CS480/680: Introduction to Machine Learning

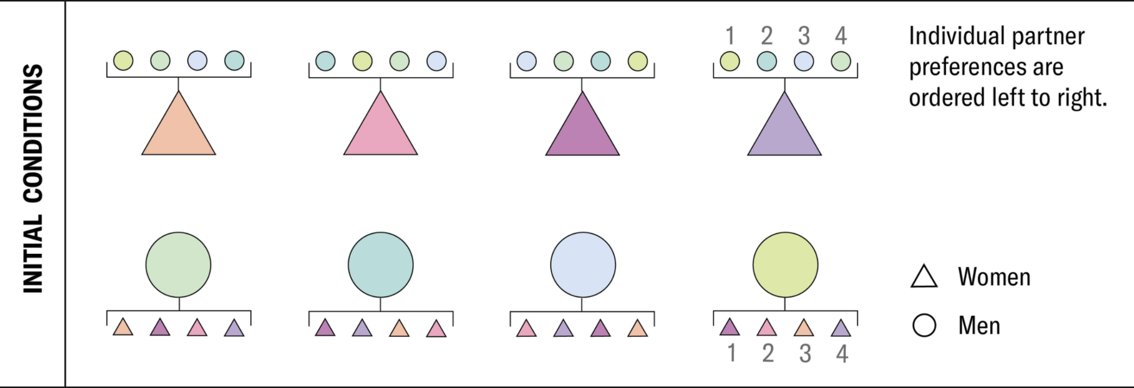
Lec 17: Optimal Transport

Yaoliang Yu

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

March 13, 2025

The Matching Problem



https://tinyurl.com/ej74dbsu

- Matching co-op students with companies, organ donors with patients, etc.

Stable Matching

Definition: Blocking pair

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

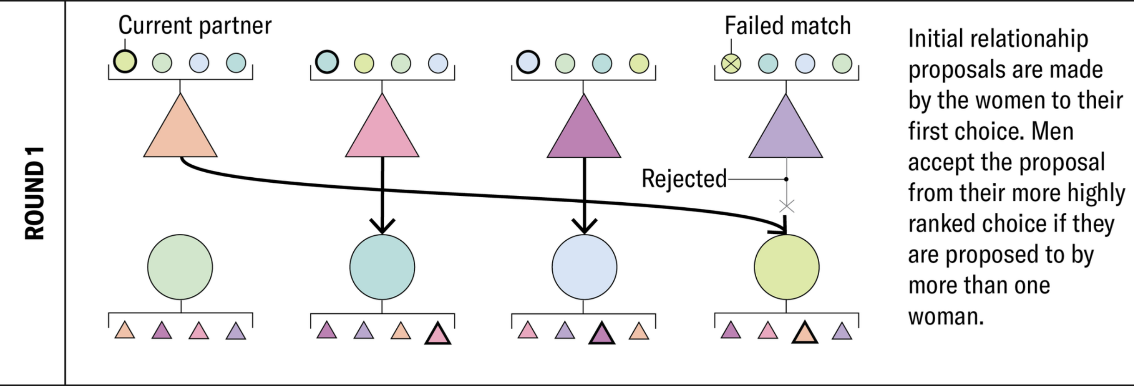
Example: Blocking pair

(△, ○) and (△, ○): everyone is better off after the swap...

- A stable matching is one when there is no blocking pair
- More generally, can define a cost $c(i, j)$ for matching i-th man with j-th woman
- A necessary condition: $c(i, j) + c(i', j') \leq c(i, j') + c(i', j)$

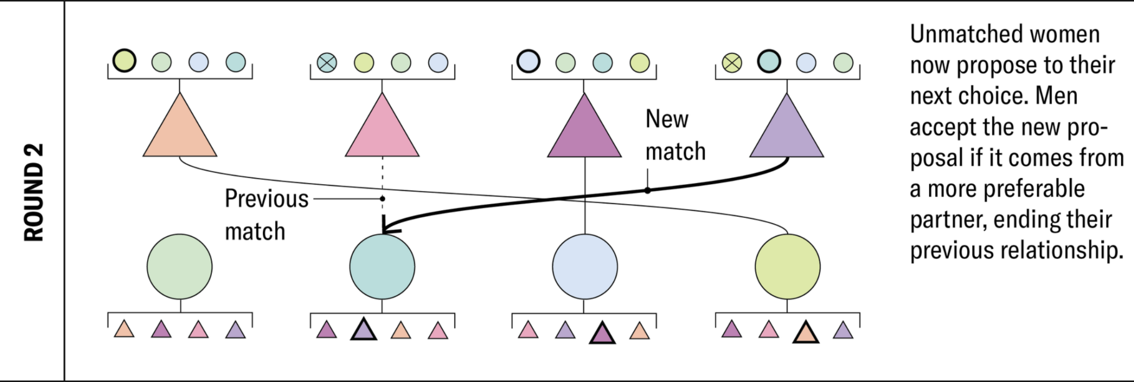
D. Gale and L. S. Shapley. "College Admissions and the Stability of Marriage". *The American Mathematical Monthly*, vol. 69, no. 1 (1962), pp. 9–15.

Gale-Shapley Algorithm



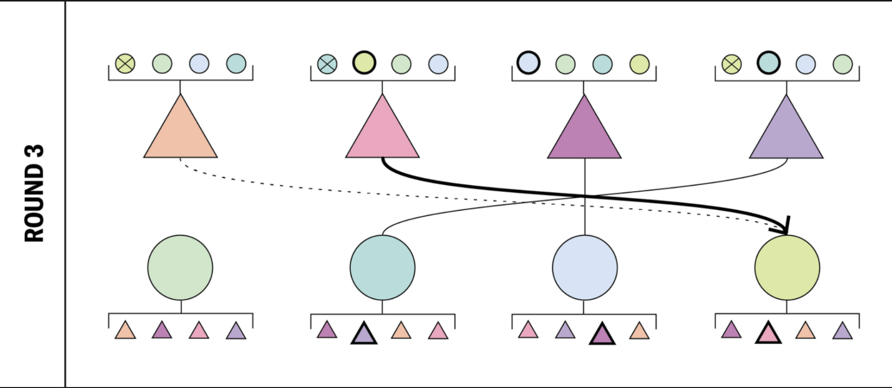
https://tinyurl.com/ej74dbsu

Gale-Shapley Algorithm



https://tinyurl.com/ej74dbsu

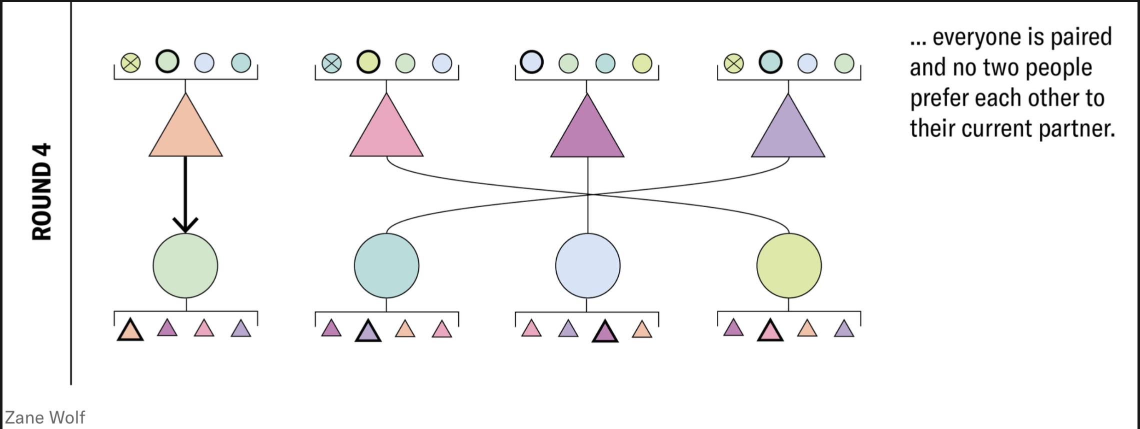
Gale-Shapley Algorithm



The process repeats until ...

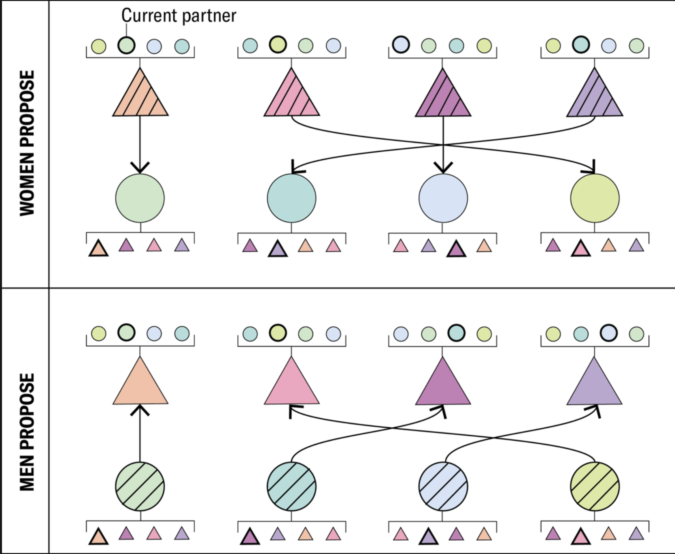
https://tinyurl.com/ej74dbsu

Gale-Shapley Algorithm



https://tinyurl.com/ej74dbsu

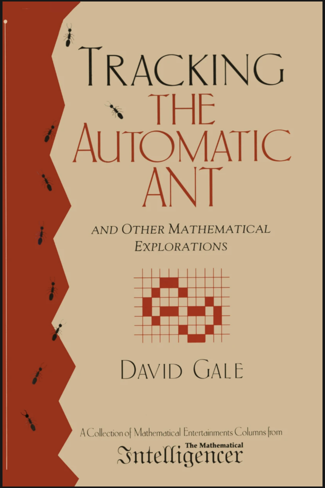
Who Should Propose?

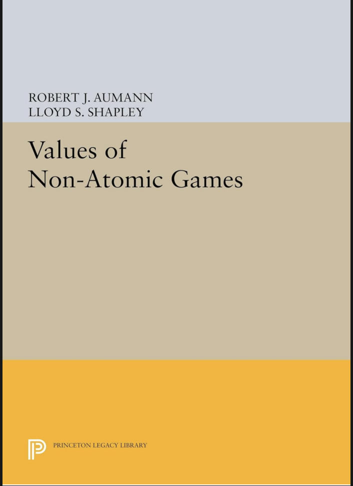


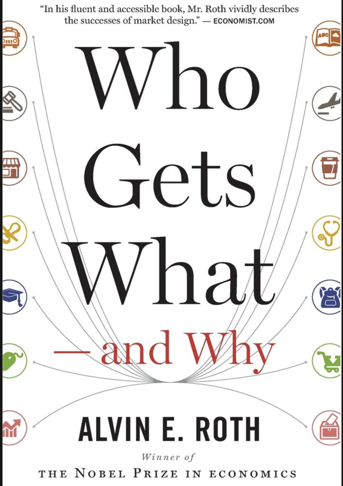
The group making the proposals is indicated by the hash pattern.

On average, the group making the proposals will finish with more desirable matches than the group receiving the proposals.

https://tinyurl.com/ej74dbsu


TRACKING THE AUTOMATIC ANT
AND OTHER MATHEMATICAL EXPLORATIONS
DAVID GALE
A Collection of Mathematical Entertainments Columns from
The Mathematical Intelligencer

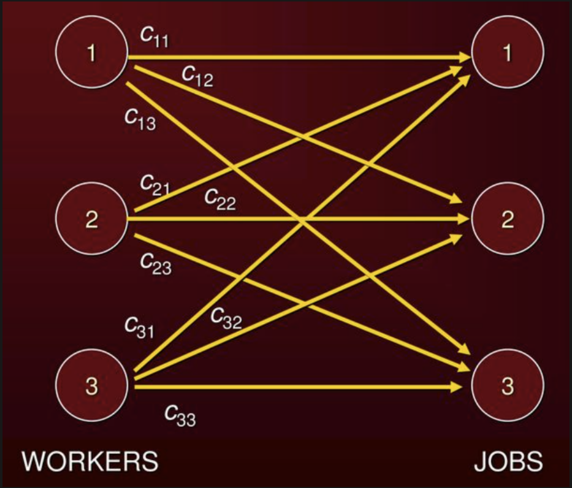

ROBERT J. AUMANN
LLOYD S. SHAPLEY
Values of Non-Atomic Games
PRINCETON LEGACY LIBRARY


"In his fluent and accessible book, Mr. Roth vividly describes the successes of market design." — ECONOMIST.COM
Who Gets What
— and Why
ALVIN E. ROTH
Winner of
THE NOBEL PRIZE IN ECONOMICS

Linear Assignment

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

- Here $\mathbf{T}$ is simply a permutation
- $c(i, \mathbf{T}(i))$ is the cost for matching i with $\mathbf{T}(i)$
- Want to minimize total cost
- Some kind of "stable" permutation



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

Discrete Optimal Transport

$$\min_{\Pi \in \mathscr{P}(\mathbf{p},\mathbf{q})} \sum_{i=1}^{m} \sum_{j=1}^{n} \pi_{ij} c(i,j)$$

- Coupling (joint distribution): $\mathscr{P}(\mathbf{p},\mathbf{q}) := \{\Pi \in \mathbb{R}_{+}^{m \times n} : \Pi \mathbf{1} = \mathbf{p}, \Pi^{\top} \mathbf{1} = \mathbf{q}\}$
  - $\mathbf{p}$ is the marginal distribution over $i = 1, \ldots, m$
  - $\mathbf{q}$ is the marginal distribution over $j = 1, \ldots, n$
  - π_{ij} is the probability of matching i with j
  - the total cost $\sum_{i=1}^{m} \sum_{j=1}^{n} \pi_{ij} c(i,j) = \mathbb{E}[c(I,J)]$, where $(I,J) \sim \Pi$
- Let $m = n$ and $\mathbf{p} = \mathbf{q} = \frac{1}{n} \cdot \mathbf{1}$: a "relaxation" of linear assignment

Iterative Proportional Fitting (IPF), a.k.a. Sinkhorn's Alg

$$\min_{\Pi \in \mathscr{P}(\mathbf{p},\mathbf{q})} \sum_{i=1}^{m} \sum_{j=1}^{n} \overbrace{[\pi_{ij} c_{ij} + \lambda \pi_{ij} \log \pi_{ij}]}^{\lambda \cdot \pi_{ij} \log \frac{\pi_{ij}}{\exp(-c_{ij}/\lambda)}} = \min_{\Pi \in \mathscr{P}(\mathbf{p},\mathbf{q})} \lambda \cdot \mathsf{KL}[\Pi \parallel \exp(-C/\lambda)]$$

- $\lambda > 0$ is a small regularization constant
- $\Gamma := \exp(-C/\lambda)$ is an unnormalized (Boltzmann-Gibbs) distribution
- $\Pi \in \mathscr{P}(\mathbf{p}, \mathbf{q})$ contains two types of constraints that we can treat separately
  - row constraint $\Pi\mathbf{1} = \mathbf{p}$: $\Gamma \leftarrow \mathrm{diag}(\mathbf{p}./(\Gamma\mathbf{1})) * \Gamma$
  - column constraint $\Pi^\top\mathbf{1} = \mathbf{q}$: $\Gamma \leftarrow \Gamma * \mathrm{diag}(\mathbf{q}./(\Gamma^\top\mathbf{1}))$

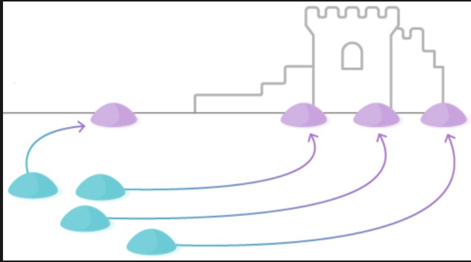
R. Sinkhorn. "A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices". *The Annals of Mathematical Statistics*, vol. 35, no. 2 (1964), pp. 876–879, W. E. Deming and F. F. Stephan. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". *The Annals of Mathematical Statistics*, vol. 11, no. 4 (1940), pp. 427–444.

E. Schrödinger. "Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique". *Annales de l'institut Henri Poincaré*, vol. 2, no. 4 (1932), pp. 269–310.

Monge's Problem

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

- A distribution p of soil and a distribution q of holes to fill
- Let $\mathsf{X} \sim p$ be a random pile of soil
- $\mathbf{T}(\mathsf{X})$ moves X to a hole Y
- Require $\mathbf{T}_{\#}p = q$ to match the mass
  - *a priori*, it is not even clear if such a $\mathbf{T}$ exists!
- Want to minimize expected cost c (effort)



https://tinyurl.com/28ue4c5k

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

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

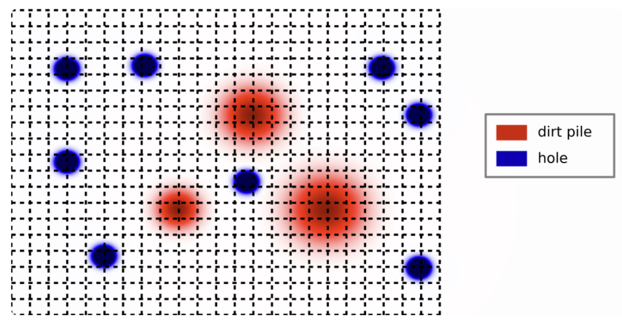
Definition: Coupling

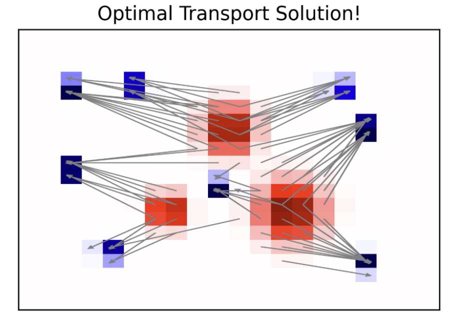
$(\mathsf{X}, \mathsf{Y}) \sim \pi$, where the joint coupling π has marginals p and q

- Deterministic pairing: $\mathbf{x}$ is matched with some $\mathbf{y} = \mathbf{T}\mathbf{x}$
- 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!

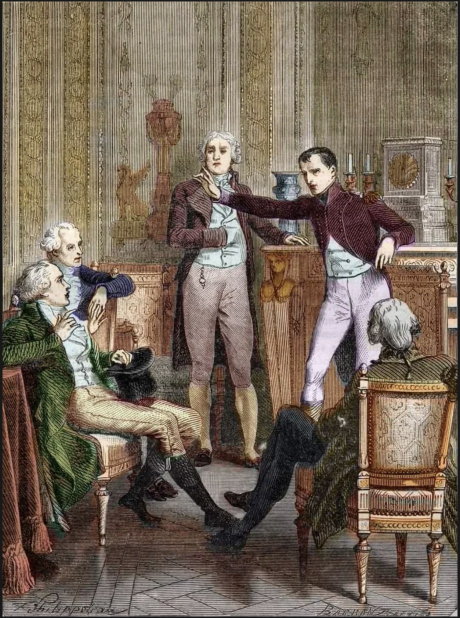
L. V. Kantorovich. "On the Translocation of Masses". *Journal of Mathematical Sciences*, vol. 133, no. 4 (2006). Originally published in Dokl. Akad. Nauk SSSR, vol. 37, No. 7–8, 227–229 (1942)., pp. 1381–1382, L. V. Kantorovich. "On a Problem of Monge". *Journal of Mathematical Sciences*, vol. 133, no.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

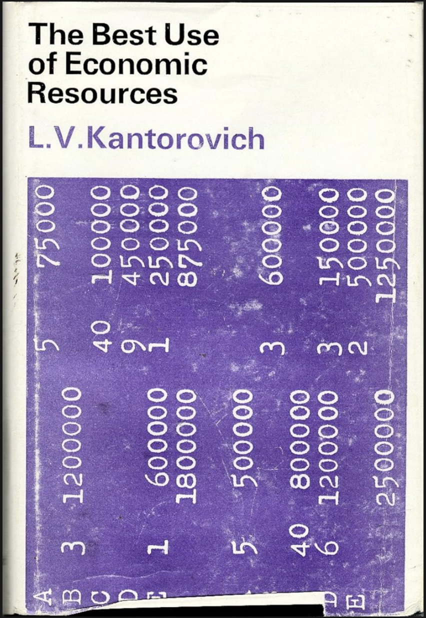
Back to Discrete





https://tinyurl.com/2n7ujhmd




The Best Use of Economic Resources
L.V.Kantorovich

Duality

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

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

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

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

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

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

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

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

- $u(\mathbf{x}) \le [\min_{\mathbf{y}}\ c(\mathbf{x}, \mathbf{y}) - v(\mathbf{y})] =: v^c(\mathbf{x})$
- $v(\mathbf{y}) \le [\min_{\mathbf{x}}\ c(\mathbf{x}, \mathbf{y}) - u(\mathbf{x})] =: u^c(\mathbf{y})$
- 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)


HUUUUUU!!
memegenerator.es

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: $\forall \mathbf{x}, \mathbf{z}, \; v^c(\mathbf{x}) - v^c(\mathbf{z}) \leq d(\mathbf{x}, \mathbf{z})$
  - 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

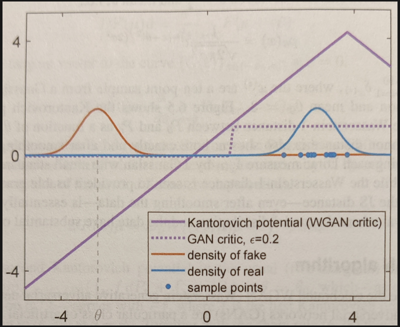
$$\min_{\mathbf{T}} \mathbb{W}_1(q, \mathbf{T}_{\#}r) = \min_{\mathbf{T}} \max_{u} \ \mathbb{E}_{\mathsf{X}\sim q}[u(\mathsf{X})] - \mathbb{E}_{\mathsf{Z}\sim r}[u(\mathbf{T}(\mathsf{Z}))], \quad \text{s.t.} \quad u \text{ is Lipschitz}$$
$$\approx \min_{\mathbf{T}} \max_{u} \ \hat{\mathbb{E}}_{\mathsf{X}\sim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

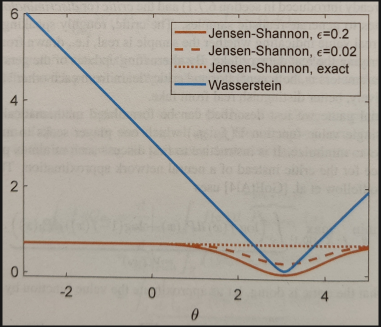
- r is the noise density, e.g., standard normal
- q is the data density: only a training sample is available
- $\mathbf{T}$ is the generator network: maps noise to data
- u is the discriminator network: maps data to a real scalar
  - u is Lipschitz iff $\|\nabla u\| \leq 1 \implies$ penalty on network weights

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

Wasserstein vs. JS/KL

- Wasserstein is a *bona fide* distance; JS/KL is not
- JS/KL enjoys data processing inequality; Wasserstein does not
- Wasserstein difficult to compute; JS/KL can become "flat"





G. Friesecke. "Optimal transport: A comprehensive introduction to modeling, analysis, simulation, applications". SIAM, 2024.

2-Wasserstein Distance

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

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

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

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

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

$$\boxed{q = (\nabla f)_{\#}p}, \quad i.e. \ \ \mathsf{X} \sim p \implies \nabla f(\mathsf{X}) \sim q$$

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.

Mirror Mirror on the Wall

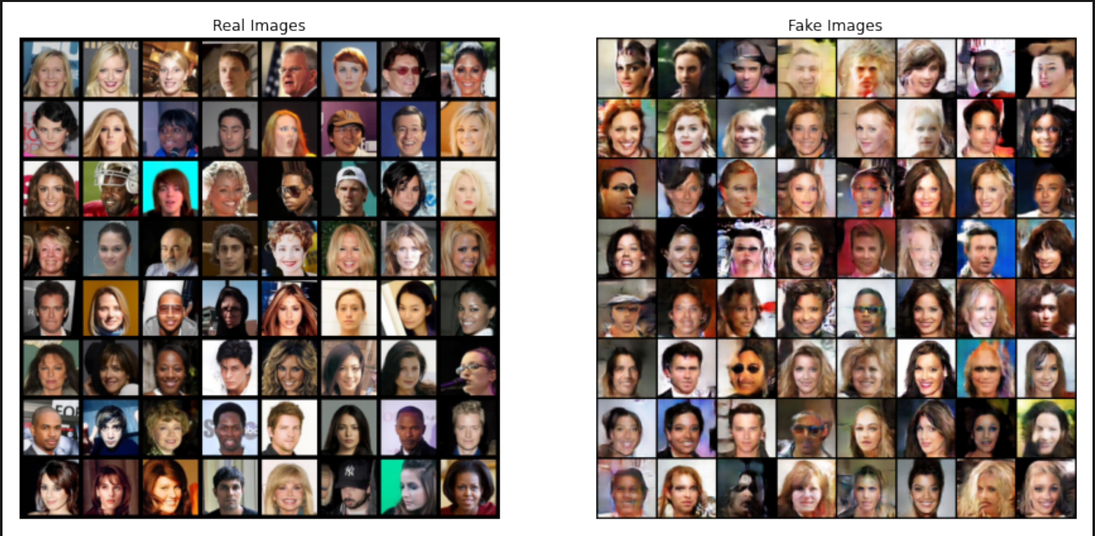
Fréchet Inception Distance (FID)

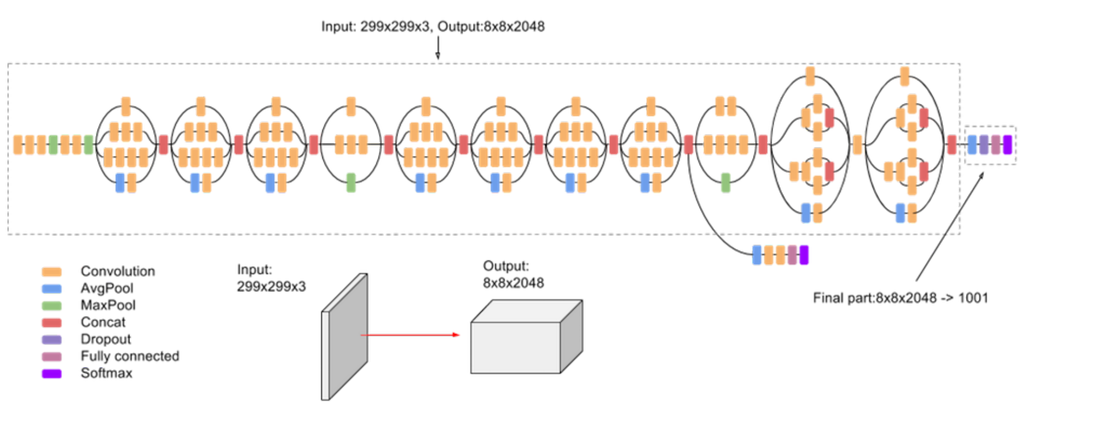
$$\mathbb{W}_2^2(\mathcal{N}(\mathbf{m}_1, \Sigma_1), \mathcal{N}(\mathbf{m}_2, \Sigma_2)) = \|\mathbf{m}_1 - \mathbf{m}_2\|^2 + \operatorname{tr}\left[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right]$$

- Consider the mapping $\mathbf{T}\mathbf{x} = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}(\mathbf{x} - \mathbf{m}_1) + \mathbf{m}_2$
  - $\mathbf{T} = \nabla f$ for some convex function f
- Plug into $\mathbb{E}\|\mathsf{X} - \mathbf{T}\mathsf{X}\|^2$ where $\mathsf{X} \sim \mathcal{N}(\mathbf{m}_1, \Sigma_1)$
- A valid lower bound on $\mathbb{W}_2^2(\mu_1, \mu_2)$
  - provided that μ_i has mean $\mathbf{m}_i$ and covariance Σ_i

M. Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Advances in Neural Information Processing Systems*. 2017.

M. Gelbrich. "On a Formula for the L_2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces". *Mathematische Nachrichten*, vol. 147, no. 1 (1990), pp. 185–203.


Real Images
Fake Images


Input: 299x299x3, Output:8x8x2048
Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmax
Input:
299x299x3
Output:
8x8x2048
Final part:8x8x2048 -> 1001

Potential GAN

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

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

- 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

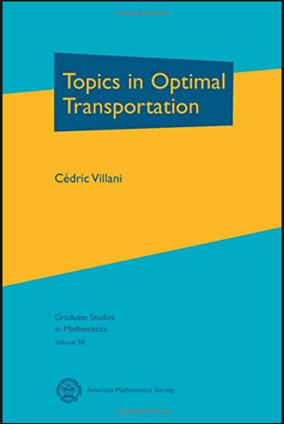
C.-W. Huang, R. T. Q. Chen, C. Tsirigotis, and A. Courville. "Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization". In: *International Conference on Learning Representations*. 2021.

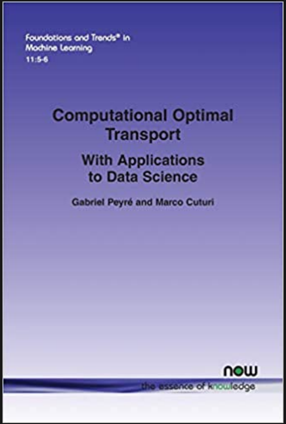
Triangular vs. Potential

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

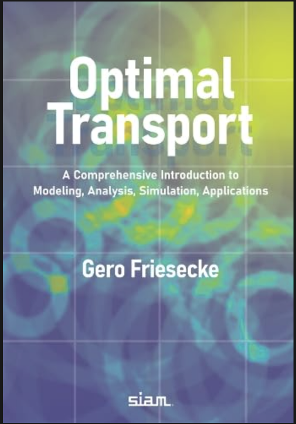
- $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$, $\mathbf{T}_{\#} p = q$
- $\mathbf{T} = \nabla f$ for convex $f : \mathbb{R}^d \to \mathbb{R}$
- $\nabla \mathbf{T} = \nabla^2 f$ is symmetric PSD
- composition fails
- rotationally equivariant

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


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


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


Cambridge studies in advanced mathematics 207
Optimal Mass Transport on Euclidean Spaces
FRANCESCO MAGGI


Optimal Transport
A Comprehensive Introduction to
Modeling, Analysis, Simulation, Applications
Gero Friesecke
siam

Complementarity

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

- $\pi(\mathbf{x},\mathbf{y}) > 0 \implies u(\mathbf{x}) + v(\mathbf{y}) = c(\mathbf{x},\mathbf{y})$
- 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!

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

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

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

- A kind of stability: any rematch cannot further reduce cost!

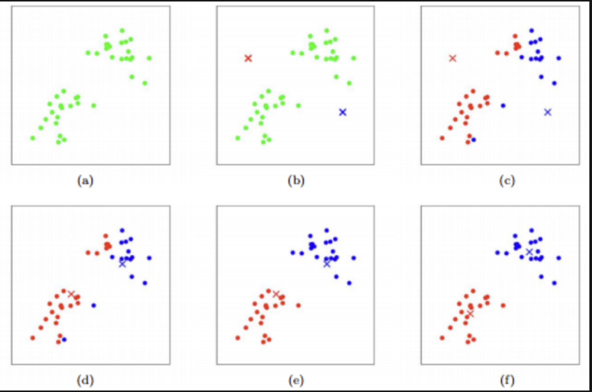
Theorem: Optimal coupling

Let $c : \mathbb{X} \times \mathbb{Y} \to [0, \infty)$ and $\pi \in \mathscr{P}(p, q)$ is a coupling with finite transport cost. If $\operatorname{supp} \pi$ is c-cyclically monotone, then π is optimal.

M. Beiglböck. "Cyclical monotonicity and the ergodic theorem". *Ergodic Theory and Dynamical Systems*, vol. 35, no. 3 (2015), pp. 710–713.

K-means Clustering

$$\min_{\mathbf{z}_1,\ldots,\mathbf{z}_k} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\|\mathbf{x}_i - \mathbf{z}_k\right\|^2$$



S. P. Lloyd. "Least squares quantization in PCM". *IEEE Transactions on Information Theory*, vol. 28, no. 2 (1982). orignally appeared in 1957, pp. 129–137.

K-means as Wasserstein Projection

- Let $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}$ be our empirical distribution
  - $\hat{\mu}_n(A) = \frac{1}{n}\sum_{i=1}^{n} [\![\mathbf{x}_i \in A]\!]$
- Can show k-means solves:

$$\min_{\nu \in \mathscr{P}_k} \mathbb{W}_2(\nu, \hat{\mu}_n)$$

  - $\mathscr{P}_k$ denotes all discrete distributions supported on at most k points
- Gaussian mixture models correspond to entropy regularization:

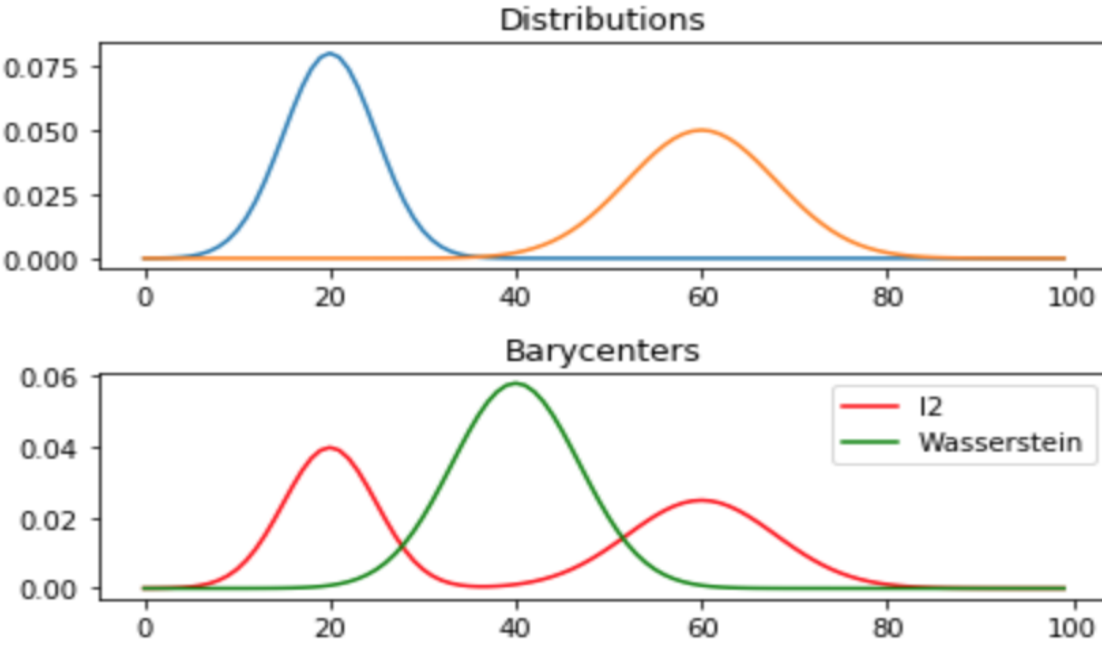
$$\min_{\nu \in \mathscr{P}_k} \mathbb{W}_2^2(\nu, \hat{\mu}_n) - \lambda \cdot \mathrm{entropy}(\nu)$$

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!

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

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