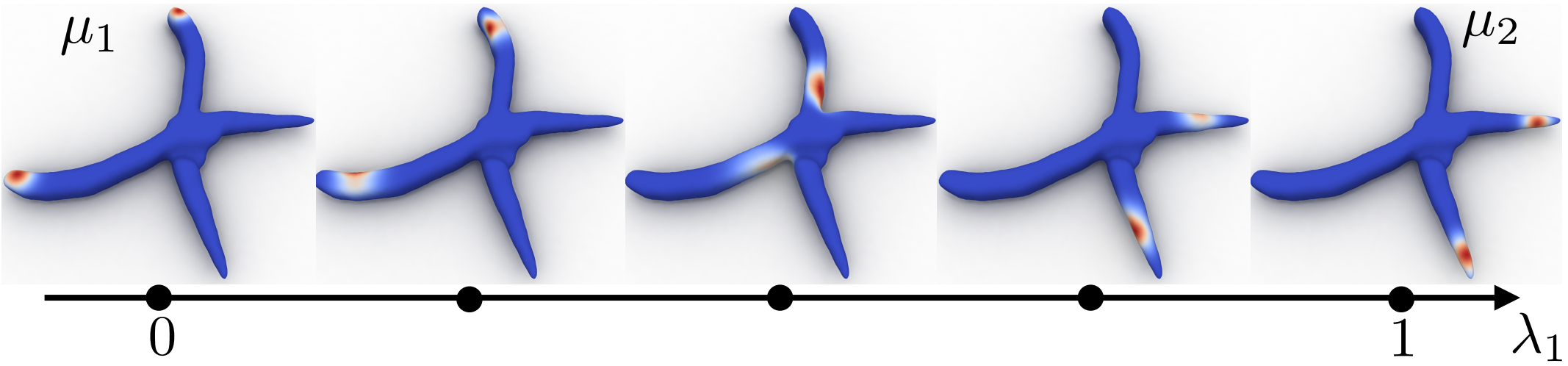


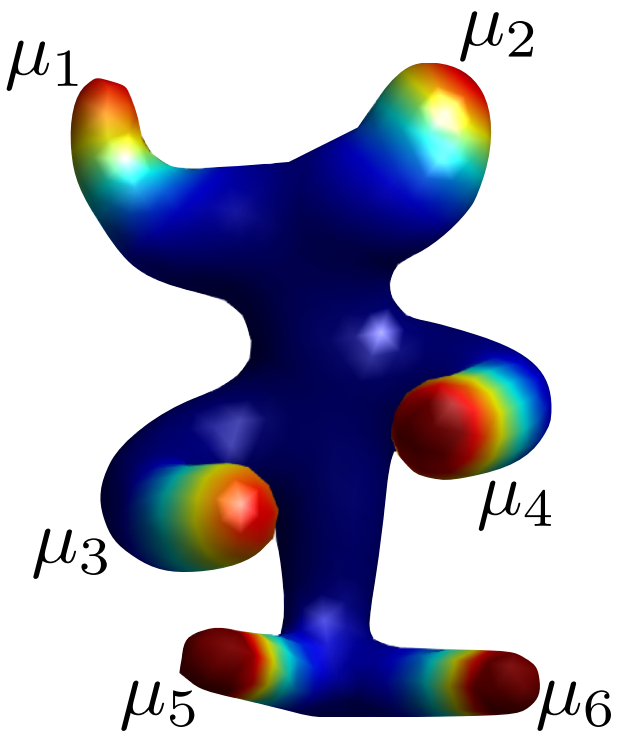
Barycenter on a Surface

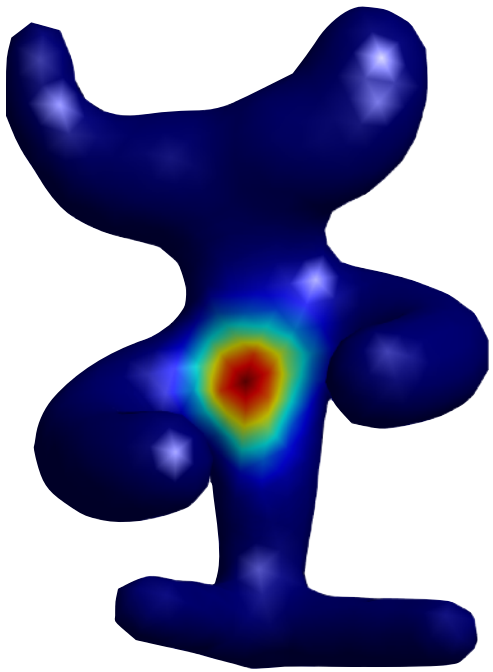




Barycenter on a Surface







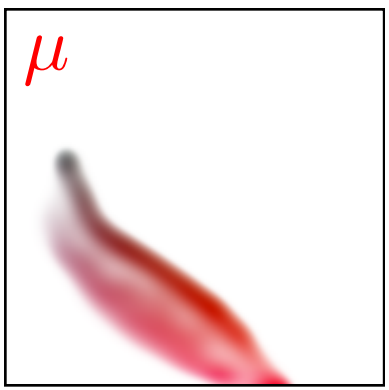
$\lambda = (1, \ldots, 1)/6$

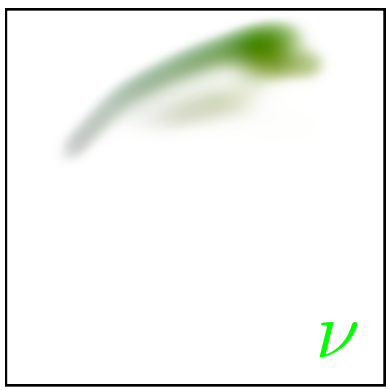
Color Transfer

Input images: (f, g) (chrominance components)

Input measures: $\mu(A) = \mathcal{U}(f^{-1}(A)), \nu(A) = \mathcal{U}(g^{-1}(A))$

f





g

Color Transfer

Input images: (f, g) (chrominance components)

Input measures: $\mu(A) = \mathcal{U}(f^{-1}(A)), \nu(A) = \mathcal{U}(g^{-1}(A))$

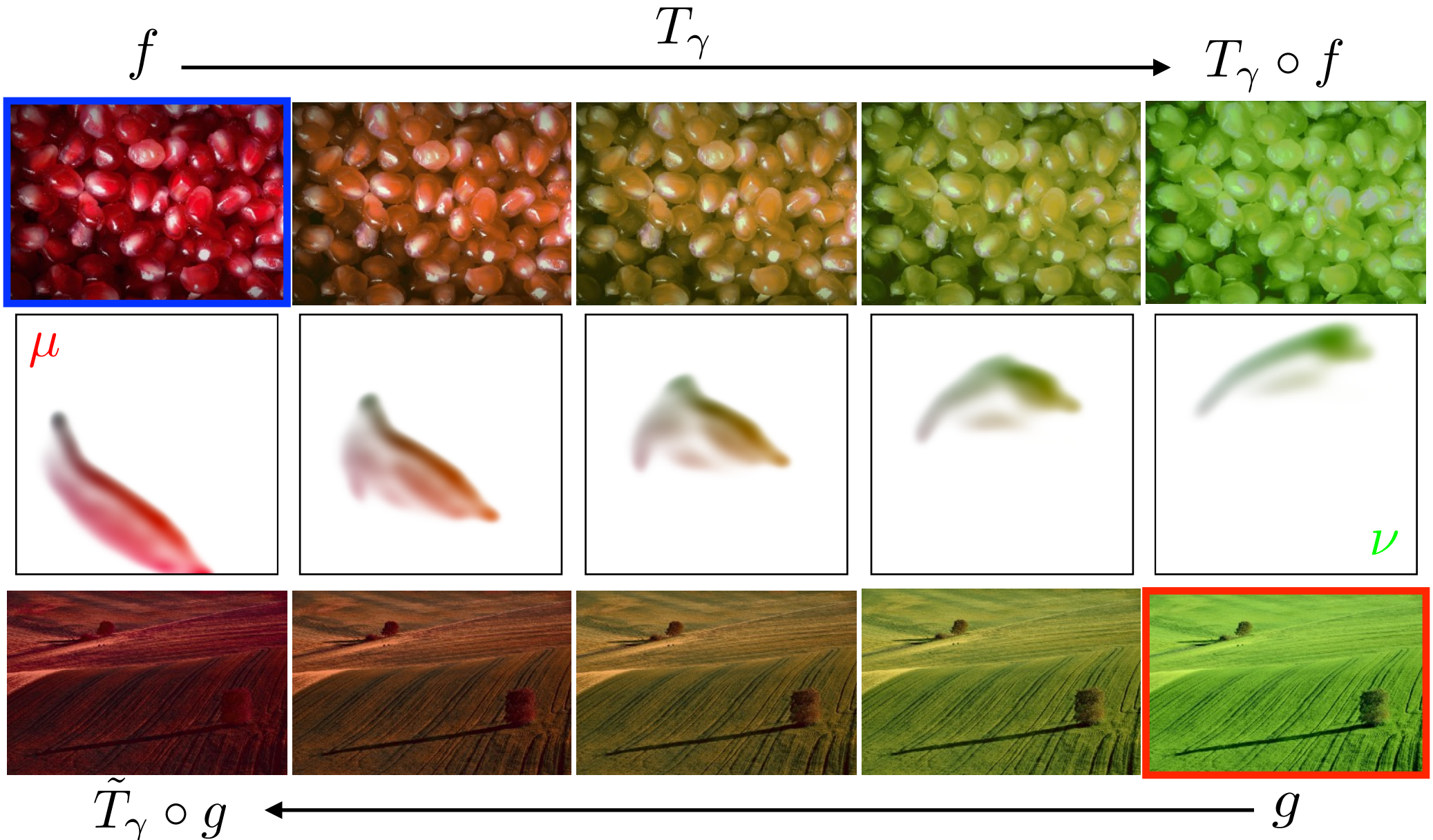
f

g

Color Transfer

Input images: (f, g) (chrominance components)

Input measures: $\mu(A) = \mathcal{U}(f^{-1}(A)), \nu(A) = \mathcal{U}(g^{-1}(A))$



Topic Models

[Rolet'16]

Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- Entropic Regularization and Sinkhorn
- Barycenters
- **Unbalanced OT and Gradient Flows**
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein

Unbalanced Transport

$$(\xi, \mu) \in \mathcal{M}_+(X)^2, \quad \mathrm{KL}(\xi|\mu) \stackrel{\text{def.}}{=} \int_X \log\left(\frac{\mathrm{d}\xi}{\mathrm{d}\mu}\right) \mathrm{d}\mu + \int_X (\mathrm{d}\mu - \mathrm{d}\xi)$$

$$W F_c(\mu, \nu) \stackrel{\text{def.}}{=} \min_{\pi} \langle c, \pi \rangle + \lambda \mathrm{KL}(P_{1\sharp}\pi|\mu) + \lambda \mathrm{KL}(P_{2\sharp}\pi|\nu)$$

[Liereo, Mielke, Savaré 2015]

Unbalanced Transport

$$(\xi, \mu) \in \mathcal{M}_+(X)^2, \quad \mathrm{KL}(\xi|\mu) \stackrel{\text{def.}}{=} \int_X \log\left(\frac{\mathrm{d}\xi}{\mathrm{d}\mu}\right) \mathrm{d}\mu + \int_X (\mathrm{d}\mu - \mathrm{d}\xi)$$

$$WF_c(\mu, \nu) \stackrel{\text{def.}}{=} \min_{\pi} \langle c, \pi \rangle + \lambda \mathrm{KL}(P_{1\sharp}\pi|\mu) + \lambda \mathrm{KL}(P_{2\sharp}\pi|\nu)$$

[Liereo, Mielke, Savaré 2015]

Proposition: If $c(x, y) = -\log(\cos(\min(d(x, y), \frac{\pi}{2}))$ then $WF_c^{1/2}$ is a distance on $\mathcal{M}_+(X)$.

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

Unbalanced Transport

$$(\xi, \mu) \in \mathcal{M}_+(X)^2, \quad \mathrm{KL}(\xi|\mu) \stackrel{\text{def.}}{=} \int_X \log\left(\frac{\mathrm{d}\xi}{\mathrm{d}\mu}\right) \mathrm{d}\mu + \int_X (\mathrm{d}\mu - \mathrm{d}\xi)$$

$$WF_c(\mu, \nu) \stackrel{\text{def.}}{=} \min_{\pi} \langle c, \pi \rangle + \lambda \mathrm{KL}(P_{1\sharp}\pi|\mu) + \lambda \mathrm{KL}(P_{2\sharp}\pi|\nu)$$

[Liereo, Mielke, Savaré 2015]

Proposition: If $c(x, y) = -\log(\cos(\min(d(x, y), \frac{\pi}{2}))$ then $WF_c^{1/2}$ is a distance on $\mathcal{M}_+(X)$.

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

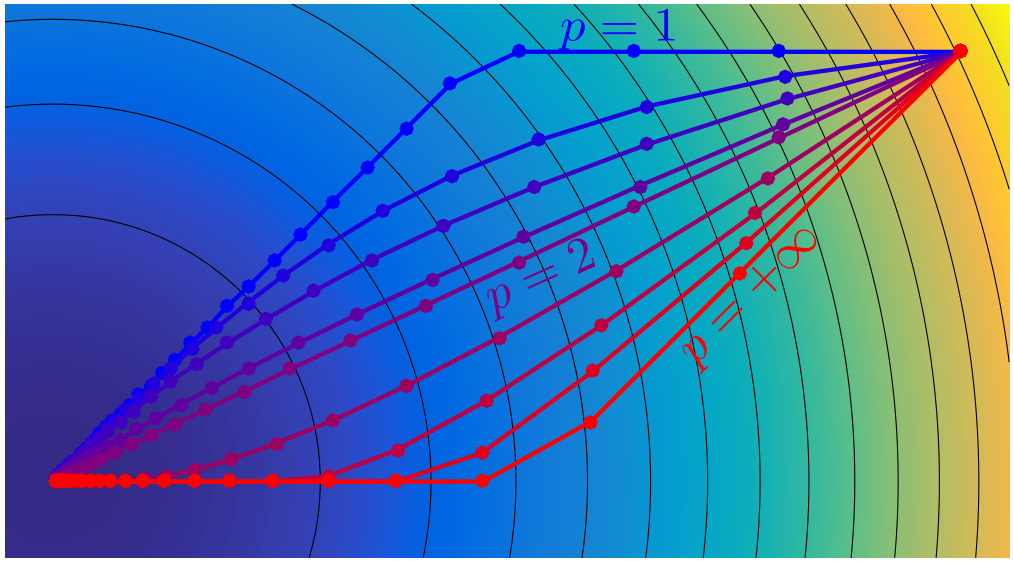
$\rightarrow$ "Dynamic" Benamou-Brenier formulation.

[Liereo, Mielke, Savaré 2015] [Kondratyev, Monsaingeon, Vorotnikov, 2015]

[Chizat, Schmitzer, P, Vialard 2015]

Implicit Euler Stepping

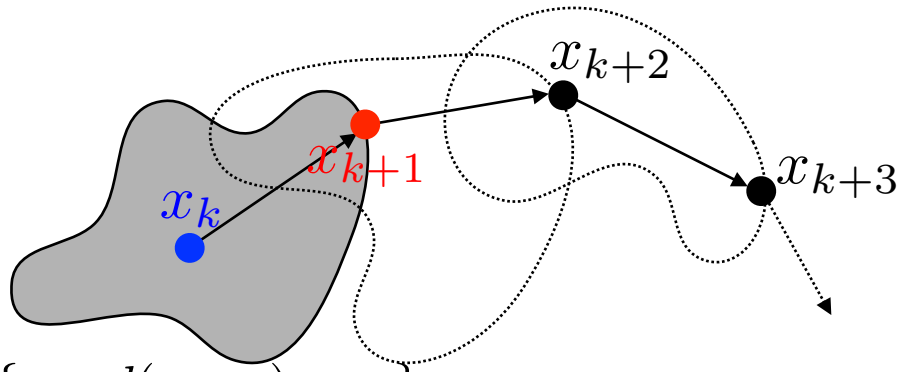
Metric space $(\mathcal{X}, d)$, minimize $F(x)$ on $\mathcal{X}$.



$F(x) = \|x\|^2$ on $(\mathcal{X} = \mathbb{R}^2, \|\cdot\|_p)$

Implicit Euler step:

$$x_{k+1} \overset{\text{def.}}{=} \operatorname*{argmin}_{x \in \mathcal{X}} d(x_k, x)^2 + \tau F(x)$$



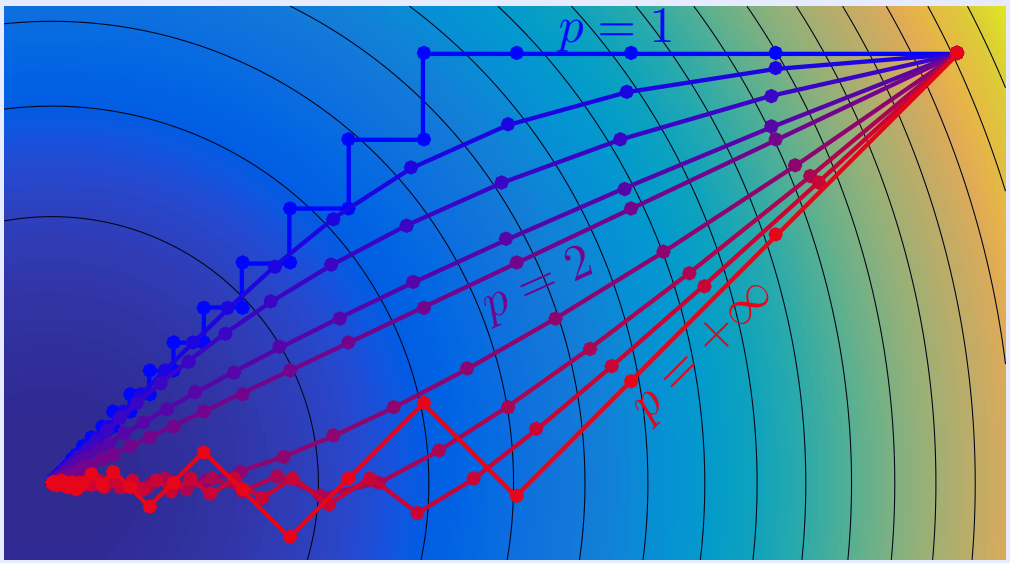
$\{x \; ; \; d(x_k, x) \sim \tau\}$

Implicit vs. Explicit Stepping

Metric space $(\mathcal{X}, d)$, minimize $F(x)$ on $\mathcal{X}$.

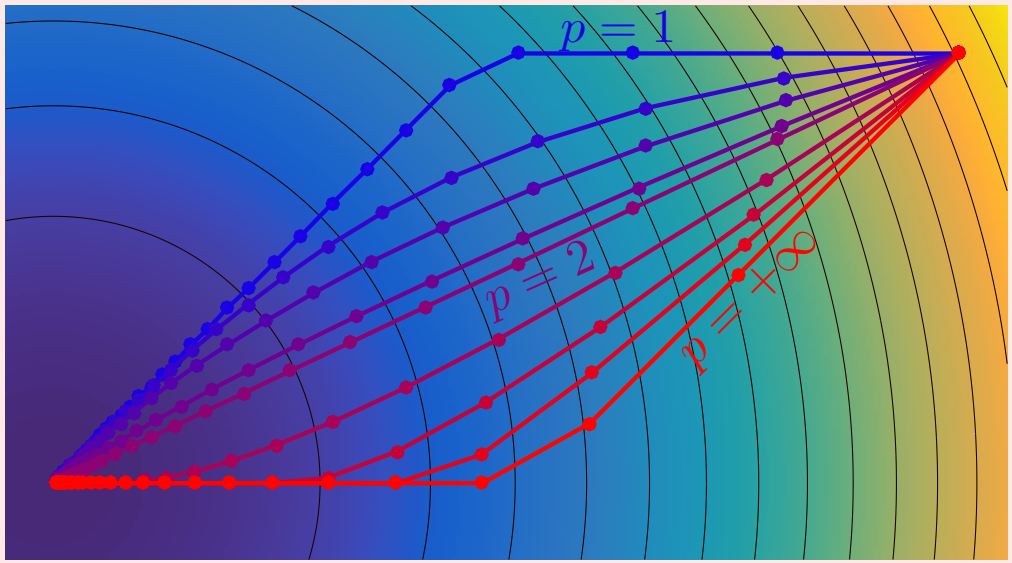
Explicit

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} d(x_k, x)^2 + \tau \langle \nabla F(x_k),\, x \rangle$$



Implicit

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} d(x_k, x)^2 + \tau F(x)$$



$F(x) = \|x\|^2$ on $(\mathcal{X} = \mathbb{R}^2, \|\cdot\|_p)$

Wasserstein Gradient Flows

Implicit Euler step: [Jordan, Kinderlehrer, Otto 1998]

$$\mu_{t+1} = \mathrm{Prox}^W_{\tau f}(\mu_t) \overset{\text{def.}}{=} \underset{\mu \in \mathcal{M}_+(X)}{\mathrm{argmin}} \; W_2^2(\mu_t, \mu) + \tau f(\mu)$$

Wasserstein Gradient Flows

Implicit Euler step: [Jordan, Kinderlehrer, Otto 1998]

$$\mu_{t+1} = \mathrm{Prox}^W_{\tau f}(\mu_t) \overset{\text{def.}}{=} \operatorname*{argmin}_{\mu \in \mathcal{M}_+(X)} W_2^2(\mu_t, \mu) + \tau f(\mu)$$

Formal limit $\tau \to 0$: $\partial_t \mu = \mathrm{div}\left(\mu \nabla(f'(\mu))\right)$

Wasserstein Gradient Flows

Implicit Euler step: [Jordan, Kinderlehrer, Otto 1998]

$$\mu_{t+1} = \mathrm{Prox}^W_{\tau f}(\mu_t) \overset{\text{def.}}{=} \underset{\mu \in \mathcal{M}_+(X)}{\mathrm{argmin}} \; W_2^2(\mu_t, \mu) + \tau f(\mu)$$

Formal limit $\tau \to 0$: $\quad \partial_t \mu = \mathrm{div}\left(\mu \nabla(f'(\mu))\right)$

$$f(\mu) = \int \log(\tfrac{\mathrm{d}\mu}{\mathrm{d}x}) \mathrm{d}\mu \longrightarrow \partial_t \mu = \Delta \mu \qquad \text{(heat diffusion)}$$

Wasserstein Gradient Flows

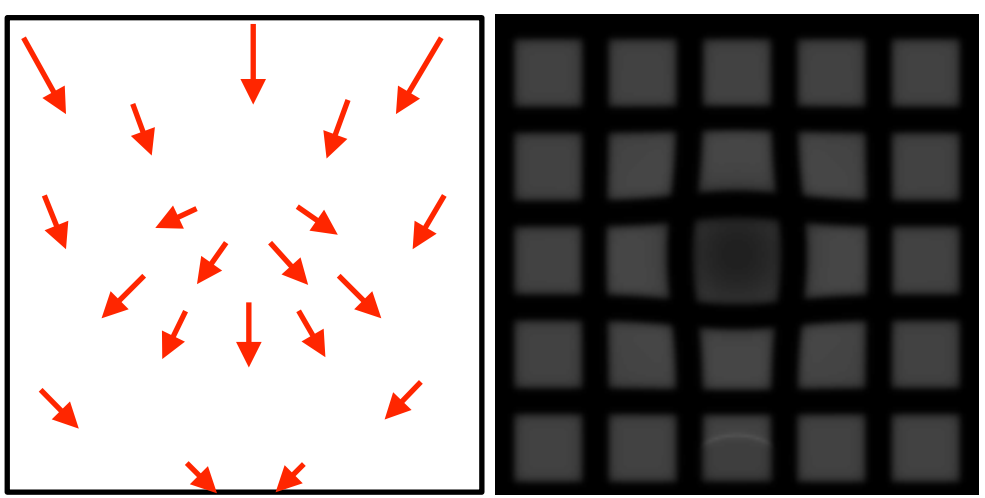
Implicit Euler step: [Jordan, Kinderlehrer, Otto 1998]

$$\mu_{t+1} = \mathrm{Prox}^{W}_{\tau f}(\mu_t) \overset{\text{def.}}{=} \underset{\mu \in \mathcal{M}_+(X)}{\mathrm{argmin}} \; W_2^2(\mu_t, \mu) + \tau f(\mu)$$

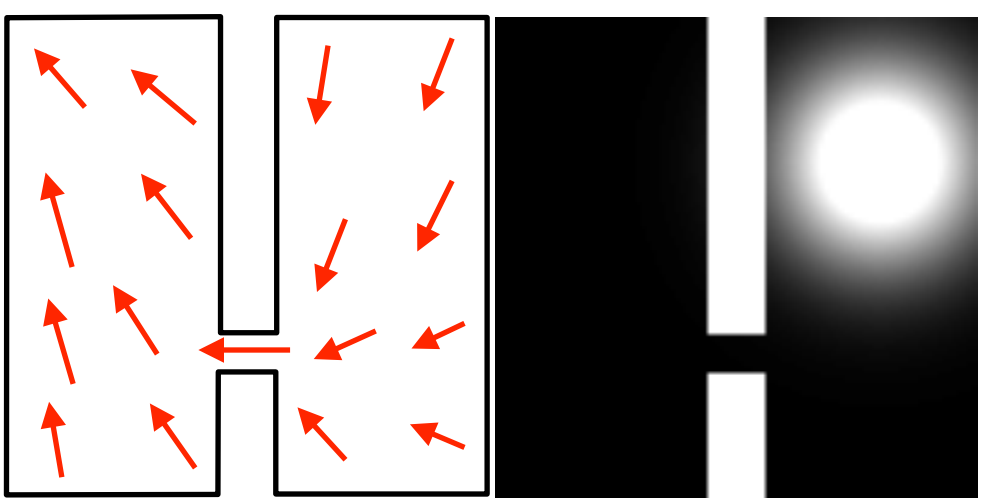
Formal limit $\tau \to 0$: $\partial_t \mu = \mathrm{div}\left(\mu \nabla (f'(\mu))\right)$

$$f(\mu) = \int \log(\tfrac{\mathrm{d}\mu}{\mathrm{d}x}) \mathrm{d}\mu \longrightarrow \partial_t \mu = \Delta \mu \qquad \text{(heat diffusion)}$$

$$f(\mu) = \int w \mathrm{d}\mu \longrightarrow \partial_t \mu = \mathrm{div}(\mu \nabla w) \qquad \text{(advection)}$$



∇w Evolution μ_t



∇w Evolution μ_t

Wasserstein Gradient Flows

Implicit Euler step: [Jordan, Kinderlehrer, Otto 1998]

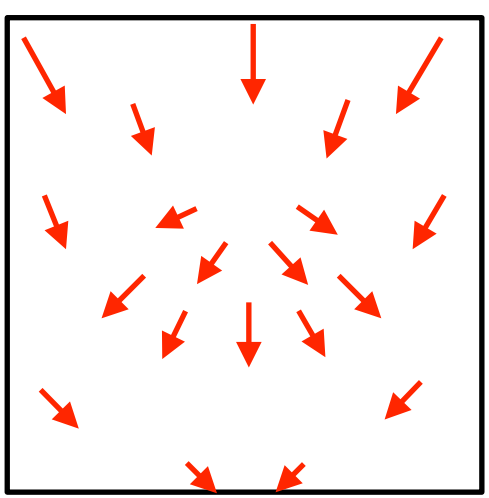
$$\mu_{t+1} = \mathrm{Prox}^{W}_{\tau f}(\mu_t) \overset{\text{def.}}{=} \underset{\mu \in \mathcal{M}_+(X)}{\mathrm{argmin}} \; W_2^2(\mu_t, \mu) + \tau f(\mu)$$

Formal limit $\tau \to 0$: $\partial_t \mu = \mathrm{div}\left(\mu \nabla (f'(\mu))\right)$

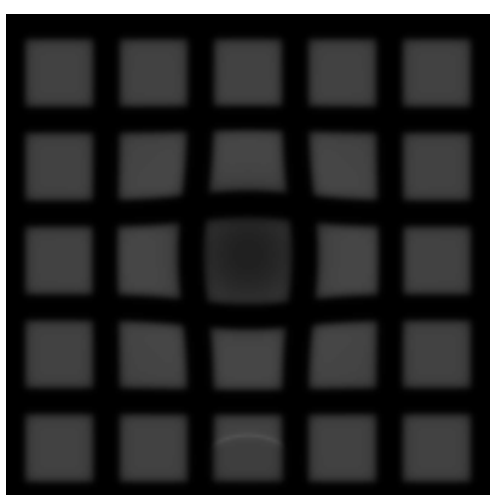
$f(\mu) = \int \log(\frac{\mathrm{d}\mu}{\mathrm{d}x}) \mathrm{d}\mu \longrightarrow \partial_t \mu = \Delta \mu$ (heat diffusion)

$f(\mu) = \int w \mathrm{d}\mu \longrightarrow \partial_t \mu = \mathrm{div}(\mu \nabla w)$ (advection)

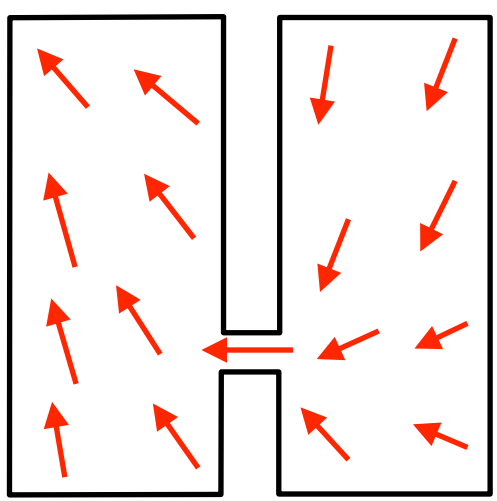
$f(\mu) = \frac{1}{m-1} \int (\frac{\mathrm{d}\mu}{\mathrm{d}x})^{m-1} \mathrm{d}\mu \longrightarrow \partial_t \mu = \Delta \mu^m$ (non-linear diffusion)



∇w



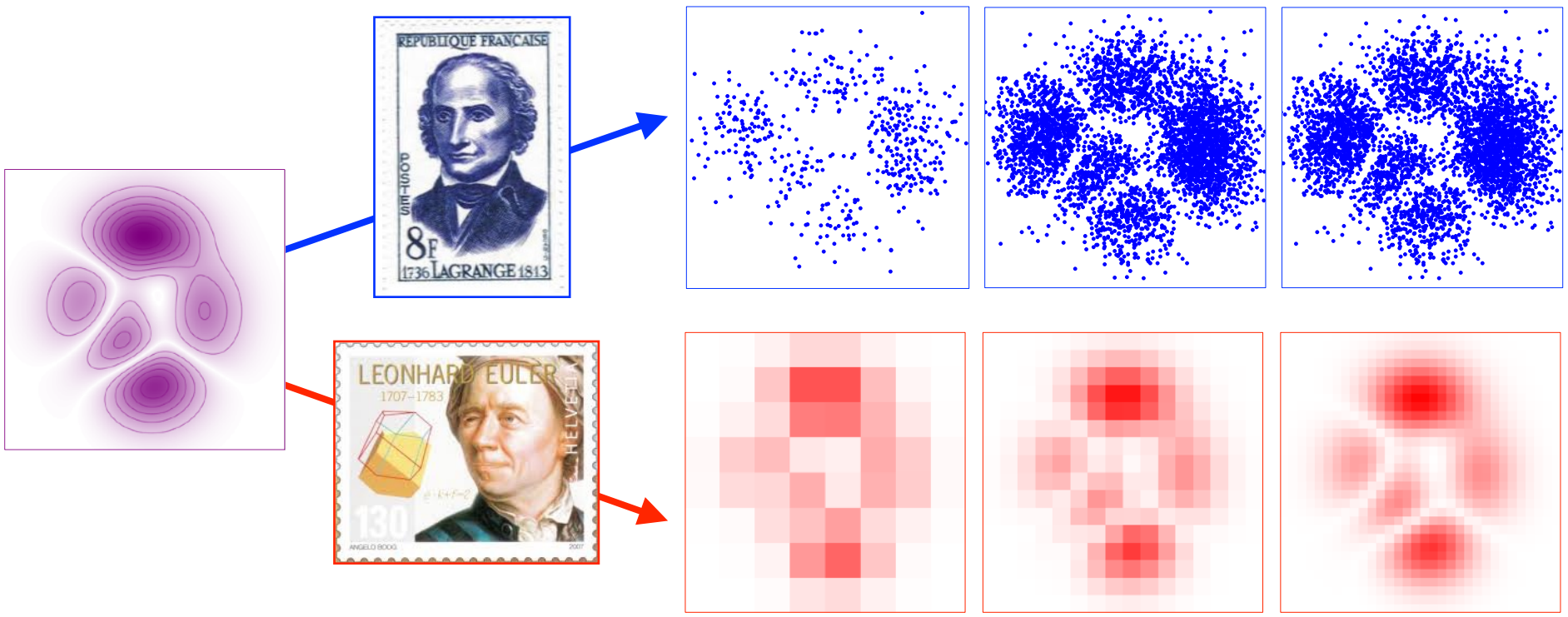
Evolution μ_t



∇w

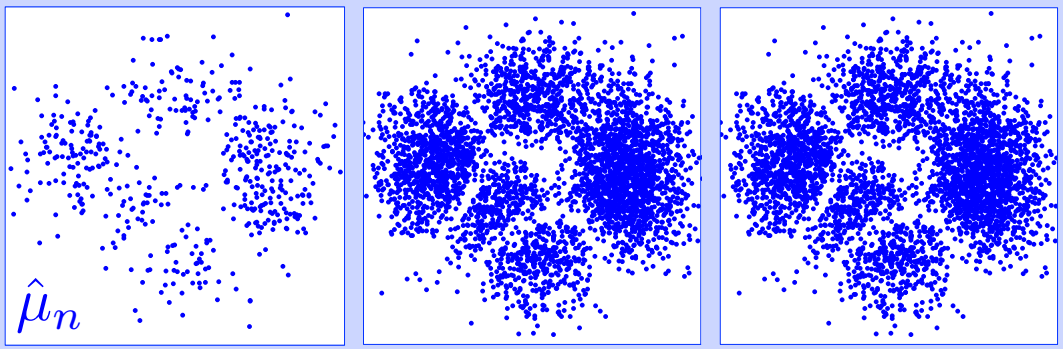
Evolution μ_t

Eulerian vs. Lagrangian Discretization



Lagrangian Discretization of Entropy

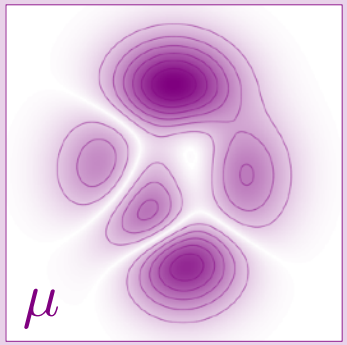
$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$



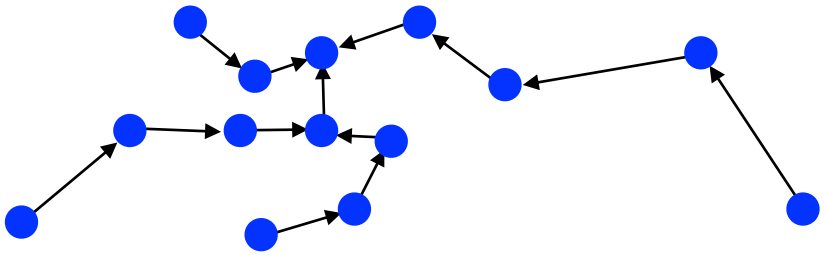
$$\hat{H}(\hat{\mu}_n) \stackrel{\text{def.}}{=} \sum_i \log(\min_{j \neq i} \|x_i - x_j\|)$$

$n \to +\infty$

$n \to +\infty$



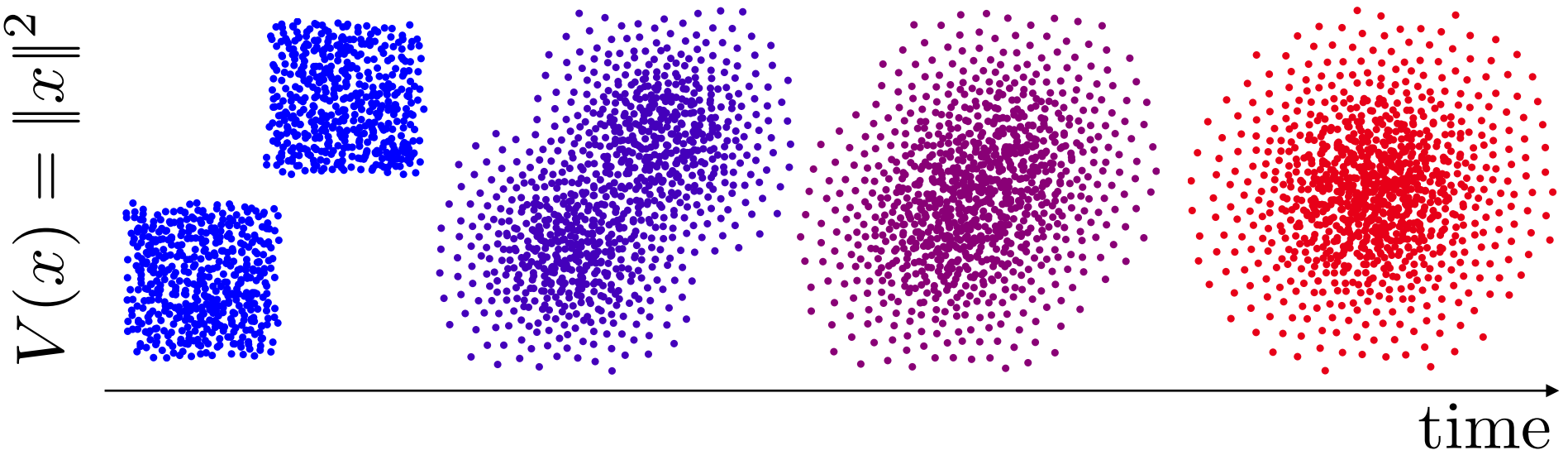
$$H(\mu) \stackrel{\text{def.}}{=} -\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}(x)\right)\mathrm{d}\mu(x)$$



Lagrangian Discretization of Gradient Flows

$$\min_{\rho} E(\rho) \overset{\text{def.}}{=} \int V(x)\rho(x)\mathrm{d}x + \int \rho(x)\log(\rho(x))\mathrm{d}x$$

Wasserstein flow of E: $\dfrac{\mathrm{d}\rho_t}{\mathrm{d}t} = \Delta\rho_t + \nabla(V\rho_t)$



Generalized Entropic Regularization

$$\textit{Primal: } \min_{\pi} \langle d^p, \pi \rangle + f_1(P_{1\sharp}\pi) + f_2(P_{2\sharp}\pi) + \varepsilon \mathrm{KL}(\pi | \pi_0)$$

Generalized Entropic Regularization

Primal: $\min_{\pi} \langle d^p, \pi \rangle + f_1(P_{1\sharp}\pi) + f_2(P_{2\sharp}\pi) + \varepsilon \mathrm{KL}(\pi|\pi_0)$

Dual: $\max_{u,v} \; -f_1^*(u) - f_2^*(v) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, K e^{-\frac{v}{\varepsilon}} \rangle$

$\pi(x,y) = a(x)K(x,y)b(y) \qquad (a,b) \overset{\text{def.}}{=} (e^{-\frac{u}{\varepsilon}}, e^{-\frac{v}{\varepsilon}})$

Generalized Entropic Regularization

Primal: $\min_{\pi} \langle d^p, \pi \rangle + f_1(P_{1\sharp}\pi) + f_2(P_{2\sharp}\pi) + \varepsilon \mathrm{KL}(\pi|\pi_0)$

Dual: $\max_{u,v} \; -f_1^*(u) - f_2^*(v) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, K e^{-\frac{v}{\varepsilon}} \rangle$

$\pi(x,y) = a(x)K(x,y)b(y) \qquad (a,b) \overset{\text{def.}}{=} (e^{-\frac{u}{\varepsilon}}, e^{-\frac{v}{\varepsilon}})$

Block coordinates relaxation:

$\max_{u} - f_1^*(u) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, K e^{-\frac{v}{\varepsilon}} \rangle \qquad (\mathcal{I}_u)$

$\max_{v} - f_2^*(v) - \varepsilon \langle e^{-\frac{v}{\varepsilon}}, K^* e^{-\frac{u}{\varepsilon}} \rangle \qquad (\mathcal{I}_v)$

Generalized Entropic Regularization

Primal: $\min_{\pi} \langle d^p, \pi \rangle + f_1(P_{1\sharp}\pi) + f_2(P_{2\sharp}\pi) + \varepsilon \mathrm{KL}(\pi|\pi_0)$

Dual: $\max_{u,v} -f_1^*(u) - f_2^*(v) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, K e^{-\frac{v}{\varepsilon}} \rangle$

$\pi(x,y) = a(x)K(x,y)b(y) \qquad (a,b) \overset{\text{def.}}{=} (e^{-\frac{u}{\varepsilon}}, e^{-\frac{v}{\varepsilon}})$

Block coordinates relaxation:

$$\max_u -f_1^*(u) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, K e^{-\frac{v}{\varepsilon}} \rangle \qquad (\mathcal{I}_u)$$

$$\max_v -f_2^*(v) - \varepsilon \langle e^{-\frac{v}{\varepsilon}}, K^* e^{-\frac{u}{\varepsilon}} \rangle \qquad (\mathcal{I}_v)$$

Proposition: the solutions of $(\mathcal{I}_u)$ and $(\mathcal{I}_v)$ read:

$$a = \frac{\mathrm{Prox}^{\mathrm{KL}}_{f_1/\varepsilon}(Kb)}{Kb} \qquad b = \frac{\mathrm{Prox}^{\mathrm{KL}}_{f_2/\varepsilon}(K^*a)}{K^*a}$$

$$\mathrm{Prox}^{\mathrm{KL}}_{f_1/\varepsilon}(\mu) \overset{\text{def.}}{=} \mathrm{argmin}_\nu f_1(\nu) + \varepsilon \mathrm{KL}(\nu|\mu)$$

Generalized Entropic Regularization

Primal: $\min_{\pi} \langle d^p, \pi \rangle + f_1(P_{1\sharp}\pi) + f_2(P_{2\sharp}\pi) + \varepsilon \mathrm{KL}(\pi|\pi_0)$

Dual: $\max_{u,v} \; -f_1^*(u) - f_2^*(v) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, K e^{-\frac{v}{\varepsilon}} \rangle$

$\pi(x,y) = a(x)K(x,y)b(y) \qquad (a,b) \overset{\text{def.}}{=} (e^{-\frac{u}{\varepsilon}}, e^{-\frac{v}{\varepsilon}})$

Block coordinates relaxation:

$$\max_u \; -f_1^*(u) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, K e^{-\frac{v}{\varepsilon}} \rangle \qquad (\mathcal{I}_u)$$

$$\max_v \; -f_2^*(v) - \varepsilon \langle e^{-\frac{v}{\varepsilon}}, K^* e^{-\frac{u}{\varepsilon}} \rangle \qquad (\mathcal{I}_v)$$

Proposition: the solutions of $(\mathcal{I}_u)$ and $(\mathcal{I}_v)$ read:

$$a = \frac{\mathrm{Prox}^{\mathrm{KL}}_{f_1/\varepsilon}(Kb)}{Kb} \qquad b = \frac{\mathrm{Prox}^{\mathrm{KL}}_{f_2/\varepsilon}(K^*a)}{K^*a}$$

$$\mathrm{Prox}^{\mathrm{KL}}_{f_1/\varepsilon}(\mu) \overset{\text{def.}}{=} \mathrm{argmin}_{\nu} f_1(\nu) + \varepsilon \mathrm{KL}(\nu|\mu)$$

$\rightarrow$ Only matrix-vector multiplications. $\rightarrow$ Highly parallelizable.

$\rightarrow$ On regular grids: only convolutions! Linear time iterations.

Gradient Flows: Crowd Motion

$$\mu_{t+1} \overset{\text{def.}}{=} \operatorname{argmin}_{\mu} W_{\alpha}^{\alpha}(\mu_t, \mu) + \tau f(\mu)$$

Congestion-inducing function:

$$f(\mu) = \iota_{[0,\kappa]}(\mu) + \langle w,\, \mu \rangle$$

[Maury, Roudneff-Chupin, Santambrogio 2010]

Gradient Flows: Crowd Motion

$$\mu_{t+1} \overset{\text{def.}}{=} \operatorname{argmin}_{\mu} W_{\alpha}^{\alpha}(\mu_t, \mu) + \tau f(\mu)$$

Congestion-inducing function:

$$f(\mu) = \iota_{[0,\kappa]}(\mu) + \langle w,\, \mu \rangle$$

[Maury, Roudneff-Chupin, Santambrogio 2010]

Proposition: $\operatorname{Prox}_{\frac{1}{\varepsilon} f}(\mu) = \min(e^{-\varepsilon w}\mu, \kappa)$

Gradient Flows: Crowd Motion

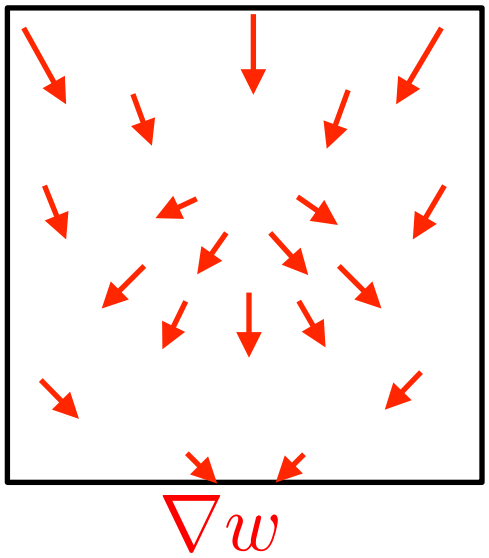
$$\mu_{t+1} \overset{\text{def.}}{=} \operatorname{argmin}_{\mu} W_{\alpha}^{\alpha}(\mu_t, \mu) + \tau f(\mu)$$

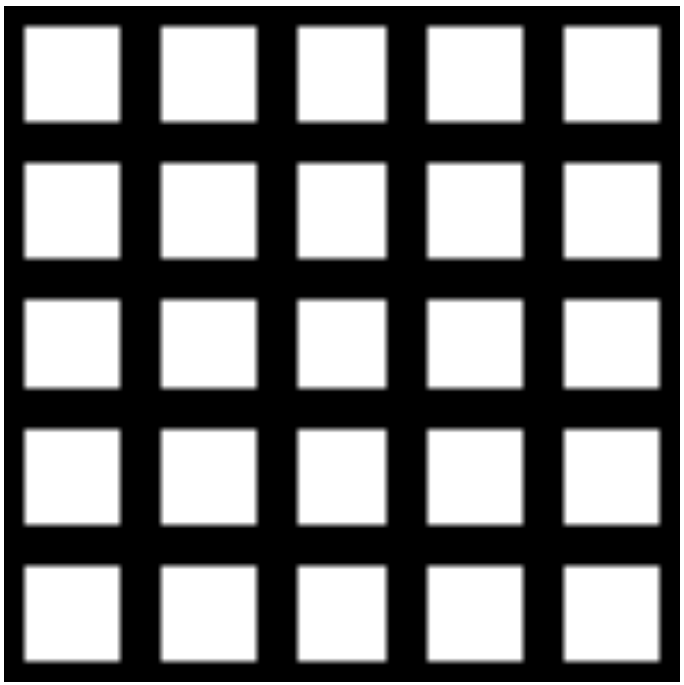
Congestion-inducing function:

$$f(\mu) = \iota_{[0,\kappa]}(\mu) + \langle w,\, \mu \rangle$$

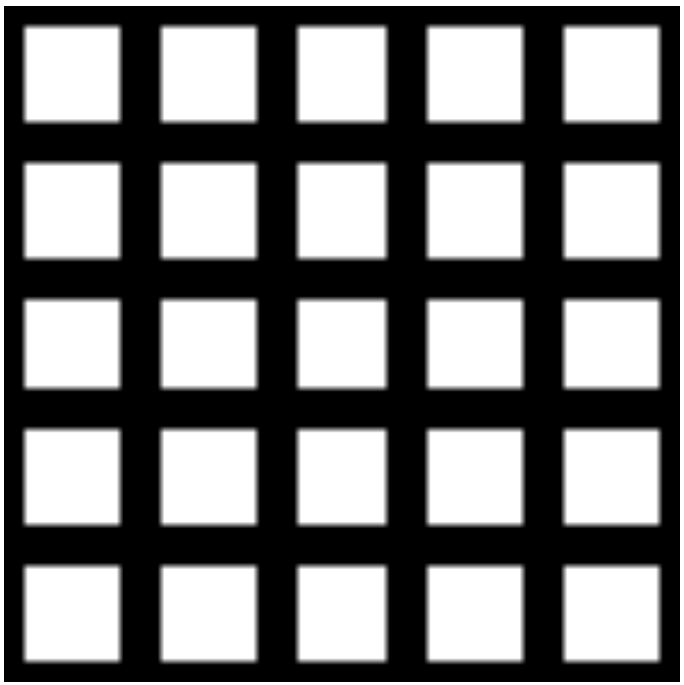
[Maury, Roudneff-Chupin, Santambrogio 2010]

Proposition: $\operatorname{Prox}_{\frac{1}{\varepsilon} f}(\mu) = \min(e^{-\varepsilon w}\mu, \kappa)$

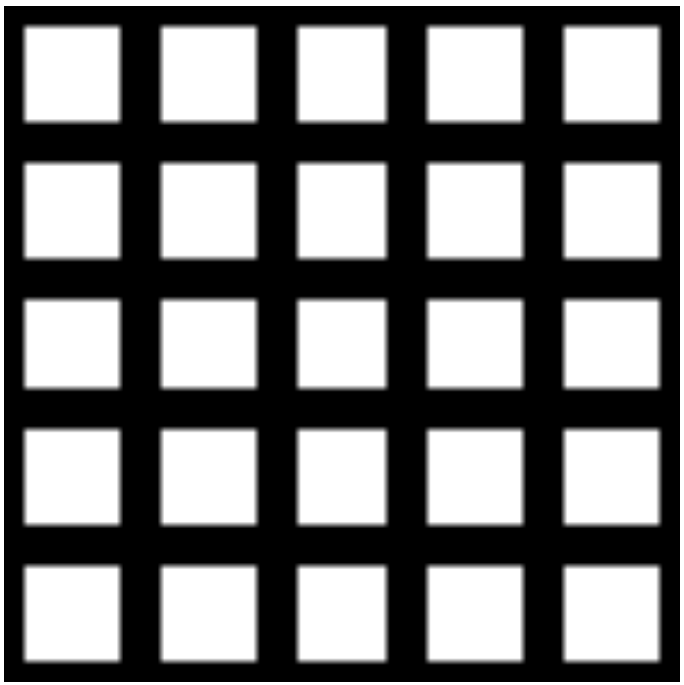




$\kappa = \|\mu_{t=0}\|_{\infty}$



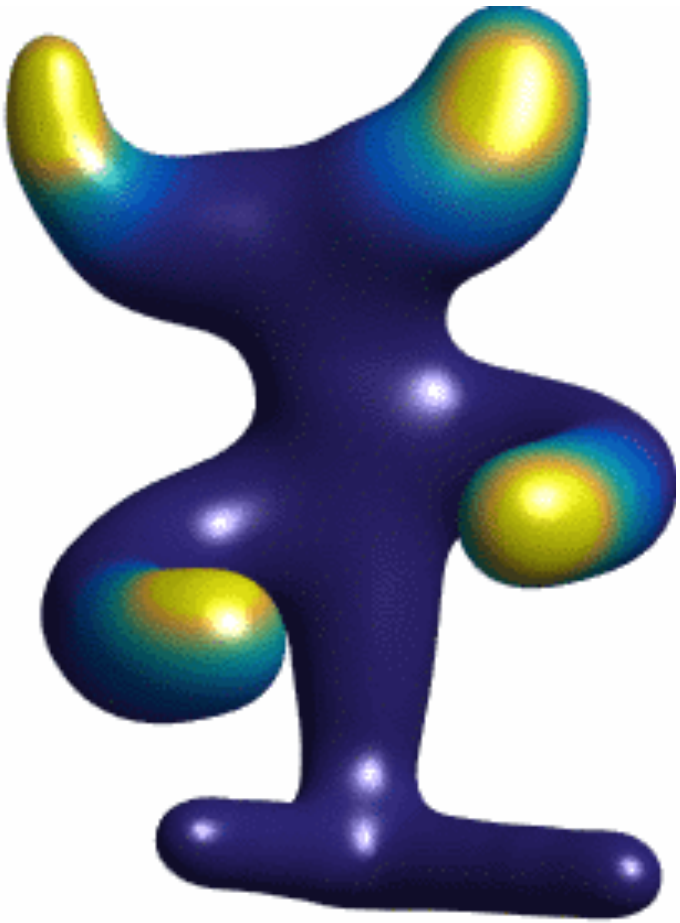
$\kappa = 2\|\mu_{t=0}\|_{\infty}$



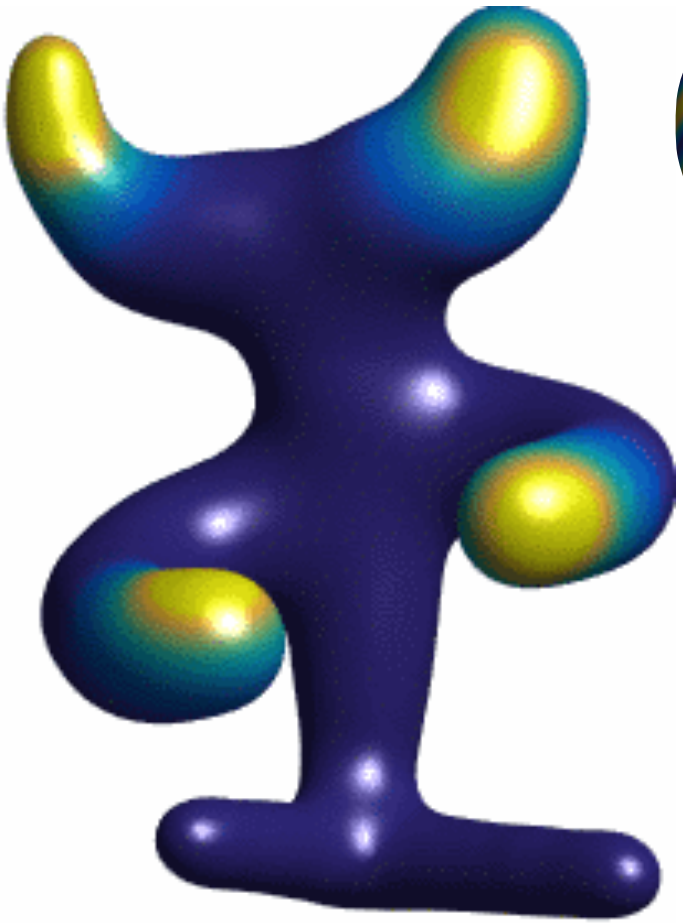
$\kappa = 4\|\mu_{t=0}\|_{\infty}$

Crowd Motion on a Surface

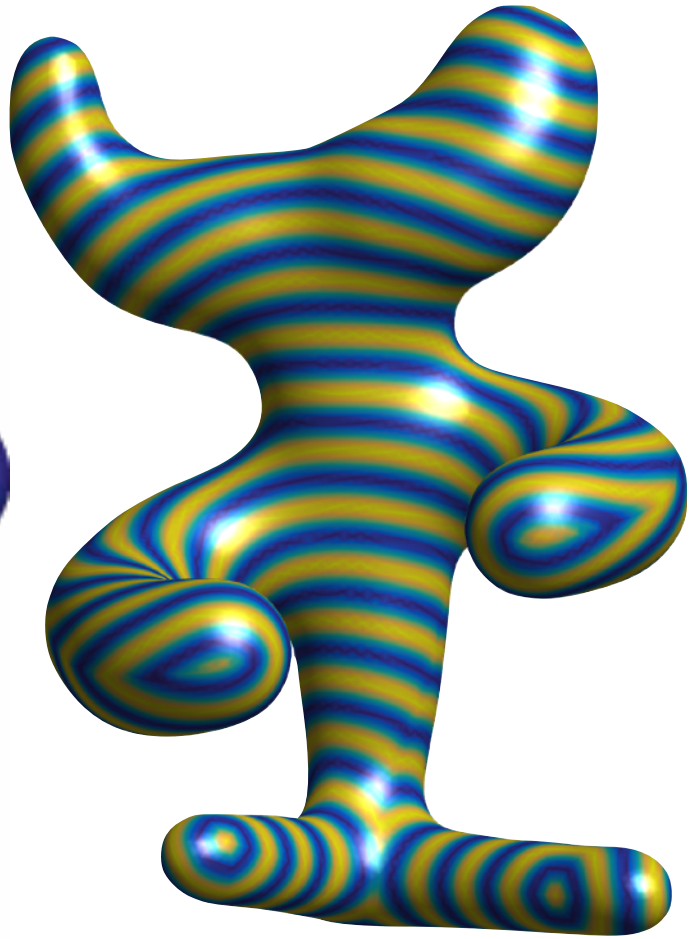
$X =$ triangulated mesh.



$\kappa = \|\mu_{t=0}\|_\infty$



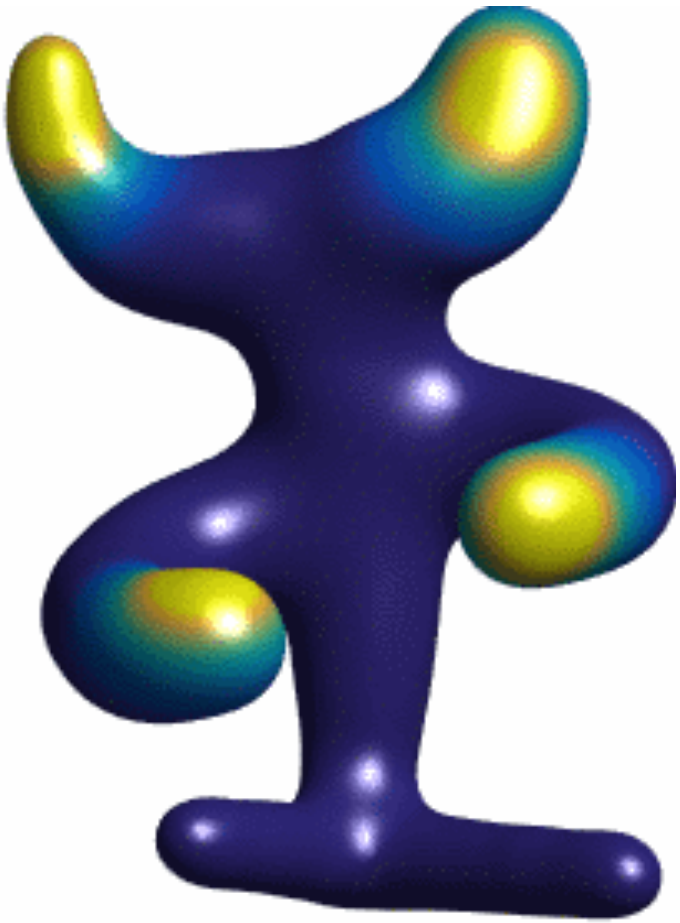
$\kappa = 6\|\mu_{t=0}\|_\infty$



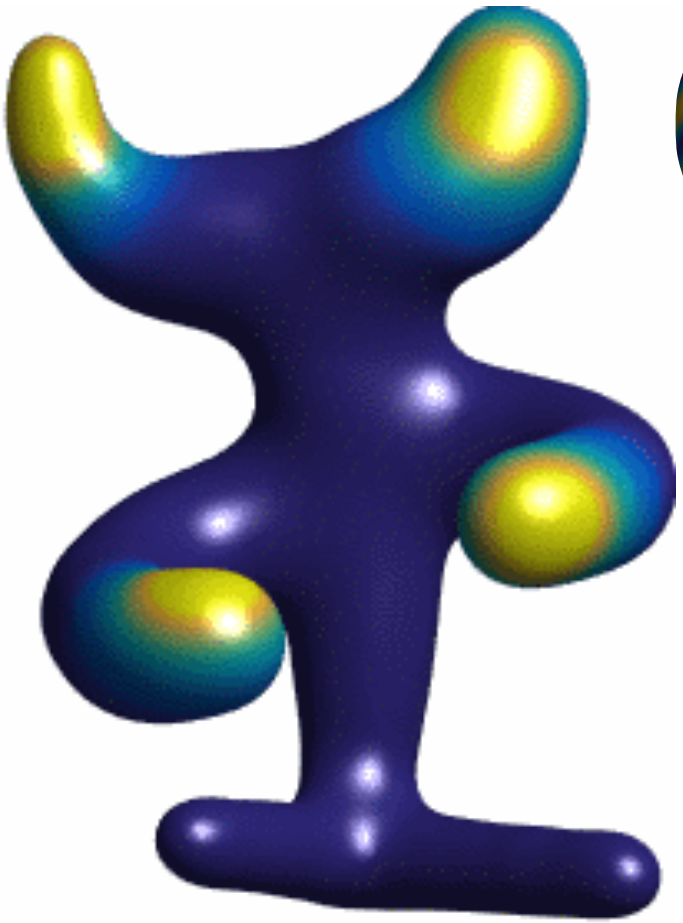
Potential $\cos(w)$

Crowd Motion on a Surface

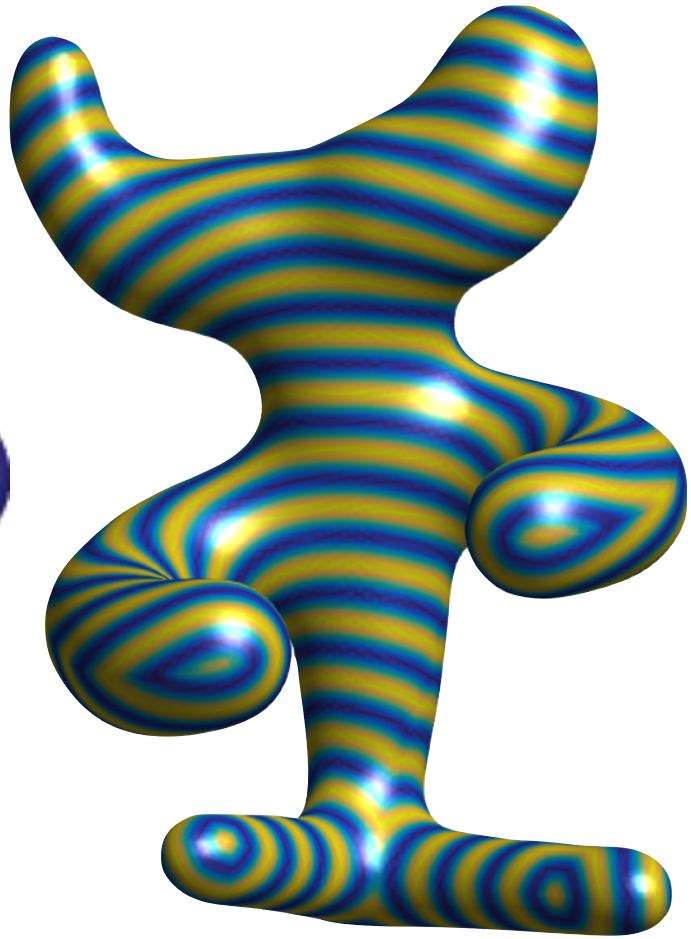
$X =$ triangulated mesh.



$\kappa = \|\mu_{t=0}\|_\infty$



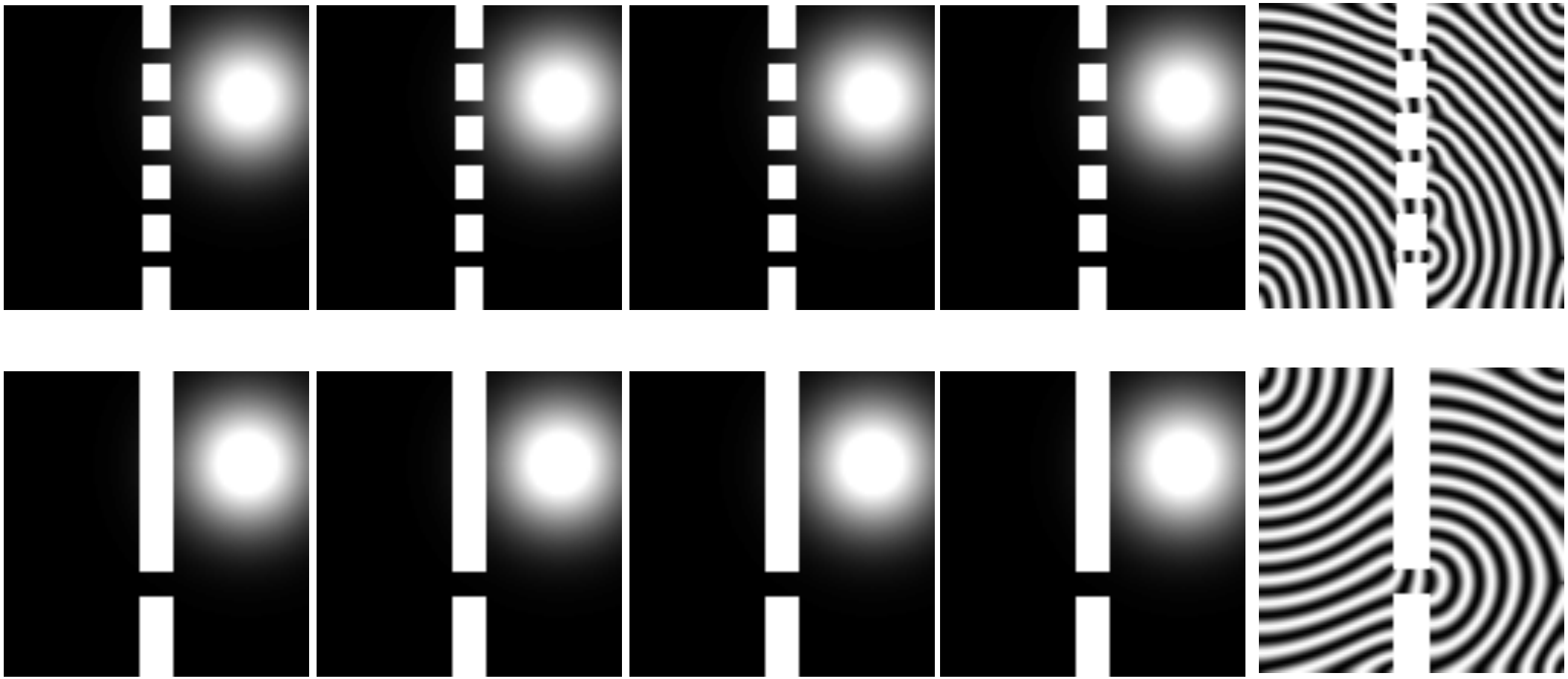
$\kappa = 6\|\mu_{t=0}\|_\infty$



Potential $\cos(w)$

Gradient Flows: Crowd Motion with Obstacles

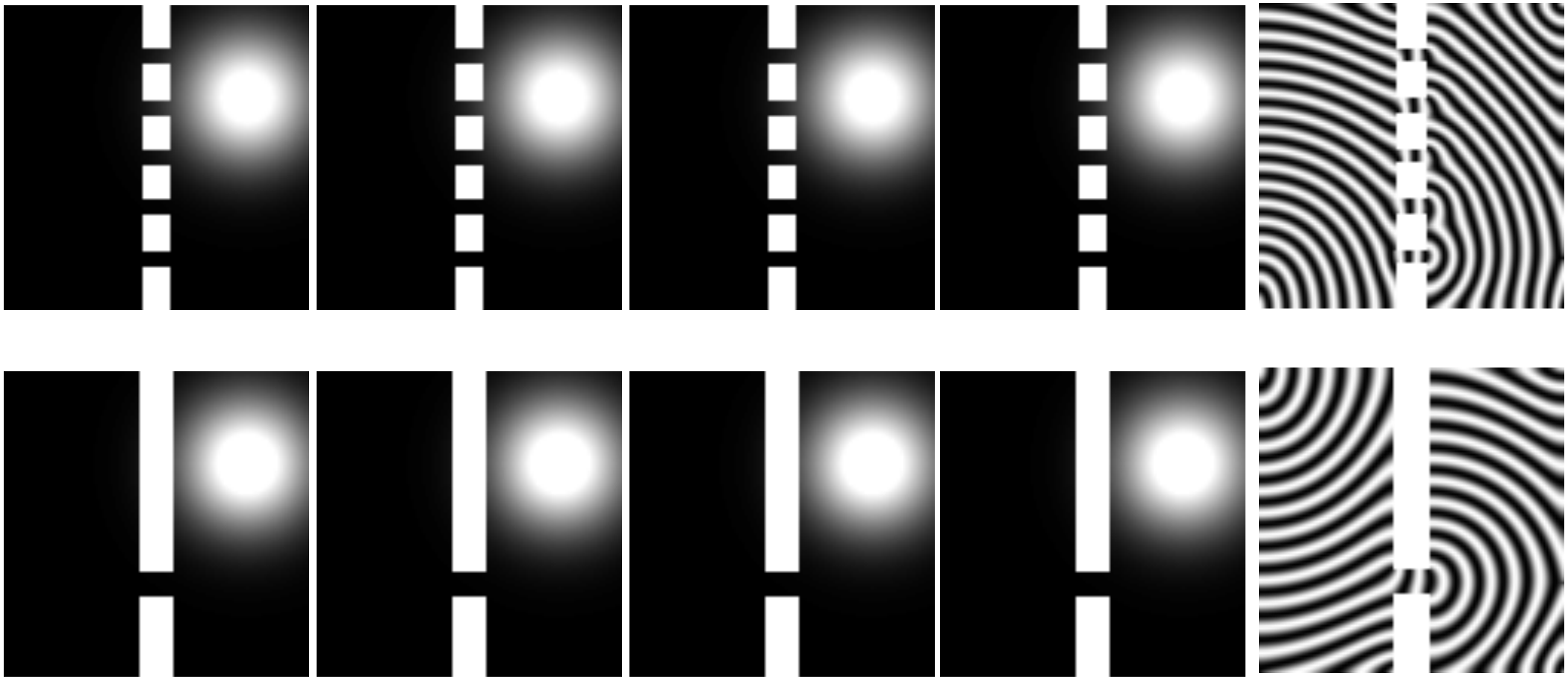
$X = \text{sub-domain of } \mathbb{R}^2.$



$\kappa = \|\mu_{t=0}\|_\infty$ $\kappa = 2\|\mu_{t=0}\|_\infty$ $\kappa = 4\|\mu_{t=0}\|_\infty$ $\kappa = 6\|\mu_{t=0}\|_\infty$ Potential $\cos(w)$

Gradient Flows: Crowd Motion with Obstacles

$X = \text{sub-domain of } \mathbb{R}^2.$



$\kappa = \|\mu_{t=0}\|_\infty$ $\kappa = 2\|\mu_{t=0}\|_\infty$ $\kappa = 4\|\mu_{t=0}\|_\infty$ $\kappa = 6\|\mu_{t=0}\|_\infty$ Potential $\cos(w)$

Multiple-Density Gradient Flows

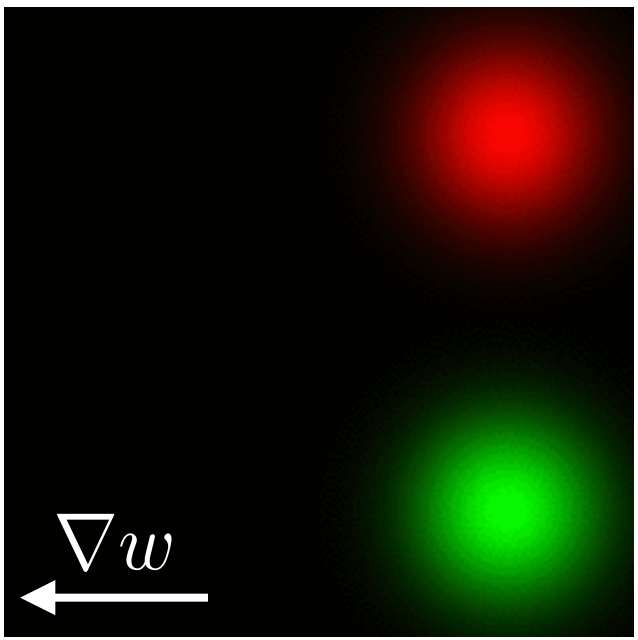
$$(\mu_{1,t+1}, \mu_{2,t+1}) \overset{\text{def.}}{=} \underset{(\mu_1,\mu_2)}{\operatorname{argmin}} W_\alpha^\alpha(\mu_{1,t}, \mu_1) + W_\alpha^\alpha(\mu_{2,t}, \mu_2) + \tau f(\mu_1, \mu_2)$$

Multiple-Density Gradient Flows

$$(\mu_{1,t+1}, \mu_{2,t+1}) \overset{\text{def.}}{=} \underset{(\mu_1,\mu_2)}{\operatorname{argmin}} W_\alpha^\alpha(\mu_{1,t}, \mu_1) + W_\alpha^\alpha(\mu_{2,t}, \mu_2) + \tau f(\mu_1, \mu_2)$$

Wasserstein attraction:

$$f(\mu_1, \mu_2) = W_\alpha^\alpha(\mu_1, \mu_2) + h_1(\mu_1) + h_2(\mu_2)$$



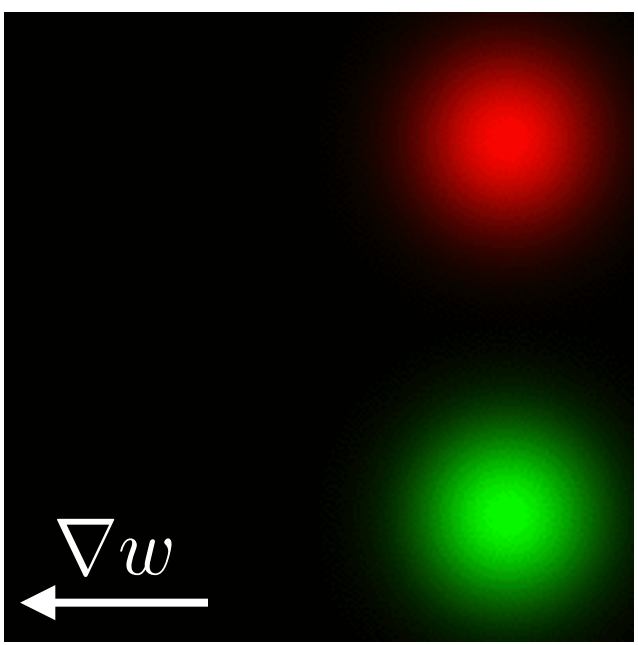
$$h_i(\mu) = \langle w, \mu \rangle$$

Multiple-Density Gradient Flows

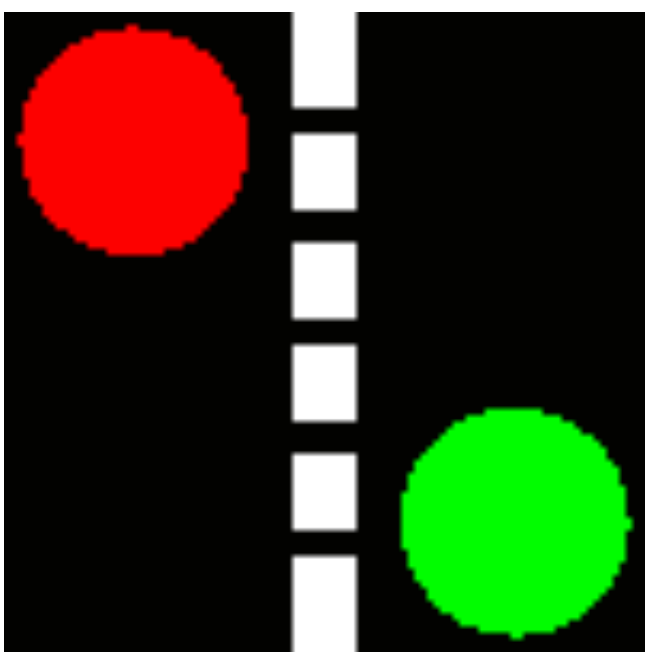
$$(\mu_{1,t+1}, \mu_{2,t+1}) \overset{\text{def.}}{=} \underset{(\mu_1,\mu_2)}{\operatorname{argmin}} W_\alpha^\alpha(\mu_{1,t}, \mu_1) + W_\alpha^\alpha(\mu_{2,t}, \mu_2) + \tau f(\mu_1, \mu_2)$$

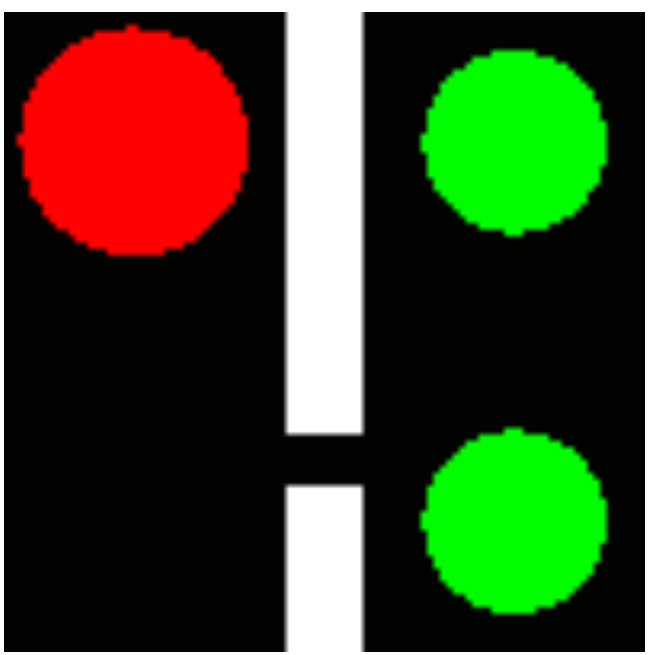
Wasserstein attraction:

$$f(\mu_1, \mu_2) = W_\alpha^\alpha(\mu_1, \mu_2) + h_1(\mu_1) + h_2(\mu_2)$$



$h_i(\mu) = \langle w, \mu \rangle$



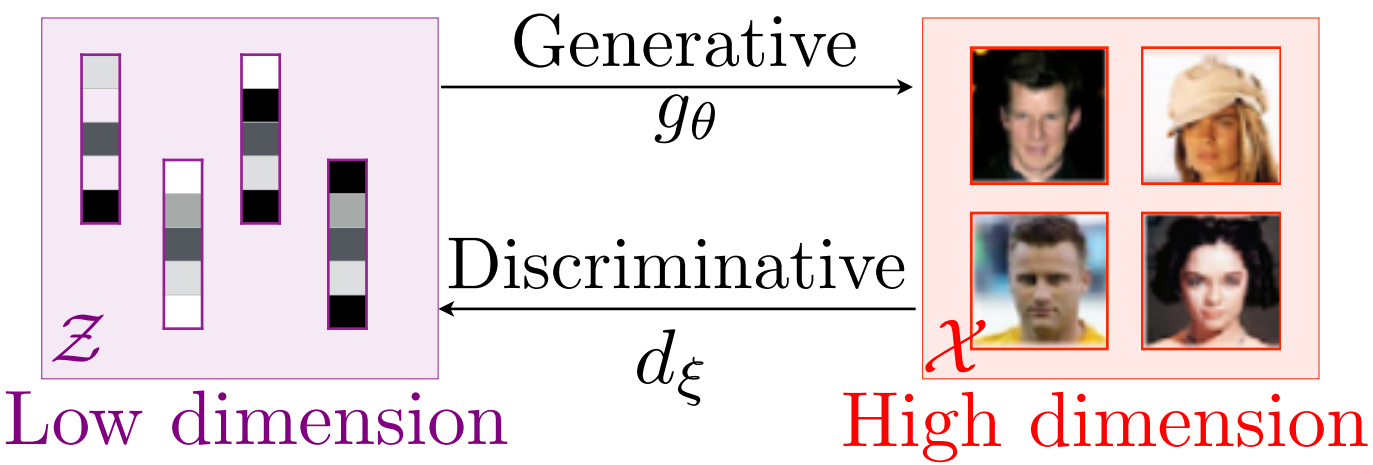


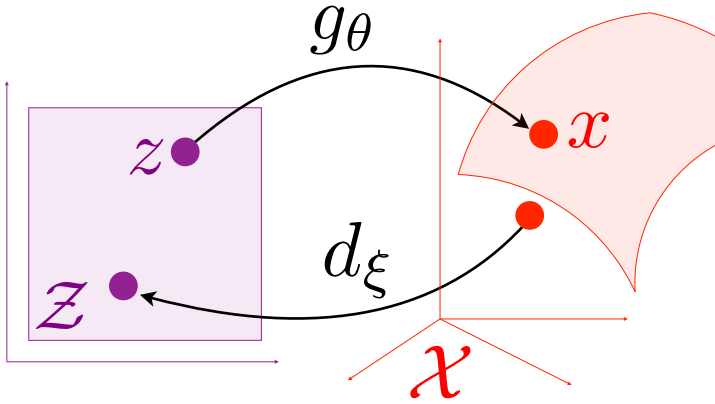
$h_i(\mu) = \iota_{[0,\kappa]}(\mu).$

Overview

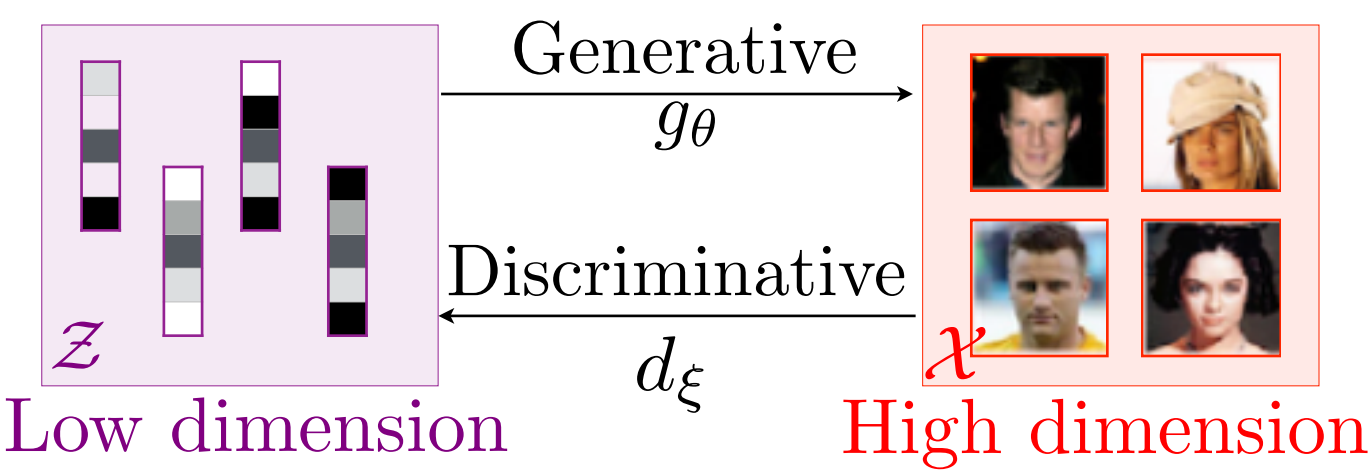
- Measures and Histograms
- From Monge to Kantorovitch Formulations
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- **Minimum Kantorovitch Estimators**
- Gromov-Wasserstein

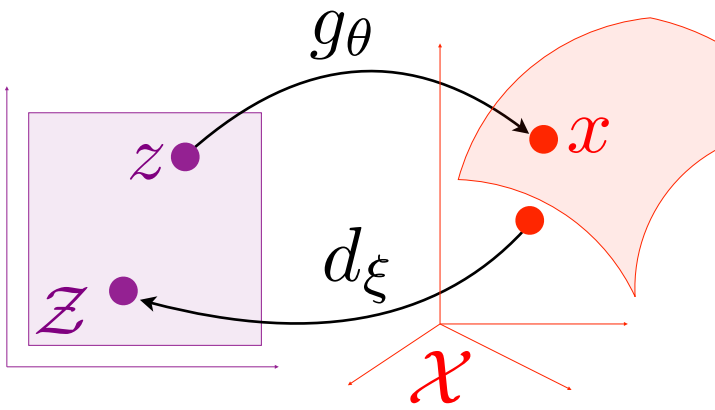
Discriminative vs Generative Models





Discriminative vs Generative Models

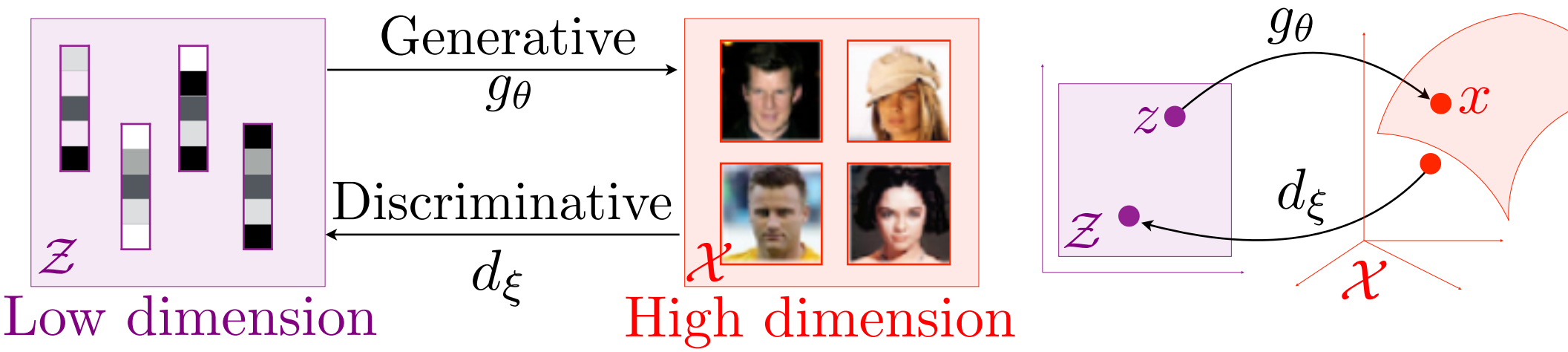




Supervised: classification, z =class probability

$\rightarrow$ Learn d_ξ from labeled data $(x_i, z_i)_i$.

Discriminative vs Generative Models



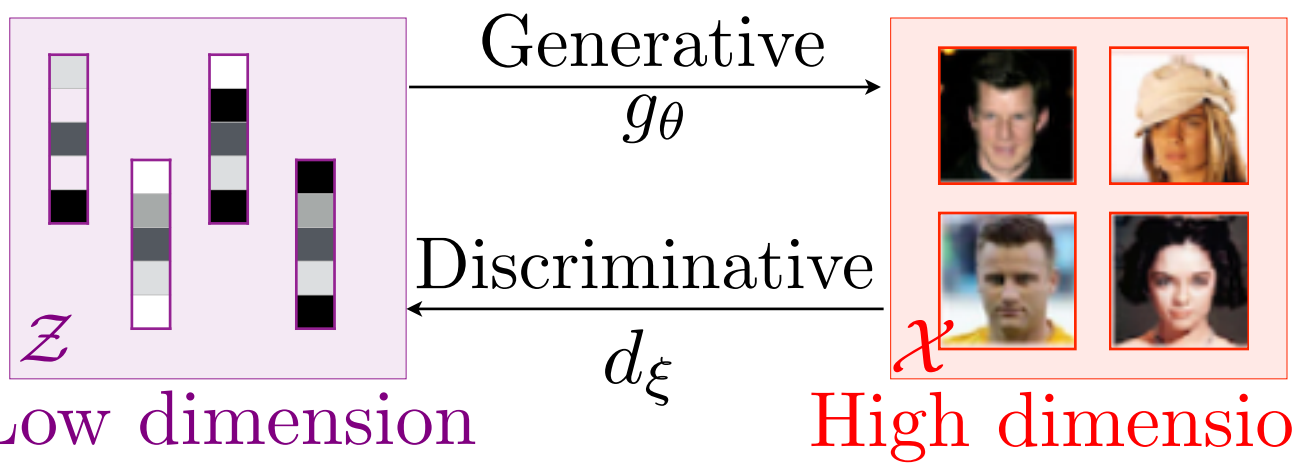
Supervised: classification, z =class probability

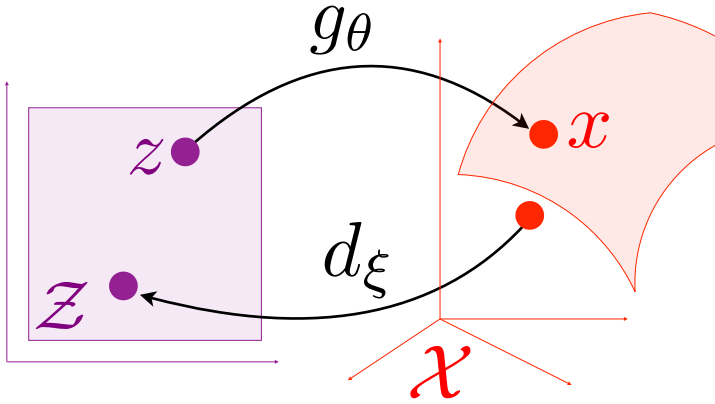
$\rightarrow$ Learn d_ξ from labeled data $(x_i, z_i)_i$.

Un-supervised: Compression: $z = d_\xi(x)$ is a representation.
Generation: $x = g_\theta(z)$ is a synthesis.

$\rightarrow$ Learn (g_θ, d_ξ) from data $(x_i)_i$.

Discriminative vs Generative Models





Supervised: classification, z =class probability

$\rightarrow$ Learn d_ξ from labeled data $(x_i, z_i)_i$.

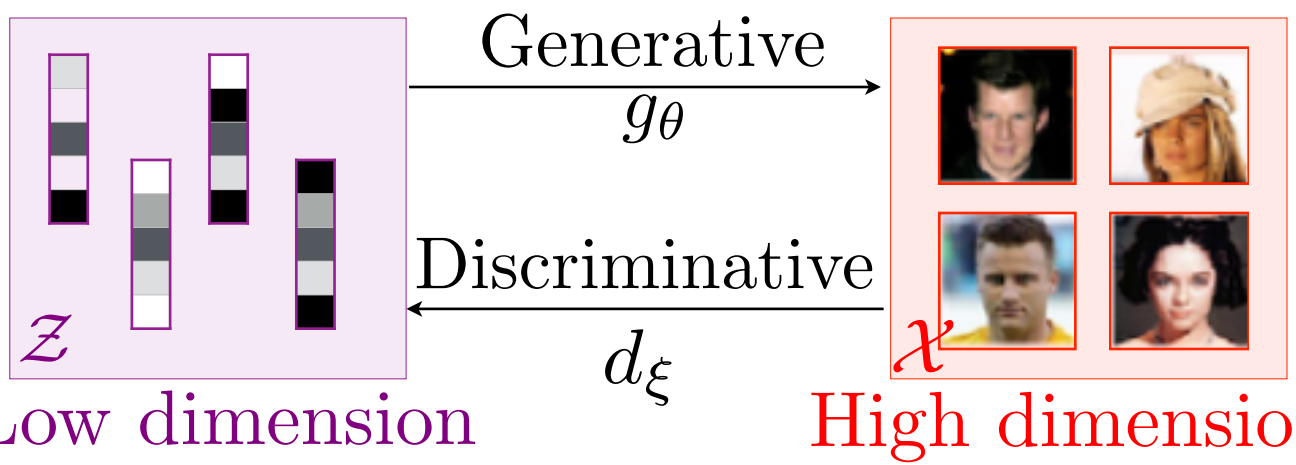
Un-supervised: Compression: $z = d_\xi(x)$ is a representation.
Generation: $x = g_\theta(z)$ is a synthesis.

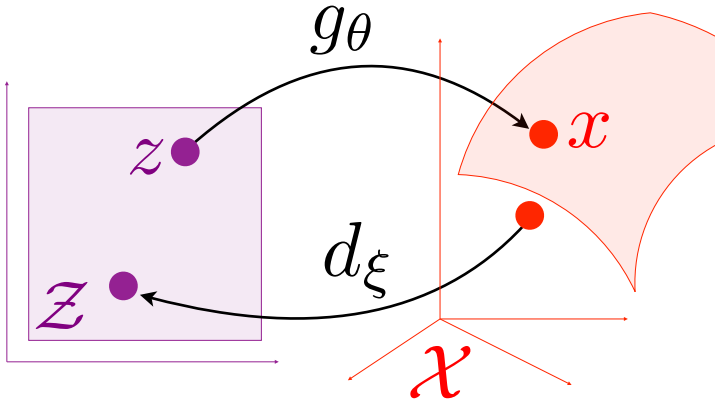
$\rightarrow$ Learn (g_θ, d_ξ) from data $(x_i)_i$.

Density fitting
$g_\theta(\{z_i\}_i) \approx \{x_i\}_i$

Auto-encoders
$g_\theta(d_\xi(x_i)) \approx x_i$

Discriminative vs Generative Models





Supervised: classification, z =class probability

$\rightarrow$ Learn d_ξ from labeled data $(x_i, z_i)_i$.

Un-supervised: Compression: $z = d_\xi(x)$ is a representation.
Generation: $x = g_\theta(z)$ is a synthesis.

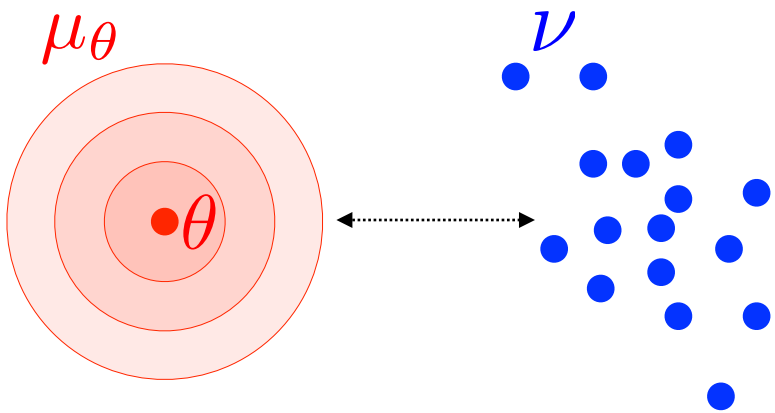
$\rightarrow$ Learn (g_θ, d_ξ) from data $(x_i)_i$.

Density fitting $g_\theta(\{z_i\}_i) \approx \{x_i\}_i$	Optimal transport map d_ξ	Auto-encoders $g_\theta(d_\xi(x_i)) \approx x_i$

Density Fitting and Generative Models

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

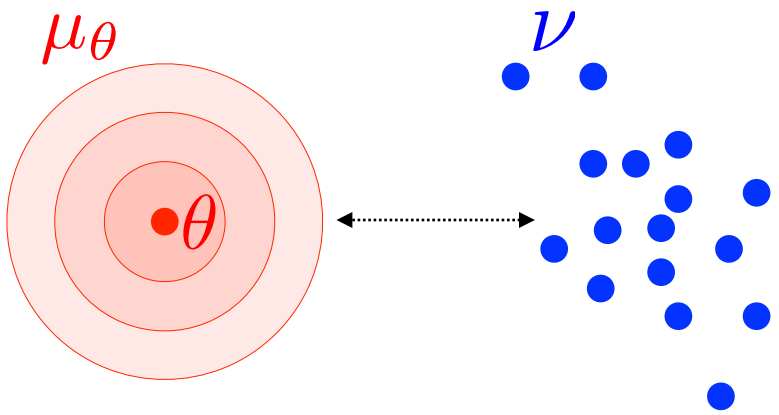
Parametric model: $\theta \mapsto \mu_\theta$



Density Fitting and Generative Models

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

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



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

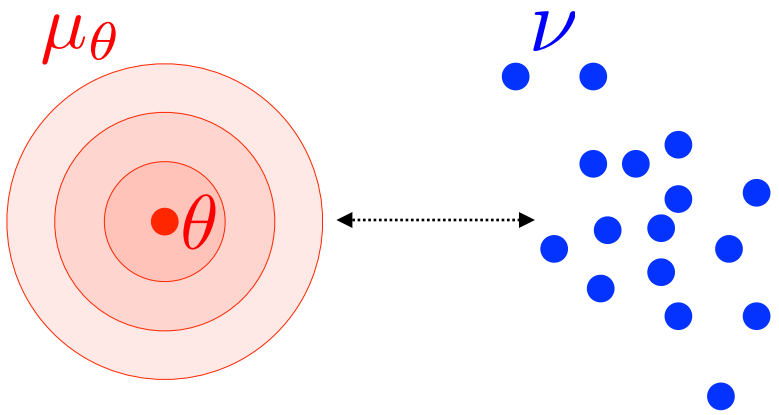
$$\min_\theta \widehat{\mathrm{KL}}(\nu|\mu_\theta) \overset{\text{def.}}{=} -\sum_j \log(f_\theta(y_j))$$

Maximum likelihood (MLE)

Density Fitting and Generative Models

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

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



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

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

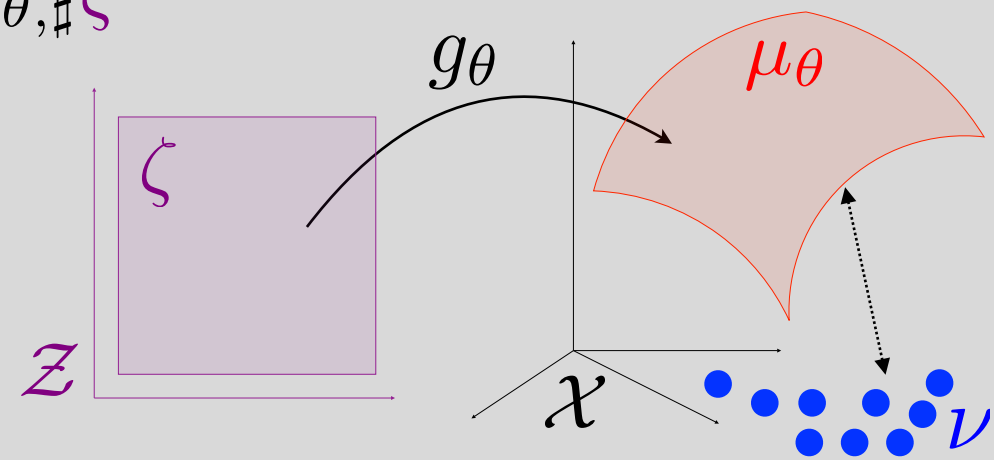
Maximum likelihood (MLE)

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

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

$\rightarrow$ MLE undefined.

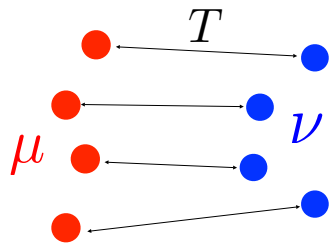
$\rightarrow$ Need a weaker metric.



Loss Functions for Measures

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

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



Optimal Transport Distances

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

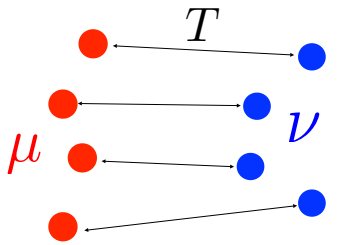
Loss Functions for Measures

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

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

Optimal Transport Distances

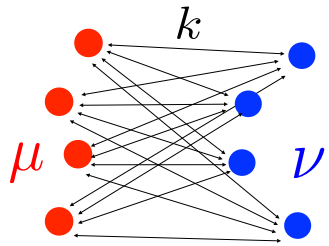
$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j}\|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

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

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

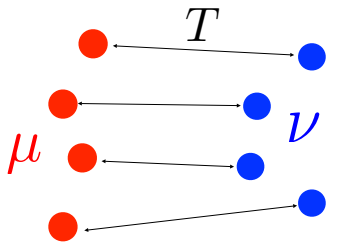


Loss Functions for Measures

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

Optimal Transport Distances

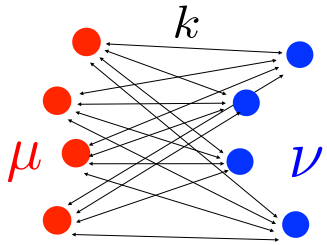
$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j}\|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

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

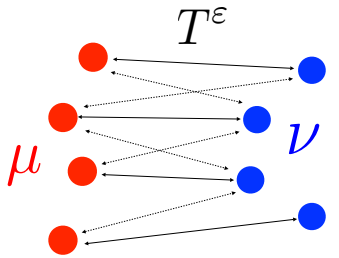
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Cuturi 13]

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

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

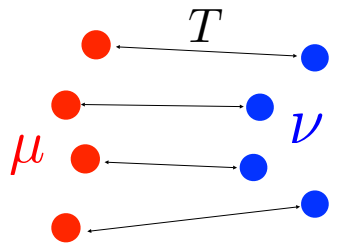


Loss Functions for Measures

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

Optimal Transport Distances

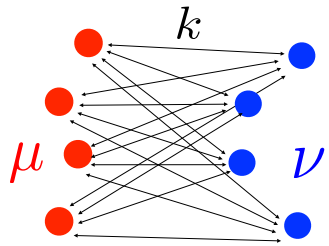
$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

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

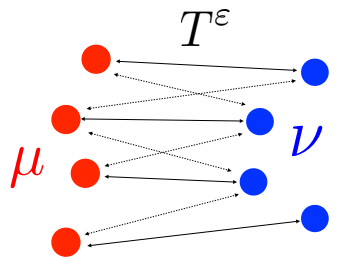
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Cuturi 13]

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

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



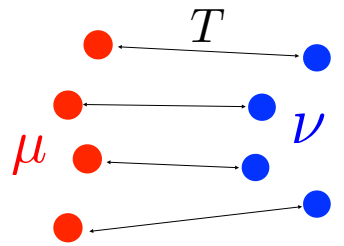
Theorem: [Ramdas, G.Trillos, Cuturi 17] $\quad \bar{W}_\varepsilon(\mu,\nu)^p \overset{\varepsilon\to 0}{\underset{\varepsilon\to+\infty}{\longrightarrow}} \begin{matrix} W(\mu,\nu)^p \\ \|\mu-\nu\|_k^2 \end{matrix}$ for $k(x,y) = -\|x-y\|^p$

Loss Functions for Measures

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

Optimal Transport Distances

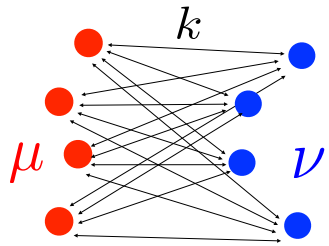
$$W(\mu,\nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu,\nu}} \sum_{i,j} T_{i,j}\|x_i - y_j\|^p$$



Maximum Mean Discrepancy (MMD)

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

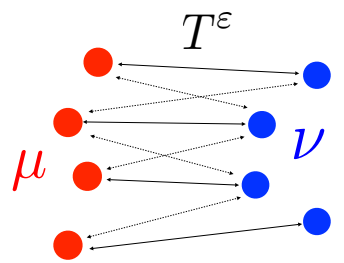
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x,y) = -\|x-y\|^2$.



Sinkhorn divergences [Cuturi 13]

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

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



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

Best of both worlds:

$\rightarrow$ cross-validate ε

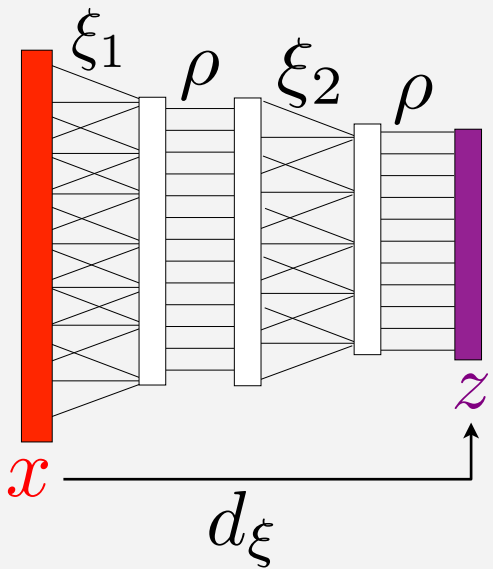
– Scale free (no σ, no heavy tail kernel).
– Non-Euclidean, arbitrary ground distance.
– Less biased gradient.
– No curse of dimension (low sample complexity).

Deep Discriminative vs Generative Models

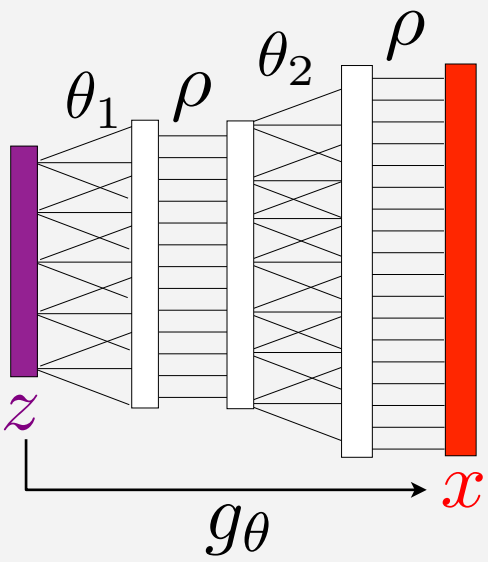
Deep networks:

$$d_\xi(x) = \rho(\xi_K(\dots\rho(\xi_2(\rho(\xi_1(x)\dots)$$
$$g_\theta(z) = \rho(\theta_K(\dots\rho(\theta_2(\rho(\theta_1(z)\dots)$$

Discriminative



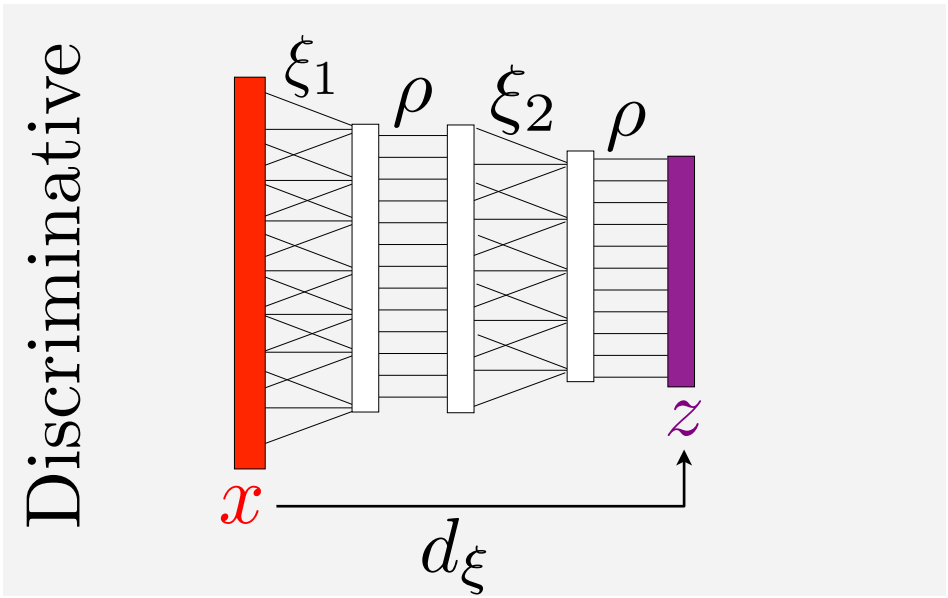
Generative

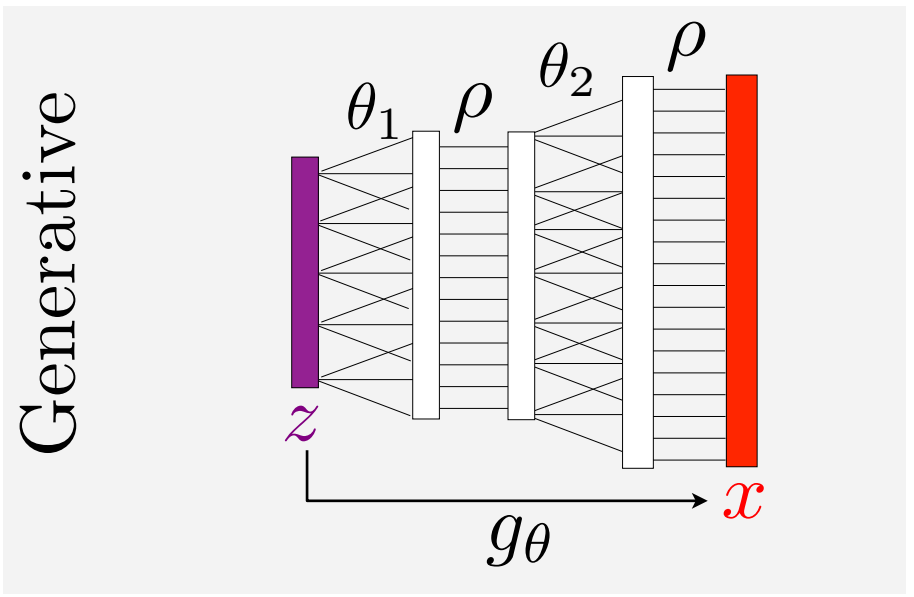


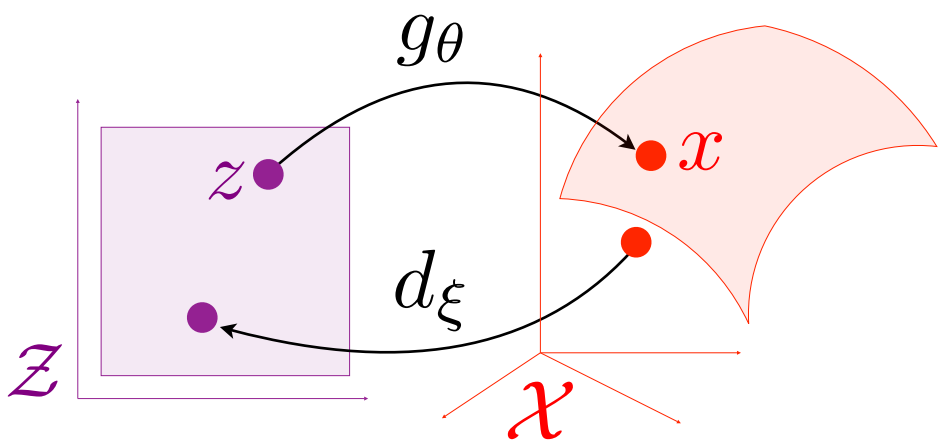
Deep Discriminative vs Generative Models

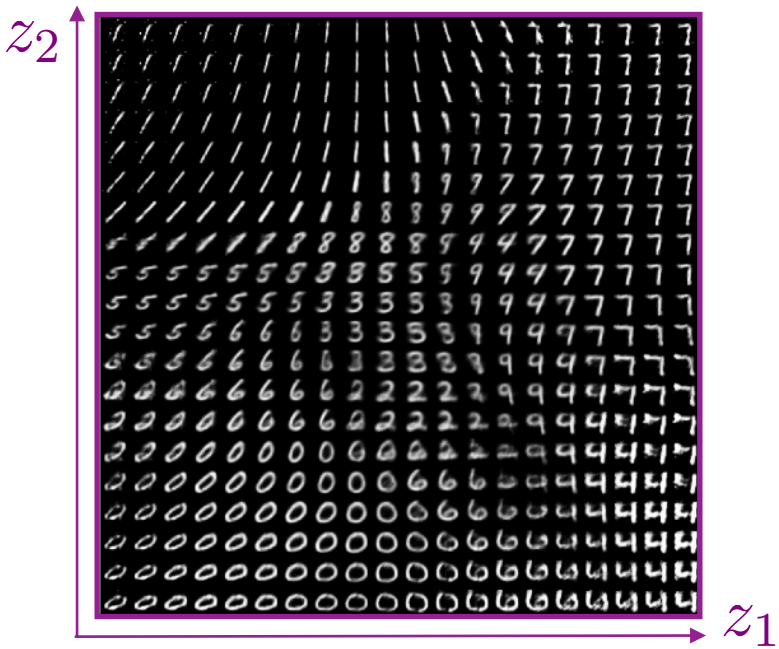
Deep networks:

$$d_\xi(x) = \rho(\xi_K(\dots\rho(\xi_2(\rho(\xi_1(x)\dots)$$
$$g_\theta(z) = \rho(\theta_K(\dots\rho(\theta_2(\rho(\theta_1(z)\dots)$$

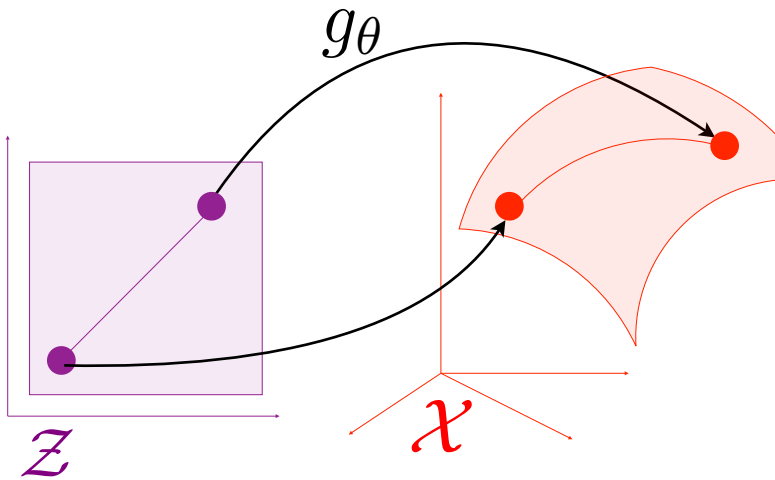








Examples of Image Generation



[Credit ArXiv:1511.06434]

Overview

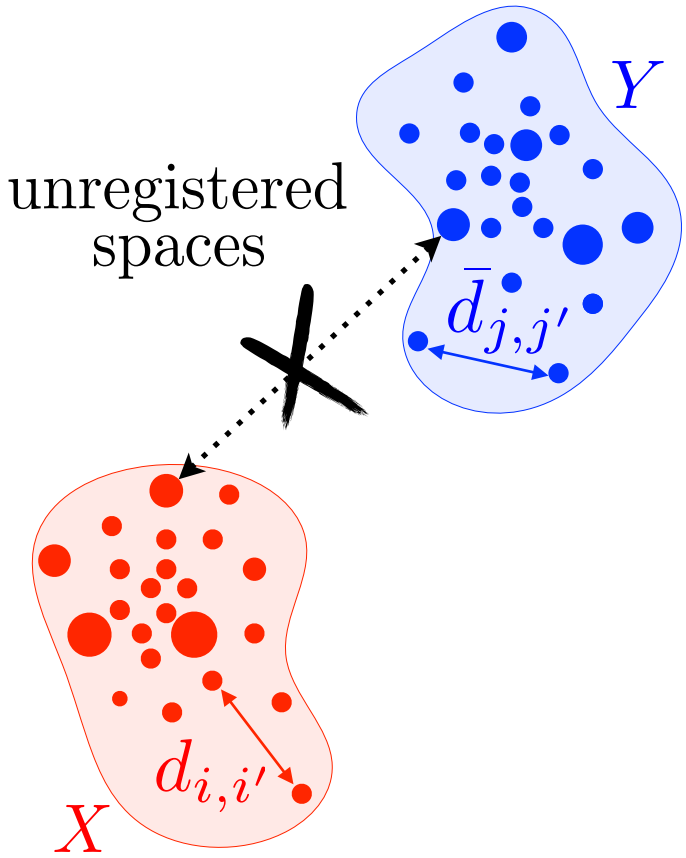
- Measures and Histograms
- From Monge to Kantorovitch Formulations
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- **Gromov-Wasserstein**

Gromov-Wasserstein

Inputs: {(similarity/kernel matrix, histogram)}

(d, μ) $\quad \mu = \sum_i \mu_i \delta_{x_i}$ $\quad d_{i,i'} = d(x_i, x_{i'})$

$(\bar{d}, \nu)$ $\quad \nu = \sum_j \nu_j \delta_{y_j}$ $\quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})$



Gromov-Wasserstein

Inputs: {(similarity/kernel matrix, histogram)}

(d, μ) $\quad \mu = \sum_i \mu_i \delta_{x_i}$ $\quad d_{i,i'} = d(x_i, x_{i'})$

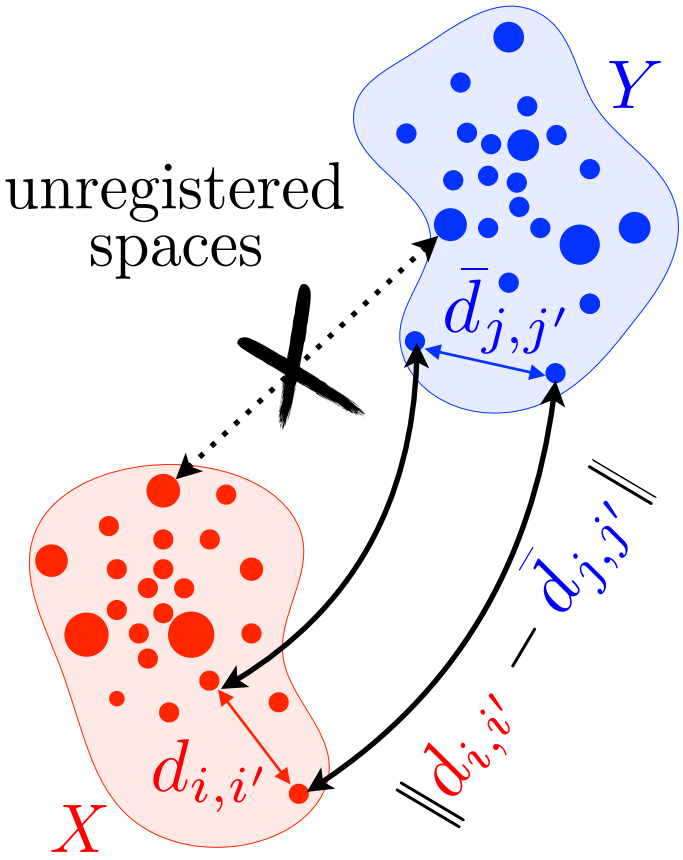
$(\bar{d}, \nu)$ $\quad \nu = \sum_j \nu_j \delta_{y_j}$ $\quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})$

Def. *Gromov-Wasserstein distance:*

$$\mathrm{GW}_p^p(d, \mu, \bar{d}, \nu) \overset{\text{def.}}{=} \min_{T \in C_{\mu,\nu}} \mathcal{E}^p_{d,\bar{d}}(T)$$

$$\mathcal{E}^p_{d,\bar{d}}(T) \overset{\text{def.}}{=} \sum_{i,i',j,j'} |d_{i,i'} - \bar{d}_{j,j'}|^p \, T_{i,j} T_{i',j'}$$

[Memoli 2011] [Sturm 2012]



Gromov-Wasserstein

Inputs: {(similarity/kernel matrix, histogram)}

(d, μ) $\quad \mu = \sum_i \mu_i \delta_{x_i}$ $\quad d_{i,i'} = d(x_i, x_{i'})$

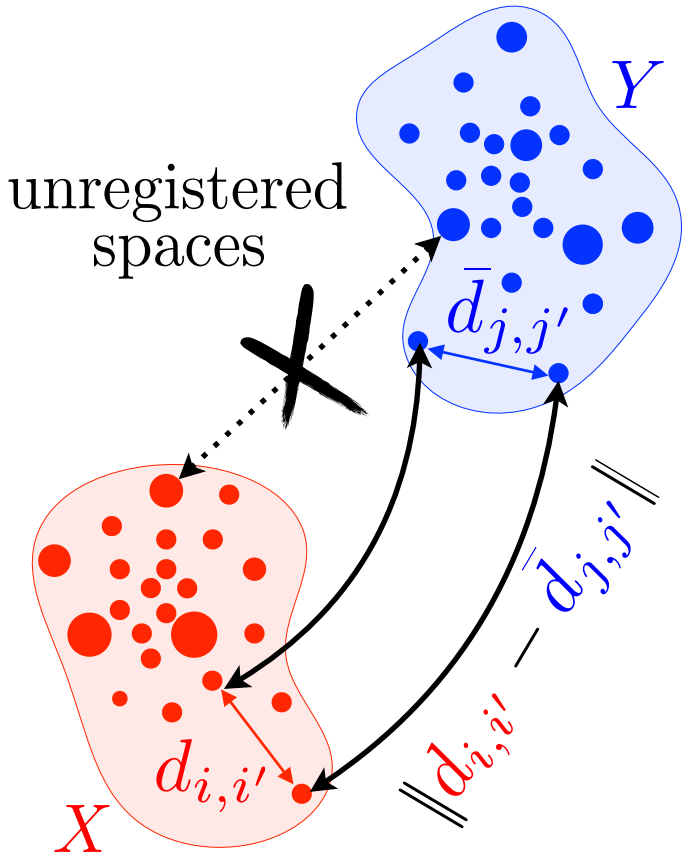
$(\bar{d}, \nu)$ $\quad \nu = \sum_j \nu_j \delta_{y_j}$ $\quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})$

Def. *Gromov-Wasserstein distance:*

$$\mathrm{GW}_p^p(d, \mu, \bar{d}, \nu) \overset{\text{def.}}{=} \min_{T \in C_{\mu,\nu}} \mathcal{E}^p_{d,\bar{d}}(T)$$

$$\mathcal{E}^p_{d,\bar{d}}(T) \overset{\text{def.}}{=} \sum_{i,i',j,j'} |d_{i,i'} - \bar{d}_{j,j'}|^p \, T_{i,j} T_{i',j'}$$

[Memoli 2011] [Sturm 2012]



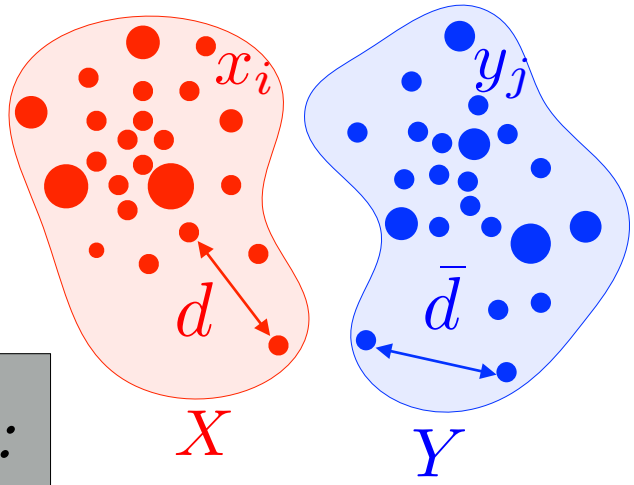
Computation of GW is a QAP:

$\rightarrow$ NP-hard in general.

$\rightarrow$ need for a fast approximate solver.

Gromov-Wasserstein as a Metric

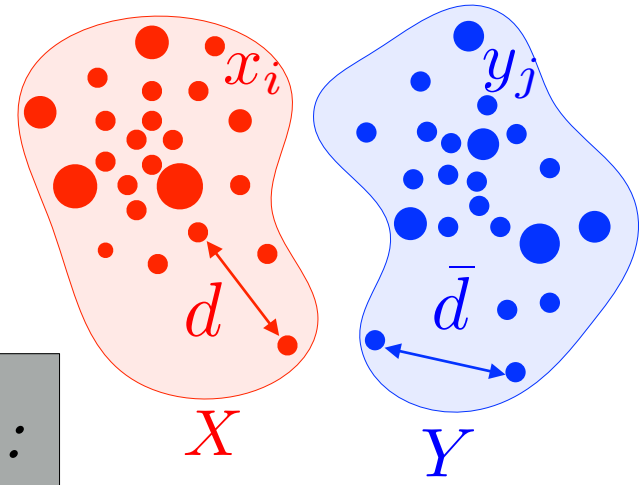
$$\mu = \sum_i \mu_i \delta_{x_i} \in \mathcal{M}_+^1(X) \qquad d_{i,i'} = d(x_i, x_{i'})$$
$$\nu = \sum_j \nu_j \delta_{y_j} \in \mathcal{M}_+^1(Y) \qquad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})$$



Def. *Metric-measured spaces* $(X, \mu, d) \in \mathbb{M}$*:*
$\mu \in \mathcal{M}_+^1(X)$ and d is a distance on X

Gromov-Wasserstein as a Metric

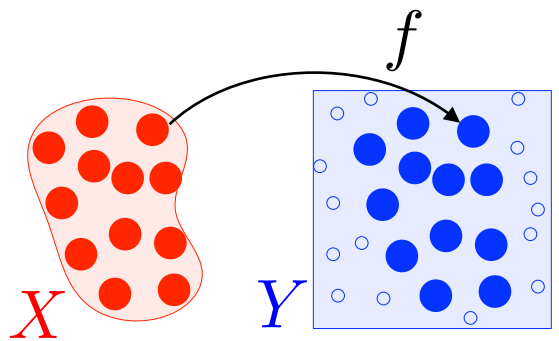
$$\mu = \sum_i \mu_i \delta_{x_i} \in \mathcal{M}_+^1(X) \qquad d_{i,i'} = d(x_i, x_{i'})$$
$$\nu = \sum_j \nu_j \delta_{y_j} \in \mathcal{M}_+^1(Y) \qquad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})$$

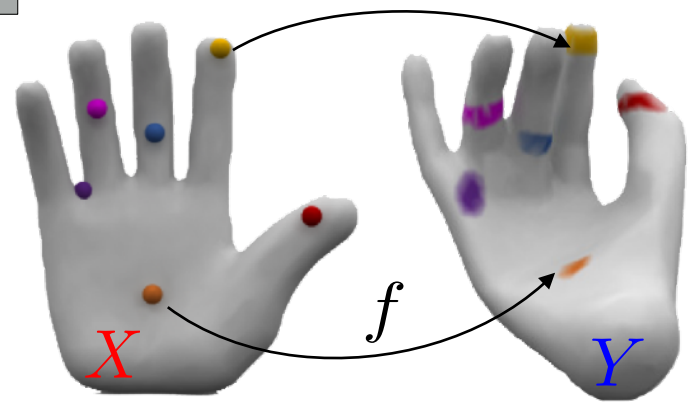


Def. *Metric-measured spaces* $(X, \mu, d) \in \mathbb{M}$:
$\mu \in \mathcal{M}_+^1(X)$ and d is a distance on X

Def. *Isometries on* $\mathbb{M}$: $(\mu, d) \sim (\nu, \bar{d})$

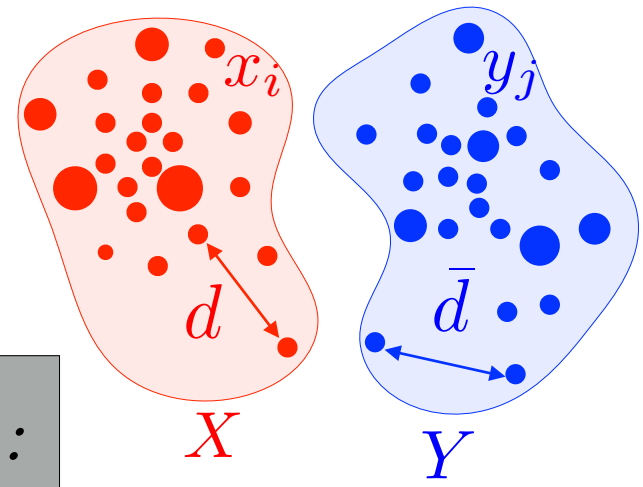
$$\iff \exists f : X \to Y, \begin{cases} f_\sharp \mu = \nu, \\ d(x, x') = \bar{d}(f(x), f(x')). \end{cases}$$





Gromov-Wasserstein as a Metric

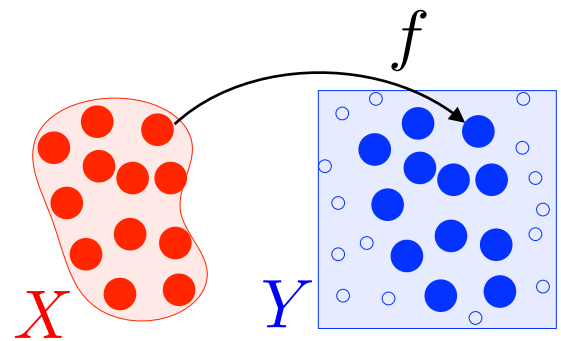
$$\mu = \sum_i \mu_i \delta_{x_i} \in \mathcal{M}_+^1(X) \qquad d_{i,i'} = d(x_i, x_{i'})$$
$$\nu = \sum_j \nu_j \delta_{y_j} \in \mathcal{M}_+^1(Y) \qquad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})$$



Def. *Metric-measured spaces* $(X, \mu, d) \in \mathbb{M}$:
$\mu \in \mathcal{M}_+^1(X)$ and d is a distance on X

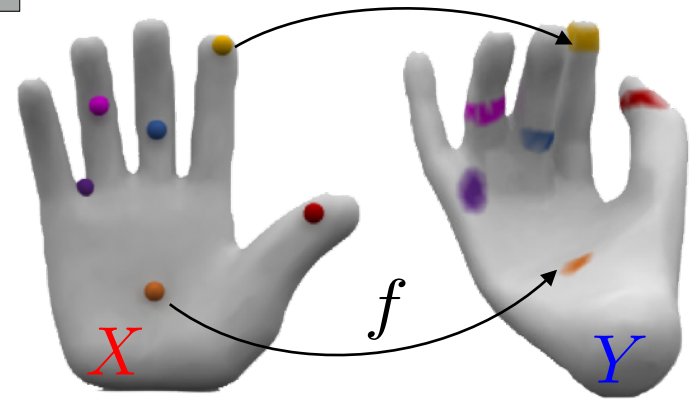
Def. *Isometries on* $\mathbb{M}$: $(\mu, d) \sim (\nu, \bar{d})$

$$\iff \exists f : X \to Y, \begin{cases} f_\sharp \mu = \nu, \\ d(x, x') = \bar{d}(f(x), f(x')). \end{cases}$$



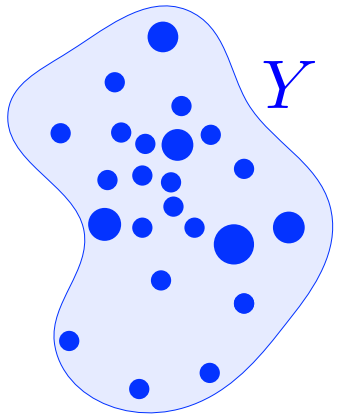
Prop. GW defines a distance on $\mathbb{M}/\sim$.
[Memoli 2011]

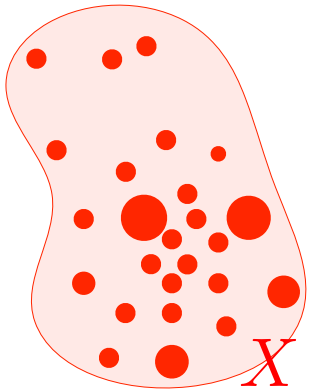
$\longrightarrow$ "bending-invariant" objects recognition.



For Arbitrary Spaces

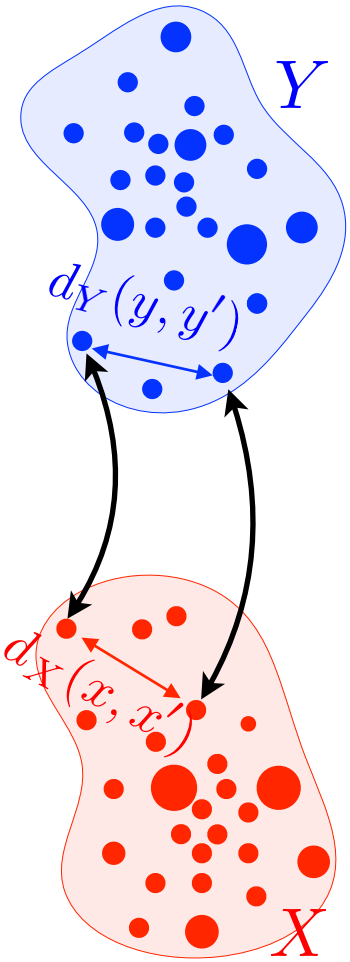
Metric-measure spaces (X, Y): $(d_X, \mu), (d_Y, \nu)$





For Arbitrary Spaces

Metric-measure spaces (X, Y): $(d_X, \mu), (d_Y, \nu)$



Def. *Gromov-Wasserstein distance:*

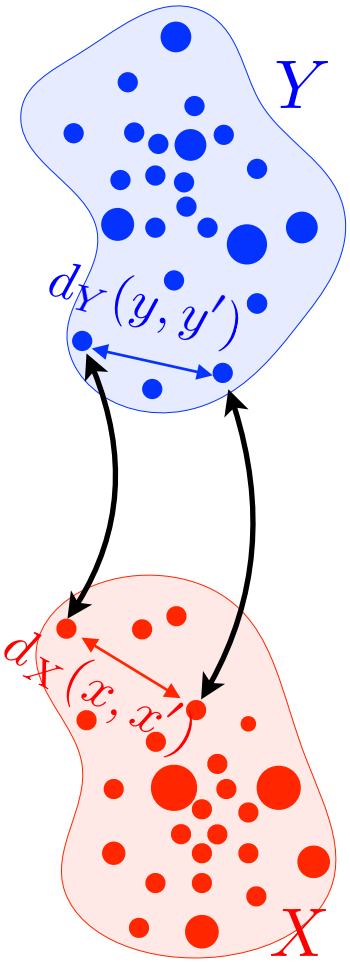
$$\mathrm{GW}_2^2(d_X, \mu, d_Y, \nu) \overset{\text{def.}}{=}$$

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X^2 \times Y^2} |d_X(x, x') - d_Y(y, y')|^2 \mathrm{d}\pi(x, y)\mathrm{d}\pi(x', y')$$

[Sturm 2012] [Memoli 2011]

For Arbitrary Spaces

Metric-measure spaces (X, Y): $(d_X, \mu), (d_Y, \nu)$



Def. *Gromov-Wasserstein distance:*

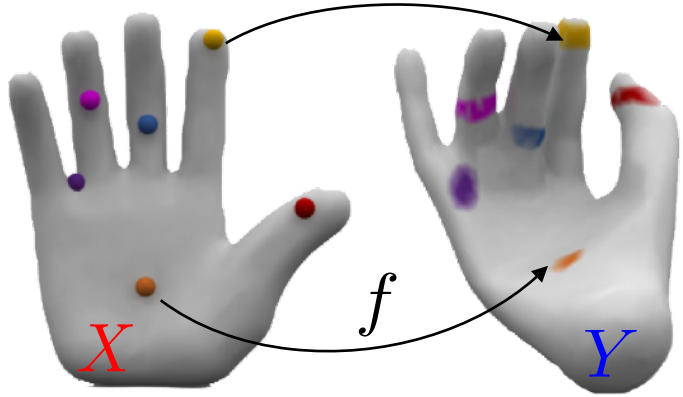
$$\mathrm{GW}_2^2(d_X, \mu, d_Y, \nu) \overset{\text{def.}}{=}$$

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X^2 \times Y^2} |d_X(x, x') - d_Y(y, y')|^2 \mathrm{d}\pi(x, y)\mathrm{d}\pi(x', y')$$

[Sturm 2012] [Memoli 2011]

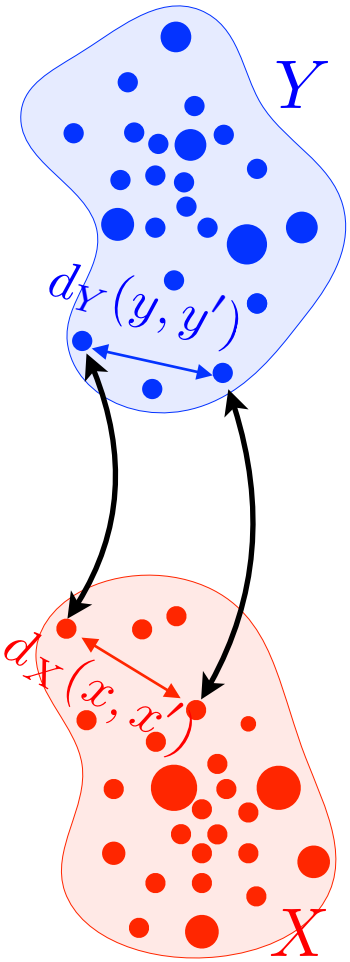
Prop. GW is a distance on mm-spaces/isometries.

$\rightarrow$ "bending-invariant" objects recognition.



For Arbitrary Spaces

Metric-measure spaces (X, Y): $(d_X, \mu), (d_Y, \nu)$



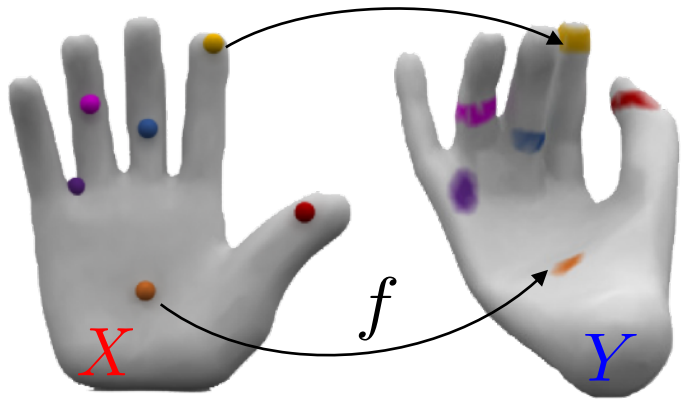
Def. *Gromov-Wasserstein distance:*

$$\mathrm{GW}_2^2(d_X, \mu, d_Y, \nu) \overset{\text{def.}}{=}$$

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X^2 \times Y^2} |d_X(x, x') - d_Y(y, y')|^2 \mathrm{d}\pi(x, y)\mathrm{d}\pi(x', y')$$

[Sturm 2012] [Memoli 2011]

Prop. GW is a distance on mm-spaces/isometries.



$\rightarrow$ "bending-invariant" objects recognition.

$\rightarrow$ QAP: NP-hard in general.

$\rightarrow$ need for a fast approximate solver.

Entropic Gromov Wasserstein

Def. *Entropic Gromov-Wasserstein*

$$\mathrm{GW}^p_{p,\varepsilon}(d,\mu,\bar{d},\nu) \overset{\text{def.}}{=} \min_{T\in C_{\mu,\nu}} \mathcal{E}^p_{d,\bar{d}}(T) - \varepsilon H(T)$$

Entropic Gromov Wasserstein

Def. *Entropic Gromov-Wasserstein*

$$\mathrm{GW}^p_{p,\varepsilon}(d,\mu,\bar{d},\nu) \overset{\text{def.}}{=} \min_{T\in C_{\mu,\nu}} \mathcal{E}^p_{d,\bar{d}}(T) - \varepsilon H(T)$$

Def. *Projected mirror descent:*

$$T \leftarrow \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}_{\mu,\nu}}\left(T \odot e^{-\tau(-\nabla\mathcal{E}^p_{d,\bar{d}}(T)-\varepsilon\nabla H(T))}\right)$$

where $\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}_{\mu,\nu}}(K) \overset{\text{def.}}{=} \mathrm{argmin}_T \{\mathrm{KL}(T|K)\ ;\ T\in\mathcal{C}_{\mu,\nu}\}$

Prop. for $\tau = 1/\varepsilon$, the iteration reads

$$T \leftarrow \texttt{Sinkhorn}(\mu,\nu,-d\times T\times \bar{d})$$

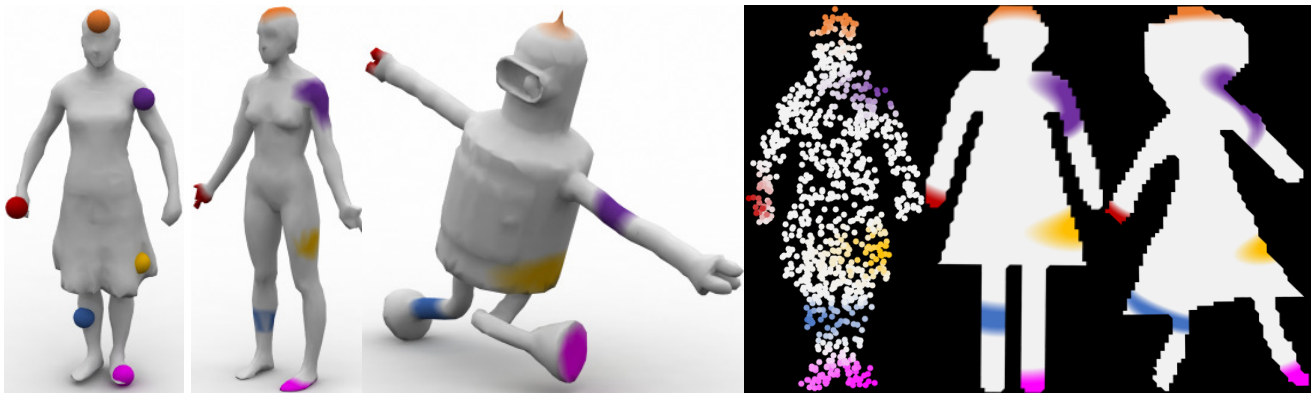
Prop. T converges to a stationary point.

$\texttt{func}\ T = \texttt{GW}(C,\bar{C},p,q)$
initialization:
$T \leftarrow \mu\nu^\top$
repeat:
$D \leftarrow -d\times T\times d$
$T \leftarrow \texttt{Sinkhorn}(\mu,\nu,D)$
until convergence.
return T

Applications of GW: Shapes Analysis

Use T to define registration between:

Shape ⟷ Shape



Source

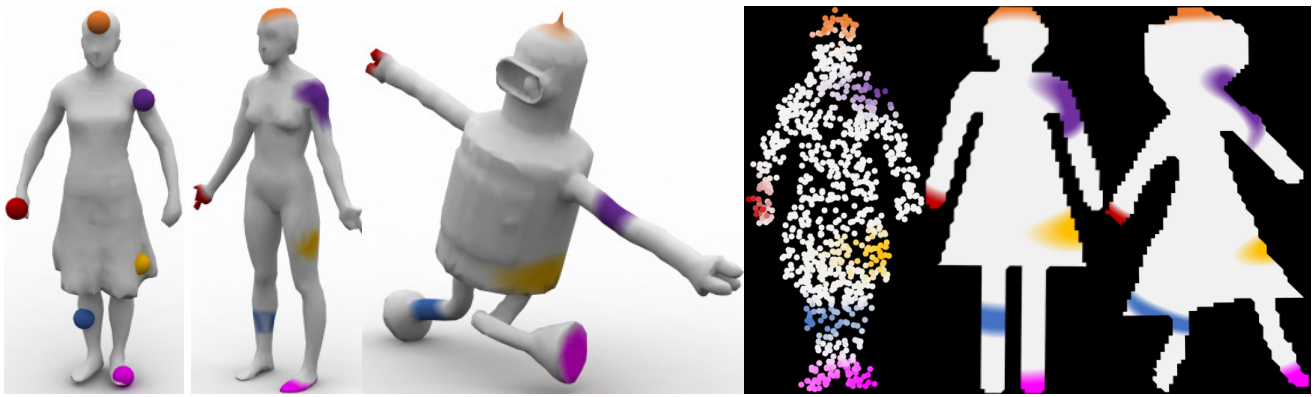
Targets

Colors distribution ⟷ Shape

Applications of GW: Shapes Analysis

Use T to define registration between:

Shape ⟷ Shape

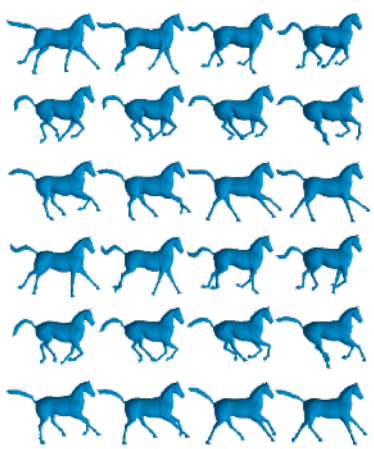


Source

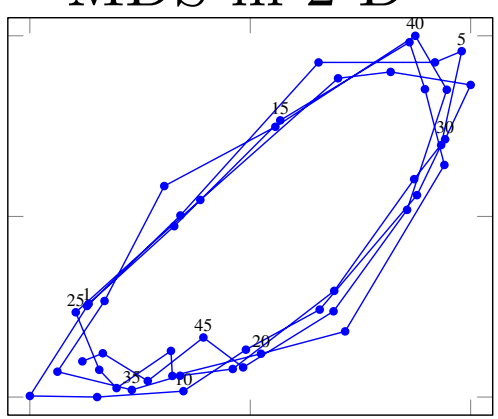
Targets

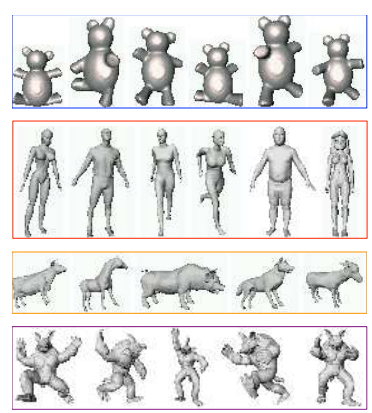
Colors distribution ⟷ Shape

Shapes $(X_s)_s$ → Geodesic distances $d_s = (D_{X_s}(x_i, x_{i'}))_{i,i'}$ → GW distances $(\mathrm{GW}_\varepsilon(d_s, d_{s'}))_{s,s'}$ → MDS Vizualization

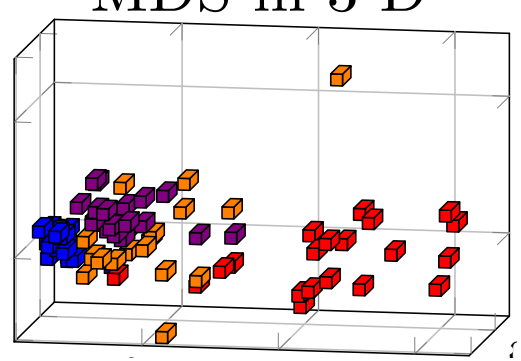


MDS in 2-D





MDS in 3-D

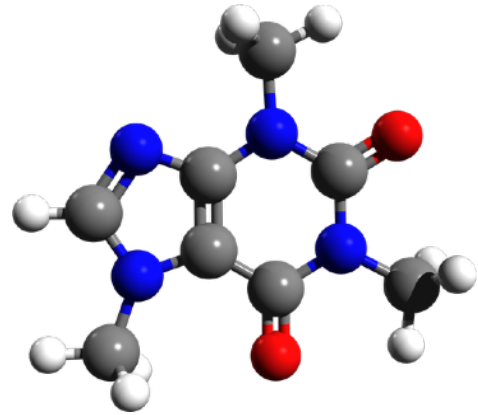


Applications of GW: Quantum Chemistry

Input: Molecules with positions and charges $\mu = \sum_i \mu_i \delta_{x_i}$.

Regression problem: approximate ground state energy $\mu \mapsto f(\mu)$.

$\rightarrow f$ by solving DFT approximation is too costly.



Applications of GW: Quantum Chemistry

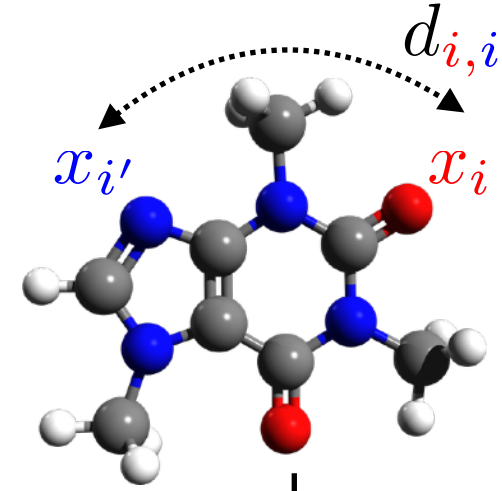
Input: Molecules with positions and charges $\mu = \sum_i \mu_i \delta_{x_i}$.

Regression problem: approximate ground state energy $\mu \mapsto f(\mu)$.

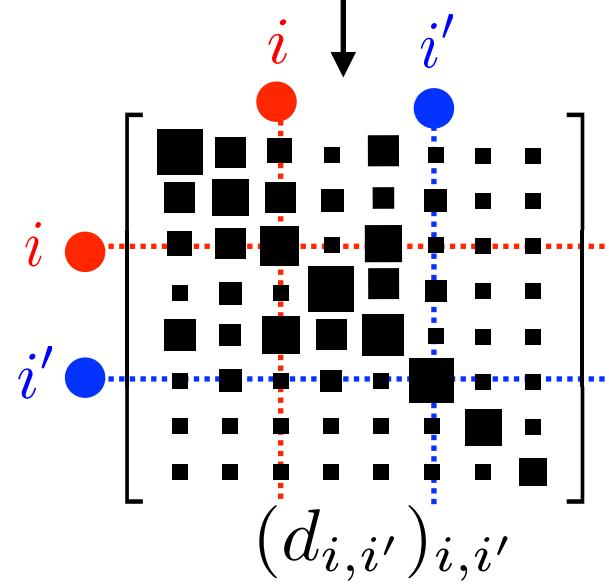
$\rightarrow f$ by solving DFT approximation is too costly.

Coulomb matrices $d = d(\mu)$:

$$d_{i,i'} \overset{\text{def.}}{=} \begin{cases} \frac{\mu_i \mu_{i'}}{\|x_i - x_{i'}\|} & \text{for } (i \neq i') \\ \frac{1}{2}\mu_i^{2.4} & \text{for } (i = i'). \end{cases}$$



[Rupp et al 2012]



Applications of GW: Quantum Chemistry

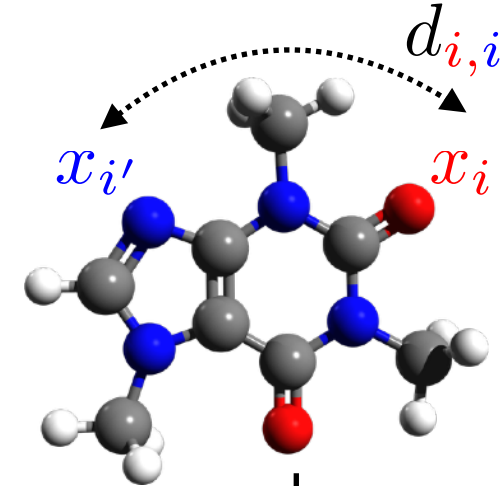
Input: Molecules with positions and charges $\mu = \sum_i \mu_i \delta_{x_i}$.

Regression problem: approximate ground state energy $\mu \mapsto f(\mu)$.

$\to f$ by solving DFT approximation is too costly.

Coulomb matrices $d = d(\mu)$:

$$d_{i,i'} \overset{\text{def.}}{=} \begin{cases} \frac{\mu_i \mu_{i'}}{\|x_i - x_{i'}\|} & \text{for } (i \neq i') \\ \frac{1}{2}\mu_i^{2.4} & \text{for } (i = i'). \end{cases}$$



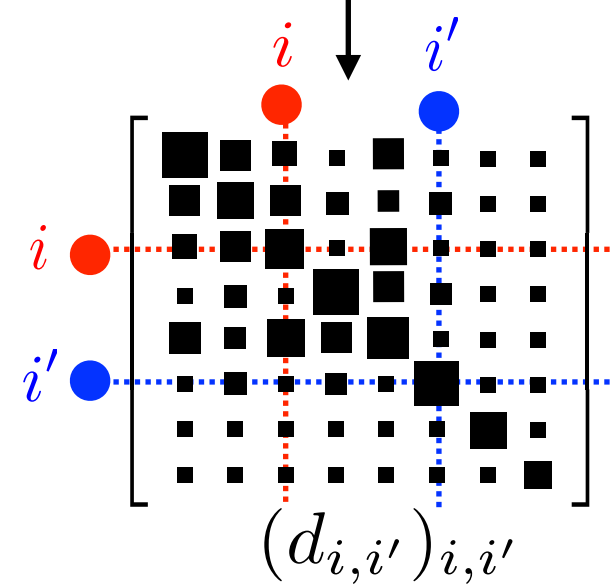
[Rupp et al 2012]

Learning: $(\mu_s, f(\mu_s))_s \to$ approximation $\tilde{f}$.

GW-interpolation: $\tilde{f}(\mu) = f(\mu_{s^\star})$

$s^\star = \operatorname{argmin}_s \mathrm{GW}(d(\mu), d(\mu_s))$

Algorithm	$\|f - \tilde{f}\|_1$
k-nearest neighbors	71.54
Linear regression	20.72
Gaussian kernel ridge regression	8.57
Laplacian kernel ridge regression (8)	3.07
Multilayer Neural Network (1000)	3.51
GW 3-nearest neighbors	10.83



Gromov-Wasserstein Geodesics

Def. *Gromov-Wasserstein Geodesic*

$$(\mu_t, d_t) \in \operatorname*{argmin}_{(\mu,d)\in\mathbb{X}} (1-t)\mathrm{GW}_2^2(\mu_0, d_0, \mu, d) + t\mathrm{GW}_2^2(\mu_1, d_1, \mu, d)$$

Gromov-Wasserstein Geodesics

Def. *Gromov-Wasserstein Geodesic*

$$(\mu_t, d_t) \in \operatorname*{argmin}_{(\mu,d)\in\mathbb{X}} (1-t)\mathrm{GW}_2^2(\mu_0, d_0, \mu, d) + t\mathrm{GW}_2^2(\mu_1, d_1, \mu, d)$$

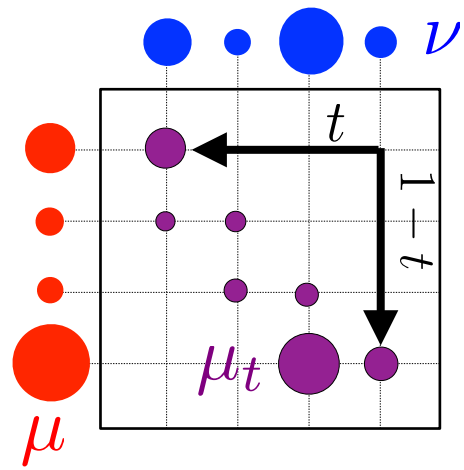
Optimal coupling $T^\star$: $\quad \mathrm{GW}_2^2(d_0, \mu_0, d_1, \mu_1) \overset{\text{def.}}{=} \mathcal{E}^2_{d_0,d_1}(T^\star)$

Prop. One can define (μ_t, d_t) on $X \times Y$ as

$$\mu_t = \sum_{i,j} T^\star_{i,j} \delta_{x_i,y_j}$$

$$d_t((x,y),(x',y')) = (1-t)d_0(x,x') + t d_1(y,y')$$

[Sturm 2012]



Gromov-Wasserstein Geodesics

Def. *Gromov-Wasserstein Geodesic*

$$(\mu_t, d_t) \in \underset{(\mu,d)\in\mathbb{X}}{\operatorname{argmin}} (1-t)\mathrm{GW}_2^2(\mu_0, d_0, \mu, d) + t\mathrm{GW}_2^2(\mu_1, d_1, \mu, d)$$

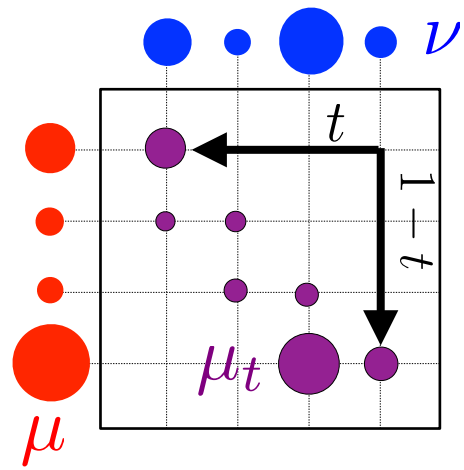
Optimal coupling $T^\star$: $\quad \mathrm{GW}_2^2(d_0, \mu_0, d_1, \mu_1) \overset{\text{def.}}{=} \mathcal{E}^2_{d_0,d_1}(T^\star)$

Prop. One can define (μ_t, d_t) on $X \times Y$ as

$$\mu_t = \textstyle\sum_{i,j} T^\star_{i,j} \delta_{x_i,y_j}$$

$$d_t((x,y),(x',y')) = (1-t)d_0(x,x') + t d_1(y,y')$$

[Sturm 2012]



$\rightarrow$ $X \times Y$ is not practical for most applications.
(need to fix the size of the geodesic embedding space)

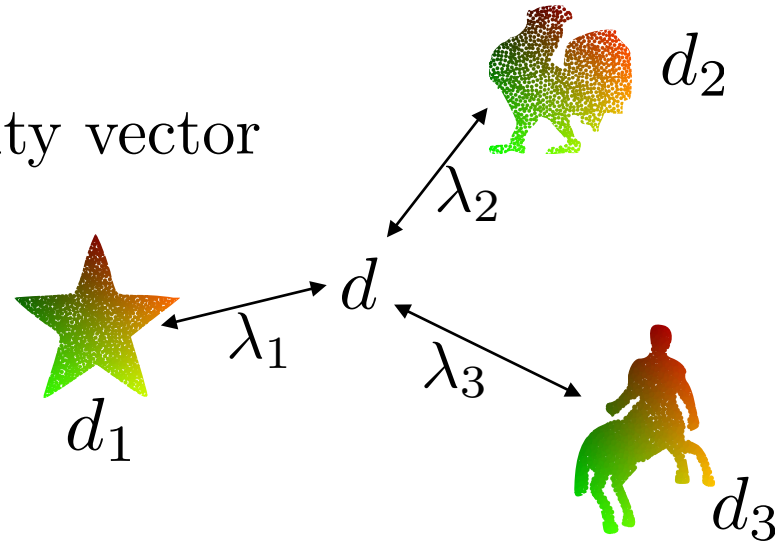
$\rightarrow$ Extension to more than 2 input spaces?

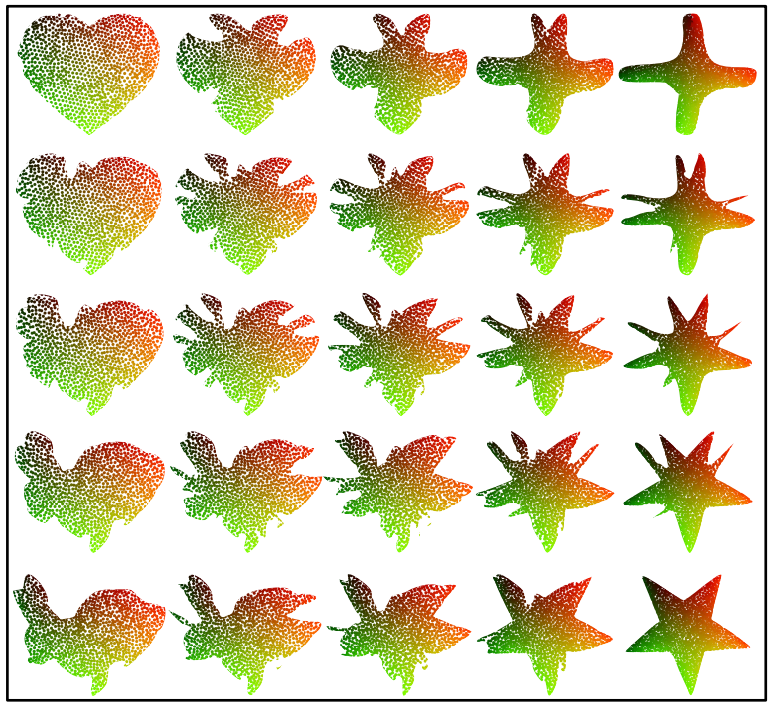
Gromov-Wasserstein Barycenters

Input: | Measures $(\mu_s)_s$, matrices $(d_s)_s$
Weights λ, size N, $\mu \in \mathbb{R}_+^N$ probability vector

Def. GW Barycenters

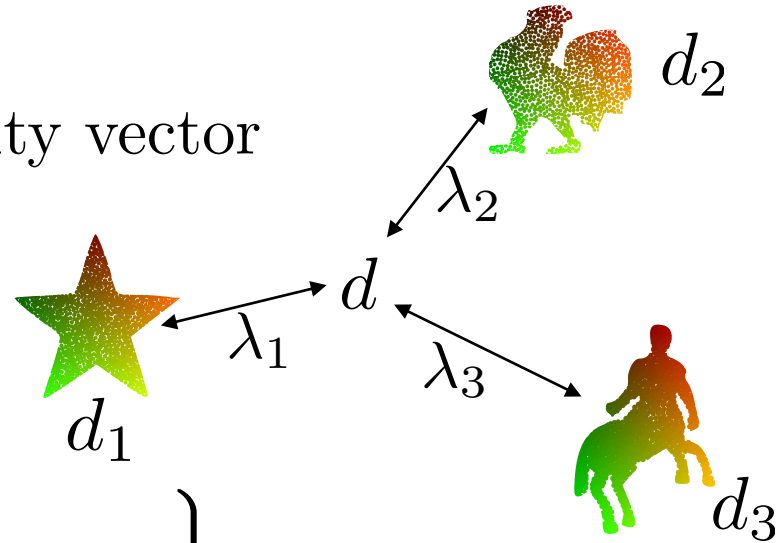
$$\min_{d \in \mathbb{R}^{N \times N}} \sum_s \lambda_s \mathrm{GW}_{2,\varepsilon}^2(d_s, \mu_s, d, \mu)$$





Gromov-Wasserstein Barycenters

Input: Measures $(\mu_s)_s$, matrices $(d_s)_s$
Weights λ, size N, $\mu \in \mathbb{R}_+^N$ probability vector



Def. GW Barycenters

$$\min_{d \in \mathbb{R}^{N \times N}} \sum_s \lambda_s \mathrm{GW}_{2,\varepsilon}^2(d_s, \mu_s, d, \mu)$$

$$\min_{d,(T_s)_s} \left\{ \sum_s \lambda_s \left(\mathcal{E}^2_{d,d_s}(T_s) - \varepsilon H(T_s) \right) \;;\; \forall\, s, T_s \in \mathcal{C}_{\mu,\mu_s} \right\}$$

Alternating minimization:

func $C = \texttt{GW-bary}(d_s, \mu_s, \mu)_s$

initialization: $C \leftarrow C_0$

repeat:

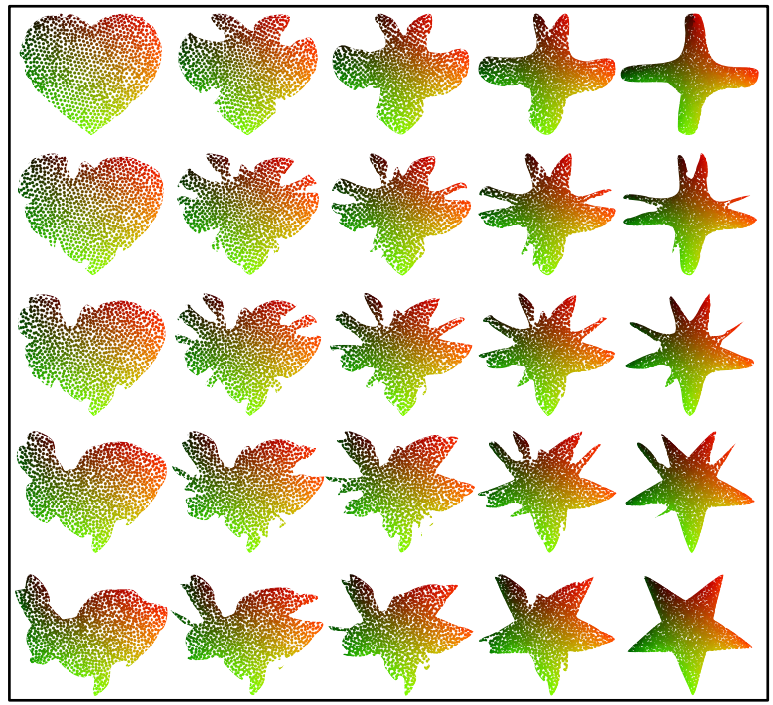
for $s = 1$ **to** S **do**

On T_s $\longrightarrow$ $T_s \leftarrow = \texttt{GW}(d, \mu, d_s, \mu_s)$

On d $\longrightarrow$ $d \leftarrow \dfrac{1}{\mu\mu^\top} \sum_s \lambda_s T_s^\top d_s T_s$

until convergence.

return C



Conclusion: Toward High-dimensional OT

Monge Kantorovich Dantzig Brenier Otto McCann Villani