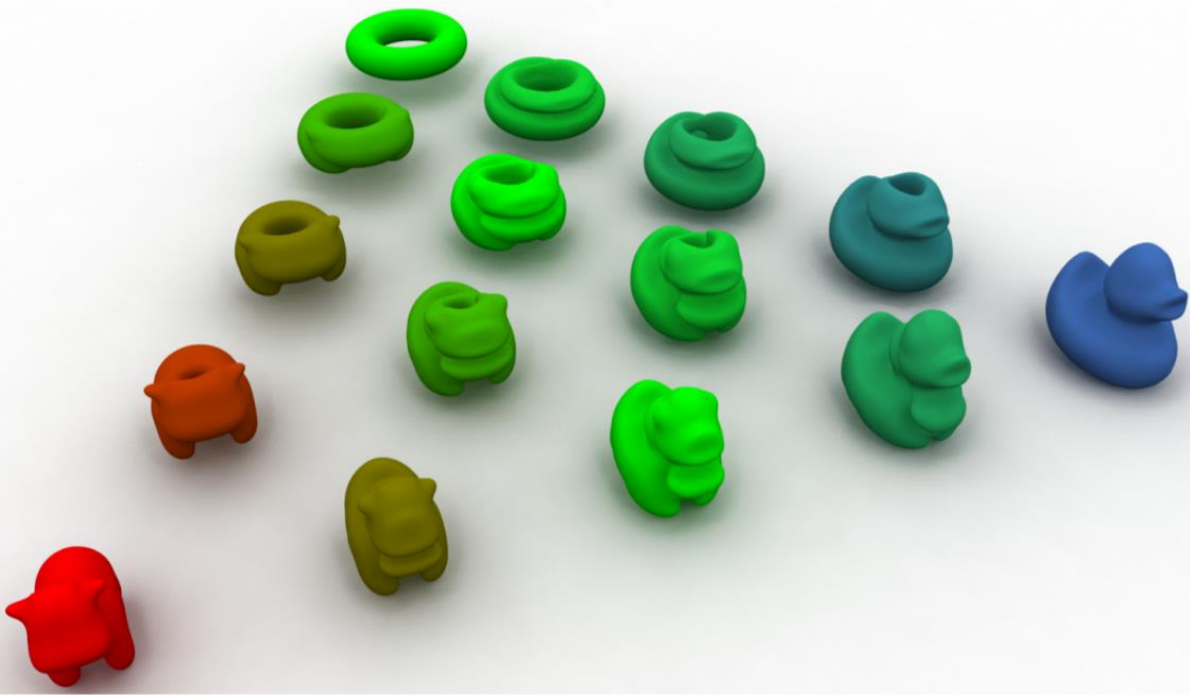
CONVOLUTIONAL WASSERSTEIN DISTANCES: EFFICIENT OPTIMAL TRANSPORTATION ON GEOMETRIC DOMAINS

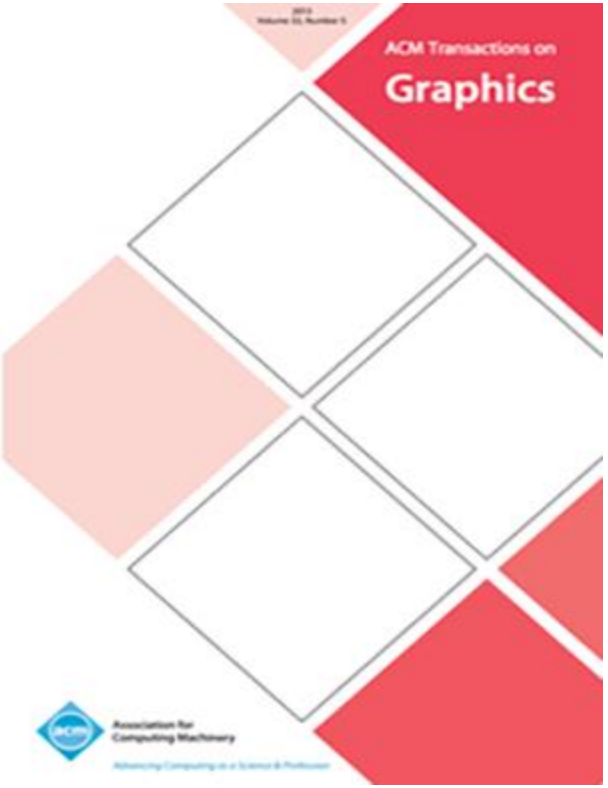
JUSTIN SOLOMON, FERNANDO DE GOES, GABRIEL PEYRE, MARCO CUTURI, ADRIAN BUTSCHER, ANDY NGUYEN, TAO DU, LEONIDAS GUIBAS

Presenter: Zhakshylyk Nurlanov
Supervisor: Marvin Eisenberger



CONVOLUTIONAL WASSERSTEIN DISTANCES

- SIGGRAPH 2015
- ACM Transactions on Graphics (TOG)
- Code available: https://github.com/gpeyre/2015-SIGGRAPH-convolutional-ot
- 239 Citations for May, 2020



- Distance between Probability Distributions
- Approach: Entropy Regularization, Heat Kernel
- Applications
- Conclusion

PLAN

PROBABILITY DISTRIBUTIONS IN GRAPHICS

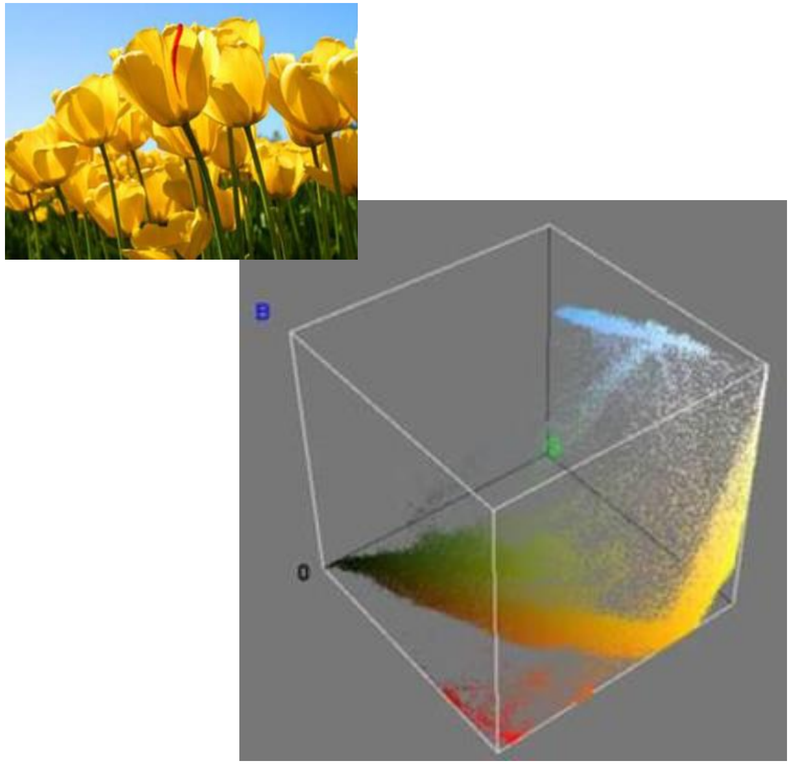
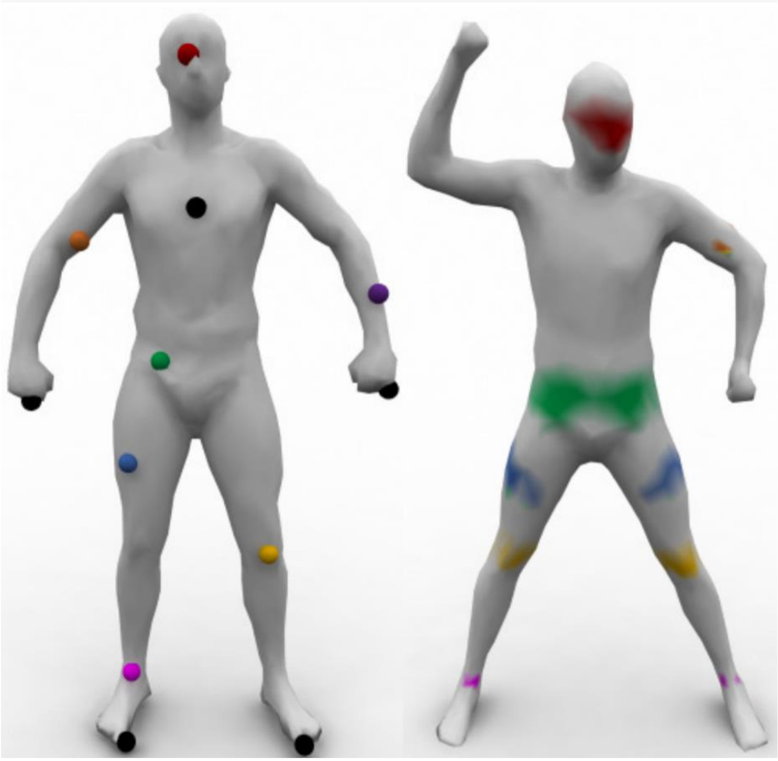
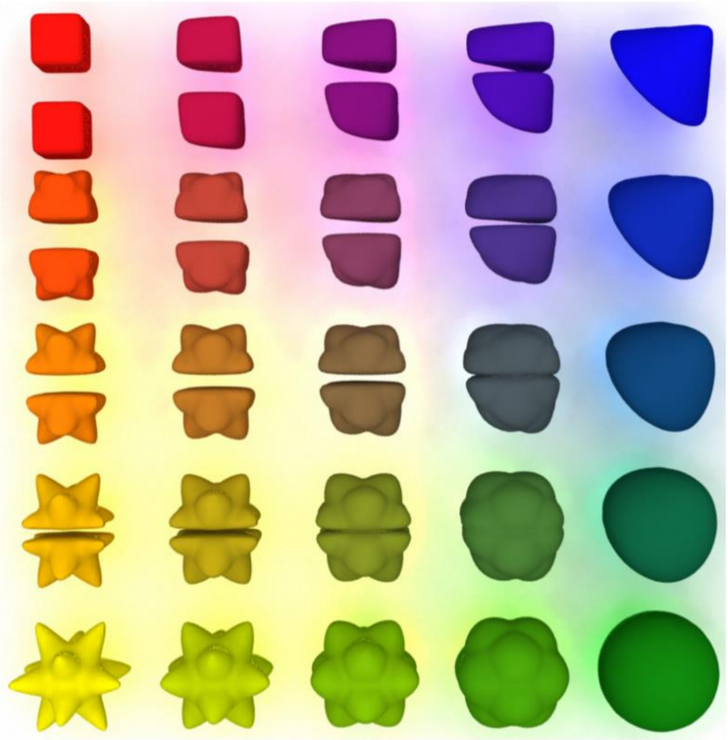


Image histograms



Correspondence maps



Shapes

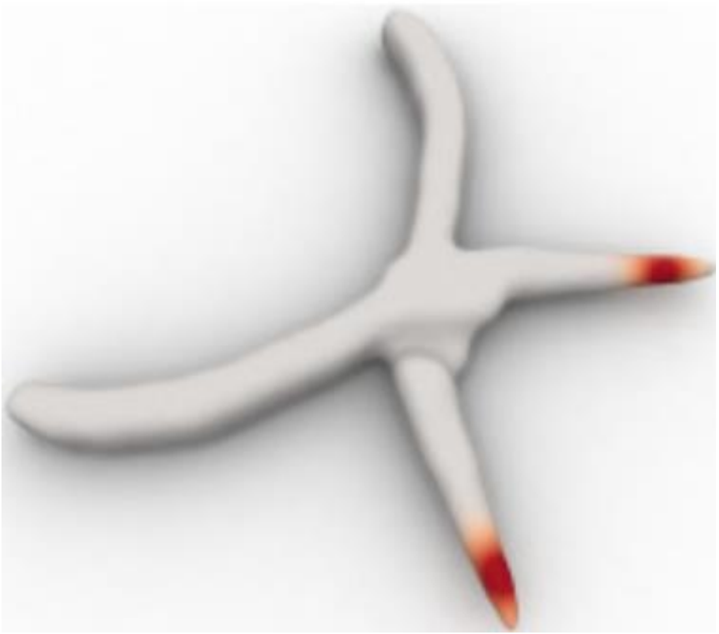
http://rsb.info.nih.gov/ij/plugins/color-inspector.html

PROBABILITY DISTRIBUTIONS IN GRAPHICS

$$p: M \rightarrow \mathbb{R}_+$$

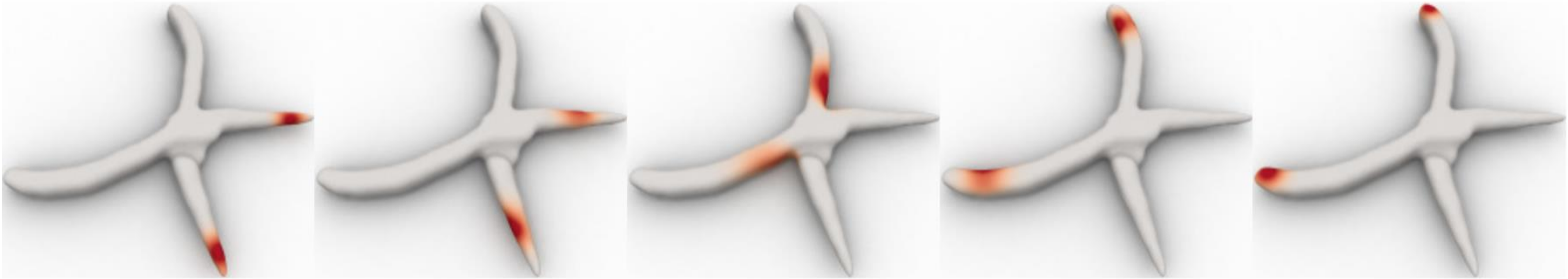
$$\int_M p(x)dx = 1$$

How is the feature distributed **over** geometric domain M?



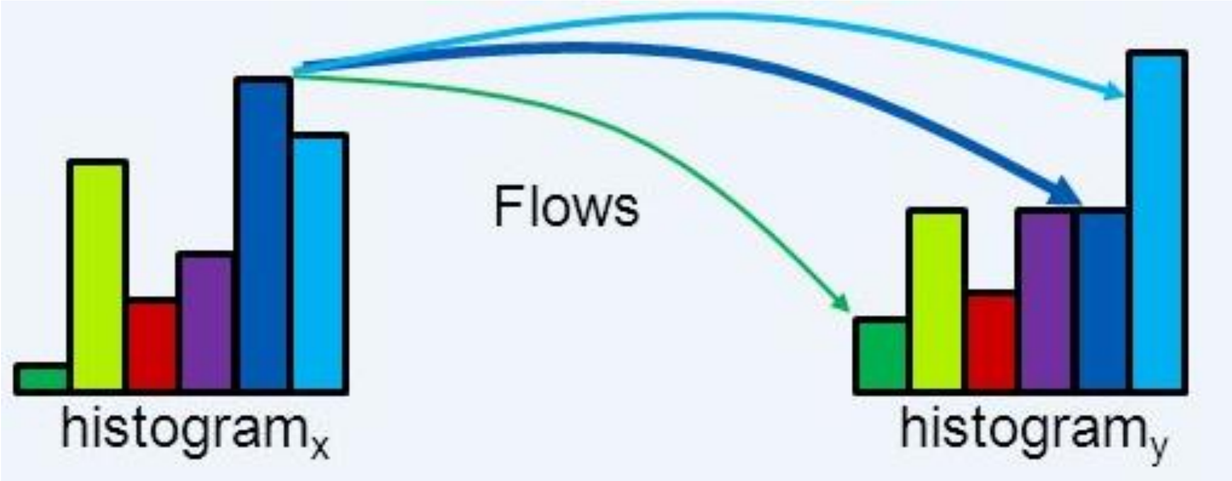
GOAL

- To manipulate distributions over geometric domains



$t = 0$ $t = 1/4$ $t = 1/2$ $t = 3/4$ $t = 1$

EARTH MOVER'S DISTANCE (EMD)



Perceptual Metrics for Image Database Navigation, Rubner, 2001

$$\min_{T\geq 0} \sum_{ij} T_{ij} d(x_i, x_j)$$

mass × distance

$$s.t. \sum_{j} T_{ij} = p_i \sum_{i} T_{ij} = q_j$$

Starts at p

Ends at q

- Transform the mass from source distribution to match the target distribution
- Minimum "mass_flow × distance" for transformation

WASSERSTEIN DISTANCE

$$W_2^2(\mu_0, \mu_1) := \inf_{T \in \Pi} \left[\iint_{M \times M} \underbrace{d(x,y)^2}_{\text{geodesic distance}} \underbrace{T(x,y)}_{\text{transportation plan}} dx\,dy \right]$$

$$\Pi(\mu_0, \mu_1) := \{T \in Prob(M \times M) : T(\cdot, M) = \mu_0, T(M, \cdot) = \mu_1\}$$

Continuous analog of EMD

http://www.sciencedirect.com/science/article/pii/S152407031200029X#

Why do we use Wasserstein distance?

WHICH DISTRIBUTION IS MORE SIMILAR?

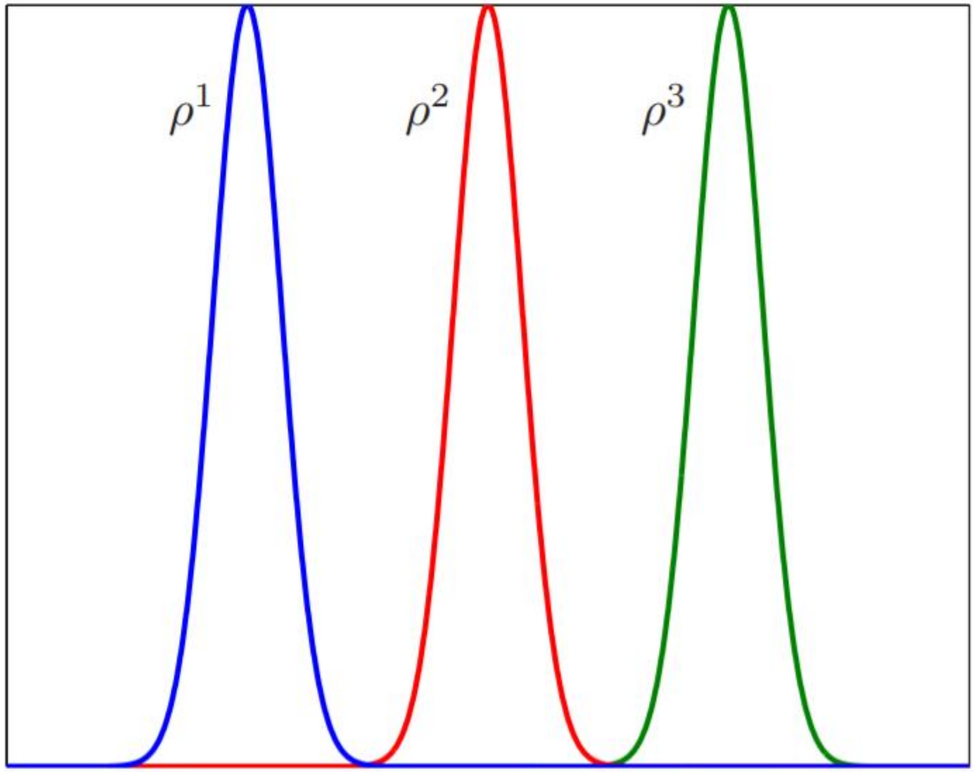


Fig. 4. L^p distance vs. Wasserstein distance.

Sensitivity analysis of the LWR model for traffic forecast on large networks using Wasserstein distance, Briani et al., 2016

$$L_1(\rho_1, \rho_2) = \mathrm{E}\left[||\rho_1 - \rho_2||_1\right] = 2$$

$$L_1(\rho_1, \rho_3) = \mathrm{E}\left[||\rho_1 - \rho_3||_1\right] = 2$$

WHICH DISTRIBUTION IS MORE SIMILAR?

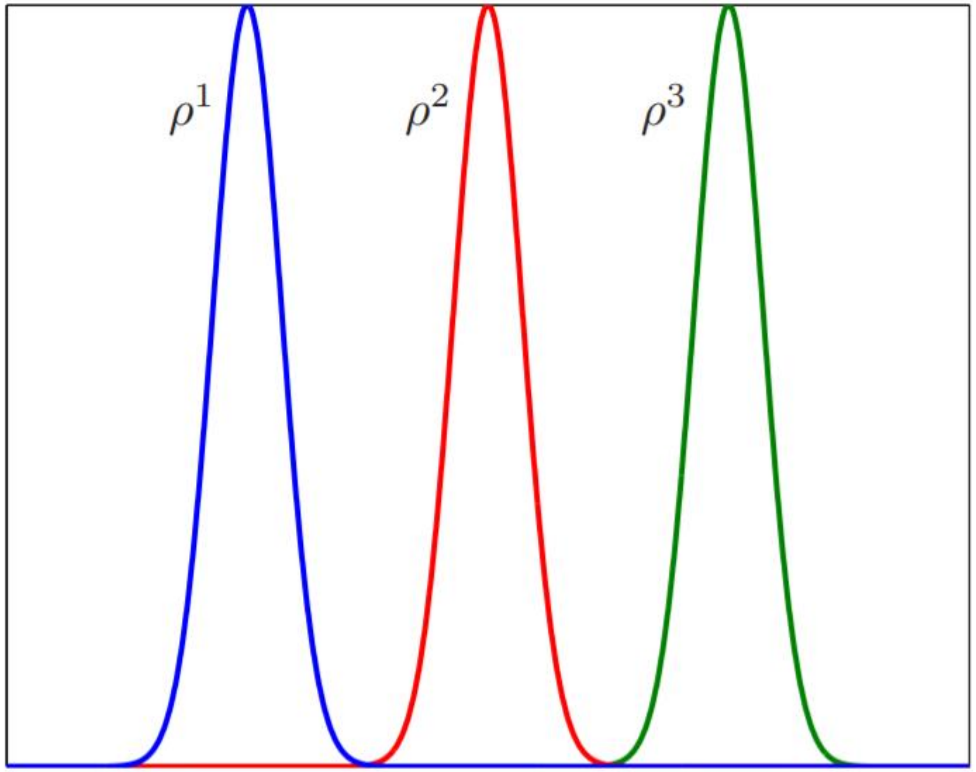


Fig. 4. L^p distance vs. Wasserstein distance.

Sensitivity analysis of the LWR model for traffic forecast on large networks using Wasserstein distance, Briani et al., 2016

$$L_1(\rho_1, \rho_2) = \mathrm{E}\left[||\rho_1 - \rho_2||_1\right] = 2$$

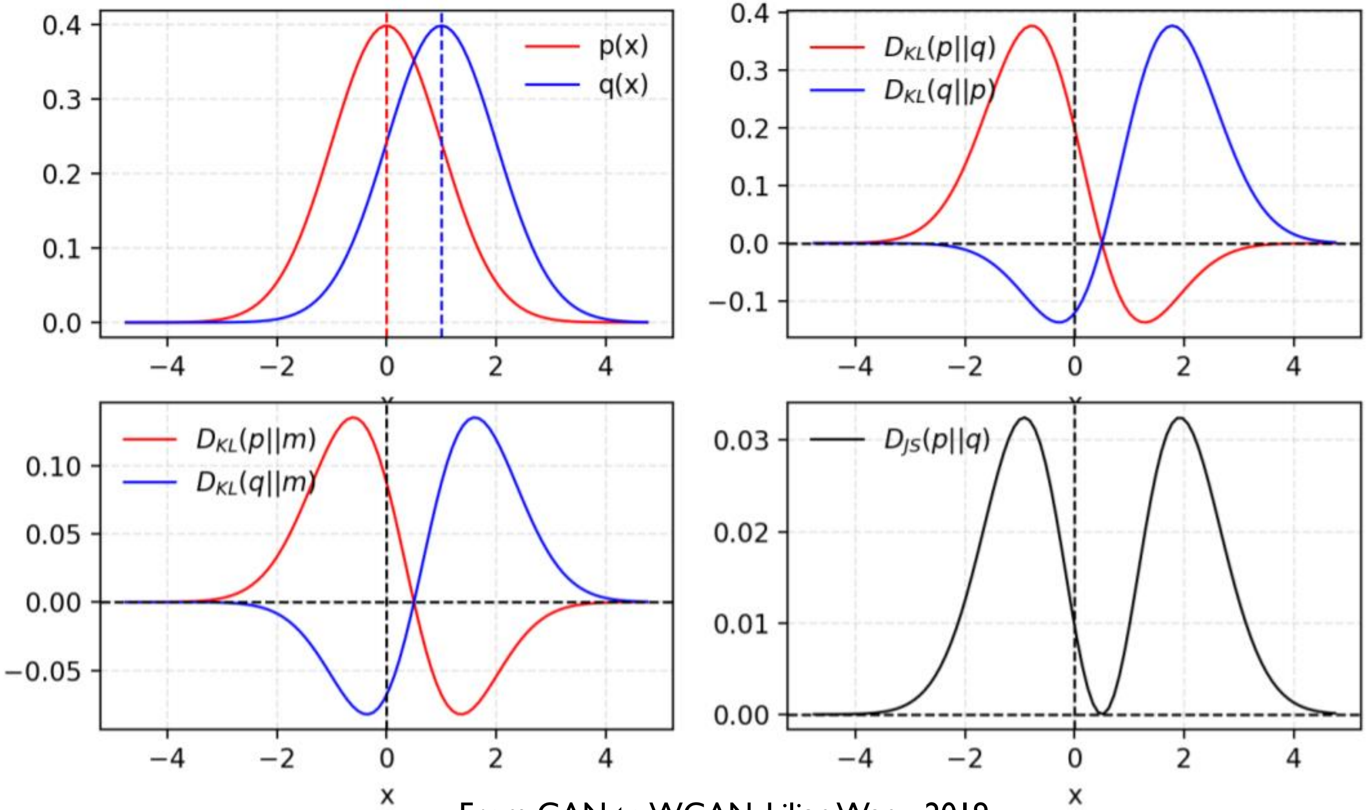
$$L_1(\rho_1, \rho_3) = \mathrm{E}\left[||\rho_1 - \rho_3||_1\right] = 2$$

but

$$W_2(\rho_1, \rho_2) < W_2(\rho_1, \rho_3)$$

KULLBACK–LEIBLER (KL), JENSEN–SHANNON (JS) DIVERGENCE

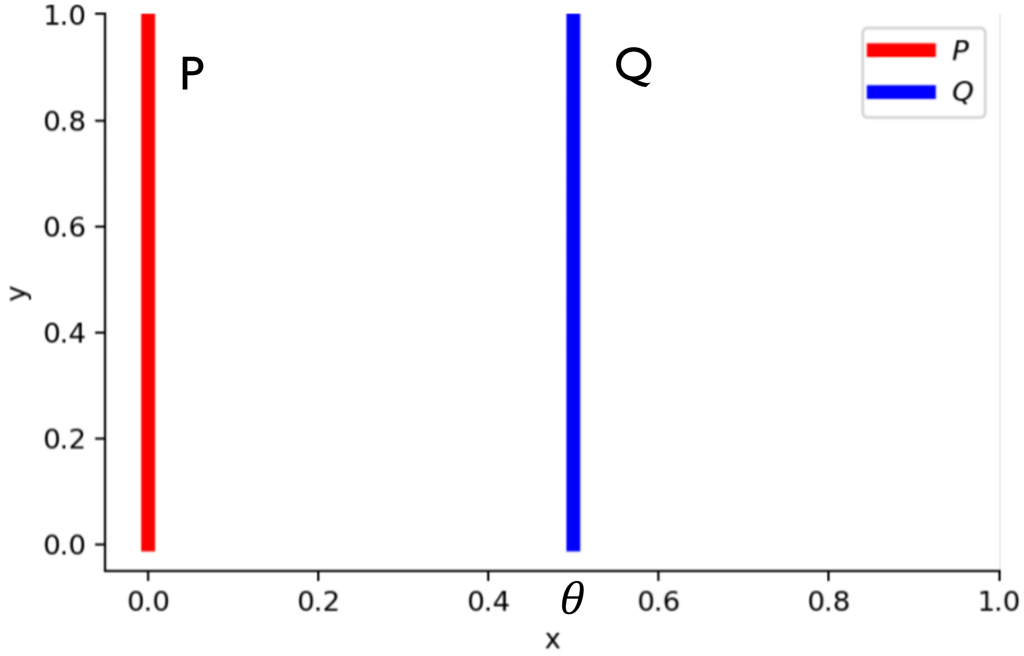
$$D_{KL}(p||q) = \int_x p(x)\log\frac{p(x)}{q(x)}dx\,, \quad D_{JS}(p,q) = \frac{1}{2}D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q||\frac{p+q}{2})$$



From GAN to WGAN, Lilian Weng, 2019

WHY W_2 IS BETTER THAN JS OR KL DIVERGENCE?

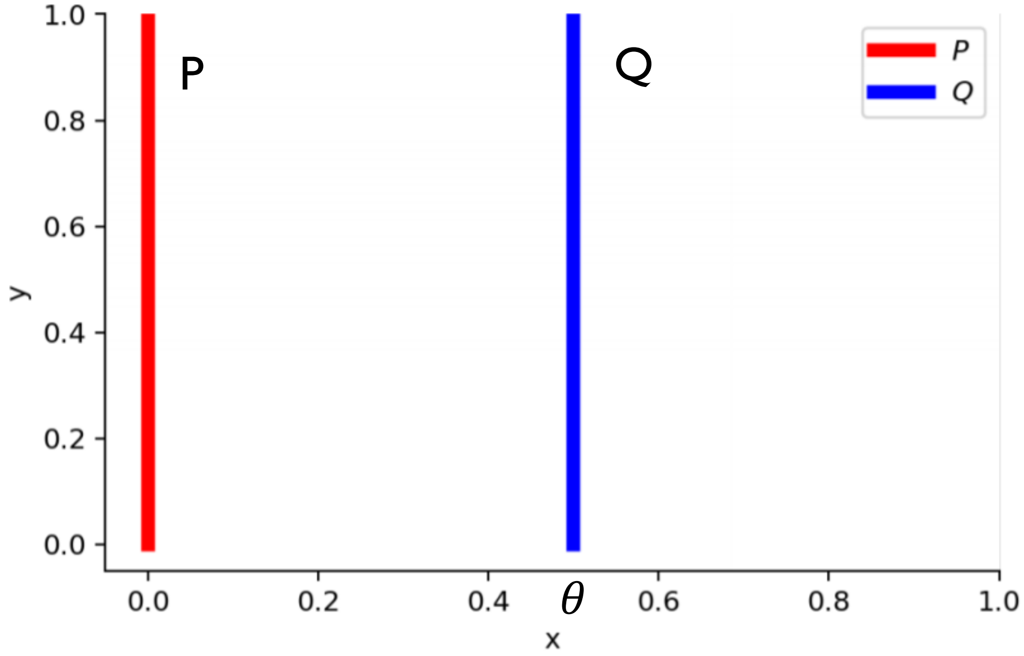
$$\forall(x,y)\in P, x=0 \text{ and } y\sim U(0,1) \forall(x,y)\in Q, x=\theta, 0\leq\theta\leq 1 \text{ and } y\sim U(0,1)$$



From GAN to WGAN, Lilian Weng, 2019

WHY W_2 IS BETTER THAN JS OR KL DIVERGENCE?

$$\forall(x,y) \in P, x = 0 \text{ and } y \sim U(0,1) \forall(x,y) \in Q, x = \theta, 0 \leq \theta \leq 1 \text{ and } y \sim U(0,1)$$



When $\theta \neq 0$:

$D_{KL}(P||Q) = +\infty, \quad D_{KL}(Q||P) = +\infty$

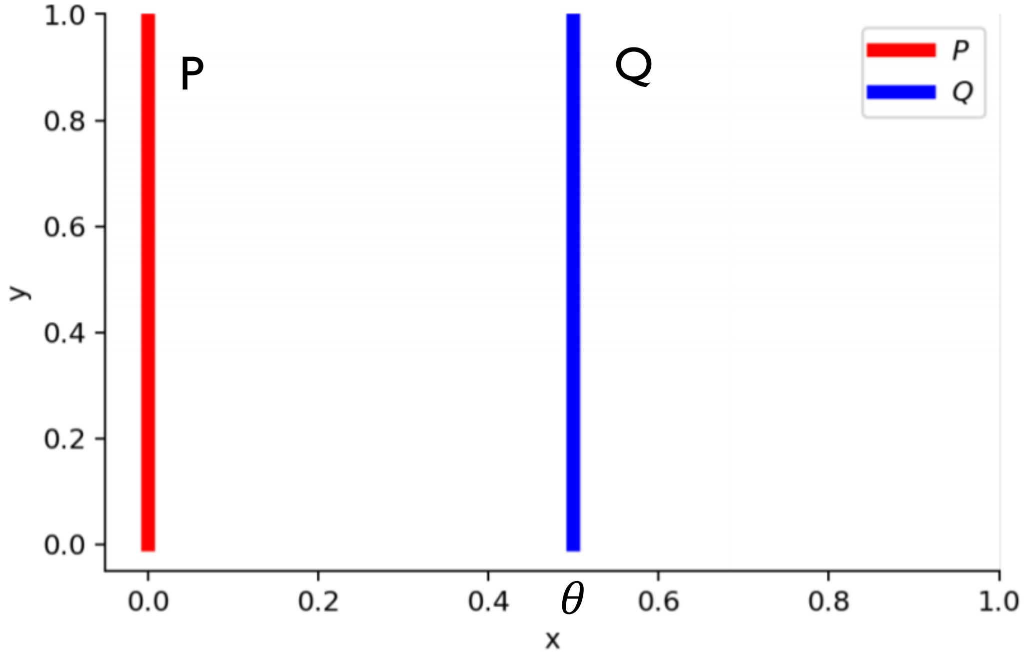
$D_{JS}(P,Q) = log2$

$W_2(P,Q) = |\theta|$

From GAN to WGAN, Lilian Weng, 2019

WHY W_2 IS BETTER THAN JS OR KL DIVERGENCE?

$$\forall(x,y) \in P, x = 0 \text{ and } y \sim U(0,1) \forall(x,y) \in Q, x = \theta, 0 \leq \theta \leq 1 \text{ and } y \sim U(0,1)$$



When $\theta \neq 0$:

$D_{KL}(P||Q) = +\infty, \quad D_{KL}(Q||P) = +\infty$

$D_{JS}(P,Q) = log2$

$W_2(P,Q) = |\theta|$

But when $\theta = 0$:

$D_{KL}(P||Q) = D_{KL}(Q||P) = D_{JS}(P,Q) = 0$

$W_2(P,Q) = |\theta| = 0$

From GAN to WGAN, Lilian Weng, 2019

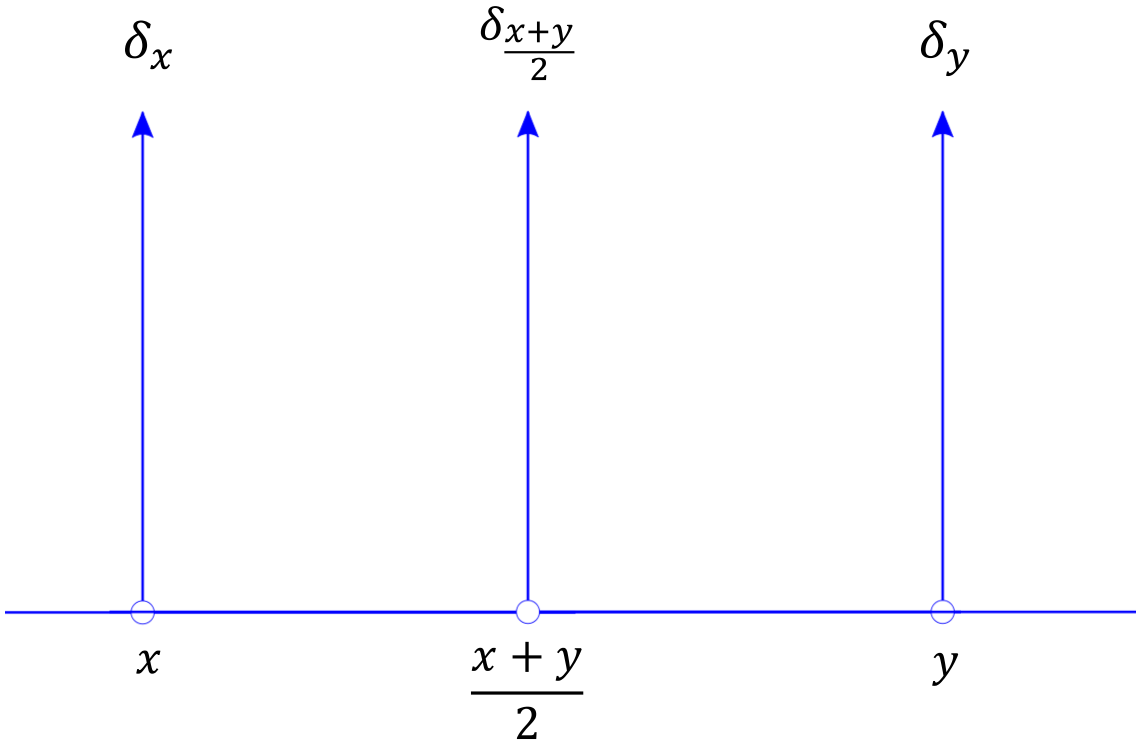
AVERAGING DISTRIBUTIONS

What is between two given distributions?

$\delta_x \quad ? \quad \delta_y$

$\frac{\delta_x + \delta_y}{2}$ is bimodal centered at x and y

AVERAGING DISTRIBUTIONS



What is between two given distributions?

$\delta_x \quad ? \quad \delta_y$

$\frac{\delta_x + \delta_y}{2}$ is bimodal centered at x *and* y

but

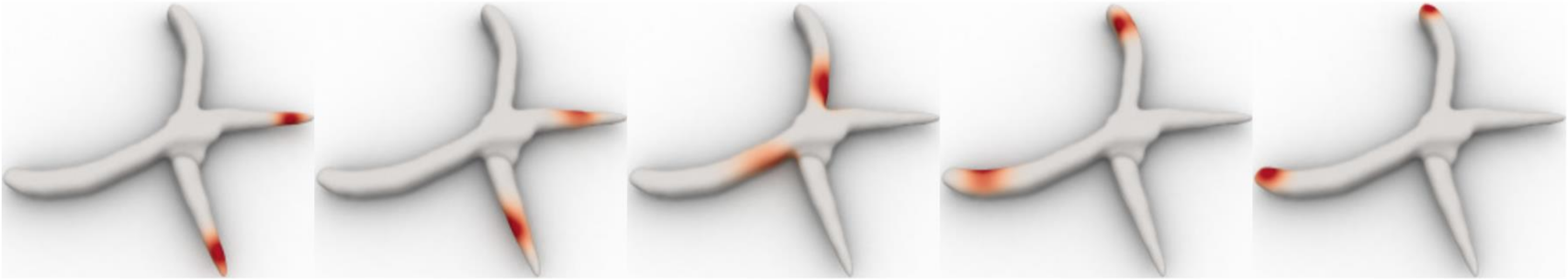
$$f^* = \underset{f}{\operatorname{argmin}}\left(W_2^2(f, \delta_x) + W_2^2(f, \delta_y)\right)$$

$f^* = \delta_{\frac{x+y}{2}}$ is a Dirac at midpoint

GOAL

- To manipulate distributions over geometric domains

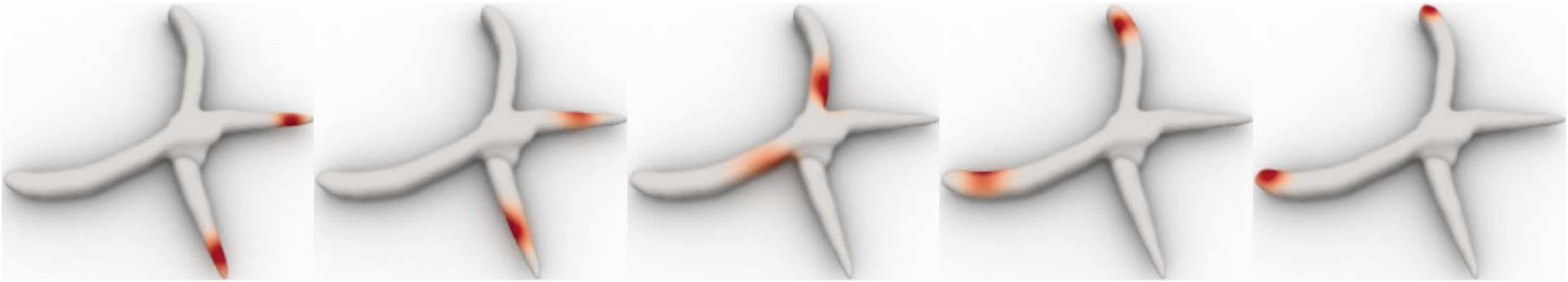
  using <u>Wasserstein Distance</u>



$t = 0$ $\quad$ $t = {}^1\!/_4$ $\quad$ $t = {}^1\!/_2$ $\quad$ $t = {}^3\!/_4$ $\quad$ $t = 1$

INTERPOLATION

$$\boxed{\mu_t = arg\min_{\mu} \left[(1-t)W_2^2(\mu,\mu_0) + tW_2^2(\mu,\mu_1)\right]}$$

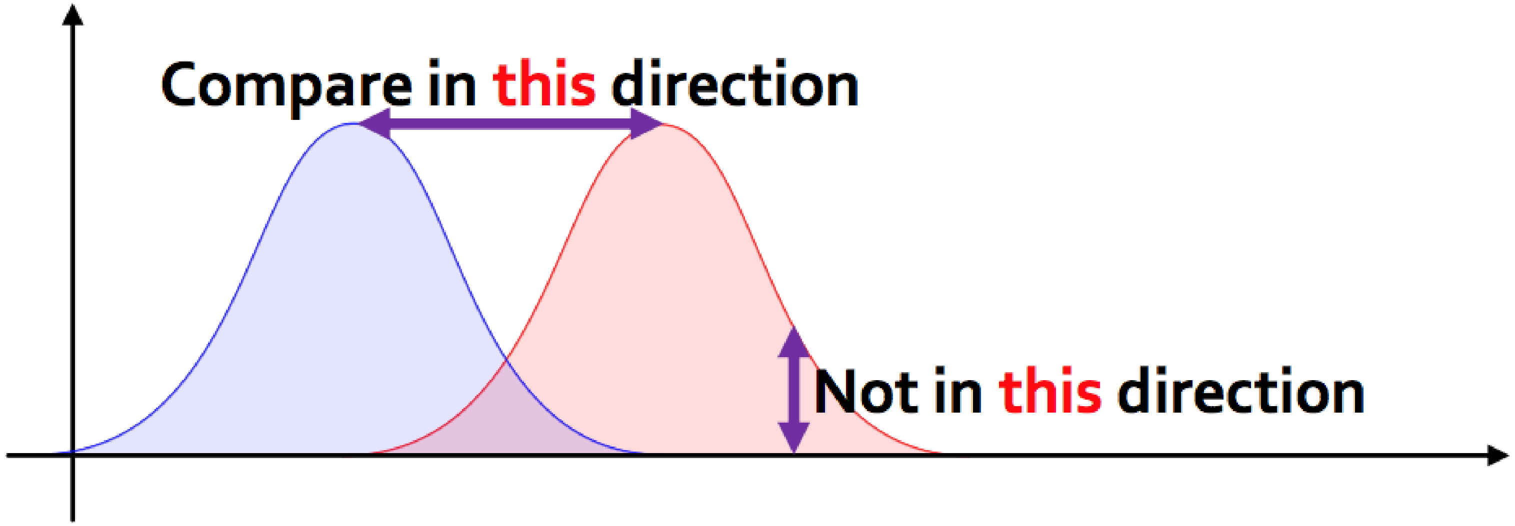


$t = 0$ $t = {}^1\!/_4$ $t = {}^1\!/_2$ $t = {}^3\!/_4$ $t = 1$

How to compute Wasserstein distance?

WASSERSTEIN DISTANCE

$$W_2^2(\mu_0, \mu_1) := \inf_{T \in \Pi} \left[\iint_{M \times M} d(x,y)^2 T(x,y) dx dy \right]$$



Looks terribly complex!

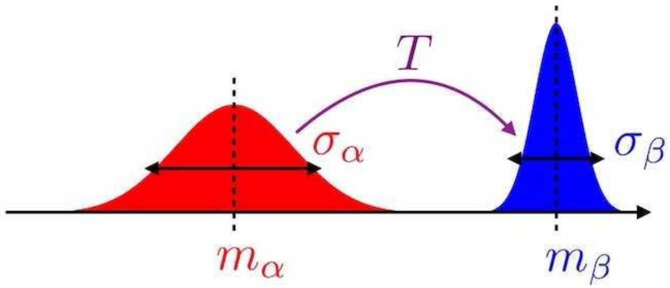
https://jeremykun.com/tag/wasserstein-metric/

WASSERSTEIN DISTANCE BETWEEN 2 GAUSSIANS

$$W_2^2\left(N(m_1,\Sigma_1);N(m_2,\Sigma_2)\right) = ||m_1 - m_2||_2^2 + Tr\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)$$

One Dimensional case:

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

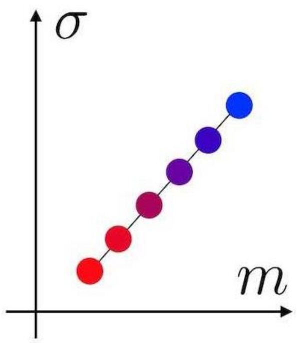


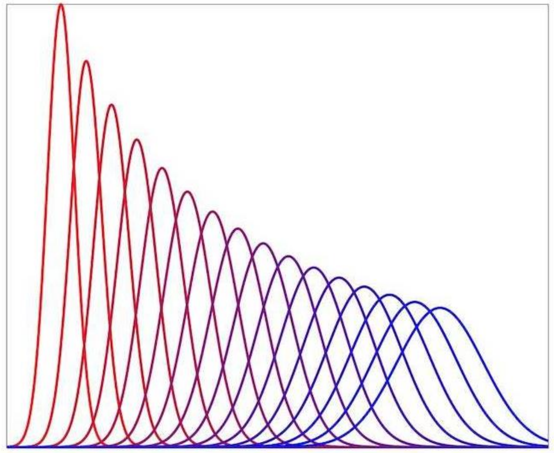
Optimal transport: $T(x) = \frac{\sigma_\beta}{\sigma_\alpha}(x - m_\alpha) + m_\beta$

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

https://twitter.com/gabrielpeyre/status/1153530233011757056/photo/1

Interpolation





Wasserstein geometry of Gaussian measures, Asuka Takatsu, 2010

COMPUTATIONAL ISSUES

$$W_2^2(\mu_0, \mu_1) := \inf_{T \in \Pi} \left[\iint_{M \times M} d(x,y)^2 T(x,y) dx dy \right]$$

- Quadratic number of unknowns
- Needs $d(x, y)$ for all (x, y) to compute

In discrete case:

$p(x) \sim \vec{p} \in \mathbb{R}^n$

$q(x) \sim \vec{q} \in \mathbb{R}^n$

$\Rightarrow$ $T(x, y) \sim T \in \mathbb{R}^{n \times n}$

$\boxed{\text{General linear programming is impractical}}$

- Approach:
  - First Approximation: Entropy Regularization
  - Second Approximation: Heat Kernel for geodesic distances
  - Alternating Projections algorithm: Sinkhorn iterations

PLAN

FIRST APPROXIMATION: REGULARIZATION

$$\min\{f\} \to \min\{f + \gamma g\}$$

FIRST APPROXIMATION: REGULARIZATION

$$\min\{f\} \to \min\{f + \gamma g\}$$

small additional term

FIRST APPROXIMATION: REGULARIZATION

$$\min\{f\} \to \min\{f + \gamma g\}$$

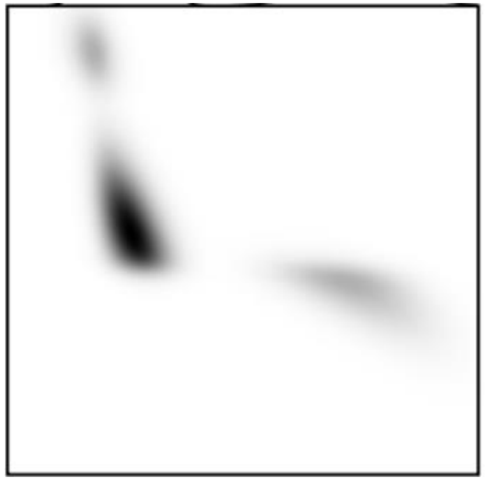
small additional term

$$f^*_\gamma \xrightarrow{\gamma \to 0} f^*$$

ENTROPY

$$H(T) := -\iint_{M\times M} T(x,y) ln T(x,y) dx dy$$

Small H



Large H

Measure of unsertainty

FIRST APPROXIMATION: ENTROPIC REGULARIZATION

$$W_{2,\gamma}^2(\mu_0, \mu_1) := \inf_{T \in \Pi} \left[\iint_{M \times M} d(x,y)^2 T(x,y) dx dy - \gamma H(T) \right]$$

FIRST APPROXIMATION: ENTROPIC REGULARIZATION

$$W_{2,\gamma}^2(\mu_0,\mu_1) := \inf_{T\in\Pi}\left[\iint_{M\times M} d(x,y)^2 T(x,y)\,dxdy - \gamma H(T)\right]$$

$$\Updownarrow$$

$$W_{2,\gamma}^2(\mu_0,\mu_1) := \inf_{T\in\Pi}\left[\iint_{M\times M} (d(x,y)^2 + \gamma lnT(x,y))T(x,y)\,dxdy\right]$$

FIRST APPROXIMATION: ENTROPIC REGULARIZATION

$$W_{2,\gamma}^2(\mu_0, \mu_1) := \inf_{T \in \Pi} \left[\iint_{M \times M} d(x,y)^2 T(x,y) dx dy - \gamma H(T) \right]$$

$$\Updownarrow$$

$$W_{2,\gamma}^2(\mu_0, \mu_1) := \inf_{T \in \Pi} \left[\iint_{M \times M} (d(x,y)^2 + \gamma ln T(x,y)) T(x,y) dx dy \right]$$

Remember KL Divergence?

$$D_{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx \,,$$

KERNEL REPRESENTATION OF DISTANCE

$$\mathcal{K}_\gamma(x, y) = e^{-\frac{d(x,y)^2}{\gamma}}, \quad d(x, y)^2 = -\gamma ln \mathcal{K}_\gamma(x, y)$$

KERNEL REPRESENTATION OF DISTANCE

$$\mathcal{K}_\gamma(x,y) = e^{-\frac{d(x,y)^2}{\gamma}}, \qquad d(x,y)^2 = -\gamma ln\mathcal{K}_\gamma(x,y)$$

$$\mathcal{K}_\gamma(x,y) \geq 0 \quad \forall x,y \in M \times M$$

$$W_{2,\gamma}^2(\mu_0,\mu_1) := \gamma \inf_{T\in\Pi}\left[\iint_{M\times M} T(x,y) ln\frac{T(x,y)}{\mathcal{K}_\gamma(x,y)} dxdy\right]$$

REGULARIZED PROBLEM: INTERPRETATION, CONVEXITY

$$W_{2,\gamma}^2(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T \| \mathcal{K}_\gamma\right) + \gamma \cdot const$$

REGULARIZED PROBLEM: INTERPRETATION, CONVEXITY

$$W^2_{2,\gamma}(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T || \mathcal{K}_\gamma\right) + \gamma \cdot const$$

The optimal transportation plan T is the projection of the distance-based kernel $\mathcal{K}_\gamma$ onto Π (w.r.t. KL div)

$$\Pi(\mu_0, \mu_1) := \{T \in Prob(M \times M) : T(\cdot, M) = \mu_0, T(M, \cdot) = \mu_1\}$$

REGULARIZED PROBLEM: INTERPRETATION, CONVEXITY

$$W_{2,\gamma}^2(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T || \mathcal{K}_\gamma\right) + \gamma \cdot const$$

The optimal transportation plan T is the projection of the distance-based kernel $\mathcal{K}_\gamma$ onto Π (w.r.t. KL div)

$$\Pi(\mu_0, \mu_1) := \{T \in Prob(M \times M) : T(\cdot, M) = \mu_0, T(M, \cdot) = \mu_1\}$$

$KL(p, q)$ is strictly convex in p for fixed q

REGULARIZED PROBLEM

$$W_2^2(\mu_0, \mu_1) := \inf_{T \in \Pi} \left[\iint_{M \times M} d(x,y)^2 T(x,y) dx dy \right]$$

$$\Big\Downarrow \quad \boxed{\text{Entropy Regularization}}$$

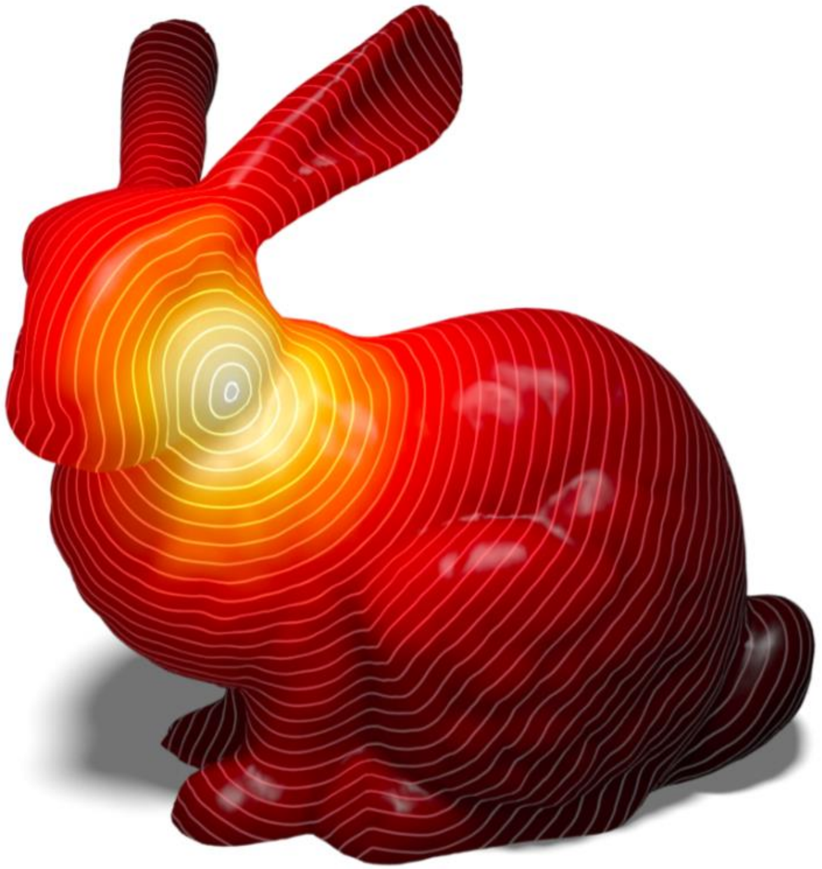
$$W_{2,\gamma}^2(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T || \mathcal{K}_\gamma\right)$$

$$d(x,y)^2 = -\gamma ln \mathcal{K}_\gamma(x,y)$$

PAIRWISE GEODESIC DISTANCE

$$\mathcal{K}_\gamma(x, y) = e^{-\frac{d(x,y)^2}{\gamma}}$$

Problem:
All geodesic distances are to be computed



Geodesics in Heat, Keenan Crane, 2013

INTRODUCING HEAT KERNEL

$$\begin{cases} \frac{\partial f(t,x)}{\partial t} = \Delta f(t,x) \\ f(0,x) = f_0(x) \end{cases} \Rightarrow f(t,x) = \int_M f_0(y) \underline{\mathcal{H}_t(x,y)} dy$$

Heat equation

Heat kernel

$\mathcal{H}_t(x,y)$ determines diffusion between $x, y \in M$ after time t

VARADHAN'S FORMULA

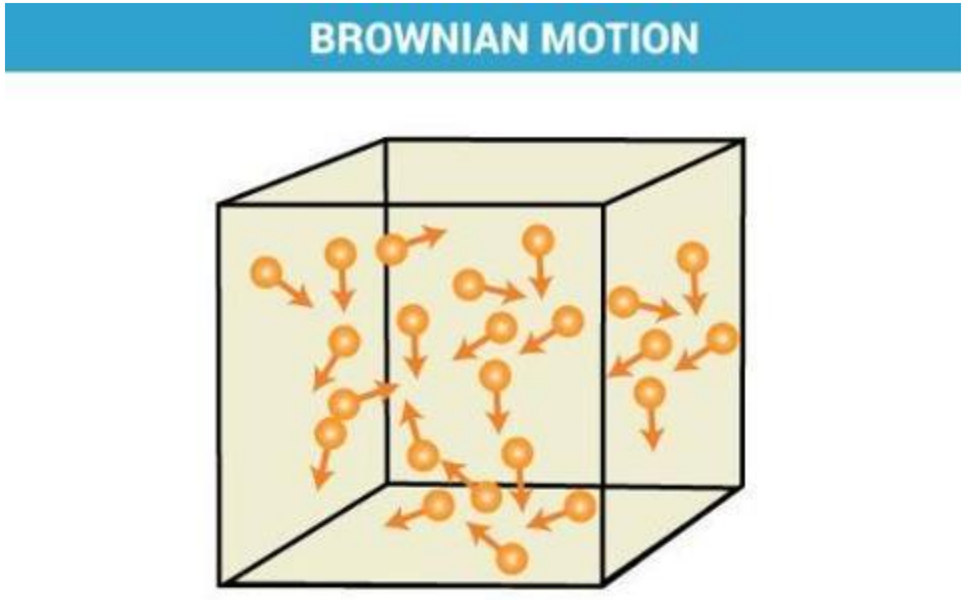
$$d(x,y)^2 = \lim_{t \to 0} \left[-2t \cdot ln\mathcal{H}_t(x,y)\right]$$

[Geodesics in Heat, Crane et al. 2013]

the distance $d(x, y)$ can be recovered by transferring the heat from x to y through manifold M over a short time interval t

VARADHAN'S FORMULA

$$d(x,y)^2 = \lim_{t \to 0} \left[-2t \cdot ln \mathcal{H}_t(x,y) \right]$$



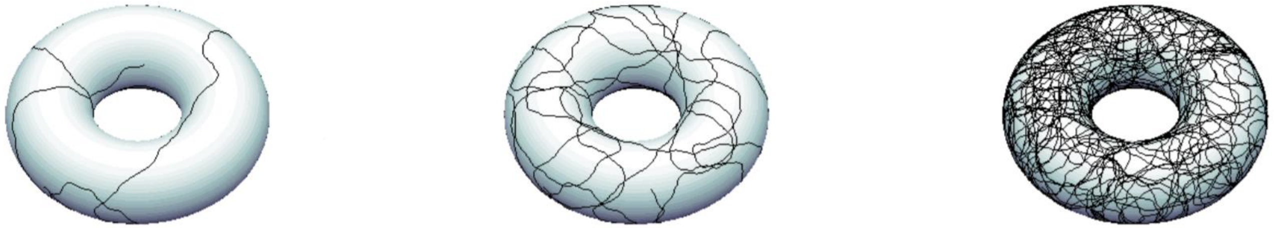


FIGURE 2. Simulations of a the kinetic Brownian motion on the torus over different time intervals.

https://byjus.com/chemistry/brownian-motion-zig-zag-motion/

Kinetic Brownian Motion in Riemannian Manifolds, Angst et al. 2015

VARADHAN'S FORMULA

$$d(x,y)^2 = \lim_{t\to 0}\left[-2t \cdot ln\mathcal{H}_t(x,y)\right]$$

Reminder: Kernel Representation of Distance

$$\mathcal{K}_\gamma(x,y) = e^{-\frac{d(x,y)^2}{\gamma}}, \qquad d(x,y)^2 = -\gamma ln\mathcal{K}_\gamma(x,y)$$

VARADHAN'S FORMULA

$$d(x,y)^2 = \lim_{t \to 0} \left[-2t \cdot ln\mathcal{H}_t(x,y) \right]$$

Reminder: Kernel Representation of Distance

$$\mathcal{K}_\gamma(x,y) = e^{-\frac{d(x,y)^2}{\gamma}}, \quad d(x,y)^2 = -\gamma ln\mathcal{K}_\gamma(x,y)$$

SECOND APPROXIMATION: HEAT KERNEL

$$t := \frac{\gamma}{2} \quad \Rightarrow \quad \mathcal{K}_\gamma(x, y) \approx \mathcal{H}_{\frac{\gamma}{2}}(x, y)$$

SECOND APPROXIMATION: HEAT KERNEL

$$t := \frac{\gamma}{2} \quad \Rightarrow \quad \mathcal{K}_\gamma(x, y) \approx \mathcal{H}_{\frac{\gamma}{2}}(x, y)$$

$$W^2_{2,\mathcal{H}_{\frac{\gamma}{2}}}(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T || \mathcal{H}_{\frac{\gamma}{2}}\right) + \gamma \cdot const$$

SECOND APPROXIMATION: HEAT KERNEL

$$t := \frac{\gamma}{2} \quad \Rightarrow \quad \mathcal{K}_\gamma(x, y) \approx \mathcal{H}_{\frac{\gamma}{2}}(x, y)$$

$$W^2_{2,\mathcal{H}_{\frac{\gamma}{2}}}(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T \| \mathcal{H}_{\frac{\gamma}{2}}\right) + \gamma \cdot const$$

$$\boxed{W_2^2 \approx W^2_{2,\gamma} \approx W^2_{2,\mathcal{H}_{\frac{\gamma}{2}}}}$$

DISCRETE SETTING

$$\begin{cases} \frac{\partial f(t,x)}{\partial t} = \Delta f(t,x) \\ f(0,x) = g(x) \end{cases} \longrightarrow \quad f(t,x) = \int_M g(y)\mathcal{H}_t(x,y)dy$$

Discretization $\Big\downarrow$

$$\begin{cases} \vec{f}_{t+1} - \vec{f}_t = L\vec{f}_t \\ \vec{f}_0 = \vec{g} \end{cases} \longrightarrow \quad \vec{f}_t = \mathcal{H}_t(\vec{a} \circ \vec{g})$$

DISCRETE SETTING

$$\begin{cases} \vec{f}_{t+1} - \vec{f}_t = L\vec{f}_t \\ \vec{f}_0 = \vec{g} \end{cases} \Longrightarrow \vec{f}_t = \mathcal{H}_t(\vec{a} \circ \vec{g})$$

$$\wr\wr$$

$$f(t,x) = \int_M g(y)\mathcal{H}_t(x,y)dy$$

$$\mathcal{H}_t \in \mathbb{R}^{n \times n}$$

"area weights" vector

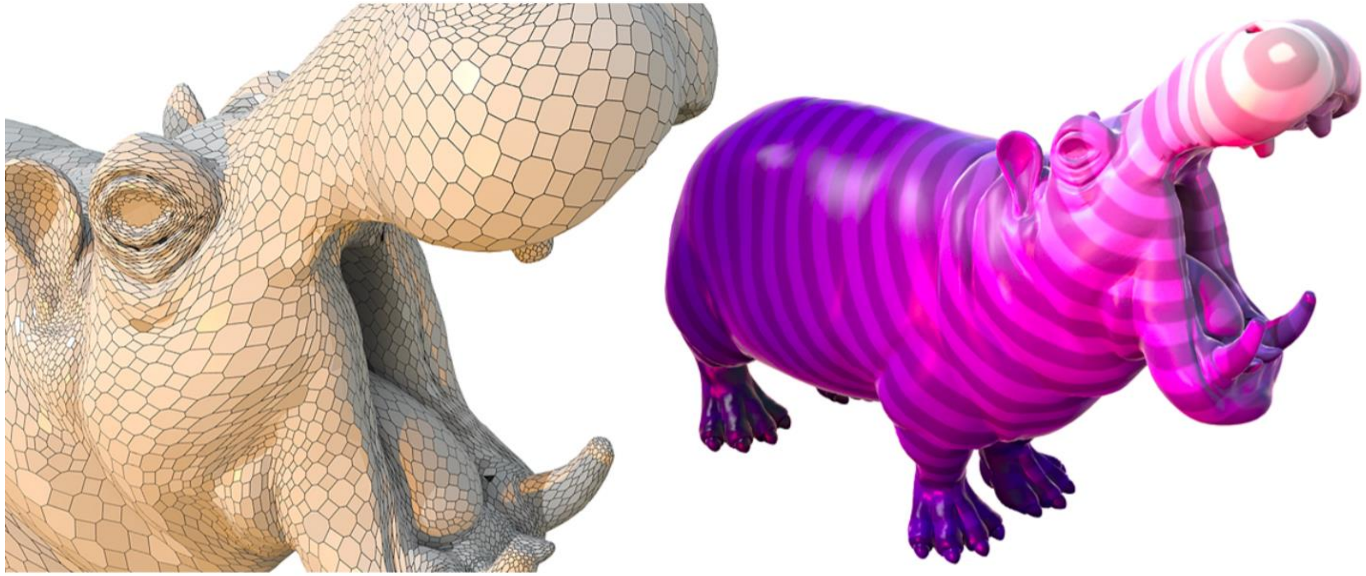
$$\vec{a} \in \mathbb{R}^n_+, \quad \vec{a}^T\vec{1} = 1$$

$$\boxed{\int_M f(x)dx \approx \vec{a}^T\vec{f}}$$

COMPUTING HEAT KERNEL: FOR IMAGES, AND MESHES

Gaussian kernel for images

https://en.wikipedia.org/wiki/Gaussian_blur



Distance on poly meshes based on heat method

Geodesics in Heat, Crane et al. 2013

CONVOLUTIONAL WASSERSTEIN DISTANCE

$$W_2^2(\mu_0, \mu_1) := \inf_{T \in \Pi} \left[\iint_{M \times M} d(x,y)^2 T(x,y)\, dx\, dy \right]$$

Entropy Regularization

$$W_{2,\gamma}^2(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T \| \mathcal{K}_\gamma\right)$$

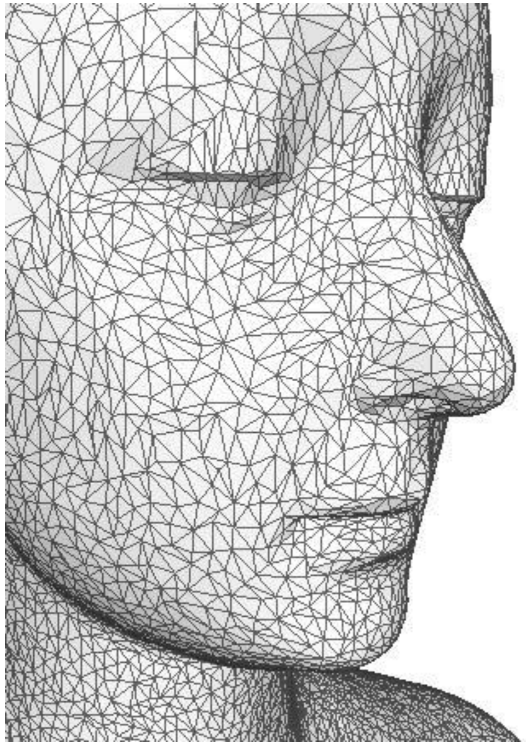
Heat Kernel

$$W_{2,\mathcal{H}_{\frac{\gamma}{2}}}^2(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T \| \mathcal{H}_{\frac{\gamma}{2}}\right)$$

PROBLEM

$$W^2_{2,\mathcal{H}_{\frac{\gamma}{2}}}(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T \| \mathcal{H}_{\frac{\gamma}{2}}\right)$$

Still quadratic number of unknowns



Mesh Generation and Cleaver, CIBC

PROPOSITION 1: LINEAR NUMBER OF VARIABLES

$$W^2_{2,\mathcal{H}_{\frac{\gamma}{2}}}(\mu_0, \mu_1) := \gamma \inf_{T \in \Pi} KL\left(T \| \mathcal{H}_{\frac{\gamma}{2}}\right) \quad (*)$$

Proposition 1:

The transportation plan $T \in \Pi(\mu_0, \mu_1)$ minimizing $(*)$ is of the form:

$$T = diag(\vec{v})\mathcal{H}_t diag(\vec{w}) = D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}$$

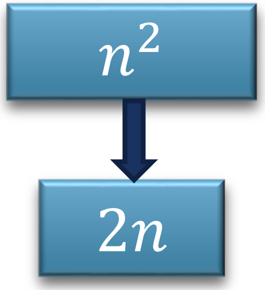
with unique vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$, satisfying:

$$\begin{cases} T\vec{a} = \mu_0 \\ T^T\vec{a} = \mu_1 \end{cases} \Longleftrightarrow \begin{cases} D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}\vec{a} = \mu_0 \\ D_{\vec{w}}\mathcal{H}_t D_{\vec{v}}\vec{a} = \mu_1 \end{cases}$$

PROPOSITION 1: LINEAR NUMBER OF VARIABLES

$$W^2_{2,\mathcal{H}_{\frac{\gamma}{2}}}(\mu_0,\mu_1) := \gamma \inf_{T\in\Pi} KL\left(T \| \mathcal{H}_{\frac{\gamma}{2}}\right) \quad (*)$$

Proposition 1:



The transportation plan $T \in \Pi(\mu_0,\mu_1)$ minimizing $(*)$ is of the form:

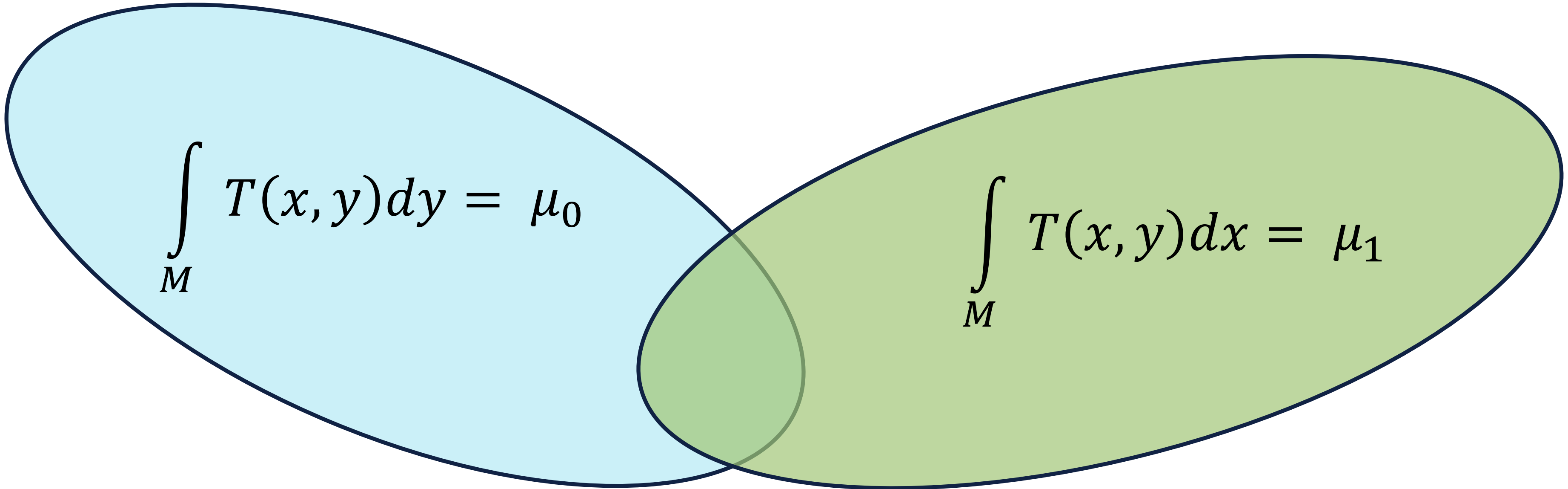
$$T = diag(\vec{v})\mathcal{H}_t diag(\vec{w}) = D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}$$

with unique vectors $\vec{v},\vec{w} \in \mathbb{R}^n$, satisfying:

$$\begin{cases} T\vec{a} = \mu_0 \\ T^T\vec{a} = \mu_1 \end{cases} \Longleftrightarrow \begin{cases} D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}\vec{a} = \mu_0 \\ D_{\vec{w}}\mathcal{H}_t D_{\vec{v}}\vec{a} = \mu_1 \end{cases}$$

ALTERNATING PROJECTIONS ALGORITHM

$$\boxed{T = diag(\vec{v})\mathcal{H}_t diag(\vec{w}) = D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}}$$

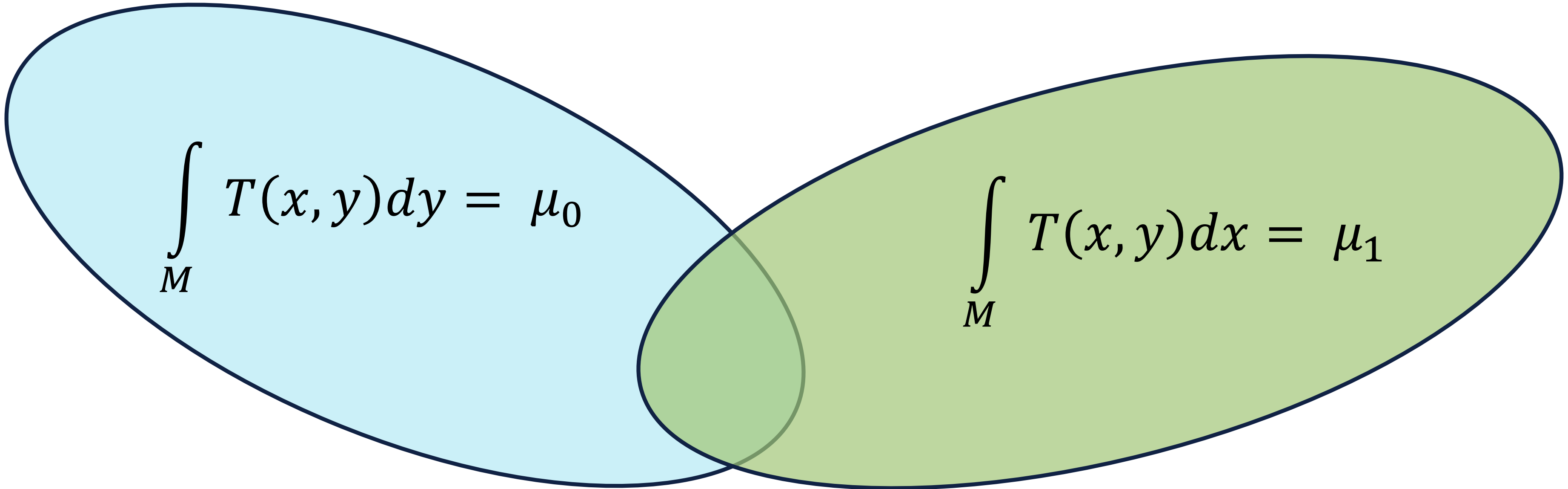


ALTERNATING PROJECTIONS ALGORITHM

$$T^{(0)} = \mathcal{H}_t$$

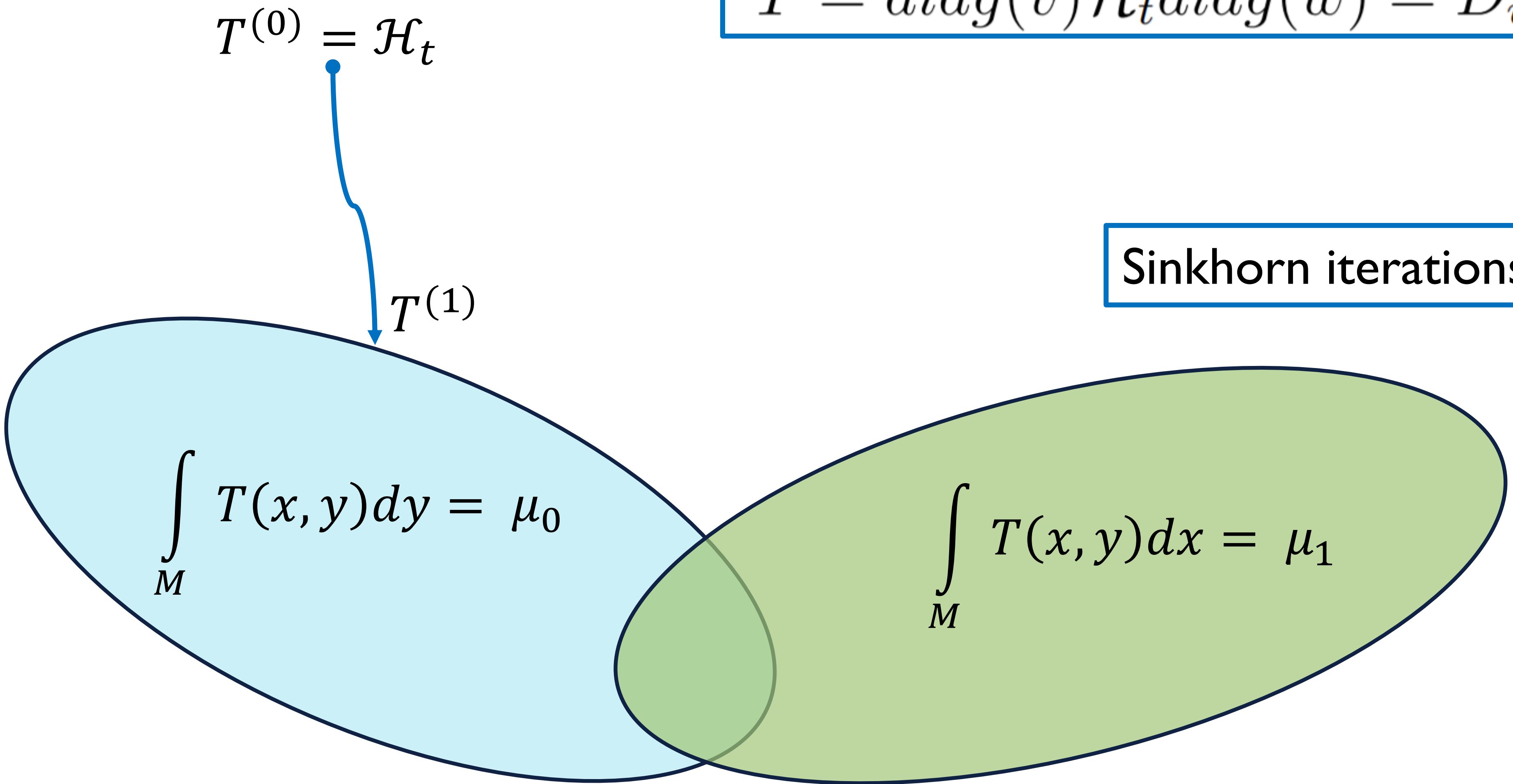
$$\boxed{T = diag(\vec{v})\mathcal{H}_t diag(\vec{w}) = D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}}$$

Sinkhorn iterations



ALTERNATING PROJECTIONS ALGORITHM

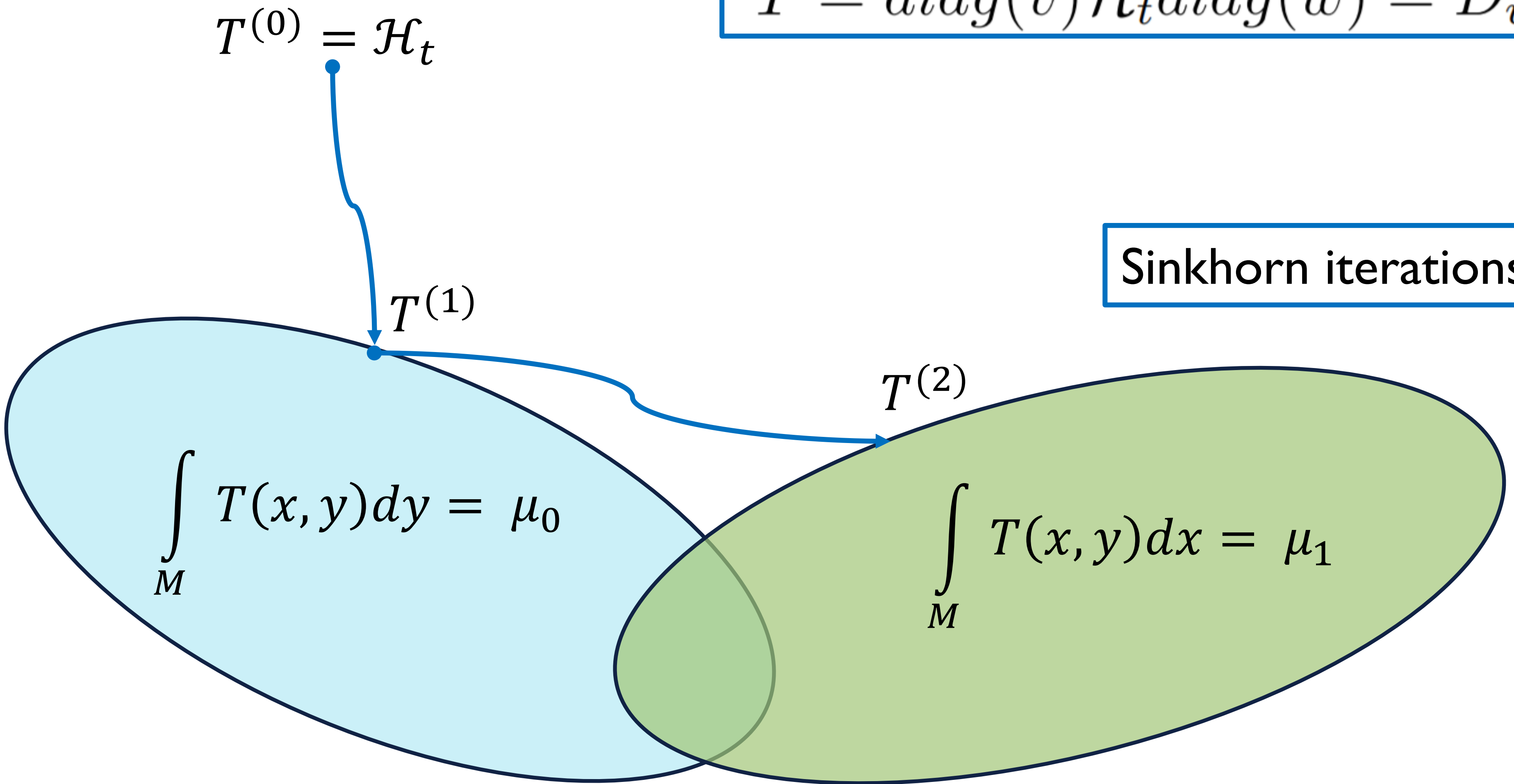
$$\boxed{T = diag(\vec{v})\mathcal{H}_t diag(\vec{w}) = D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}}$$



Sinkhorn iterations

ALTERNATING PROJECTIONS ALGORITHM

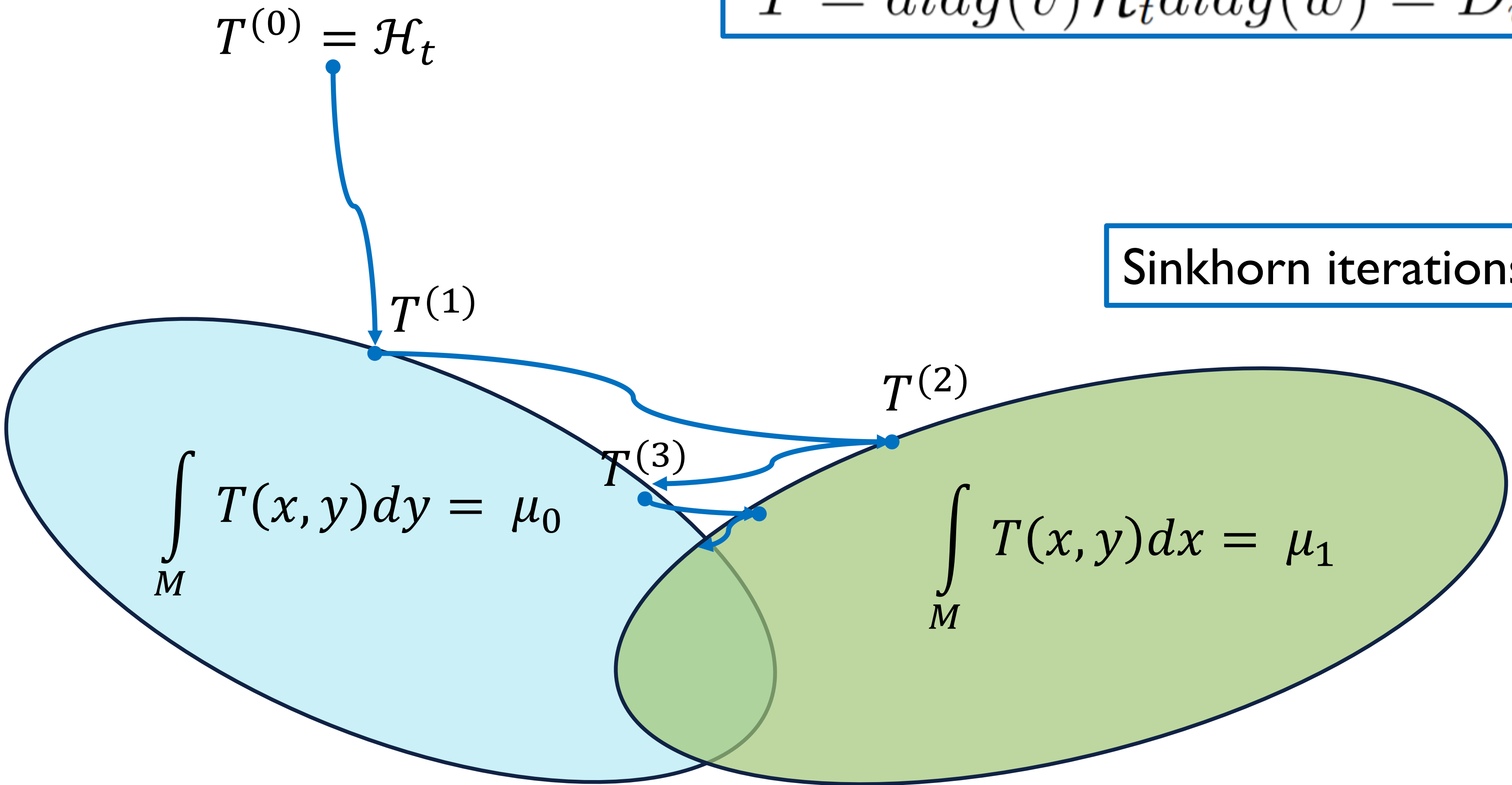
$$\boxed{T = diag(\vec{v})\mathcal{H}_t diag(\vec{w}) = D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}}$$



Sinkhorn iterations

ALTERNATING PROJECTIONS ALGORITHM

$$T = diag(\vec{v})\mathcal{H}_t diag(\vec{w}) = D_{\vec{v}}\mathcal{H}_t D_{\vec{w}}$$



Sinkhorn iterations

ALTERNATING PROJECTIONS ALGORITHM: SINKHORN ITERATIONS

function CONVOLUTIONAL-WASSERSTEIN($\boldsymbol{\mu}_0, \boldsymbol{\mu}_1; \mathbf{H}_t, \mathbf{a}$)

// Sinkhorn iterations

$\mathbf{v}, \mathbf{w} \leftarrow \mathbf{1}$

for $i = 1, 2, 3, \ldots$

$\quad \mathbf{v} \leftarrow \boldsymbol{\mu}_0 \oslash \mathbf{H}_t(\mathbf{a} \otimes \mathbf{w})$

$\quad \mathbf{w} \leftarrow \boldsymbol{\mu}_1 \oslash \mathbf{H}_t(\mathbf{a} \otimes \mathbf{v})$

// KL *divergence*

return $\gamma\, \mathbf{a}^\top [(\boldsymbol{\mu}_0 \otimes \ln \mathbf{v}) + (\boldsymbol{\mu}_1 \otimes \ln \mathbf{w})]$

Algorithm 1: *Sinkhorn iteration for convolutional Wasserstein distances. $\otimes, \oslash$ denote elementwise multiplication and division, resp.*

ALTERNATING PROJECTIONS ALGORITHM: SINKHORN ITERATIONS

function CONVOLUTIONAL-WASSERSTEIN($\boldsymbol{\mu}_0, \boldsymbol{\mu}_1; \mathbf{H}_t, \mathbf{a}$)

// Sinkhorn iterations

$\mathbf{v}, \mathbf{w} \leftarrow \mathbf{1}$

for $i = 1, 2, 3, \ldots$

$\quad \mathbf{v} \leftarrow \boldsymbol{\mu}_0 \oslash \mathbf{H}_t(\mathbf{a} \otimes \mathbf{w})$

$\quad \mathbf{w} \leftarrow \boldsymbol{\mu}_1 \oslash \mathbf{H}_t(\mathbf{a} \otimes \mathbf{v})$

// KL divergence

return $\gamma\, \mathbf{a}^\top [(\boldsymbol{\mu}_0 \otimes \ln \mathbf{v}) + (\boldsymbol{\mu}_1 \otimes \ln \mathbf{w})]$

Solve linear equation

Algorithm 1: *Sinkhorn iteration for convolutional Wasserstein distances. $\otimes, \oslash$ denote elementwise multiplication and division, resp.*

- Results:
  - Timing comparison
  - Applications

PLAN

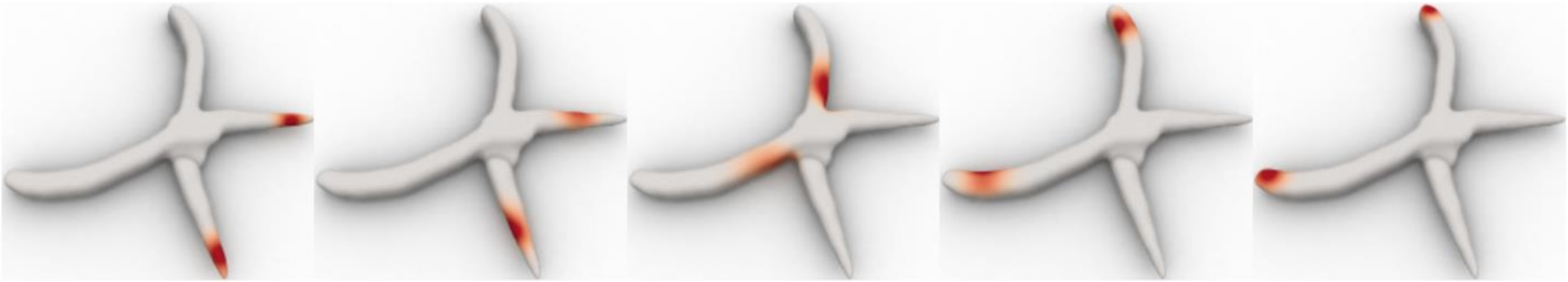
RESULTS

$\lvert V \rvert$	$\lvert T \rvert$	PD	LP	[Cuturi 2013]	PF	$\mathcal{W}^2_{2,\mathbf{H}_t}$
693	1382	0.10	9.703	0.625	0.00	1.564
1150	2296	0.28	36.524	1.284	0.00	0.571
1911	3818	0.79	*	2.725	0.02	1.010
3176	6348	2.15	*	5.435	0.03	1.553
5278	10552	6.47	*	10.490	0.06	2.477
8774	17544	18.55	*	23.326	0.10	4.516
14584	29164	53.41	*	*	0.17	8.152

Table 1: *Timing (in sec.) for approximating $\mathcal{W}_2$ between random distributions on triangle meshes, averaged over 10 trials. An asterisk * denotes time-out after one minute. Pairwise distance (PD) computation is needed for the linear program (LP) and [Cuturi 2013]; timing for this step is written separately. Cholesky pre-factorization (PF) is needed for convolutional distance and is similarly separated.*

INTERPOLATION

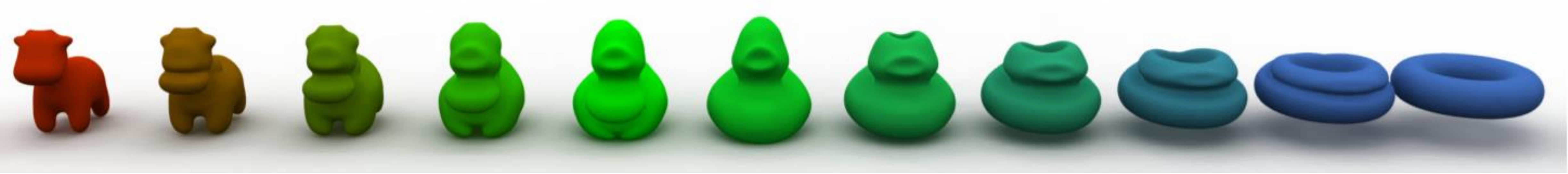
$$\boxed{\mu_t = arg\min_{\mu} \left[(1-t)W_2^2(\mu,\mu_0) + tW_2^2(\mu,\mu_1)\right]}$$



$t = 0$ $t = 1/4$ $t = 1/2$ $t = 3/4$ $t = 1$

INTERPOLATION

$$\boxed{\mu_t = arg\min_{\mu}\left[(1-t)W_2^2(\mu,\mu_0) + tW_2^2(\mu,\mu_1)\right]}$$



BARYCENTER

$$\mu^* = arg\min_{\mu} \sum_{i=1}^{N} \alpha_i W_2^2(\mu, \mu_i)$$

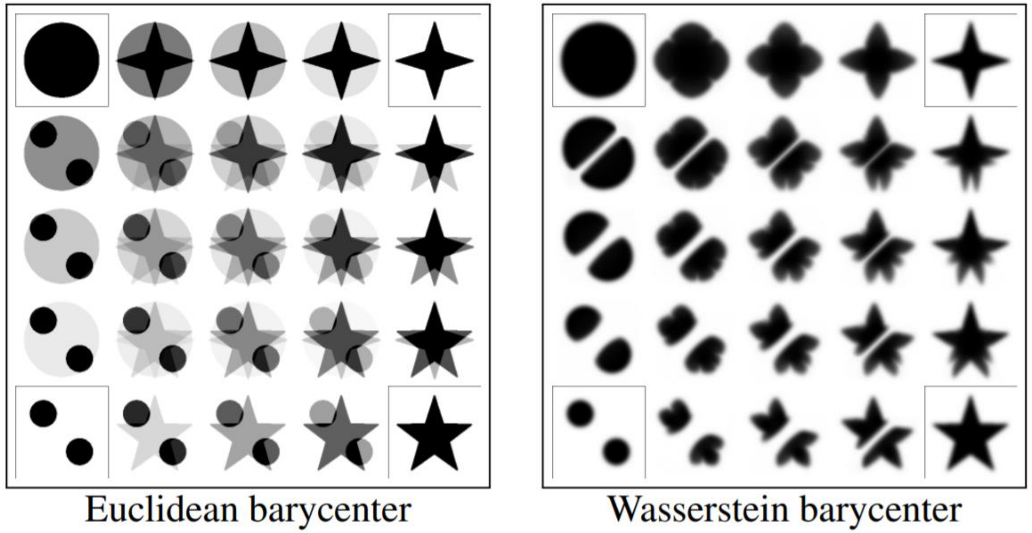
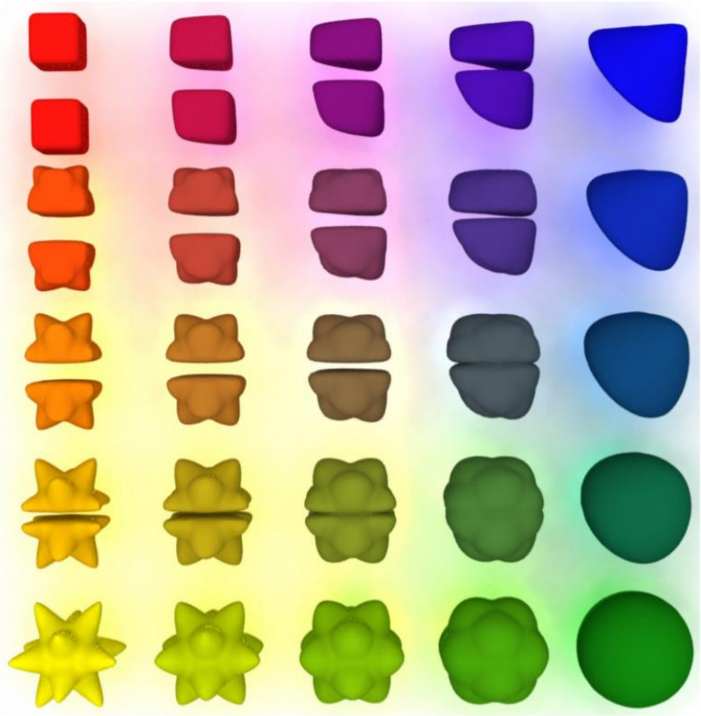


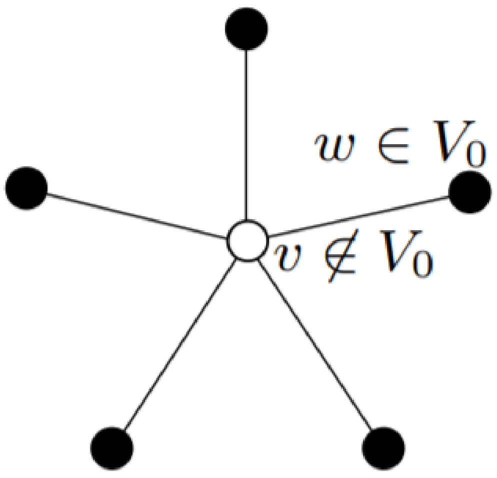
Figure 12: *Interpolating indicators using linear combinations (left) is ineffective for shape interpolation, but convolutional Wasserstein barycenters (right) move features by matching mass of the underlying distributions.*



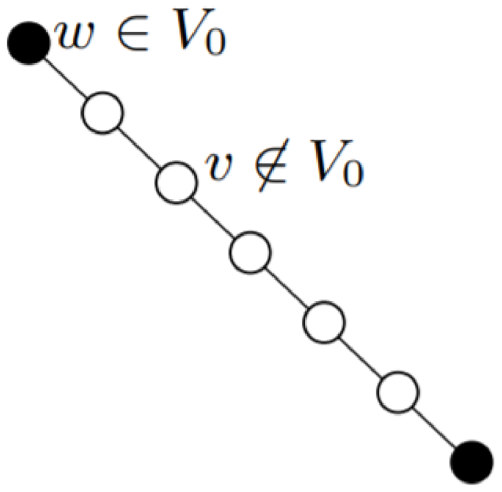
PROPAGATION

$$\{\mu_v\} = arg \min_{\mu_v, v \notin V_0} \sum_{(v,w)\in E} \alpha_{(v,w)} W_2^2(\mu_v, \mu_w)$$

$$s.t. \mu_v \quad fixed \quad \forall v \in V_0$$



Barycenter



Displacement interpolation

PROPAGATION

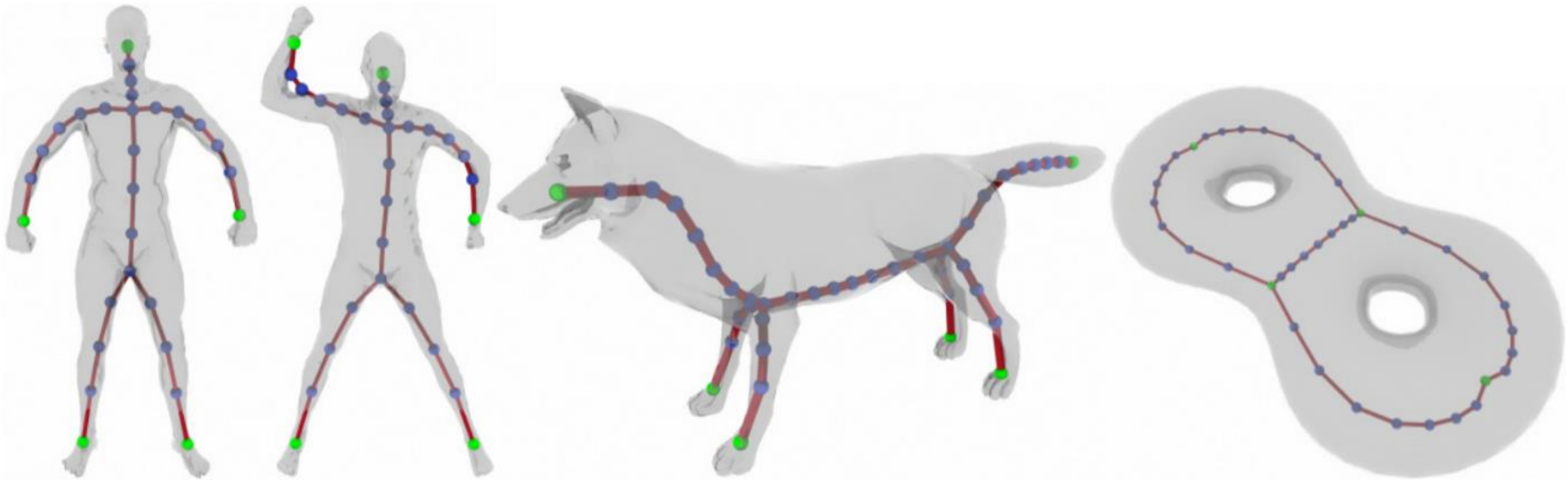
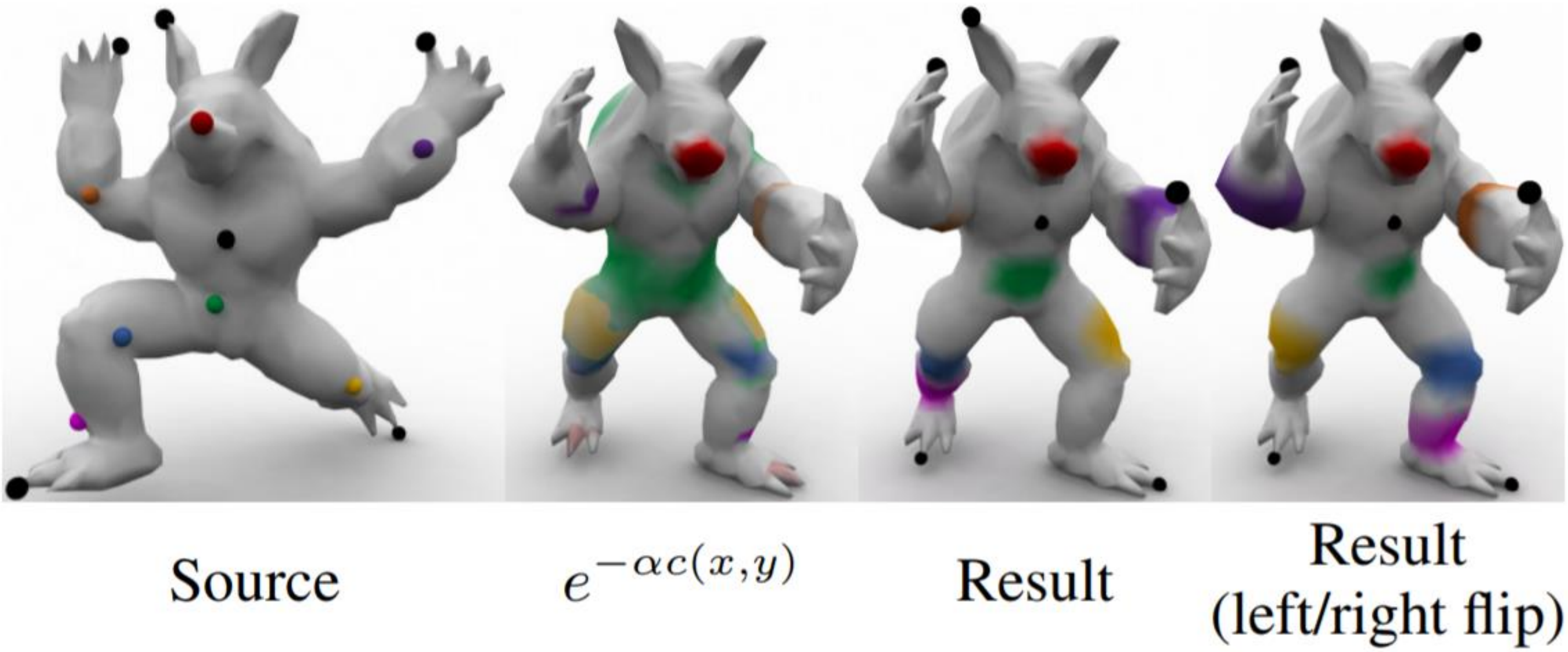


Figure 9: *Embeddings of skeletons computed using Wasserstein propagation; the positions of the blue vertices are computed automatically using the fixed green vertices and topology of the graph.*

PROPAGATION



Soft maps: Colored points on the source are mapped to the colored distributions on the target, where black points are fixed input correspondences.

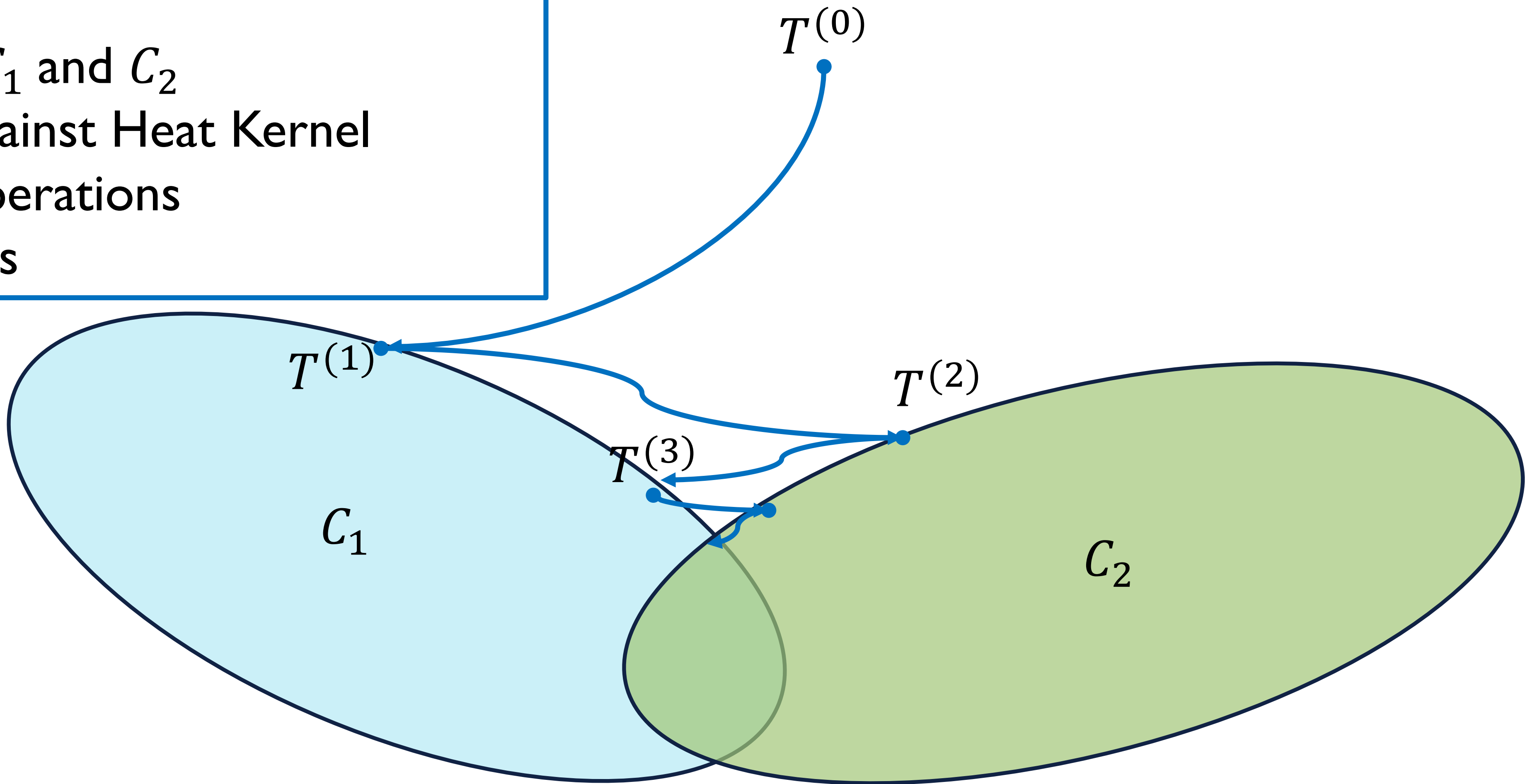
COMMONALITY OF ALGORITHMS

for $i = 1,2,3,\ldots$

Projections onto C_1 and C_2

- Convolutions against Heat Kernel
- Element-wise operations

- Compute final values



- Conclusion:
  - Main contribution of the work
  - Drawbacks
  - Questions

PLAN

MAIN CONTRIBUTION

- First Approximation: Entropy Regularization
- Second Approximation: Heat Kernel
- Alternating Projections algorithm

MAIN CONTRIBUTION

- First Approximation: Entropy Regularization → Convexity, Projection wrt KL
- Second Approximation: Heat Kernel
- Alternating Projections algorithm

MAIN CONTRIBUTION

- First Approximation: Entropy Regularization ⟶ Convexity, Projection wrt KL
- Second Approximation: Heat Kernel ⟶ Geodesic distances by linear equation
- Alternating Projections algorithm

MAIN CONTRIBUTION

- First Approximation: Entropy Regularization → Convexity, Projection wrt KL
- Second Approximation: Heat Kernel → Geodesic distances by linear equation
- Alternating Projections algorithm → Linear number of unknowns, the result by Iterations of element-wise operations and linear equation solutions

DRAWBACKS

- Solution is unstable when $\gamma \to 0$

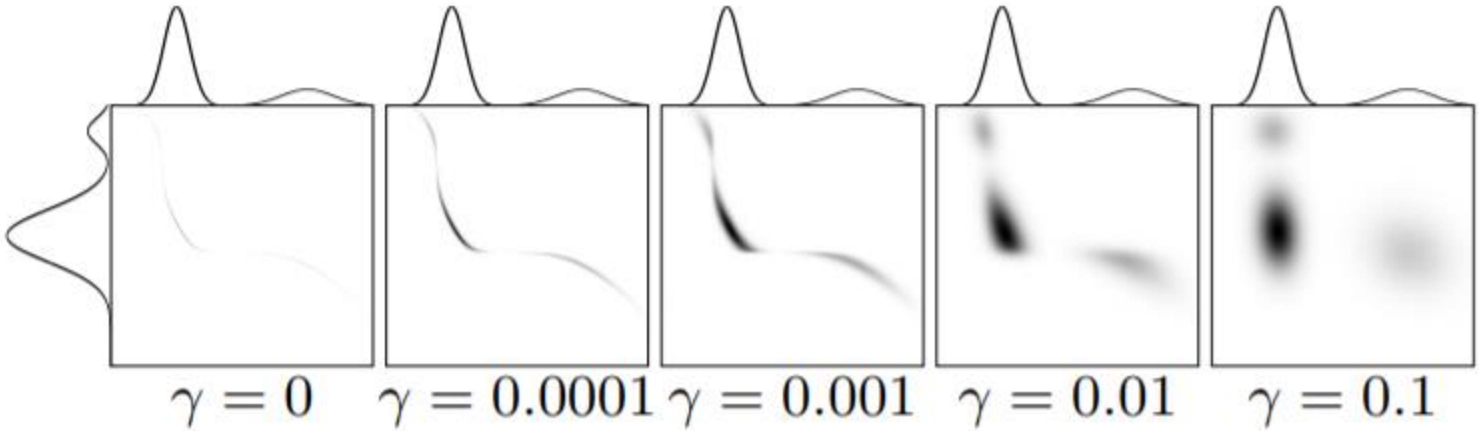


Figure 2: *Transportation plans with different values of γ, with 1D quadratic costs; $\mu_0, \mu_1 \in \mathrm{Prob}([0,1])$ are shown on the axes.*

CONVOLUTIONAL WASSERSTEIN DISTANCES: EFFICIENT OPTIMAL TRANSPORTATION ON GEOMETRIC DOMAINS

JUSTIN SOLOMON, FERNANDO DE GOES, GABRIEL PEYRE, MARCO CUTURI, ADRIAN BUTSCHER, ANDY NGUYEN, TAO DU, LEONIDAS GUIBAS

Questions?

Presenter: Zhakshylyk Nurlanov
Supervisor: Marvin Eisenberger