Optimal Transport

Justin Solomon

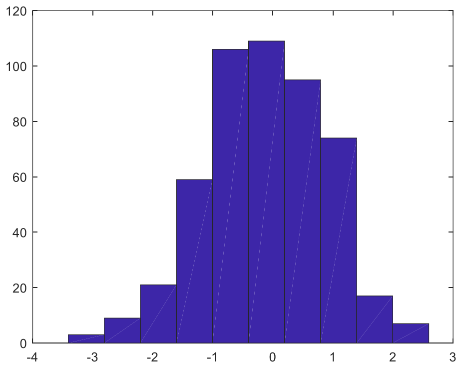
6.838: Shape Analysis

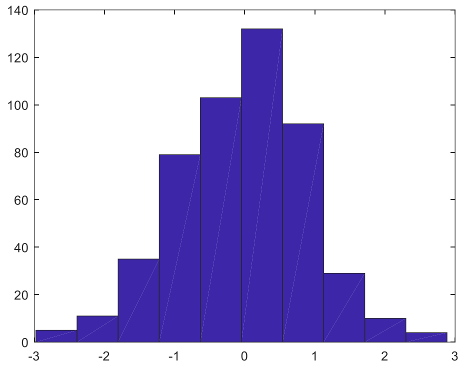
Spring 2021

MIT EECS

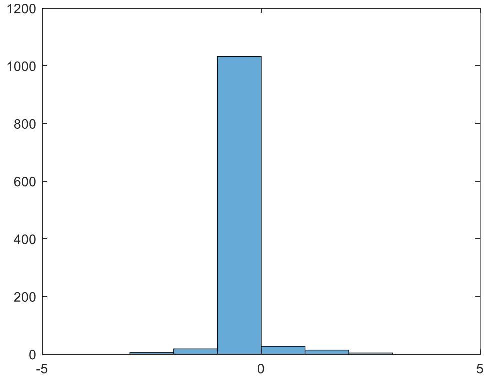
Motivation

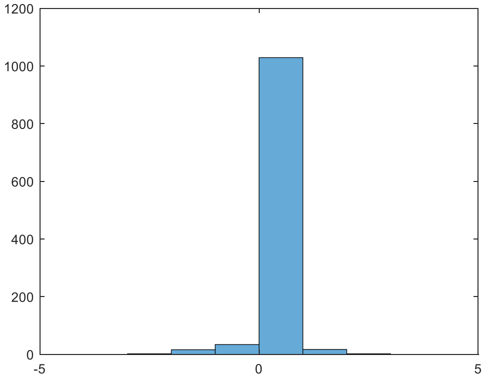
Theory

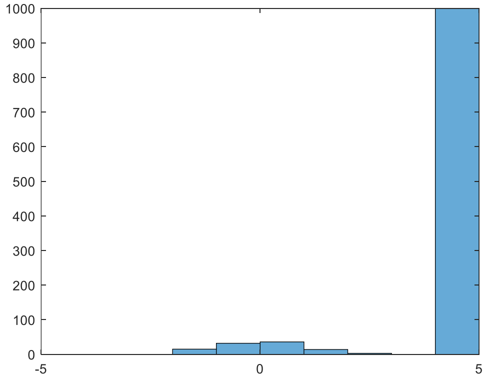




Practice







What is Optimal Transport?

A geometric way
to compare
probability measures.

Nobel prize

Fields medal
(and French politician)


Monge


Kantorovich


Dantzig


Wasserstein


Brenier


Otto

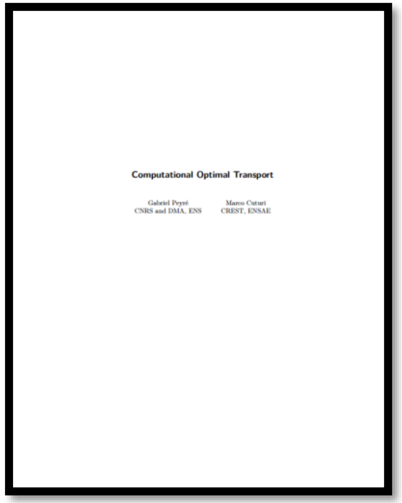

McCann

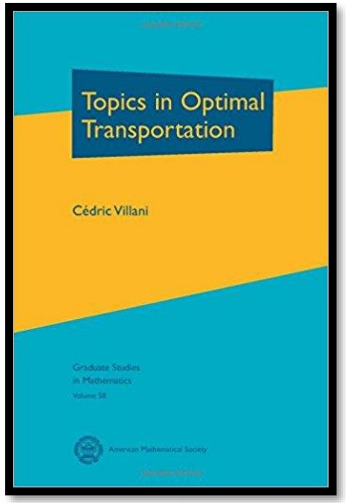

Villani

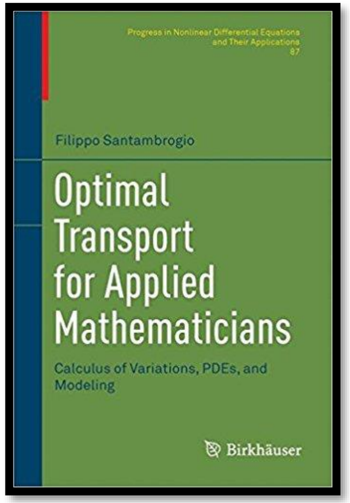
Plan For Today

1. Introduction to optimal transport
   - Construction
   - Many formulas
2. Applications
3. Discrete/discretized transport
   - Entropic regularization
   - Eulerian transport
   - Semidiscrete transport
4. Extensions & frontiers

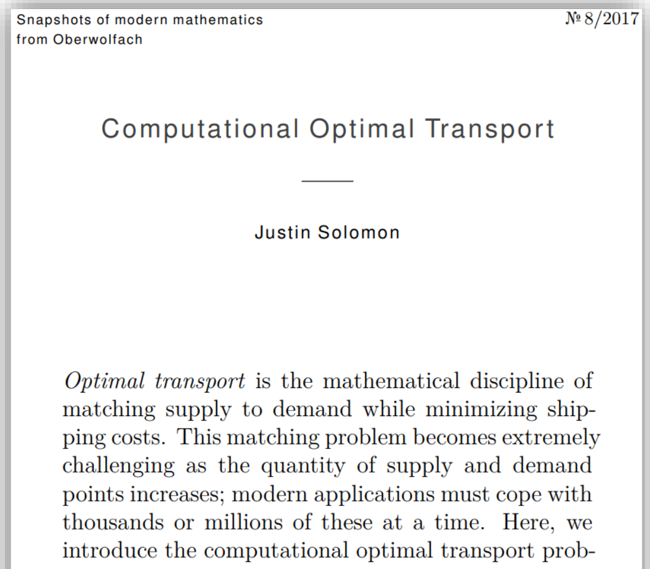
Useful References







Shameless self-promotion:

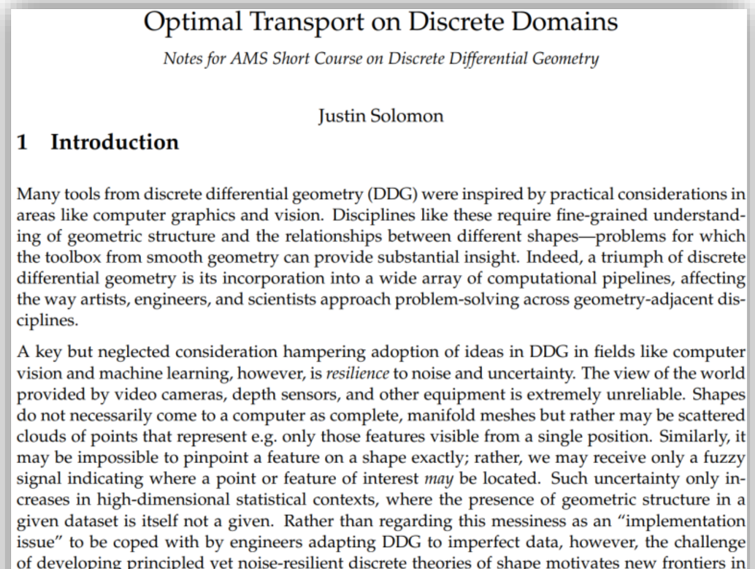

Snapshots of modern mathematics from Oberwolfach

№ 8/2017

Computational Optimal Transport

Justin Solomon

Optimal transport is the mathematical discipline of matching supply to demand while minimizing shipping costs. This matching problem becomes extremely challenging as the quantity of supply and demand points increases; modern applications must cope with thousands or millions of these at a time. Here, we introduce the computational optimal transport prob-


Optimal Transport on Discrete Domains

Notes for AMS Short Course on Discrete Differential Geometry

Justin Solomon

1 **Introduction**

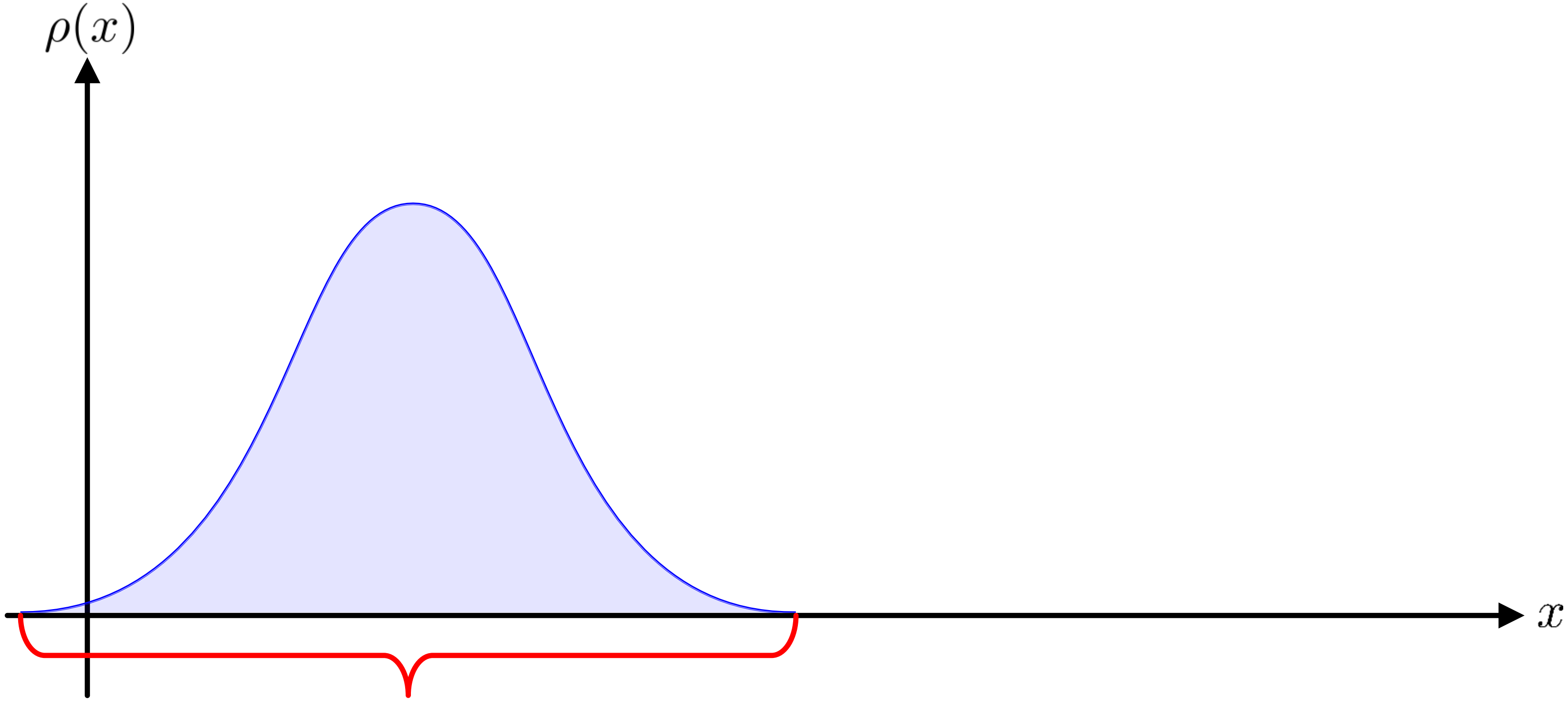
Many tools from discrete differential geometry (DDG) were inspired by practical considerations in areas like computer graphics and vision. Disciplines like these require fine-grained understanding of geometric structure and the relationships between different shapes—problems for which the toolbox from smooth geometry can provide substantial insight. Indeed, a triumph of discrete differential geometry is its incorporation into a wide array of computational pipelines, affecting the way artists, engineers, and scientists approach problem-solving across geometry-adjacent disciplines.

A key but neglected consideration hampering adoption of ideas in DDG in fields like computer vision and machine learning, however, is *resilience* to noise and uncertainty. The view of the world provided by video cameras, depth sensors, and other equipment is extremely unreliable. Shapes do not necessarily come to a computer as complete, manifold meshes but rather may be scattered clouds of points that represent e.g. only those features visible from a single position. Similarly, it may be impossible to pinpoint a feature on a shape exactly; rather, we may receive only a fuzzy signal indicating where a point or feature of interest *may* be located. Such uncertainty only increases in high-dimensional statistical contexts, where the presence of geometric structure in a given dataset is itself not a given. Rather than regarding this messiness as an "implementation issue" to be coped with by engineers adapting DDG to imperfect data, however, the challenge of developing principled yet noise-resilient discrete theories of shape motivates new frontiers in

Plan For Today

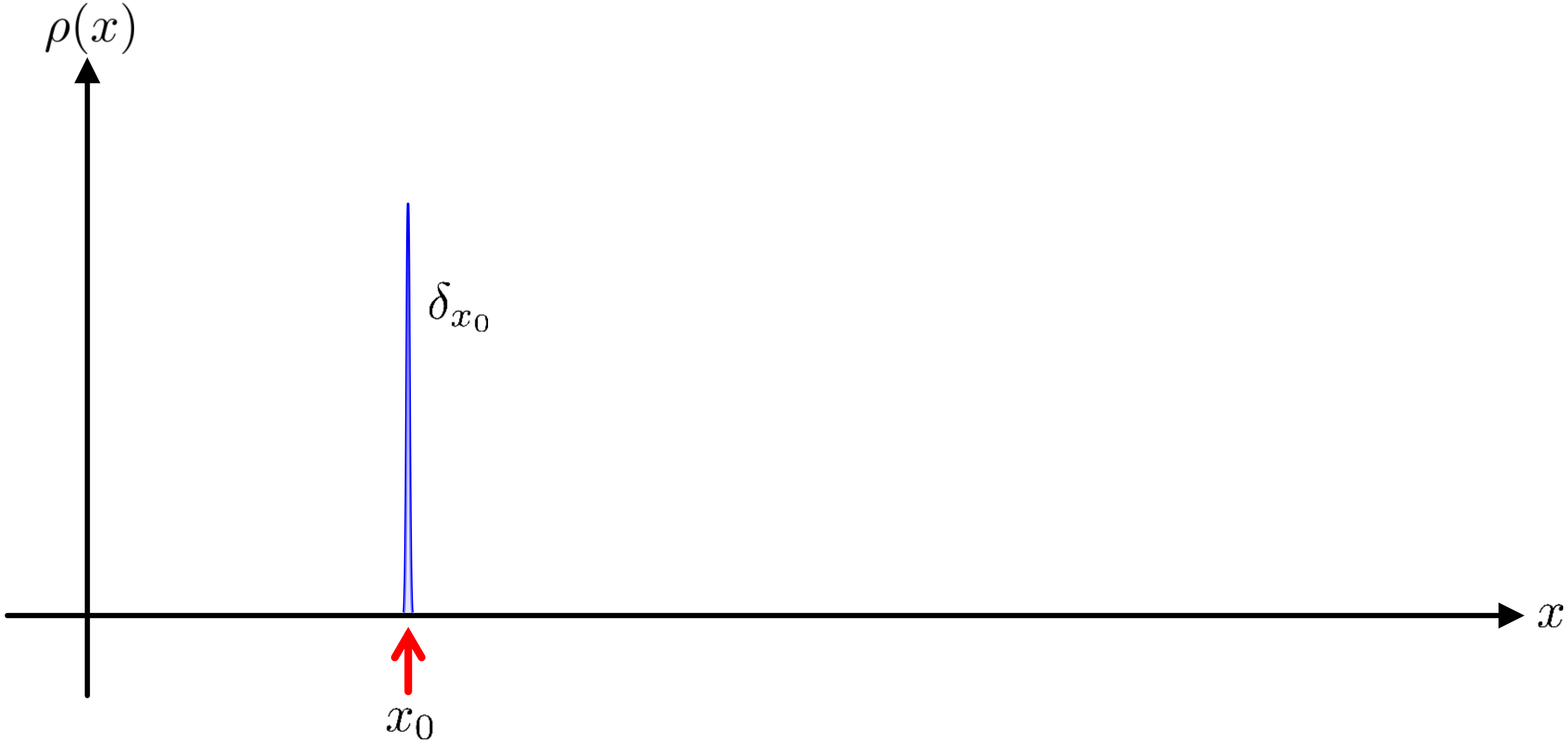
1. Introduction to optimal transport
   - Construction
   - Many formulas
2. Applications
3. Discrete/discretized transport
   - Entropic regularization
   - Eulerian transport
   - Semidiscrete transport
4. Extensions & frontiers

Probability as Geometry



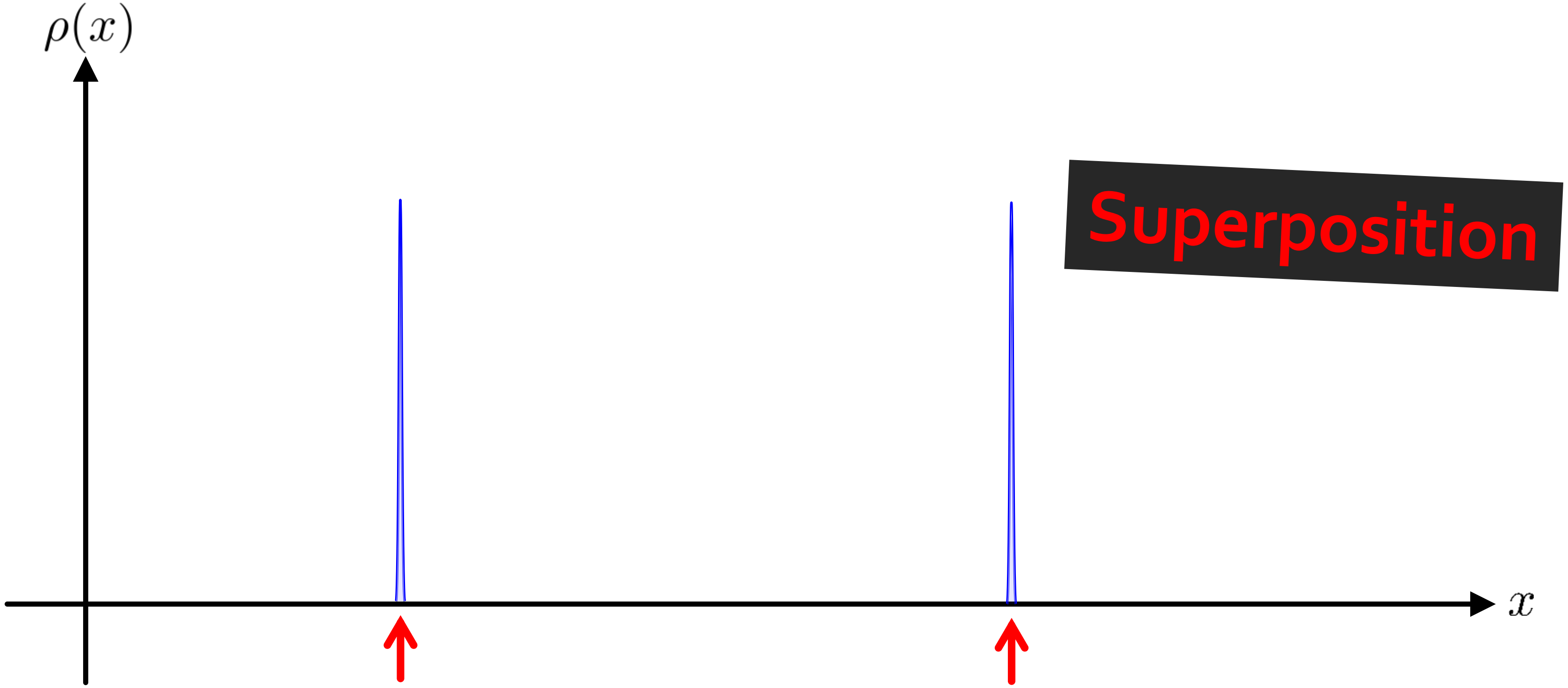
“Somewhere over here.”

Probability as Geometry



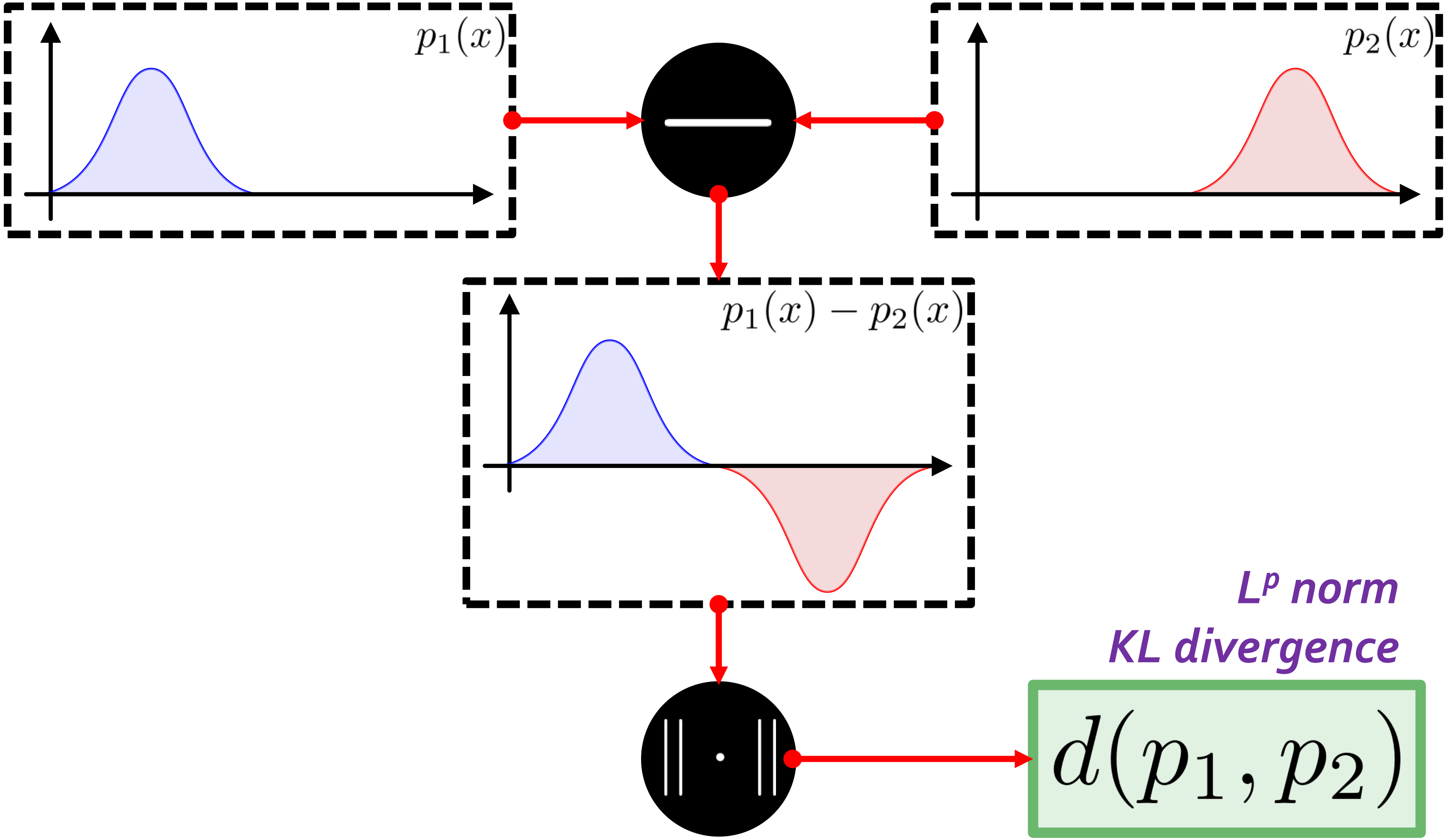
“Exactly here.”

Probability as Geometry

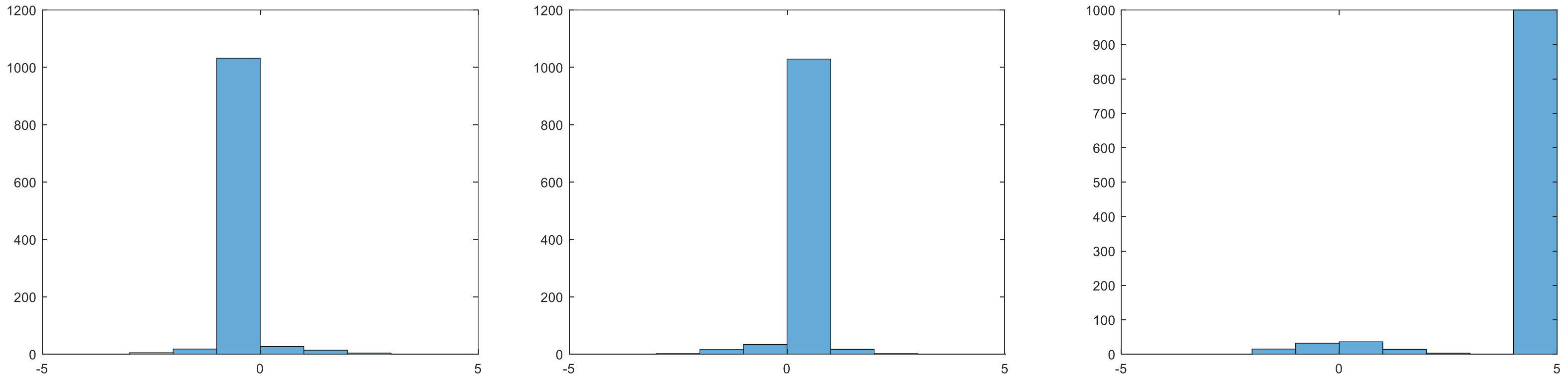


"One of these two places."

How We Compute Distances



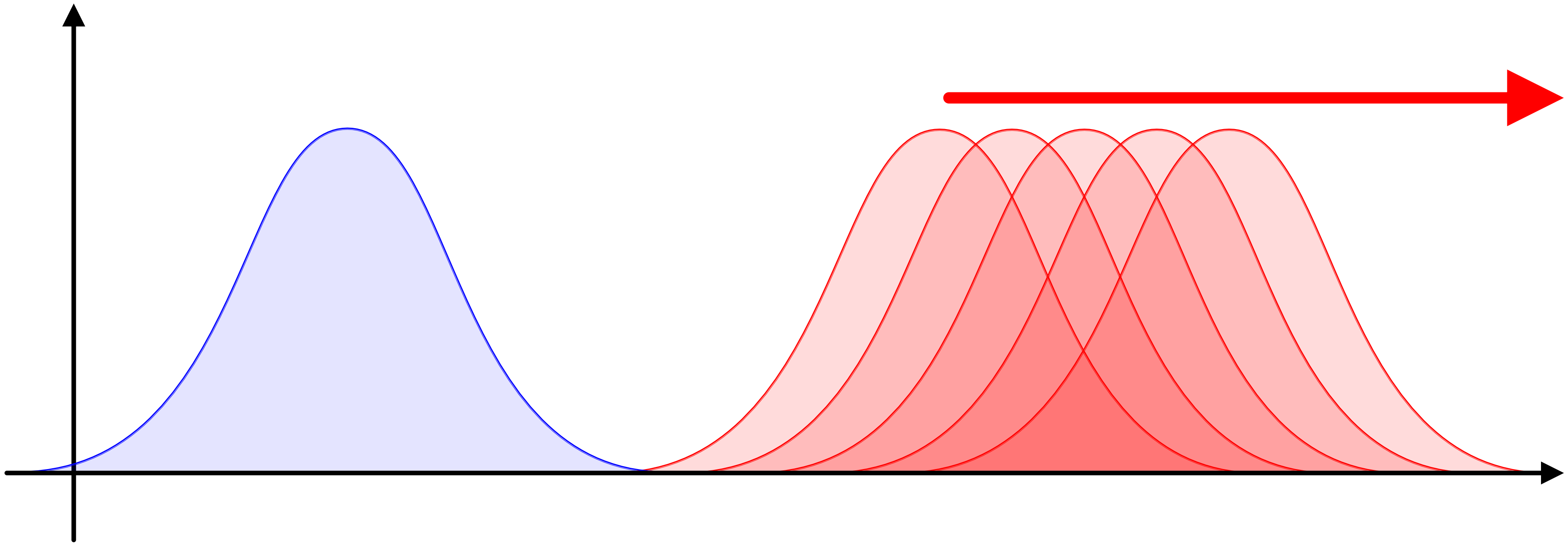
Equidistant!



$$\|p - q\|_1 = \sum_i |p_i - q_i|$$

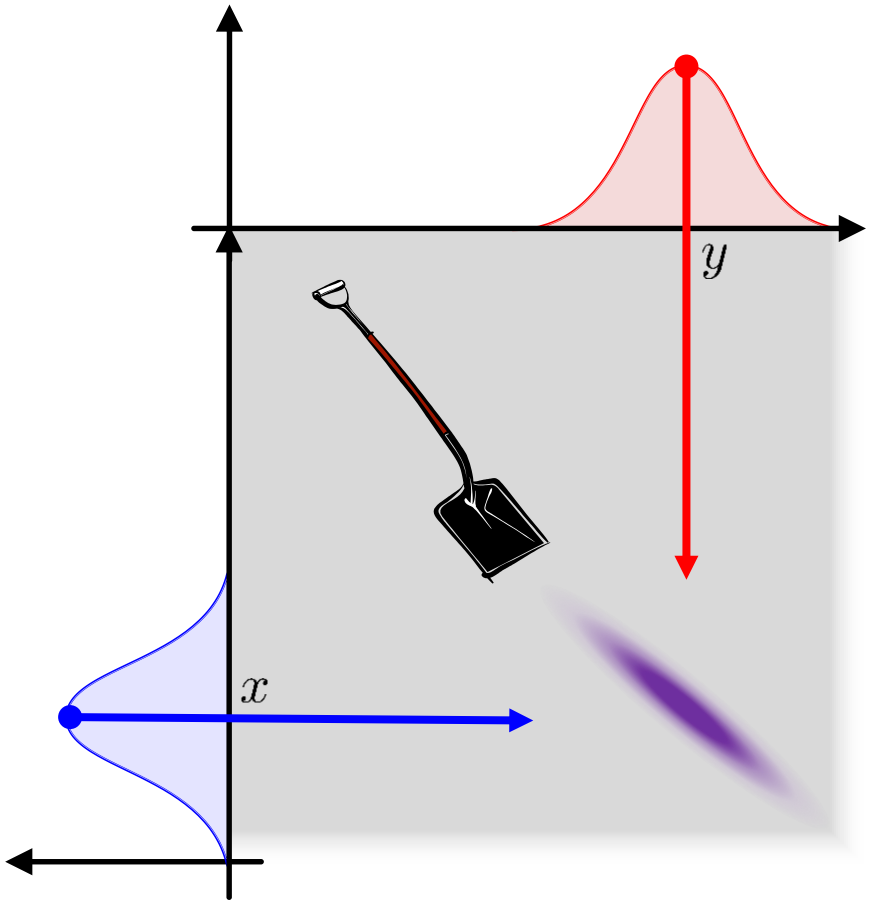
$$\mathrm{KL}(p\|q) = -\sum_i p_i \log \frac{q_i}{p_i}$$

What’s Wrong?



Measured overlap, not displacement.

Alternative Idea



Cost to move mass m from x to y:

$$m \cdot d(x, y)$$

Match mass from the distributions

Observation

Even the laziest shoveler

must do some work.

Property of the distributions themselves!

My house!

Measure Coupling

$$\pi(x,y) := \text{Amount moved from } x \text{ to } y$$

$$\pi(x,y) \geq 0 \ \forall x \in X, y \in Y$$

Mass is positive

$$\int_Y \pi(x,y)\,dy = \rho_0(x) \ \forall x \in X$$

Must scoop everything up

$$\int_X \pi(x,y)\,dx = \rho_1(y) \ \forall y \in Y$$

Must cover the target

Kantorovich Problem

$$\mathrm{OT}(\mu, \nu; c) := \min_{\pi \in \Pi(\mu,\nu)} \iint_{X \times Y} c(x, y)\, d\pi(x, y)$$

General transport problem

p-Wasserstein Distance

$$\mathcal{W}_p(\mu,\nu) \equiv \min_{\pi\in\Pi(\mu,\nu)} \left(\iint_{X\times X} d(x,y)^p \, d\pi(x,y) \right)^{1/p}$$

Shortest path distance

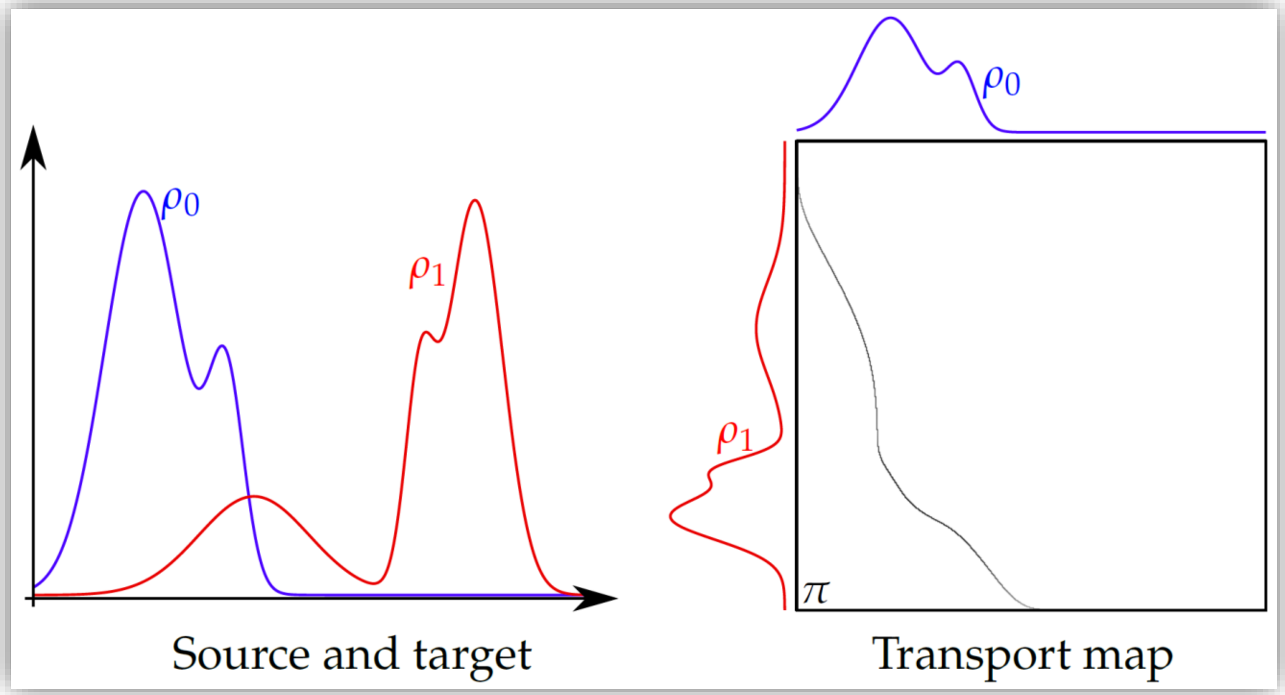
Expectation

Geodesic distance $d(x,y)$

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

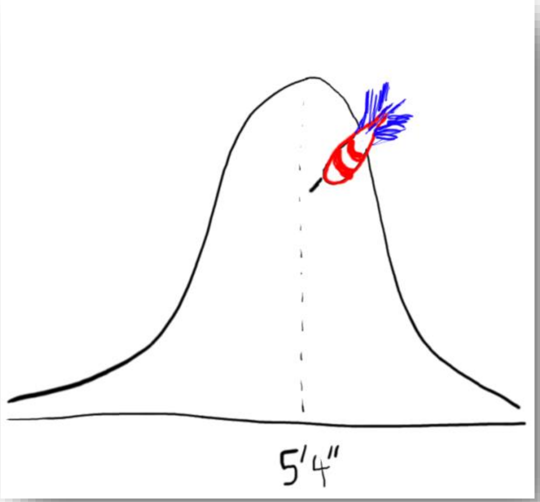
1-Wasserstein in 1D

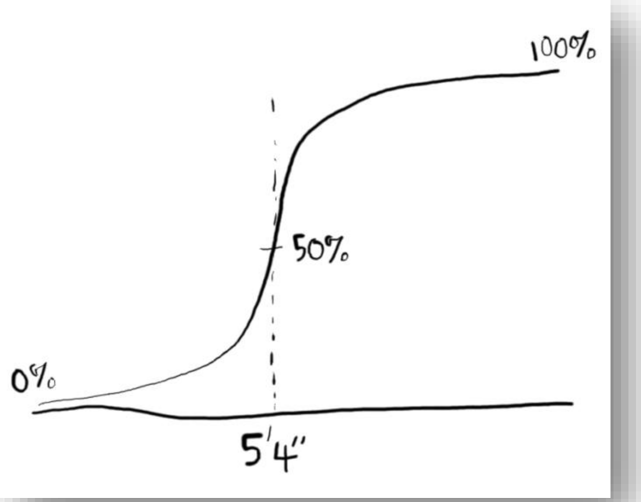
$$\mathcal{W}_1(\rho_0,\rho_1) := \begin{cases} \min_\pi & \iint_{\mathbb{R}\times\mathbb{R}} \pi(x,y)|x-y|\,dx\,dy & \text{Minimize total work} \\ \text{s.t.} & \pi \geq 0\ \forall x,y \in \mathbb{R} & \text{Nonnegative mass} \\ & \int_{\mathbb{R}} \pi(x,y)\,dy = \rho_0(x)\ \forall x \in \mathbb{R} & \text{Starts from } \rho_0 \\ & \int_{\mathbb{R}} \pi(x,y)\,dx = \rho_1(y)\ \forall y \in \mathbb{R} & \text{Ends at } \rho_1 \end{cases}$$

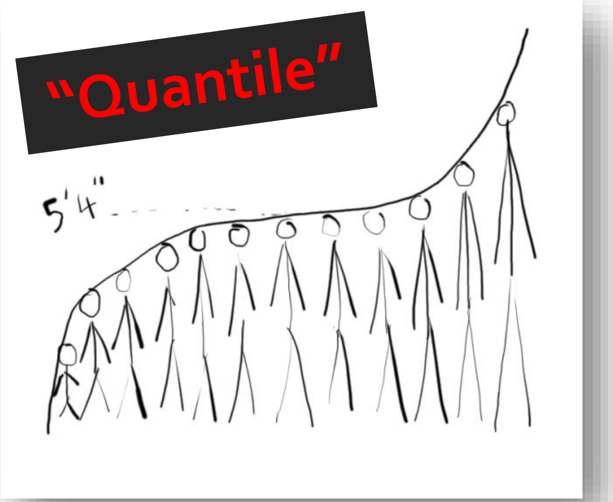


In One Dimension: Closed-Form

http://realgl.blogspot.com/2013/01/pdf-cdf-inv-cdf.html







PDF ⋯⋯▶ **[CDF]** ⋯⋯▶ **CDF^{-1}**

$$\mathcal{W}_1(\mu,\nu) = \int_{-\infty}^{\infty} |\text{CDF}(\mu) - \text{CDF}(\nu)|\, d\ell$$

Doesn't extend past 1d!

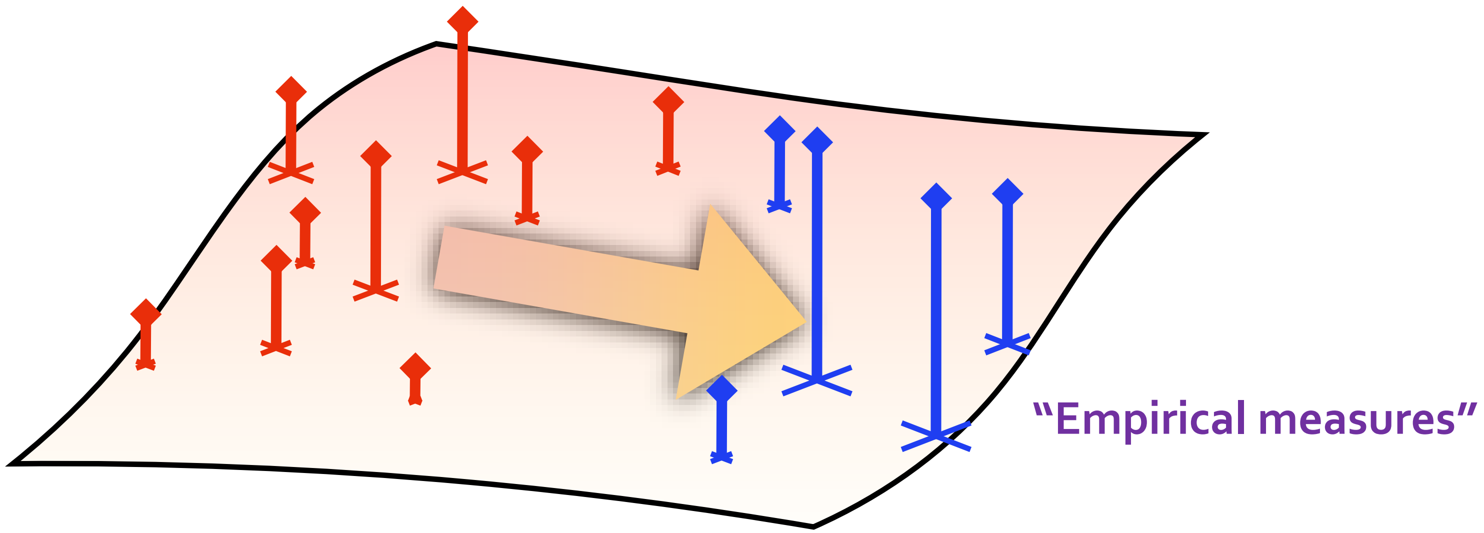
$$\mathcal{W}_2^2(\mu,\nu) = \int_{-\infty}^{\infty} \left(\text{CDF}^{-1}(\mu) - \text{CDF}^{-1}(\nu)\right)^2\, d\ell$$

Fully-Discrete Transport

$$[\mathcal{W}_p(\mu_0,\mu_1)]^p = \begin{cases} \min_{T\in\mathbb{R}^{k_0\times k_1}} & \sum_{ij} T_{ij}|x_{0i}-x_{1j}|^p \\ \text{s.t.} & T\geq 0 \\ & \sum_j T_{ij} = a_{0i} \\ & \sum_i T_{ij} = a_{1j} \end{cases}$$

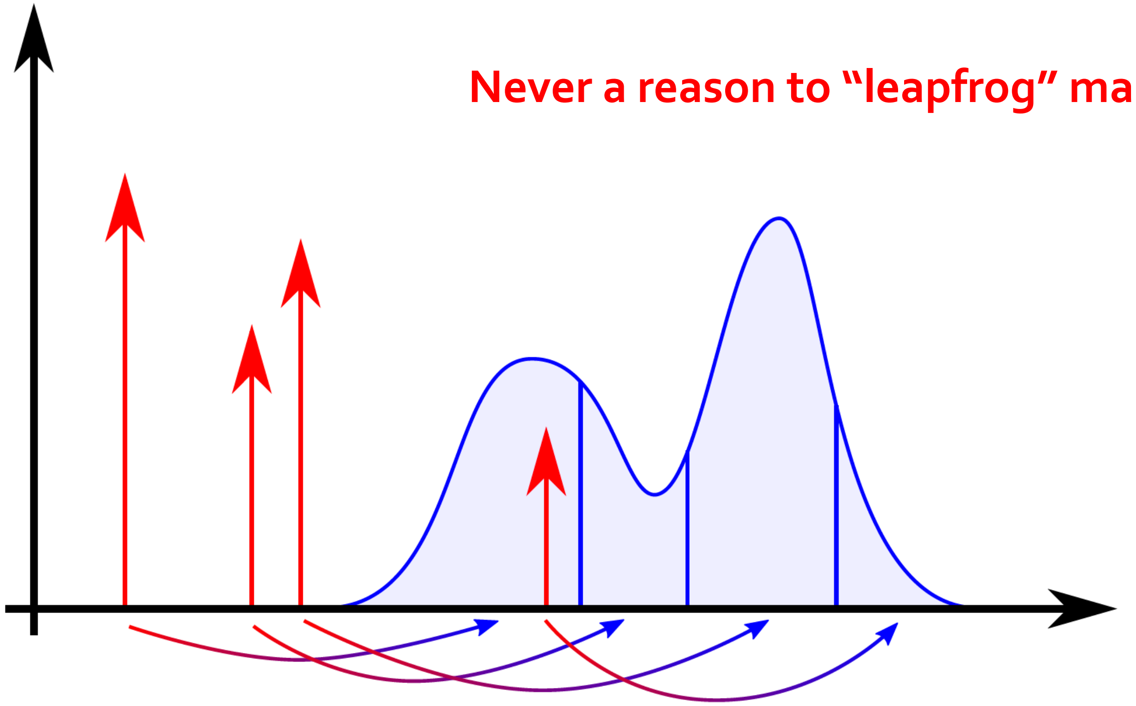
Linear program: Finite number of variables

Algorithms: Simplex, interior point, auction, …



Semidiscrete Transport

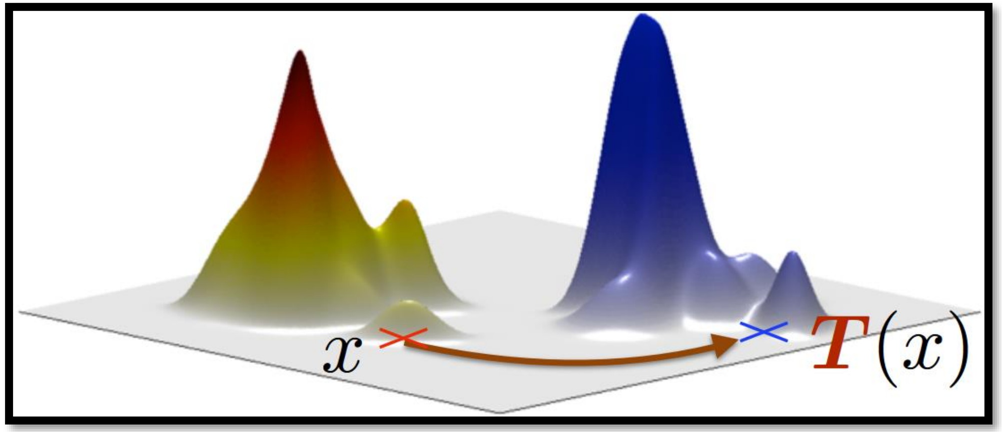
$$\mu_0 := \sum_{i=1}^{k_0} a_{0i}\delta_{x_{0i}} \qquad \mu_1(S) := \int_S \rho_1(x)\,dx$$



Never a reason to “leapfrog” mass!

Monge Formulation

$$\inf_{\phi_\sharp \rho_0 = \rho_1} \int_{-\infty}^{\infty} c(x, \phi(x))\rho_0(x)\, dx$$



[Monge 1781]; image courtesy Marco Cuturi

Not always well-posed!

Plan For Today

1. Introduction to optimal transport
   - Construction
   - Many formulas

2. Applications

3. Discrete/discretized transport
   - Entropic regularization
   - Eulerian transport
   - Semidiscrete transport

4. Extensions & frontiers

Example: Discrete Transport

$$X = \{1, 2, \ldots, k_1\}, Y = \{1, 2, \ldots, k_2\}$$

"Earth Mover's Distance"

$$\mathrm{OT}(v, w; C) = \begin{cases} \min_{T \in \mathbb{R}^{k_1 \times k_2}} & \sum_{ij} T_{ij} c_{ij} \\ \text{s.t.} & T \geq 0 \\ & \sum_j T_{ij} = v_i \ \forall i \in \{1, \ldots, k_1\} \\ & \sum_i T_{ij} = w_j \ \forall j \in \{1, \ldots, k_2\}. \end{cases}$$

Metric when d(x,y) satisfies the triangle inequality.

"The Earth Mover's Distance as a Metric for Image Retrieval"
Rubner, Tomasi, and Guibas; IJCV 40.2 (2000): 99—121.

Revised in:

"Ground Metric Learning"
Cuturi and Avis; JMLR 15 (2014)

Kantorovich Duality

$$\mathrm{OT}(\mu,\nu;c) := \begin{cases} \min_\pi \ \iint_{X\times Y} c(x,y)\, d\pi(x,y) \\ \quad \text{s.t.} \ \ \pi \in \Pi(\mu,\nu) \end{cases}$$

Primal

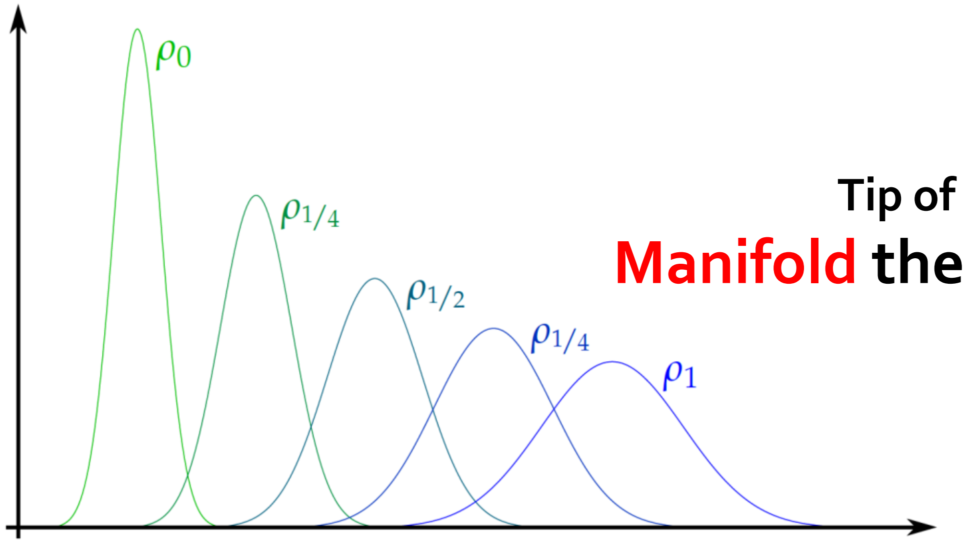
$$= \begin{cases} \max_{\phi,\psi} \ \int_X \phi(x)\, d\mu(x) + \int_Y \psi(y)\, d\nu(y) \\ \quad \text{s.t.} \ \ \phi(x) + \psi(y) \le c(x,y) \text{ for a.e. } x \in X, y \in Y \end{cases}$$

Dual

Flow-Based W_2

$$\mathcal{W}_2^2(\rho_0,\rho_1)=\begin{cases}\inf_{\rho,v} \iint_{M\times[0,1]} \frac{1}{2}\rho(x,t)\|v(x,t)\|^2\,dx\,dt\\ \quad \text{s.t. } \nabla\cdot(\rho(x,t)v(x,t)) = \frac{\partial\rho(x,t)}{dt}\\ \quad\quad v(x,t)\cdot\hat{n}(x) = 0\ \forall x\in\partial M\\ \quad\quad \rho(x,0) = \rho_0(x)\\ \quad\quad \rho(x,1) = \rho_1(x)\end{cases}$$

[Benamou & Brenier 2000]



Tip of an iceberg:

Manifold theory of transport!

Plan For Today

1. Introduction to optimal transport
    - Construction
    - Many formulas
2. Applications
3. Discrete/discretized transport
    - Entropic regularization
    - Eulerian transport
    - Semidiscrete transport
4. Extensions & frontiers

Wassersteinization

[wos-ur-stahyn-ahy-sey-sh*uh*-n]
noun.

Introduction of optimal transport into a computational problem.

cf. least-squarification, L_1ification, deep-netification, kernelization

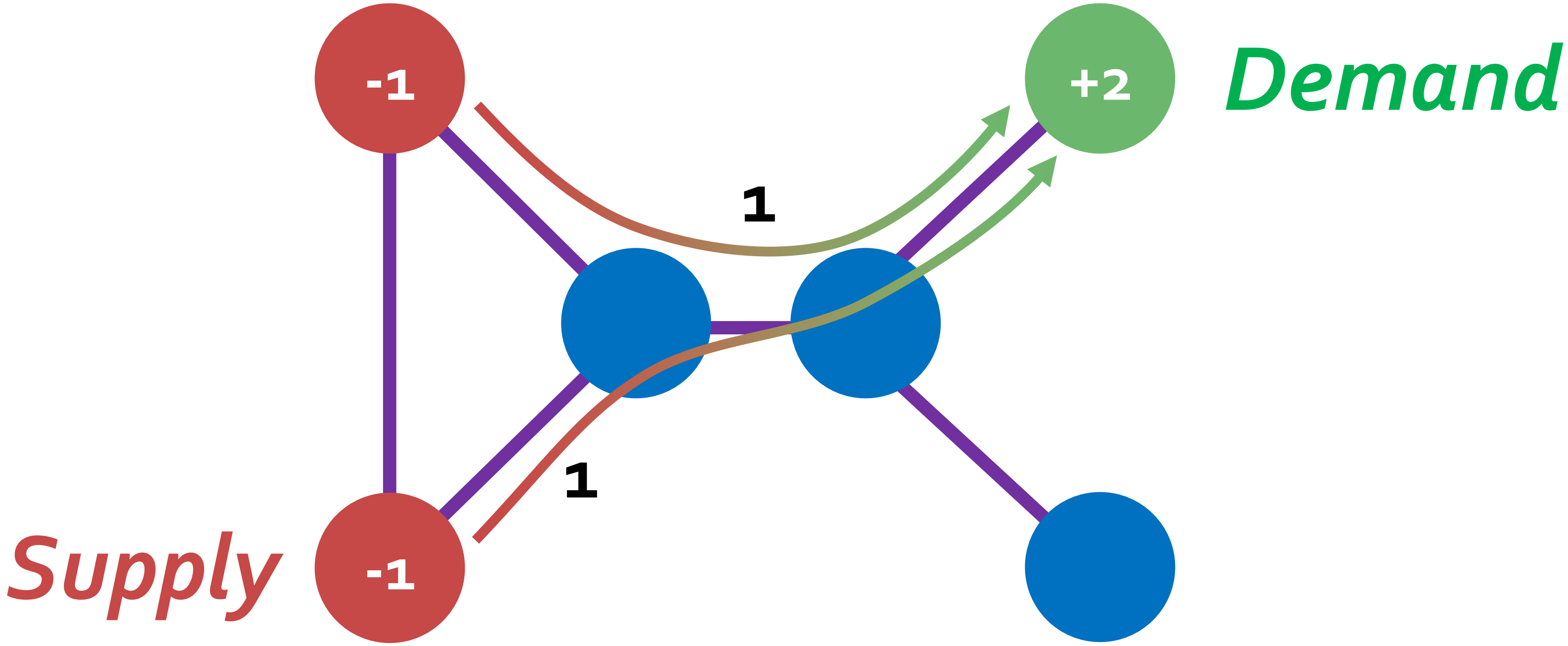
Key Ingredients

We have tools to

- Solve optimal transport problems numerically
- Differentiate transport distances in terms of their input distributions

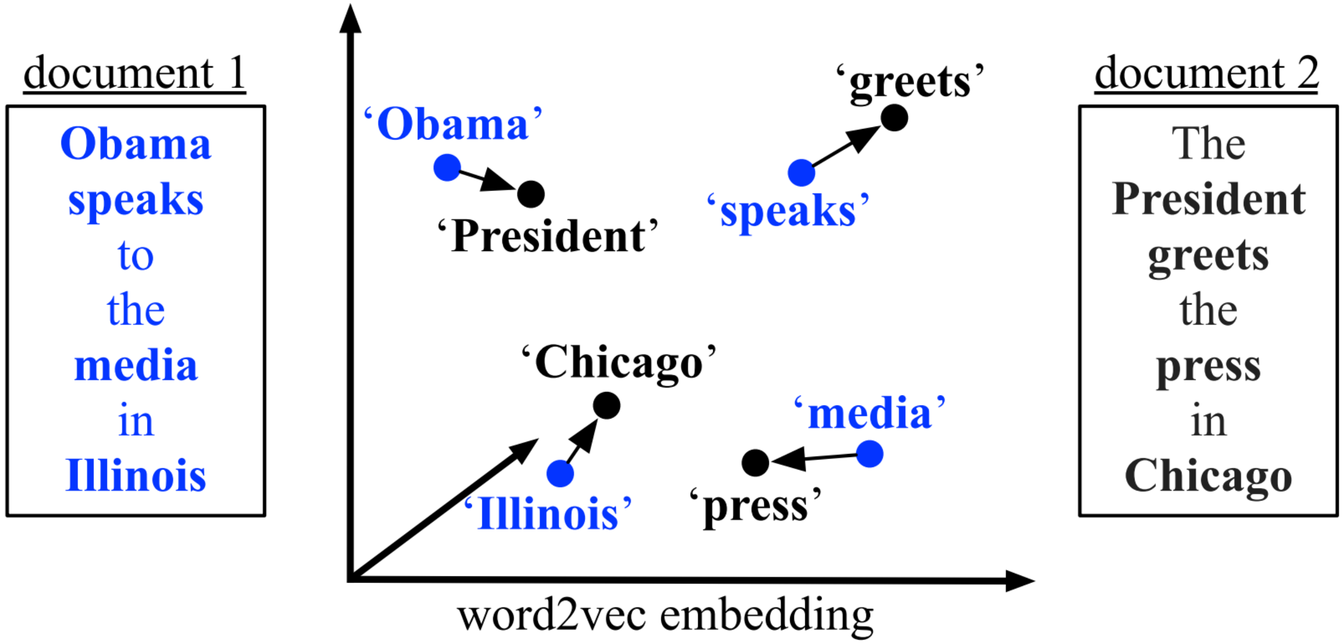
Bonus:
Transport cost from μ to ν is a convex function of μ and ν.

Operations and Logistics



Minimum-cost flow

Histograms and Descriptors



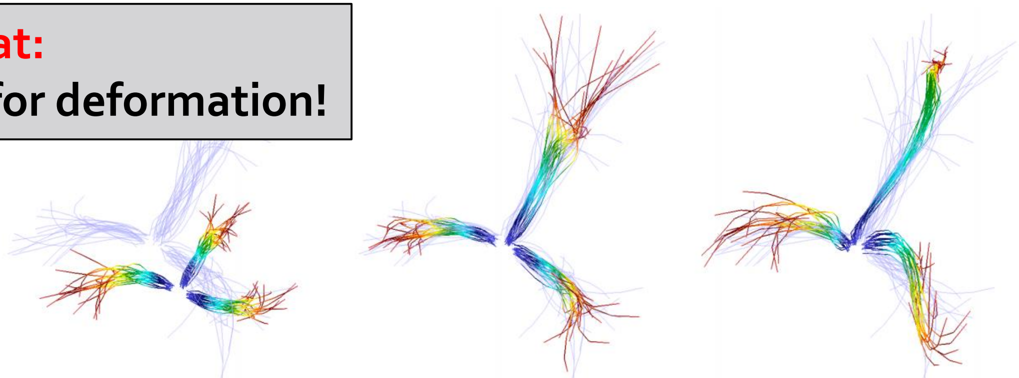
Use word embeddings

[Kusner et al. 2015]

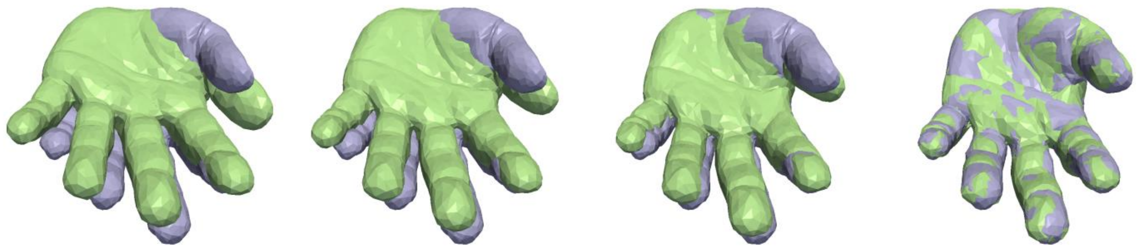
Word Mover's Distance (WMD)

Registration and Reconstruction

Caveat:
Not a good model for deformation!



(a) Dataset (b) OT fidelity (c) RKHS fidelity



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

Fig. 2. First row: Matching of fibres bundles. **Second row:** Matching of two hand surfaces using a balanced OT fidelity. Target is in purple.

[Feydy, Charlier, Vialard, and Peyré 2017]

Alignment

Engineering Design



EPFL Computer Graphics and Geometry Laboratory; Rayform SA

Interpolation

Image from [Lavenant, Claici, Chien, & Solomon 2018]

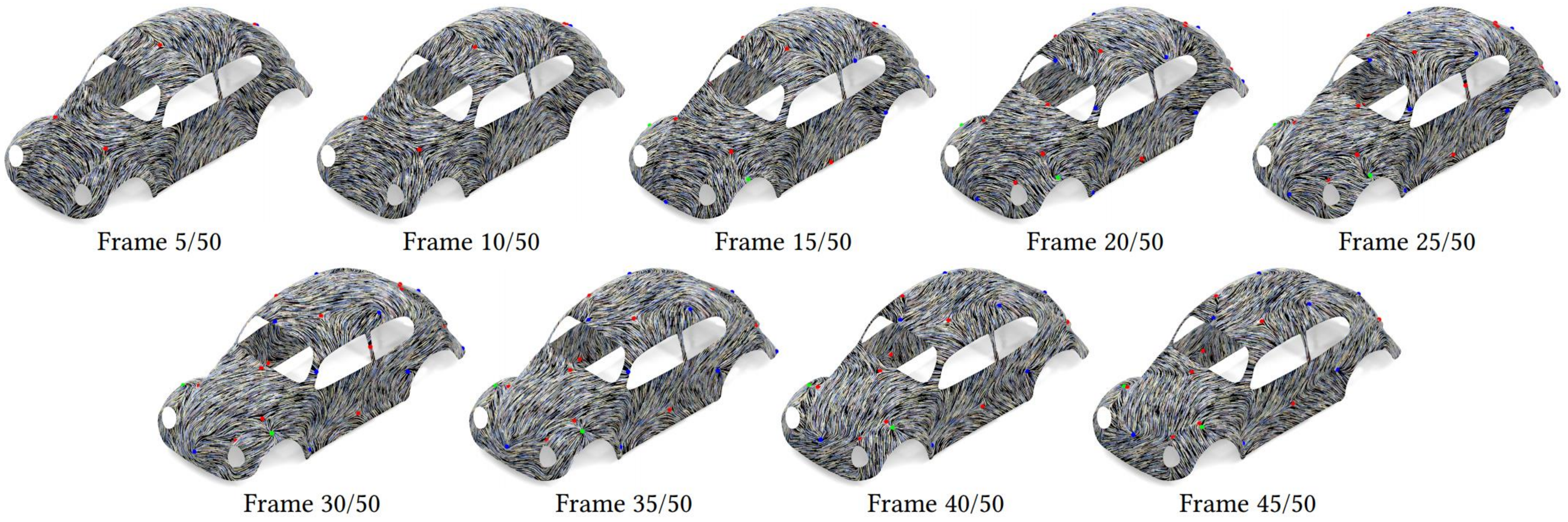
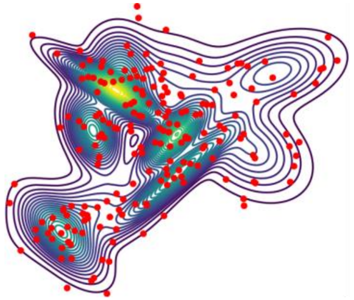
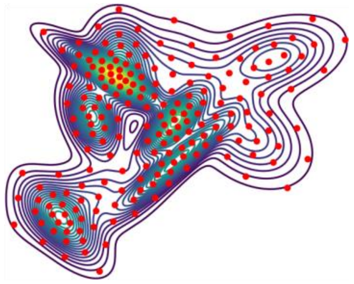


Image from [Vaxman & Solomon 2019]

Blue Noise and Distribution Approximation



Uniform samples

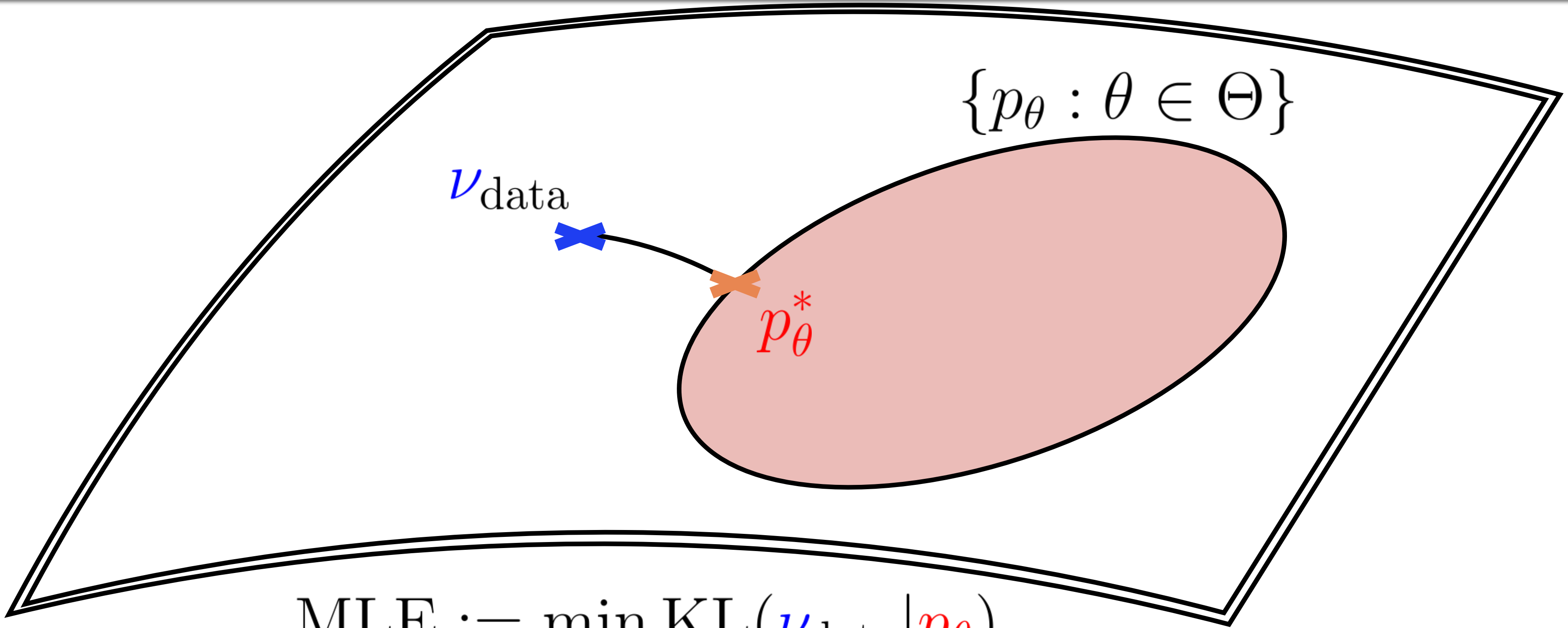


Transport

$$\min_{x_1,\ldots,x_n} \mathcal{W}_2^2\left(\mu, \frac{1}{n}\sum_i \delta_{x_i}\right)$$

Zebra image courtesy F. de Goes; photo by F. Durand; distribution image courtesy S. Claici

Statistical Estimation



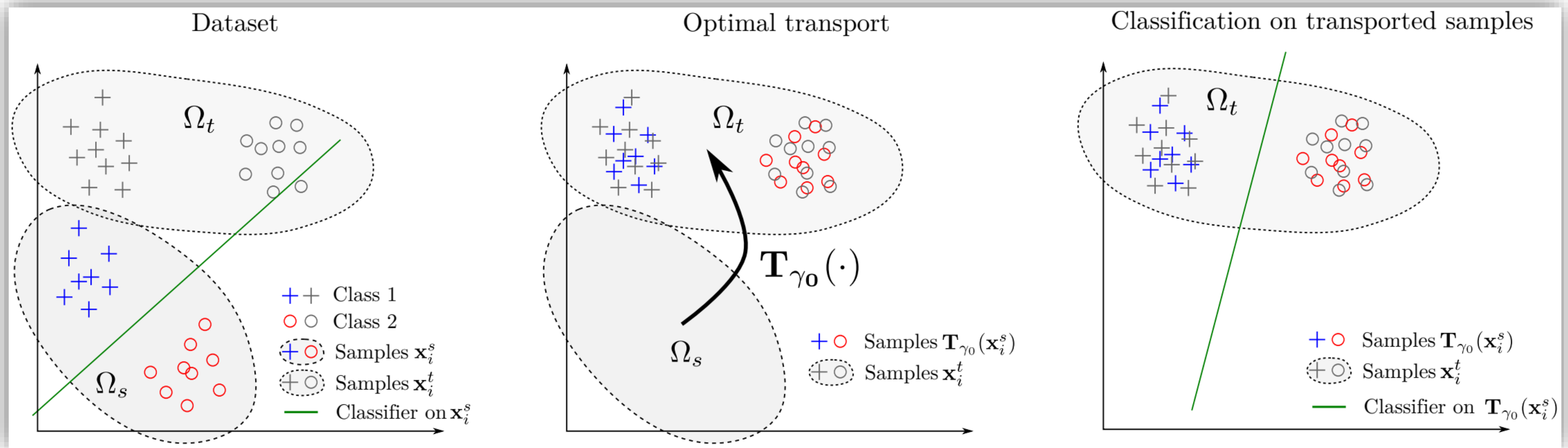
$$\text{MLE} := \min_{\theta \in \Theta} \text{KL}(\nu_{\text{data}} | p_\theta)$$

$$\longrightarrow \text{MKE} := \min_{\theta \in \Theta} \mathcal{W}_2(\nu_{\text{data}}, p_\theta)$$

[Bassetti 2006]

Minimum Kantorovich Estimator

Domain Adaptation



1. Estimate transport map
2. Transport labeled samples to new domain
3. Train classifier on transported labeled samples

[Courty et al. 2017]

Generative Adversarial Networks (GANs)

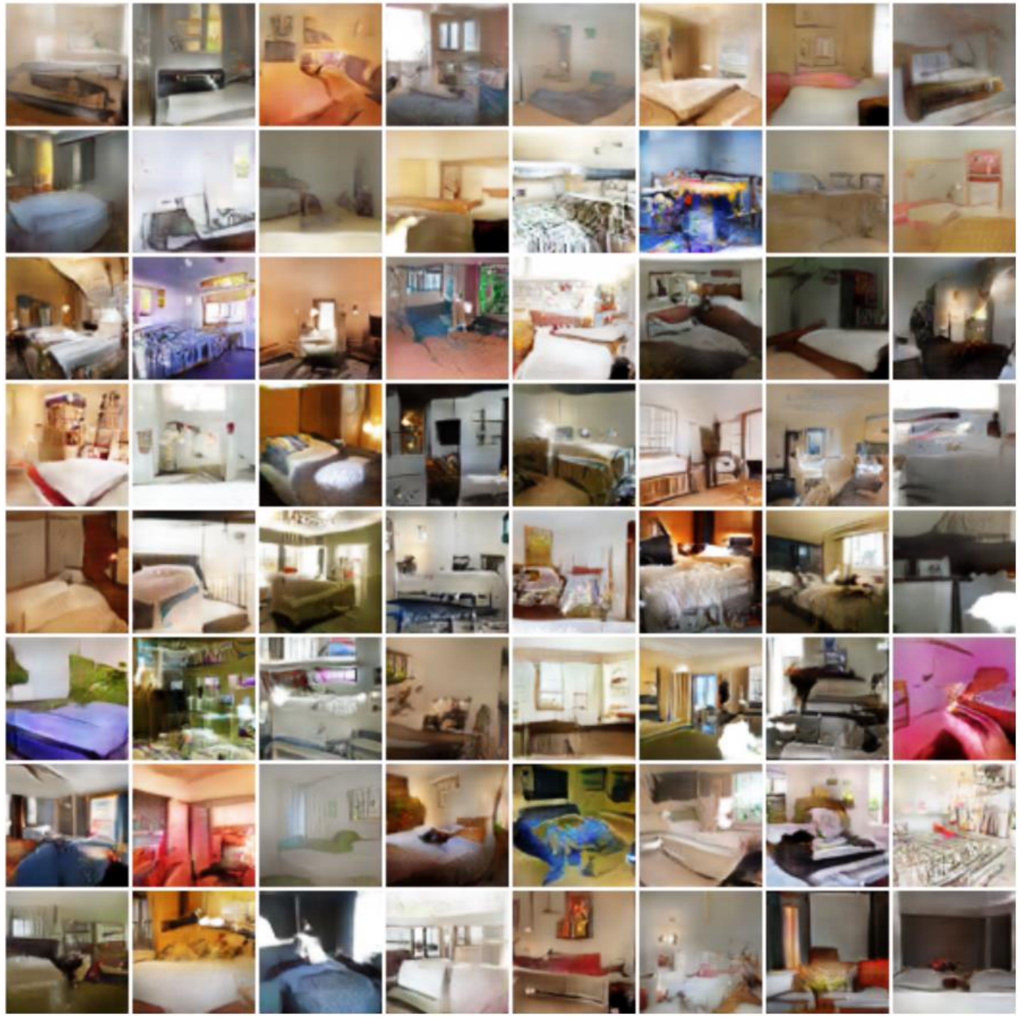


Figure 9: WGAN algorithm: generator and critic are DCGANs.

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

```
while θ has not converged do
    for t = 0, ..., n_critic do
        Sample {x^(i)}_{i=1}^m ~ P_r a batch from the real data.
        Sample {z^(i)}_{i=1}^m ~ p(z) a batch of prior samples.
        g_w ← ∇_w [ (1/m) Σ_{i=1}^m f_w(x^(i)) − (1/m) Σ_{i=1}^m f_w(g_θ(z^(i))) ]
        w ← w + α · RMSProp(w, g_w)
        w ← clip(w, −c, c)
    end for
    Sample {z^(i)}_{i=1}^m ~ p(z) a batch of prior samples.
    g_θ ← −∇_θ (1/m) Σ_{i=1}^m f_w(g_θ(z^(i)))
    θ ← θ − α · RMSProp(θ, g_θ)
end while
```

[Arjovsky et al. 2017]

Plan For Today

1. Introduction to optimal transport
    - Construction
    - Many formulas

2. Applications

3. Discrete/discretized transport
    - Entropic regularization
    - Eulerian transport
    - Semidiscrete transport

4. Extensions & frontiers

Theme in computation:

Same in theory, but different in practice

- Choose one of each:
  - Formulation
  - Discretization

Well-Known Theme

“No Free Lunch”

	SYM	LOC	LIN	POS	PSD	CON
MEAN VALUE	○	●	●	●	○	○
INTRINSIC DEL	●	○	●	●	●	?
COMBINATORIAL	●	●	○	●	●	○
COTAN	●	●	●	○	●	●

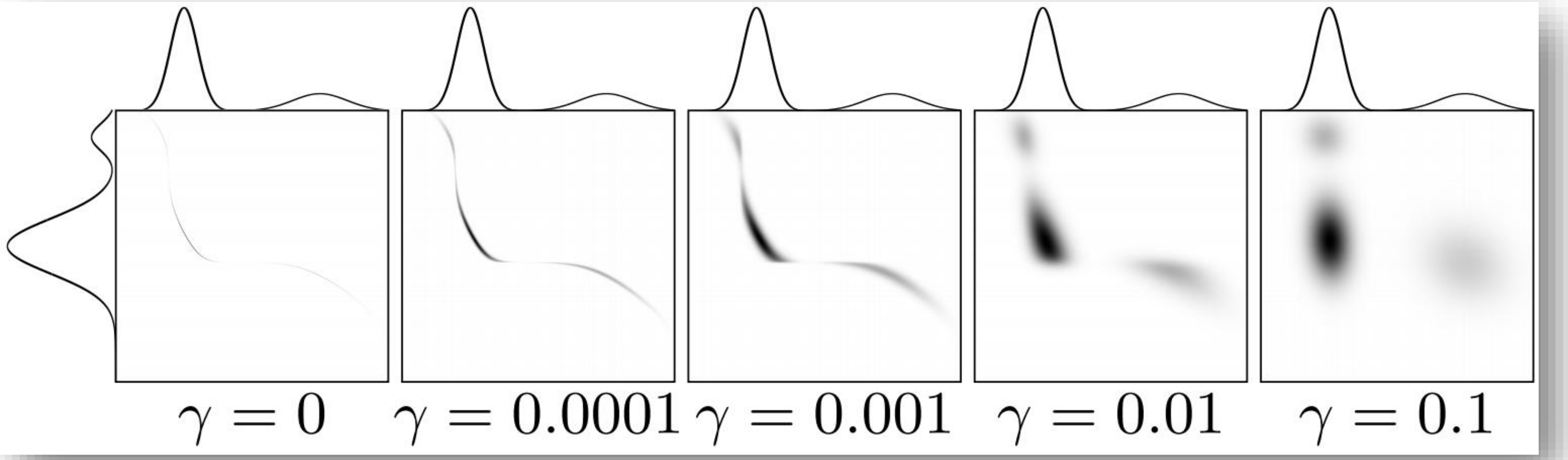
Observe that none of the Laplacians considered in graphics fulfill *all* desired properties. Even more: none of them satisfy the first four properties.

[Wardetzky et al. 2007]

Plan For Today

1. Introduction to optimal transport
    - Construction
    - Many formulas
2. Applications
3. Discrete/discretized transport
    - Entropic regularization
    - Eulerian transport
    - Semidiscrete transport
4. Extensions & frontiers

Entropic Regularization



$$\min_T \quad \sum_{ij} T_{ij} c_{ij} - \alpha H(T)$$
$$\text{s.t.} \quad \sum_j T_{ij} = p_i$$
$$\sum_i T_{ij} = q_j$$

OK to drop nonnegative constraint!

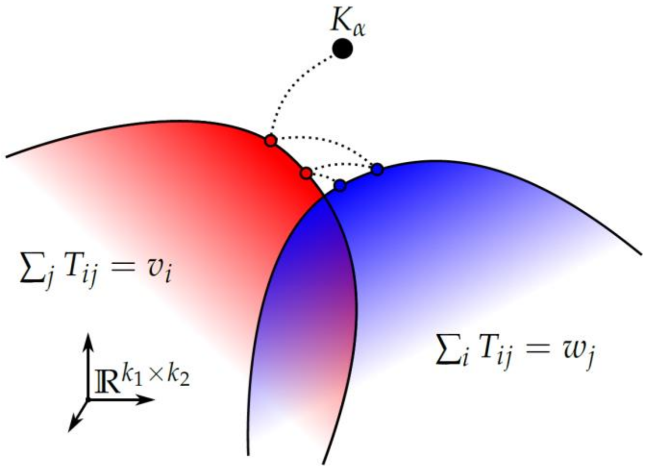
$$H(T) := -\sum_{ij} T_{ij} \log T_{ij}$$

Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport" (NIPS 2013)

Sinkhorn Algorithm

$$T = \mathrm{diag}(u) K_\alpha \mathrm{diag}(v),$$
$$\text{where } K_\alpha := \exp(-C/\alpha)$$

$$u \leftarrow p \oslash (K_\alpha v)$$
$$v \leftarrow q \oslash (K_\alpha^\top u)$$

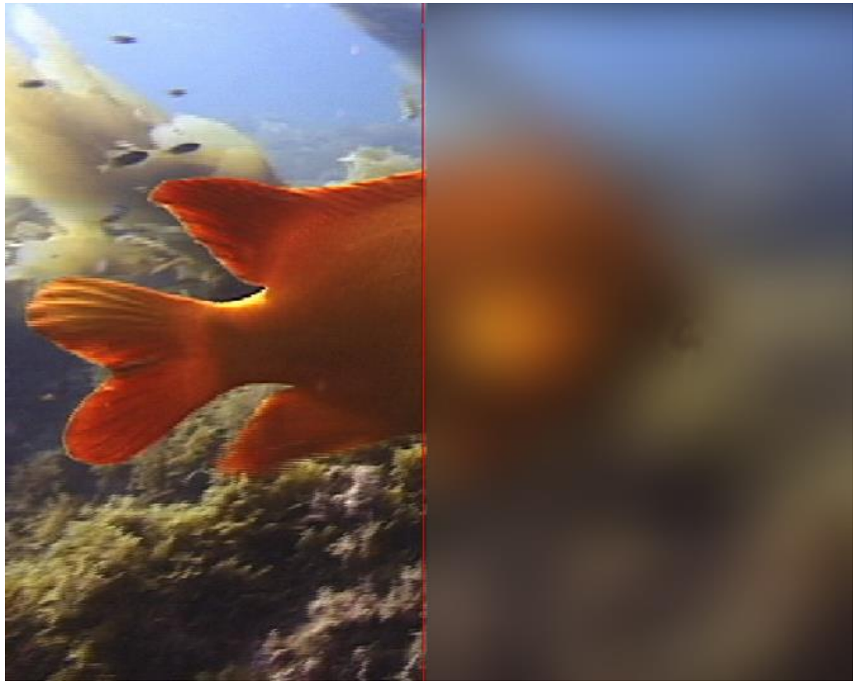


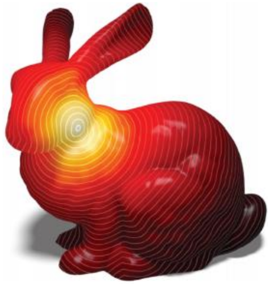
Sinkhorn & Knopp. "Concerning nonnegative matrices and doubly stochastic matrices". Pacific J. Math. 21, 343–348 (1967).

Alternating projection

Ingredients for Sinkhorn

1. Supply vector p
2. Demand vector q
3. Multiplication by K





$$K_{ij} = e^{-c_{ij}/\alpha}$$

Solomon et al. "Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains." SIGGRAPH 2015.

Sinkhorn Divergences

$$\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu) := 2\mathcal{W}_{c,\varepsilon}(\mu,\nu) - \mathcal{W}_{c,\varepsilon}(\mu,\mu) - \mathcal{W}_{c,\varepsilon}(\nu,\nu)$$

- Debiases entropy-regularized transport near zero
- Easy to compute: Three calls to Sinkhorn
- Links optimal transport to maximum mean discrepancy (MMD)

$$\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu) \xrightarrow{\varepsilon \to 0} 2\mathcal{W}_{c}(\mu,\nu)$$

$$\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu) \xrightarrow{\varepsilon \to \infty} \mathrm{MMD}_{-c}(\mu,\nu)$$

[Genevay et al. 2016]

Plan For Today

1. Introduction to optimal transport
   - Construction
   - Many formulas

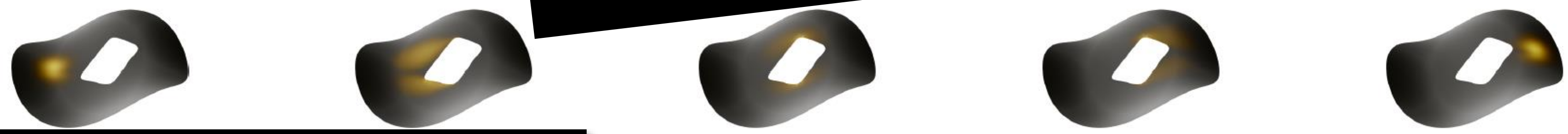
2. Applications

3. Discrete/discretized transport
   - Entropic regularization
   - Eulerian transport
   - Semidiscrete transport

4. Extensions & frontiers

Discretization

<Omit>



Unknown : $\mu : \underbrace{[0,1]}_{\text{time}} \times \underbrace{\mathcal{M}}_{\text{space}} \to \mathbb{R}_+$

$$\min_{\mu,\mathbf{m}} \left\{ \int_0^1 \int_{\mathcal{M}} \frac{|\mathbf{m}|^2}{2\mu} \right\}$$

where $\mathbf{m} = \mu\mathbf{v}$ is the momentum, under the constraints

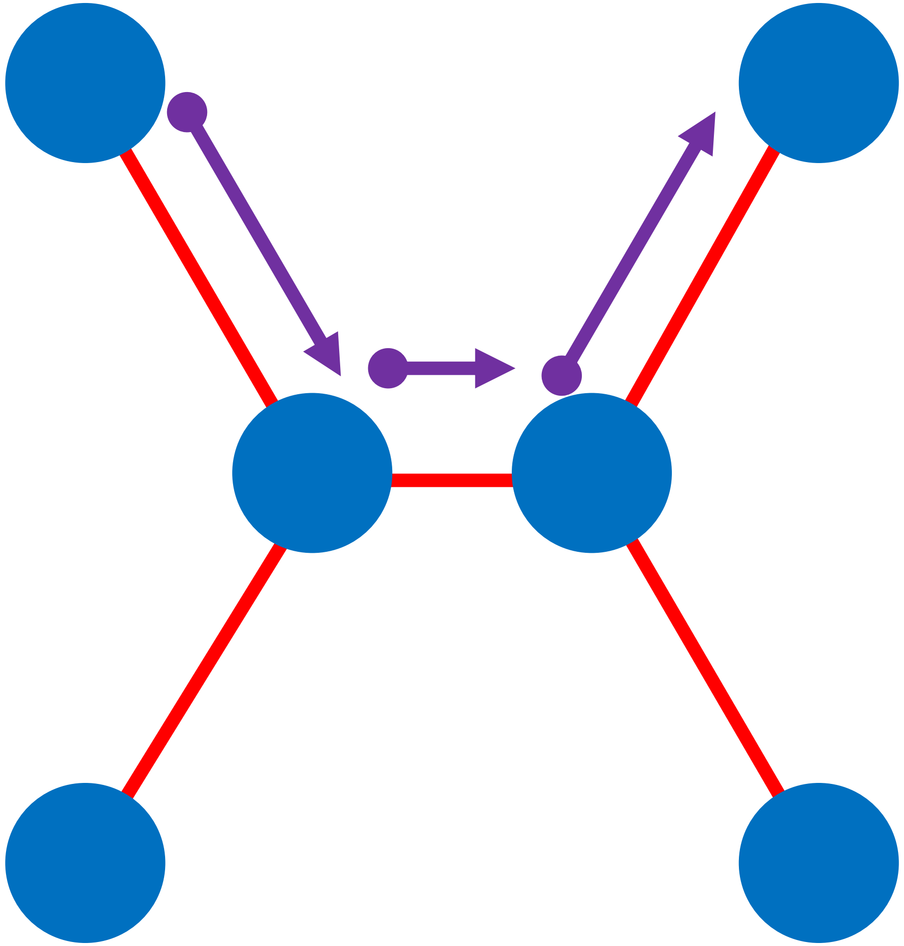
$$\begin{cases} \partial_t \mu + \boxed{\nabla \cdot \mathbf{m}} = 0, \\ \mu_0 = \bar{\mu}_0, \\ \mu_1 = \bar{\mu}_1 \end{cases}$$

Images/math from [Lavenant, Claici, Chien, and Solomon 2018]

Graph analog:

Beckmann Formulation

Better scaling for sparse graphs!



$$\begin{aligned} \min_T \quad & \textstyle\sum_e c_e |J_e| \\ \text{s.t.} \quad & D^\top J = \underbrace{p_1 - p_0}_{f} \end{aligned}$$

In computer science:

Network flow problem

Smooth PDE analog: [Solomon et al. 2014]

Plan For Today

1. Introduction to optimal transport
    - Construction
    - Many formulas
2. Applications
3. Discrete/discretized transport
    - Entropic regularization
    - Eulerian transport
    - Semidiscrete transport
4. Extensions & frontiers

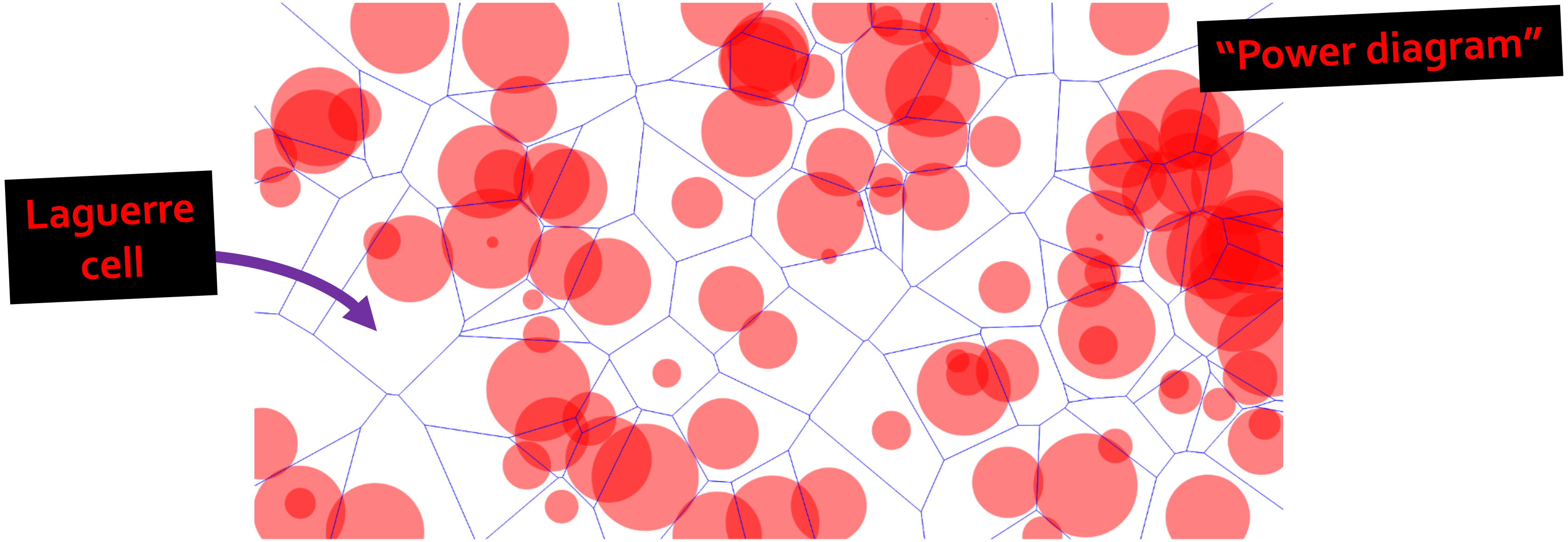
Semidiscrete General Case

$$\mu_0 := \sum_{i=1}^{k} a_i \delta_{x_i}$$

$$\nu(S) := \int_S \rho(x)\, dx$$

$$\mathcal{W}_2^2(\mu,\nu) = \sup_{\phi\in\mathbb{R}^k} \sum_i \left[a_i\phi_i + \int_{\mathrm{Lag}_\phi^c(x_i)} \rho(y)[c(x_i,y)-\phi_i]\, dA(y) \right]$$

$$\mathrm{Lag}_\phi^c(x_i) := \{y\in\mathbb{R}^n : c(x_i,y)-\phi_i \le c(x_j,y)-\phi_j\ \forall j\neq i\}$$



https://www.jasondavies.com/power-diagram/

Semidiscrete Algorithm

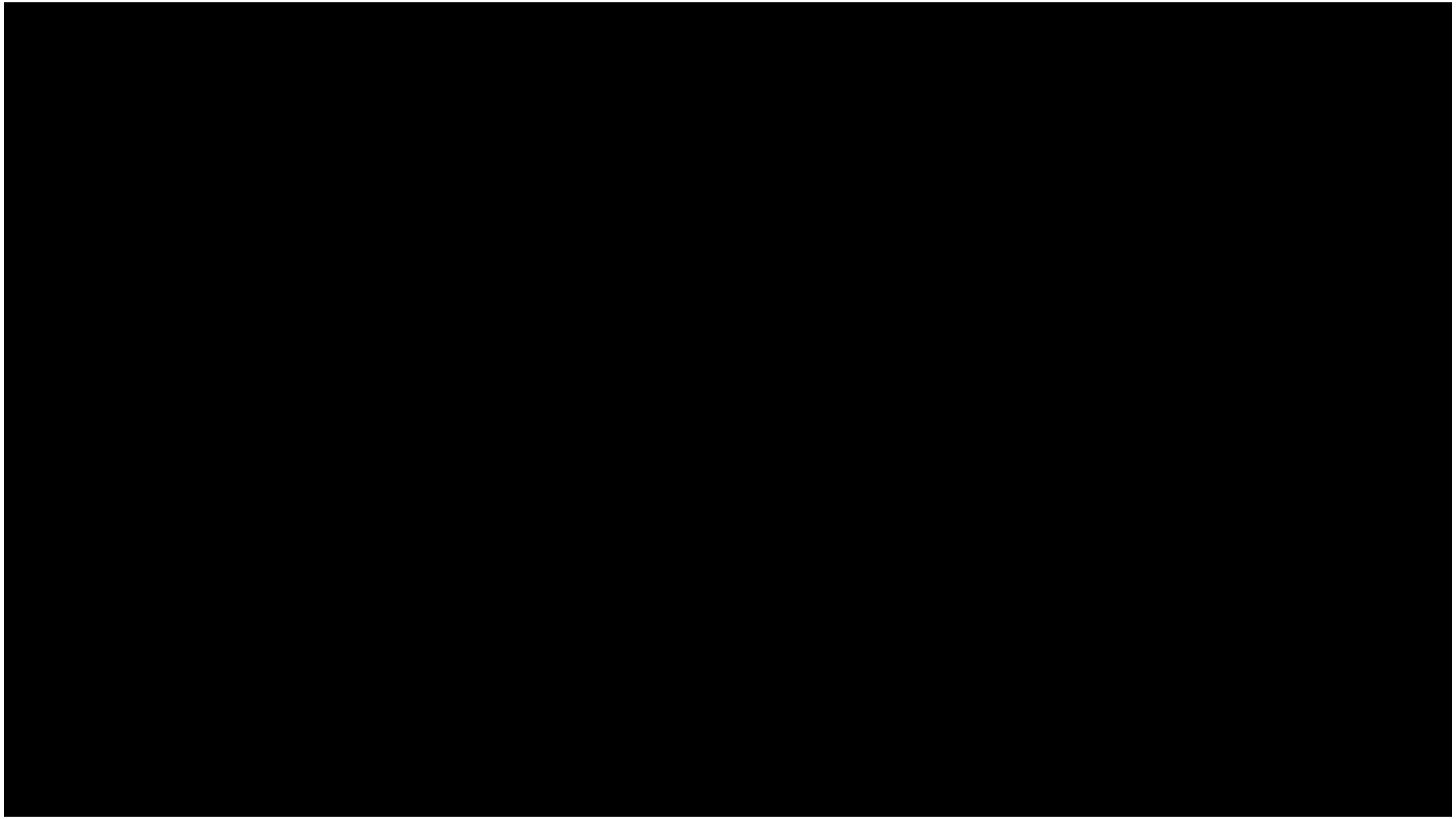
$$F(\phi) := \sum_i \left[a_i \phi_i + \int_{\mathrm{Lag}^c_\phi(x_i)} \rho(y)[c(x_i, y) - \phi_i]\, dA(y) \right]$$

$$\frac{\partial F}{\partial \phi_i} = a_i - \int_{\mathrm{Lag}^c_\phi(x_i)} \rho(y)\, dA(y)$$

Concave in ϕ!

- **Simple algorithm: Gradient ascent**
  Ingredients: Power diagram
- **More complex: Newton's method**
  Converges globally [de Goes et al. 2012; Kitagawa, Mérigot, & Thibert 2016]
- **ML setting: Stochastic optimization**
  [Genevay et al. 2016; Staib et al. 2017; Claici et al. 2018]

Application



Redux

Method	Advantages	Disadvantages
Entropic regularization	• Fast • Easy to implement • Works on mesh using heat kernel	• Blurry • Becomes singular as $\alpha \to 0$
Eulerian optimization	• Provides displacement interpolation • Connection to PDE	• Hard to optimize • Triangle mesh formulation unclear
Semidiscrete optimization	• No regularization • Connection to "classical" geometry	• Expensive computational geometry algorithms

Many others:
Stochastic transport, dual ascent, Monge-Ampère PDE, …

Plan For Today

1. Introduction to optimal transport
    - Construction
    - Many formulas

2. Applications

3. Discrete/discretized transport
    - Entropic regularization
    - Eulerian transport
    - Semidiscrete transport

4. Extensions & frontiers

Extra:
New methods in learning

Sinkhorn Autodiff

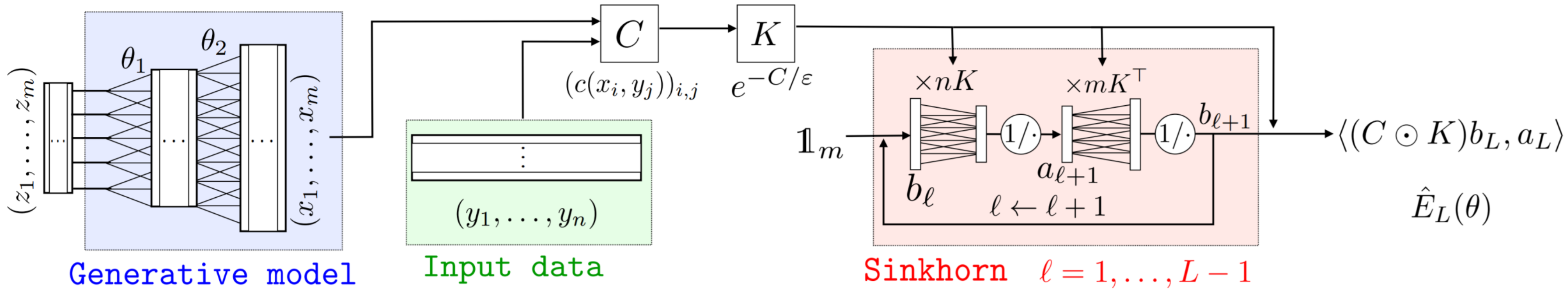


Figure 1: For a given fixed set of samples $(z_1,\ldots,z_m)$, and input data $(y_1,\ldots,y_n)$, flow diagram for the computation of Sinkhorn loss function $\theta \mapsto \hat{E}_\varepsilon^{(L)}(\theta)$. This function is the one on which automatic differentiation is applied to perform parameter learning. The display shows a simple 2-layer neural network $g_\theta : z \mapsto x$, but this applies to any generative model.

[Genevay et al. 2016]

Smoothed Dual Formulations

Proposition 19. *The dual of entropy-regularized OT between two probability measures α and β can be rewritten as the maximization of an expectation over $\alpha \otimes \beta$:*

$$W_\varepsilon^c(\alpha,\beta) = \max_{u,v \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \mathbb{E}_{\alpha\otimes\beta}[f_\varepsilon^{XY}(u,v)] + \varepsilon,$$

where

$$f_\varepsilon^{xy} \stackrel{\text{def.}}{=} u(x) + v(y) - \varepsilon \exp^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}} \quad \textit{for } \varepsilon > 0. \tag{2.2}$$

and when $\beta \stackrel{\text{def.}}{=} \sum_{j=1}^m \boldsymbol{\beta}_j \delta_{y_j}$ is discrete, the potential v is a m-dimensional vector $(\mathbf{v}_j)_j$

Algorithm 4 Averaged SGD for Semi-Discrete OT

Input: step size $C \in \mathbb{R}_+$

Output: dual potential $\bar{\mathbf{v}} \in \mathbb{R}^m$

$\mathbf{v} \leftarrow \mathbb{0}_m$ (iterates for SGD)

$\bar{\mathbf{v}} \leftarrow \mathbf{v}$ (dual potential obtained by averaging)

for $k = 1, 2, \ldots$ **do**

Sample x_k from α

$\mathbf{v} \leftarrow \mathbf{v} + \frac{C}{\sqrt{k}} \nabla_v g_\varepsilon^{x_k}(\mathbf{v})$ (gradient ascent step using $\mathbf{v}$)

$\bar{\mathbf{v}} \leftarrow \frac{1}{k}\mathbf{v} + \frac{k-1}{k}\bar{\mathbf{v}}$ (averaging step to get faster convergence of $\mathbf{v}$)

end for

Parameterize dual potentials:

- Using RKHS [Genevay et al. 2016]
- Using neural networks [Seguy et al. 2017]

[PhD thesis: Genevay 2019]

New Progress on the Monge Formulation

Input Convex Neural Networks

Brandon Amos [1] Lei Xu [2*] J. Zico Kolter [1]

Abstract

This paper presents the input convex neural network architecture. These are scalar-valued (potentially deep) neural networks with constraints on the network parameters such that the output of the network is a convex function of (some of) the inputs. The networks allow for efficient inference via optimization over some inputs to the network given others, and can be applied to settings including structured prediction, data imputation, reinforcement learning, and others. In this paper we lay the basic groundwork for these models, proposing methods for inference, optimization and learning, and analyze their representational power. We show that many existing neural network architectures can be made input-convex with a minor modification, and develop specialized optimization algorithms tailored to this setting. Finally, we highlight the performance of the methods on multi-label prediction, image completion, and reinforcement learning problems, where we show improvement over the existing state of the art in many cases.

1. Introduction

In this paper, we propose a new neural network architecture that we call the *input convex neural network* (ICNN).These are *scalar-valued* neural networks $f(x, y; \theta)$ where x and y denotes inputs to the function and θ denotes the parameters, built in such a way that the network is convex in (a subset of) *inputs* y.[3] The fundamental benefit to these IC-

y) we can globally and efficiently (because the problem is convex) solve the

Fundamentally, in the network ing predictions i ward process, w scalar function (ergy function) o ers. There are a networks.

Structured pred notation above, tured prediction. tured input and work over (x, y) for this pair, foll malisms (LeCun the $y \in \mathcal{Y}$ that is exactly the ar suming that $\mathcal{Y}$ is structured predic This is similar in networks (SPEN also use deep ne with the differen y, so the optimiz

Data imputatio

Optimal transport mapping via input convex neural networks

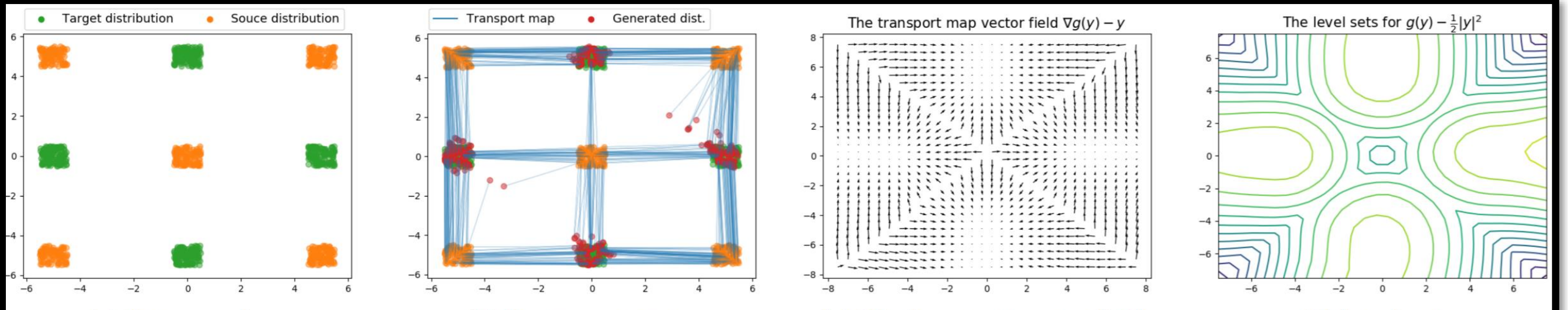
Ashok Vardhan Makkuva [*1] Amirhossein Taghvaei [*2] Jason D. Lee [3] Sewoong Oh [4]

Abstract

In this paper, we present a novel and principled approach to learn the optimal transport between two distributions, from samples. Guided by the optimal transport theory, we learn the optimal Kantorovich potential which induces the optimal

1. Introduction

Finding a mapping that transports mass from one distribution Q to another distribution P is an important task in various machine learning applications, such as deep generative models (Goodfellow et al., 2014; Kingma & Welling, 2013) and domain adaptation (Gopalan et al., 2011; Ben-



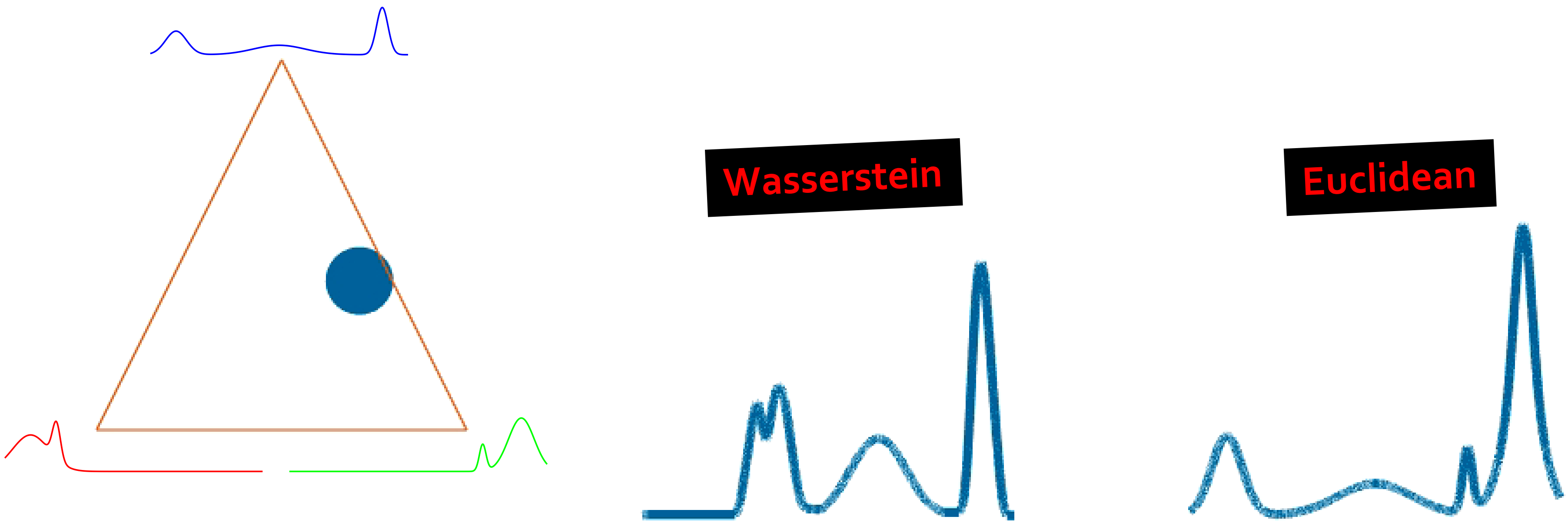
(a) Data samples (b) Our transport map (c) Displacement vector field (d) Level sets

Plan For Today

1. Introduction to optimal transport
    - Construction
    - Many formulas
2. Applications
3. Discrete/discretized transport
    - Entropic regularization
    - Eulerian transport
    - Semidiscrete transport
4. Extensions & frontiers

Extension:

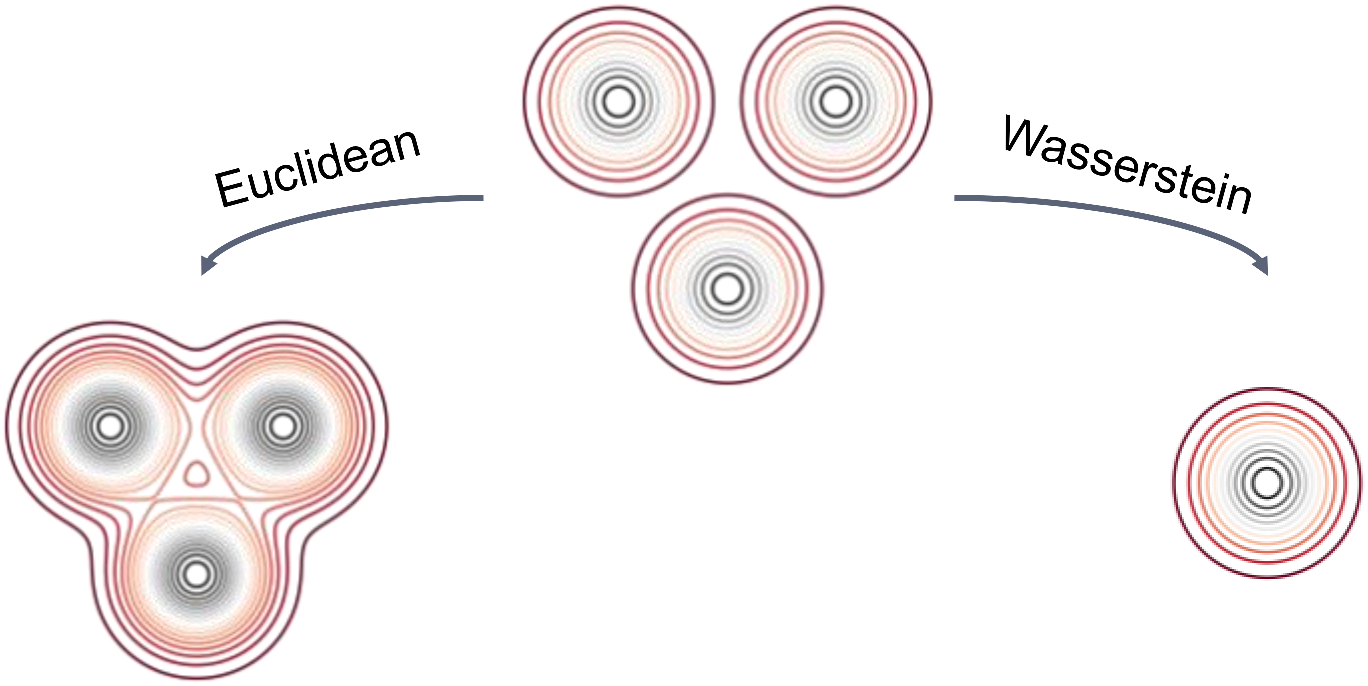
Wasserstein Barycenters



$$\text{Wasserstein:} \quad \mu^* := \left[\arg \min_{\mu \in \mathrm{Prob}(\mathbb{R}^n)} \sum_i \mathcal{W}_2^2(\mu, \mu_i) \right]$$

[Agueh and Carlier 2010]

Barycenters in Bayesian Inference



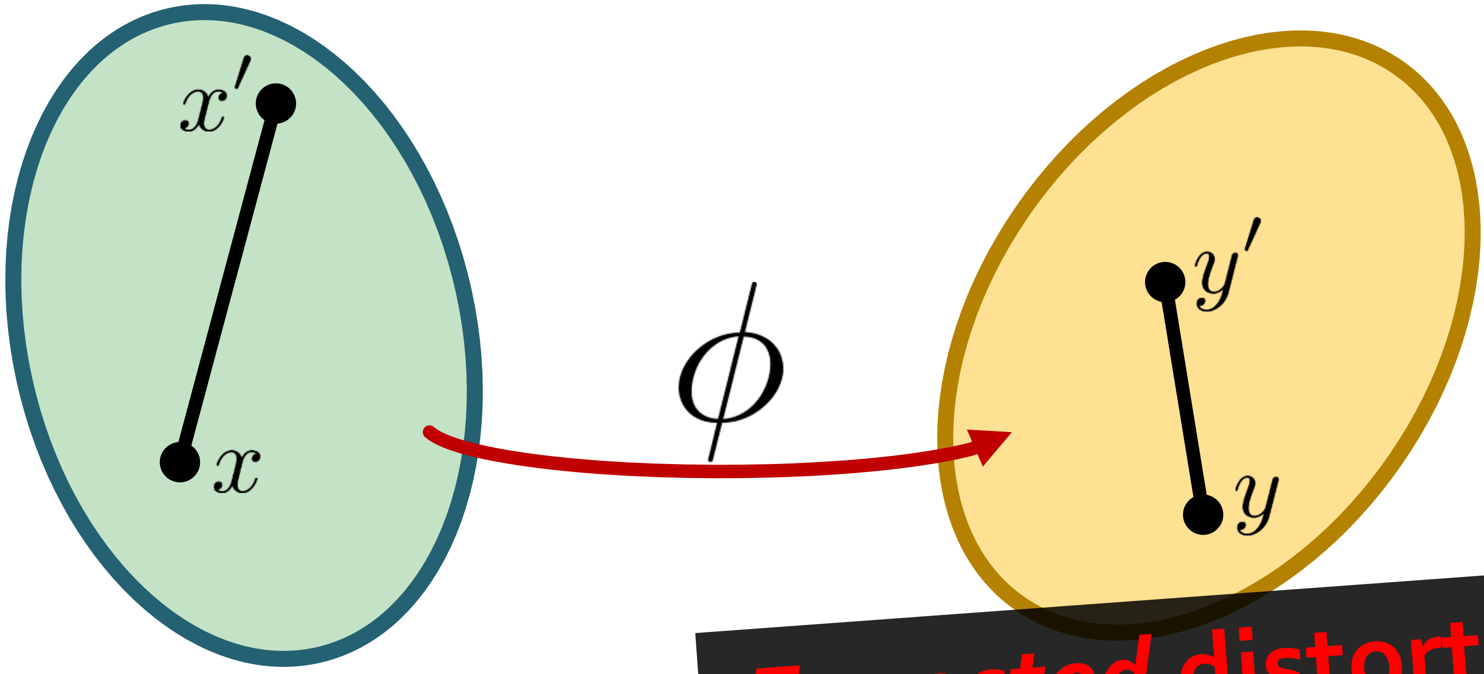
Wasserstein Subset Posterior (WASP)

[Srivastava et al. 2018]

Extension:

Quadratic Matching

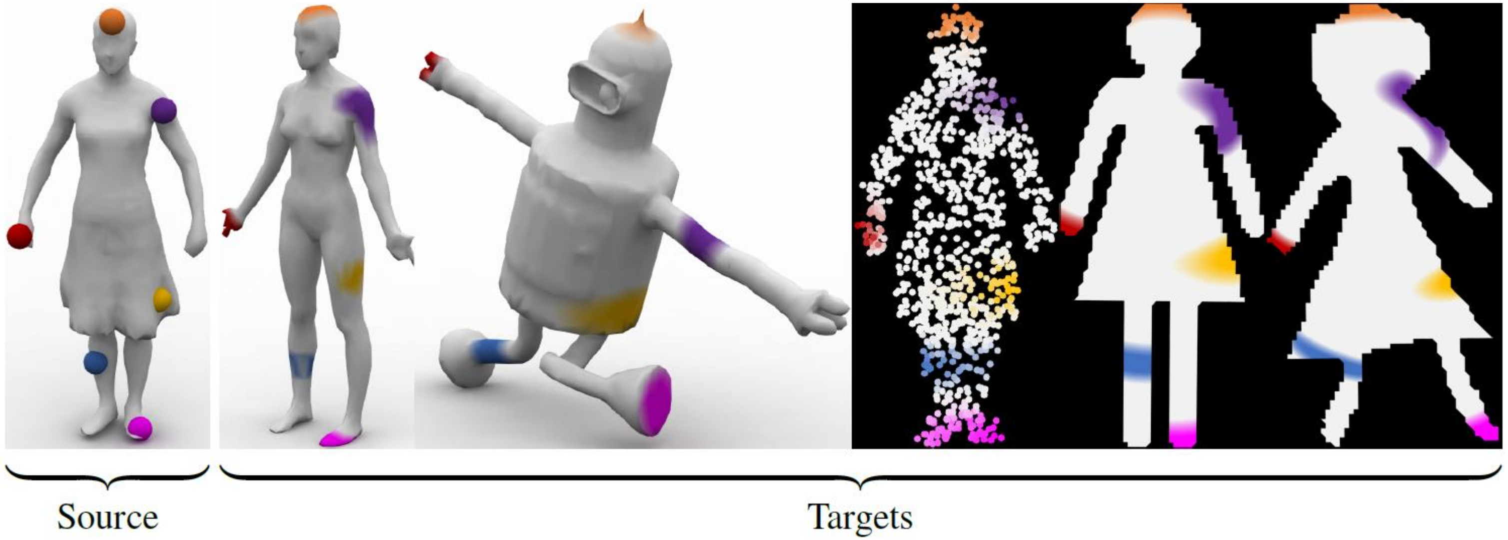
[Mémoli 2007]

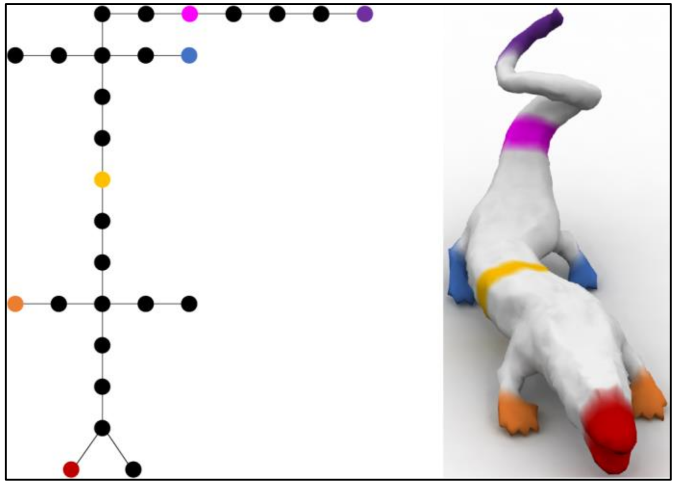


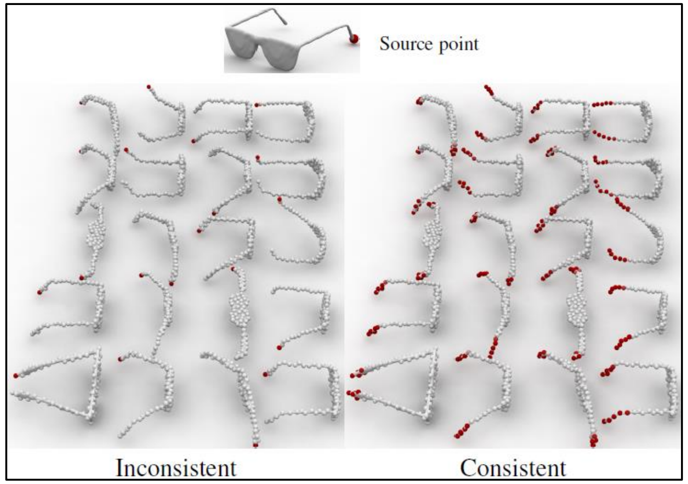
$$\mathrm{GW}_2^2((\mu_0, d_0), (\mu, d)) :=$$

$$\min_{\gamma \in \mathcal{M}(\mu_0, \mu)} \iint_{\Sigma_0 \times \Sigma} [d_0(x, x') - d(y, y')]^2 \, d\gamma(x, y) \, d\gamma(x', y')$$

Variety of Correspondence Tasks



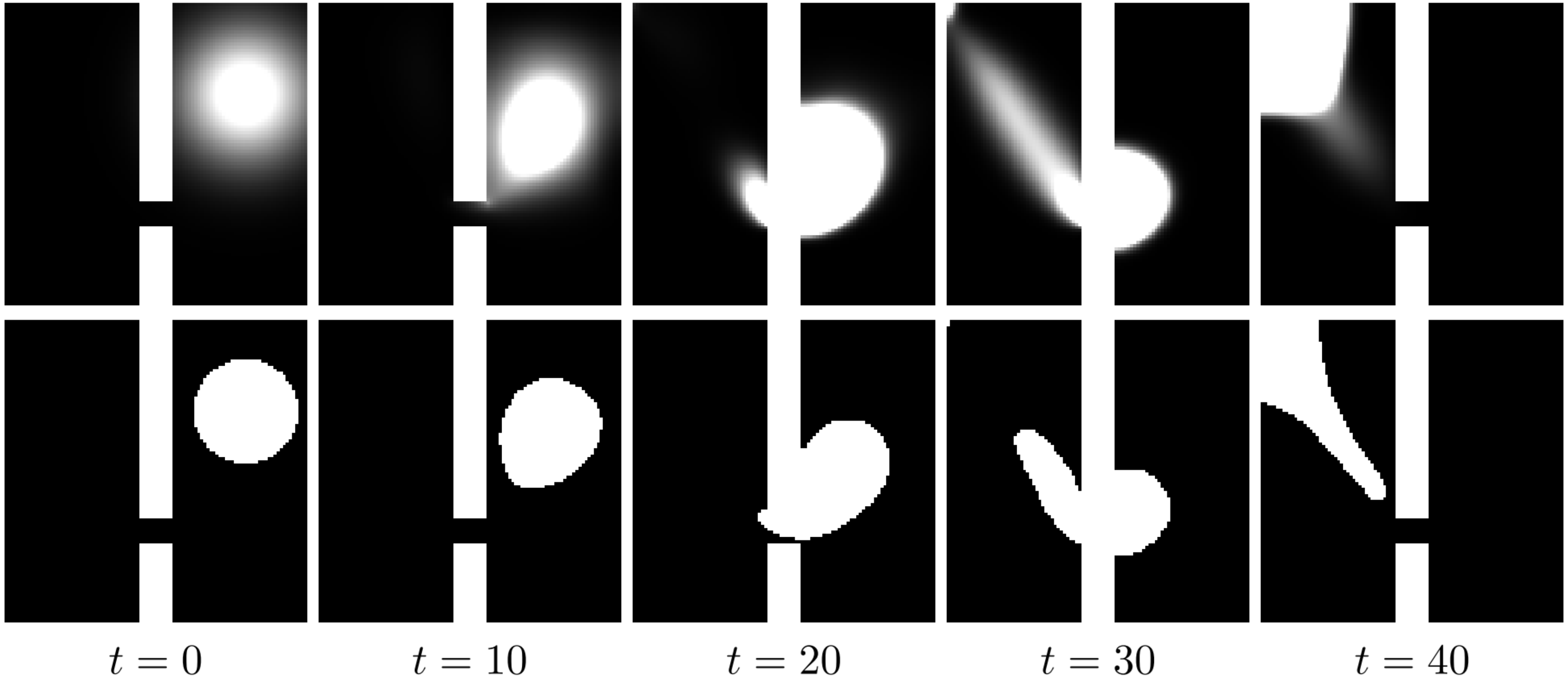




[Solomon et al. 2016]

Extension:

Gradient Flows



"Entropic Wasserstein Gradient Flows" [Peyré 2015]

Extension:

Matrix Fields and Vector Measures

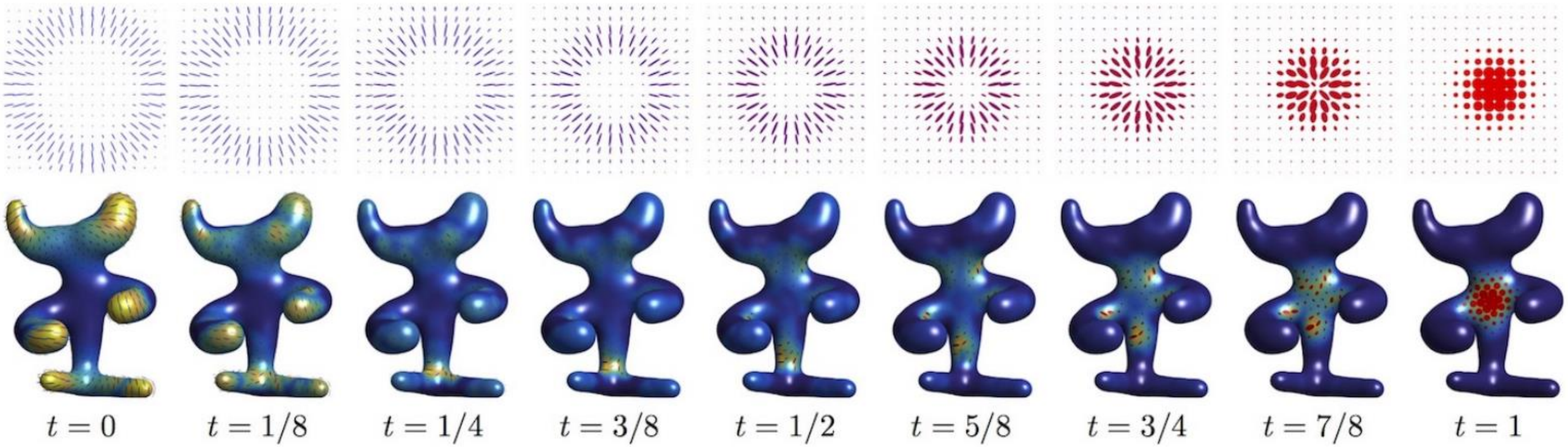


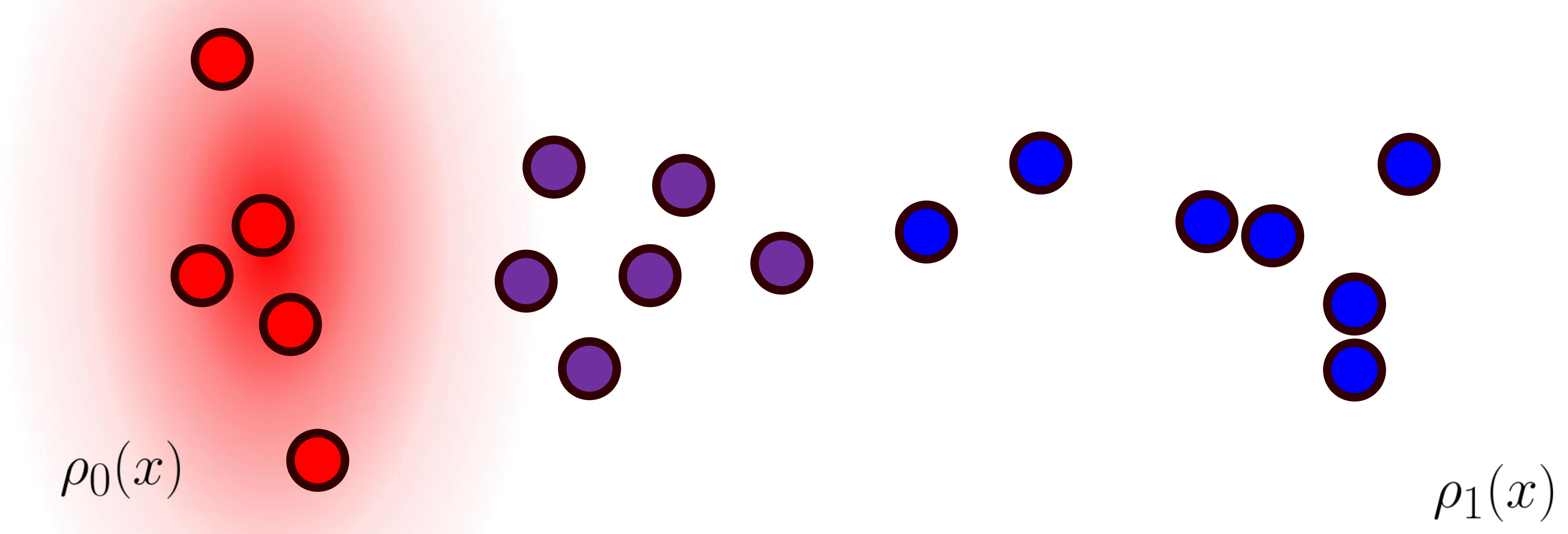
Image from
"Quantum Optimal Transport for Tensor Field Processing"
[Peyré et al. 2017]

Open problem: Dynamical version? Curved surfaces?

Extension:

Sampling Problems

$$\min_{\rho}[\mathcal{W}_2^2(\rho_0, \rho) + \mathcal{W}_2^2(\rho_1, \rho)]$$



Somewhere between semidiscrete and smooth

Wasserstein barycenter

Optimal Transport

Justin Solomon

6.838: Shape Analysis

Spring 2021