CS480/680: Introduction to Machine Learning

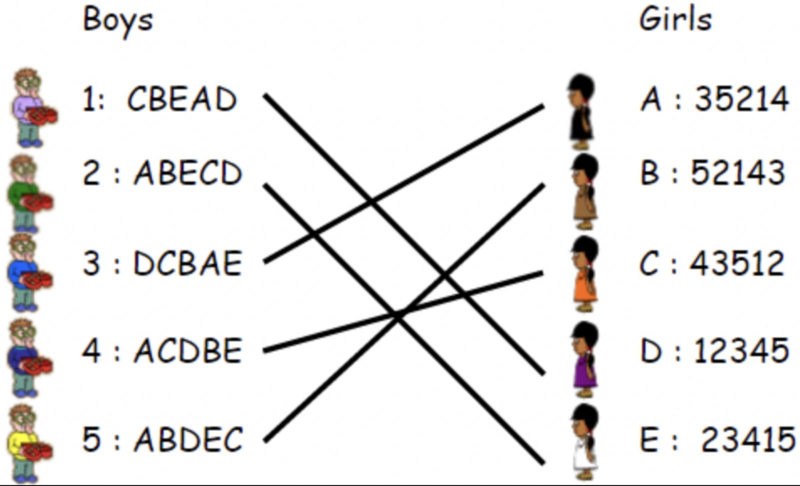
Lec 18: Optimal Transport

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

July 10, 2024

Matching Problem



https://towardsdatascience.com/stable-matching-as-a-game-a68c279d70b

- Matching co-op students/organ donors with companies/patients

Stable Matching

Definition: Blocking pair

A pair (i, j) and (i', j') where both i and j' would prefer to swap.

Example: Blocking pair

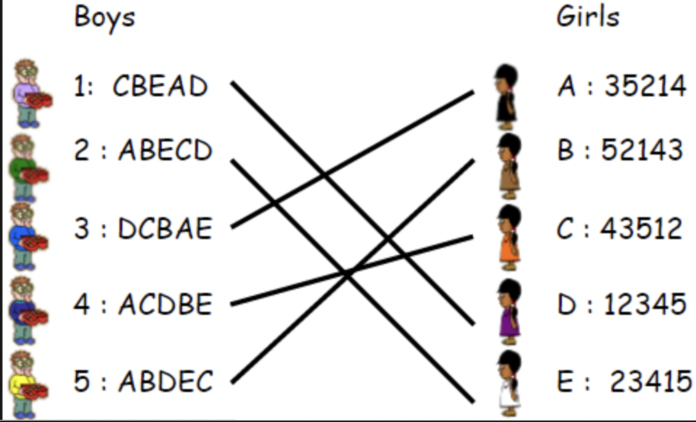
(1, E) and (2, C) —> (1, C) and (2, E)

- A stable matching is one when there is no blocking pair
  - Gale-Shapley algorithm
- More generally, can define a cost $c(i, j)$ for matching i-th boy with j-th girl
- Need $c(i, j) + c(i', j') \leq c(i, j') + c(i', j)$

D. Gale and L. S. Shapley. "College Admissions and the Stability of Marriage". *The American Mathematical Monthly*, vol. 69, no. 1 (1962), pp. 9–15.

Monge's Formulation

$$\min_{\mathbf{T}:[n]\to[n] \text{ bijective}} \sum_{i=1}^{n} c(i, \mathbf{T}(i))$$



H. W. Kuhn. "The Hungarian method for the assignment problem". *Naval Research Logistics Quarterly*, vol. 2, no. 1-2 (1955), pp. 83–97.

From Discrete to Continuous

$$\min_{\mathbf{T}_{\#}p=q} \quad \mathbb{E}[c(\mathsf{X}, \mathbf{T}(\mathsf{X}))]$$

- A distribution p of boys and a distribution q of girls
- Let $\mathsf{X} \sim p$ be a random boy
- $\mathbf{T}(\mathsf{X})$ is the matching for X
- Require $\mathbf{T}_{\#}p = q$ to preserve mass

G. Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie royale des sciences avec les mémoires de mathématique et de physique tirés des registres de cette Académie*. 1781, pp. 666–705.

Kantorovich's Relaxation

$$\min_{\mathbf{T}_{\#}p=q} \mathbb{E}[c(\mathsf{X}, \mathbf{T}(\mathsf{X}))] \geq \min_{\mathsf{X}\sim p, \mathsf{Y}\sim q} \mathbb{E}[c(\mathsf{X}, \mathsf{Y})]$$

Definition: Coupling

$(\mathsf{X}, \mathsf{Y}) \sim \pi$, where the joint coupling π has marginals p and q

- Deterministic pairing: $\mathbf{x}$ is matched with some $\mathbf{y} = \mathbf{Tx}$
- Stochastic pairing: $\mathbf{x}$ is matched to every $\mathbf{y}$ with probability $\pi(\mathbf{y}|\mathbf{x})$
- Surprisingly, at optimality, $\pi(\mathbf{y}|\mathbf{x})$ could be deterministic anyway!

L. V. Kantorovich. "On the Translocation of Masses". *Journal of Mathematical Sciences*, vol. 133, no. 4 (2006). Originally published in Dokl. Akad. Nauk SSSR, vol. 37, No. 7–8, 227–229 (1942)., pp. 1381–1382, L. V. Kantorovich. "On a Problem of Monge". *Journal of Mathematical Sciences*, vol. 133, no. 4 (2006). Originally published in Uspekhi Mat. Nauk, vol. 3, No. 2, 225-226 (1948)., pp. 1383–1383.

Duality

$$\min_{X\sim p, Y\sim q, (X,Y)\sim\pi} \mathbb{E}[c(X,Y)] = \min_{\pi\geq 0} \int c(\mathbf{x},\mathbf{y})\pi(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}$$

$$\text{s.t.} \quad \int \pi(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} = p(\mathbf{x}), \int \pi(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x} = q(\mathbf{y})$$

$$\min_{\pi\geq 0} \max_{u(\cdot),v(\cdot)} \int [c(\mathbf{x},\mathbf{y}) - u(\mathbf{x}) - v(\mathbf{y})]\pi(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \int u(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int v(\mathbf{y})p(\mathbf{y})\,\mathrm{d}\mathbf{y}$$

$$= \max_{u(\cdot),v(\cdot)} \min_{\pi\geq 0} \int [c(\mathbf{x},\mathbf{y}) - u(\mathbf{x}) - v(\mathbf{y})]\pi(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \int u(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int v(\mathbf{y})p(\mathbf{y})\,\mathrm{d}\mathbf{y}$$

$$= \max_{u(\cdot),v(\cdot)} \int u(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int v(\mathbf{y})p(\mathbf{y})\,\mathrm{d}\mathbf{y}, \quad \text{s.t.} \quad u(\mathbf{x}) + v(\mathbf{y}) \leq c(\mathbf{x},\mathbf{y}), \ \forall \mathbf{x},\mathbf{y}$$

$$\boxed{= \max_{u,v} \mathbb{E}[u(X)] + \mathbb{E}[v(Y)], \quad \text{s.t.} \quad u(\mathbf{x}) + v(\mathbf{y}) \leq c(\mathbf{x},\mathbf{y}), \ \forall \mathbf{x},\mathbf{y}}$$

L. V. Kantorovich and G. S. Rubinshtein. "On a functional space and certain extremum problems". *Dokl. Akad. Nauk SSSR*, vol. 115, no. 6 (1957), pp. 1058–1061.

Conjugacy

$$\boxed{\forall \mathbf{x}, \forall \mathbf{y},\ u(\mathbf{x}) + v(\mathbf{y}) \le c(\mathbf{x}, \mathbf{y})}$$

- $u(\mathbf{x}) \le [\inf_{\mathbf{y}}\ c(\mathbf{x}, \mathbf{y}) - v(\mathbf{y})] =: v^c(\mathbf{x})$
- $v(\mathbf{y}) \le [\inf_{\mathbf{x}}\ c(\mathbf{x}, \mathbf{y}) - u(\mathbf{x})] =: u^c(\mathbf{y})$
- Since we are maximizing $\mathbb{E}[u(\mathsf{X})] + \mathbb{E}[v(\mathsf{Y})]$, at optimality:

$$\boxed{u(\mathbf{x}) = v^c(\mathbf{x}), \qquad v(\mathbf{y}) = u^c(\mathbf{y})}$$

- $u^{cc} \ge u$ and $u^{ccc} = u^c$; similarly for v
- u is called c-concave iff $u = u^{cc}$ (or equivalently $u = v^c$ for some v)

Complementarity

$$\max_{u,v} \min_{\pi \geq 0} \int [c(\mathbf{x},\mathbf{y}) - u(\mathbf{x}) - v(\mathbf{y})]\pi(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} + \int u(\mathbf{x})p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int v(\mathbf{y})p(\mathbf{y})\, \mathrm{d}\mathbf{y}$$

- $\pi(\mathbf{x},\mathbf{y}) > 0 \implies u(\mathbf{x}) + v(\mathbf{y}) = c(\mathbf{x},\mathbf{y})$
- Recall that $u^c = v$, we define the subdifferential:

$$\boxed{\partial u(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{y}} \, [c(\mathbf{x},\mathbf{y}) - u^c(\mathbf{y})] = \{\mathbf{y} : u(\mathbf{x}) + u^c(\mathbf{y}) = c(\mathbf{x},\mathbf{y})\}}$$

  – for a c-concave u, $\boxed{\mathbf{y} \in \partial u(\mathbf{x}) \iff \mathbf{x} \in \partial u^c(\mathbf{y})}$

- Thus, $\boxed{\operatorname{supp} \pi \subseteq \operatorname{gph} \partial u}$

  – in particular, if u is differentiable, π is deterministic and the Kantorovich relaxation is tight!

Cyclic Monotonicity

Definition: Cyclic monotonicity

We call a set $\Gamma \subseteq \mathbb{X} \times \mathbb{Y}$ c-cyclically monotone if for any n and $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n) \in \Gamma$, any (cyclic) permutation $\sigma : [n] \to [n]$, we always have

$$\sum_{i=1}^{n} c(\mathbf{x}_i, \mathbf{y}_i) \leq \sum_{i=1}^{n} c(\mathbf{x}_i, \mathbf{y}_{\sigma(i)})$$

Theorem: Optimal coupling

Suppose $\mathbb{X} = \mathbb{R}^{d_x}$ and $\mathbb{Y} = \mathbb{R}^{d_y}$, $c : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is continuous, and p and q are probability densities. There exists an optimal coupling π whose support is c-cyclically monotone. Moreover, there exists a c-concave function u such that $\operatorname{supp} \pi \subseteq \operatorname{gph} \partial u$.

L18 L. Rüschendorf. "On c-optimal random variables". *Statistics & Probability Letters*, vol. 27, no. 3 (1996), pp. 267–270.

1-Wasserstein Distance

- $c(\mathbf{x}, \mathbf{y}) = d(\mathbf{x}, \mathbf{y})$ for some distance metric d:

$$\mathbb{W}_1(p, q) := \min_{(\mathsf{X},\mathsf{Y})\sim\pi, \mathsf{X}\sim p, \mathsf{Y}\sim q} \mathbb{E}[d(\mathsf{X}, \mathsf{Y})]$$
$$= \max_{u,v} \ \mathbb{E}[u(\mathsf{X})] + \mathbb{E}[v(\mathsf{Y})], \quad \text{s.t.} \quad \forall \mathbf{x}, \mathbf{y}, \ u(\mathbf{x}) + v(\mathbf{y}) \leq d(\mathbf{x}, \mathbf{y})$$

- Lipschitz envelope: $v^c(\mathbf{x}) := [\inf_{\mathbf{y}} d(\mathbf{x}, \mathbf{y}) - v(\mathbf{y})]$
  - v^c is Lipschitz continuous: $|v^c(\mathbf{x}) - v^c(\mathbf{z})| \leq d(\mathbf{x}, \mathbf{z})$
  - v^c is the largest Lipschitz continuous function majorized by $-v$
- Thus, $u = v^c = -v$ and hence

$$\boxed{\mathbb{W}_1(p, q) = \max_{u} \ \mathbb{E}[u(\mathsf{X})] - \mathbb{E}[u(\mathsf{Y})], \quad \text{s.t.} \quad \forall \mathbf{x}, \mathbf{y}, \ u(\mathbf{x}) - u(\mathbf{y}) \leq d(\mathbf{x}, \mathbf{y})}$$

Wasserstein GAN

$$\min_{\mathbf{T}} \mathrm{W}_1(q, \mathbf{T}_{\#}r) = \min_{\mathbf{T}} \max_{u} \ \mathbb{E}_{\mathsf{X}\sim q}[u(\mathsf{X})] - \mathbb{E}_{\mathsf{Z}\sim r}[u(\mathbf{T}(\mathsf{Z}))], \quad \text{s.t.} \quad u \text{ is Lipschitz}$$
$$\approx \min_{\mathbf{T}} \max_{u} \ \hat{\mathbb{E}}_{\mathsf{X}\sim q}[u(\mathsf{X})] - \hat{\mathbb{E}}_{\mathsf{Z}\sim r}[u(\mathbf{T}(\mathsf{Z}))], \quad \text{s.t.} \quad u \text{ is Lipschitz}$$

- r is the noise density, e.g., standard normal
- q is the data density: only a training sample is available
- $\mathbf{T}$ is the generator network: maps noise to data
- u is the discriminator network: maps data to a real scalar

M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. 2017.

2-Wasserstein Distance

- $c(\mathbf{x},\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ (note the square):

$$\mathbb{W}_2^2(p,q) := \min_{(\mathsf{X},\mathsf{Y})\sim\pi,\mathsf{X}\sim p,\mathsf{Y}\sim q} \mathbb{E}[\tfrac{1}{2}\|\mathsf{X}-\mathsf{Y}\|_2^2]$$

$$= \max_{u,v} \ \mathbb{E}[u(\mathsf{X})] + \mathbb{E}[v(\mathsf{Y})], \quad \text{s.t. } \ \forall \mathbf{x},\mathbf{y}, \ u(\mathbf{x}) + v(\mathbf{y}) \le \tfrac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$

- Conjugate: $v^c(\mathbf{x}) := [\inf_{\mathbf{y}} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 - v(\mathbf{y})]$
  - $\frac{1}{2}\|\mathbf{x}\|_2^2 - v^c(\mathbf{x}) = \sup_{\mathbf{y}} \langle \mathbf{x},\mathbf{y}\rangle - [\frac{1}{2}\|\mathbf{y}\|_2^2 - v(\mathbf{y})]$: convex conjugate
- Thus, $u = v^c = \frac{1}{2}\|\cdot\|_2^2 - (\frac{1}{2}\|\cdot\|_2^2 - v)^*$ and hence

$$\boxed{\mathbb{W}_2^2(p,q) = \max_f \ \mathbb{E}[\tfrac{1}{2}\|\mathsf{X}\|_2^2 - f(\mathsf{X})] + \mathbb{E}[\tfrac{1}{2}\|\mathsf{Y}\|_2^2 - f^*(\mathsf{Y})], \quad \text{s.t. } \ f \text{ is convex}}$$

$$\boxed{\pi = (\mathrm{Id} \times \partial f)_{\#} p}, \text{ in particular } q = (\partial f)_{\#} p$$

Y. Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". *Communications on Pure and Applied Mathematics*, vol. 44, no. 4 (1991), pp. 375–417, R. J. McCann. "Existence and uniqueness of monotone measure-preserving maps". *Duke Mathematical Journal*, vol. 80, no. 2 (1995), pp. 309–323.

Potential GAN

$$\min_{\mathbf{T}} \mathrm{W}_2^2(q, \mathbf{T}_{\#}r) = \min_{\mathbf{T}} \max_{f} \mathop{\mathbb{E}}_{\mathsf{X}\sim q}[\tfrac{1}{2}\|\mathsf{X}\|_2^2 - f(\mathsf{X})] + \mathop{\mathbb{E}}_{\mathsf{Z}\sim r}[\tfrac{1}{2}\|\mathbf{T}(\mathsf{Z})\|_2^2 - f^*(\mathbf{T}(\mathsf{Z}))],$$

$$\approx \min_{\mathbf{T}} \max_{f} \mathop{\hat{\mathbb{E}}}_{\mathsf{X}\sim q}[-f(\mathsf{X})] + \mathop{\hat{\mathbb{E}}}_{\mathsf{Z}\sim r}[\tfrac{1}{2}\|\mathbf{T}(\mathsf{Z})\|_2^2 - f^*(\mathbf{T}(\mathsf{Z}))], \text{ s.t. } f \text{ is convex}$$

- r is the noise density, e.g., standard normal
- q is the data density: only a training sample is available
- $\mathbf{T}$ is the generator network: maps noise to data
- f is the discriminator network: maps data to a real scalar

T. Salimans, H. Zhang, A. Radford, and D. Metaxas. "Improving GANs Using Optimal Transport". In: *International Conference on Learning Representations*. 2018, H. Liu, X. Gu, and D. Samaras. "Wasserstein GAN With Quadratic Transport Cost". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 4831–4840.

Potential Flow

$$\min_f \ \mathbb{D}\big(q, (\nabla f)_{\#} r\big), \quad \text{s.t.} \quad f \text{ is convex}$$

- r is the noise density, e.g., standard normal
- q is the data density: only a training sample is available
- ∇f is the generator network: maps noise to data
  - e.g., f is a Relu network with nonnegative weights
- $\mathbb{D}$ is some "distance" function, e.g., the KL divergence

C.-W. Huang, R. T. Q. Chen, C. Tsirigotis, and A. Courville. "Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization". In: *International Conference on Learning Representations*. 2021.

Triangular vs. Potential

- $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$, $\mathbf{T}_{\#}p = q$
- $\mathbf{T}$ is autoregressive
- $\nabla\mathbf{T}$ is always triangular
- composition holds
- no rotational equivariance

- $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$, $\mathbf{T}_{\#}p = q$
- $\mathbf{T} = \nabla f$ for convex $f : \mathbb{R}^d \to \mathbb{R}$
- $\nabla\mathbf{T} = \nabla^2 f$ is symmetric PSD
- composition fails
- rotationally equivariant

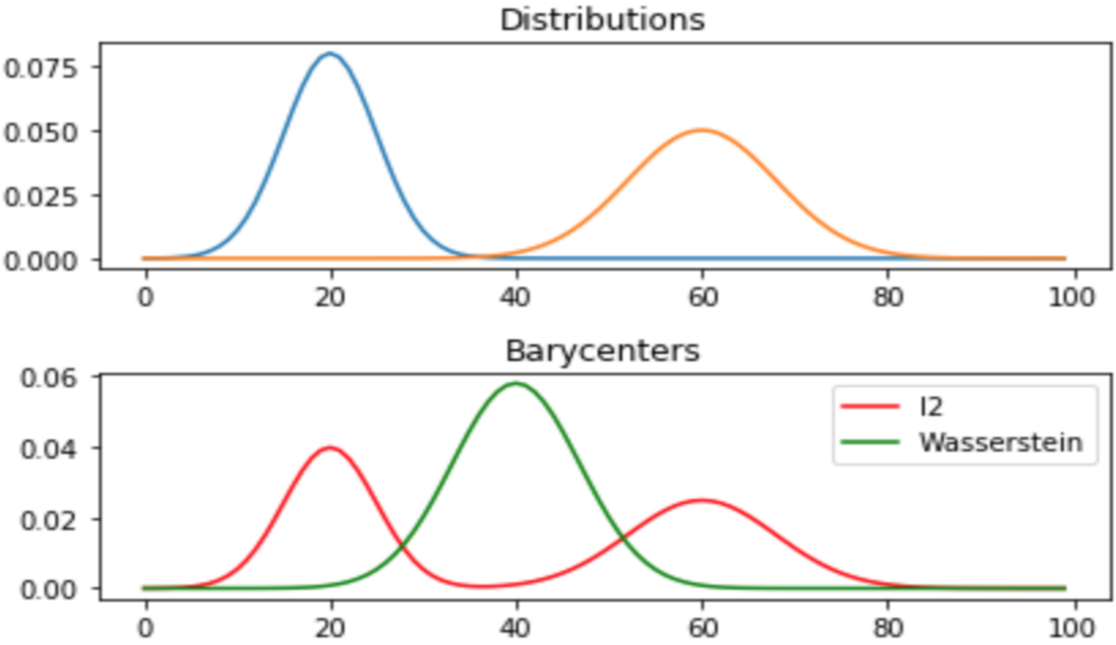
The two are equivalent iff $\mathbf{T}$ is diagonal, in particular, if $d = 1$

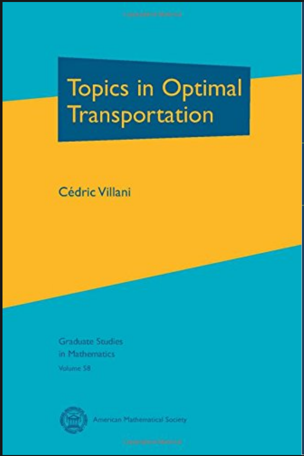
Wasserstein Barycenter

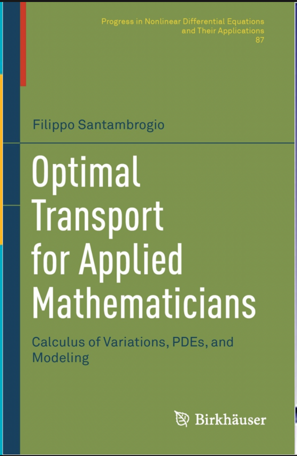
- Consider densities p_0 and p_1, say two Gaussians with different mean and variance
- How to interpolate between them?
- Exists convex f such that $p_1 = (\nabla f)_{\#} p_0$
- Obviously $p_0 = (\mathrm{Id})_{\#} p_0$
- Interpolate the push-forward maps!

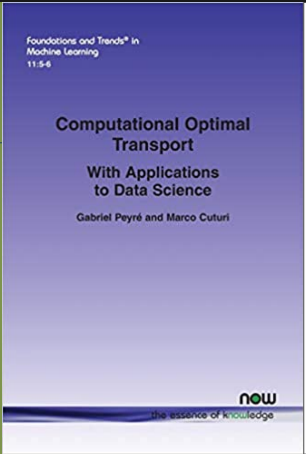
$$p_t = [(1-t)\mathrm{Id} + t\nabla f]_{\#} p_0 = \underset{p}{\operatorname{argmin}}\ (1-t)\mathbb{W}_2^2(p, p_0) + t\mathbb{W}_2^2(p, p_1)$$

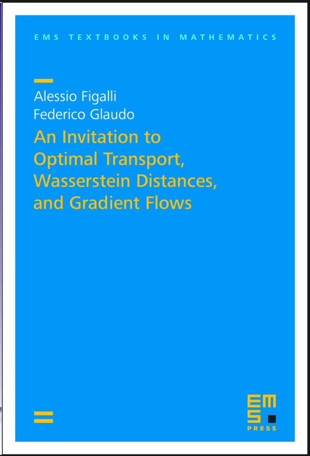
R. J. McCann. "A Convexity Principle for Interacting Gases". *Advances in Mathematics*, vol. 128, no. 1 (1997), pp. 153–179.


Distributions
0.075
0.050
0.025
0.000
0
20
40
60
80
100
Barycenters
0.06
0.04
0.02
0.00
l2
Wasserstein
0
20
40
60
80
100


Topics in Optimal Transportation
Cédric Villani
Graduate Studies in Mathematics
Volume 58
American Mathematical Society


Progress in Nonlinear Differential Equations and Their Applications
87
Filippo Santambrogio
Optimal Transport for Applied Mathematicians
Calculus of Variations, PDEs, and Modeling
Birkhäuser


Foundations and Trends® in Machine Learning
11:5-6
Computational Optimal Transport
With Applications to Data Science
Gabriel Peyré and Marco Cuturi
now
the essence of knowledge


EMS TEXTBOOKS IN MATHEMATICS
Alessio Figalli
Federico Glaudo
An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows
EMS PRESS