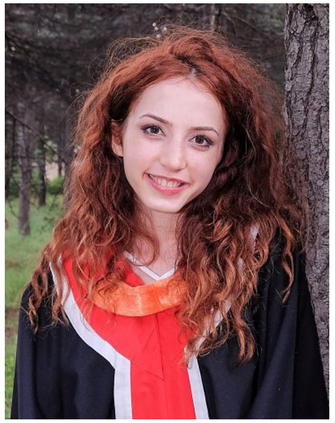
Semi-Discrete Optimal Transport: Hardness, Regularization and Numerical Solution



Soroosh Shafieezadeh Abadeh, Tepper School of Business, CMU

Optimal Transport: Old and New

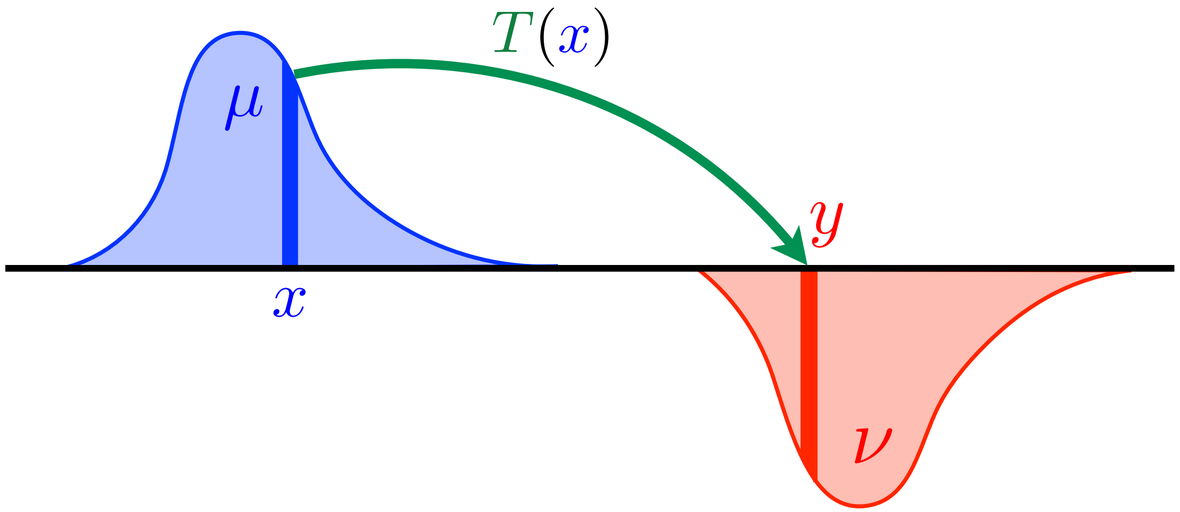
Monge **Hitchcock** **Kantorovich-Koopmans** **Vaserstein** **Brenier** **Villani** **Figalli**

Nobel ‘75

Fields ‘10

Fields ‘18

What is Optimal Transport?



$$\inf_{T:\mathbb{X}\to\mathbb{Y}} \quad \mathbb{E}_{\mu}\left[c(x, T(x)\right]$$
$$\text{s.t.} \quad T\sharp\mu = \nu$$

Nonconvex and infeasible!

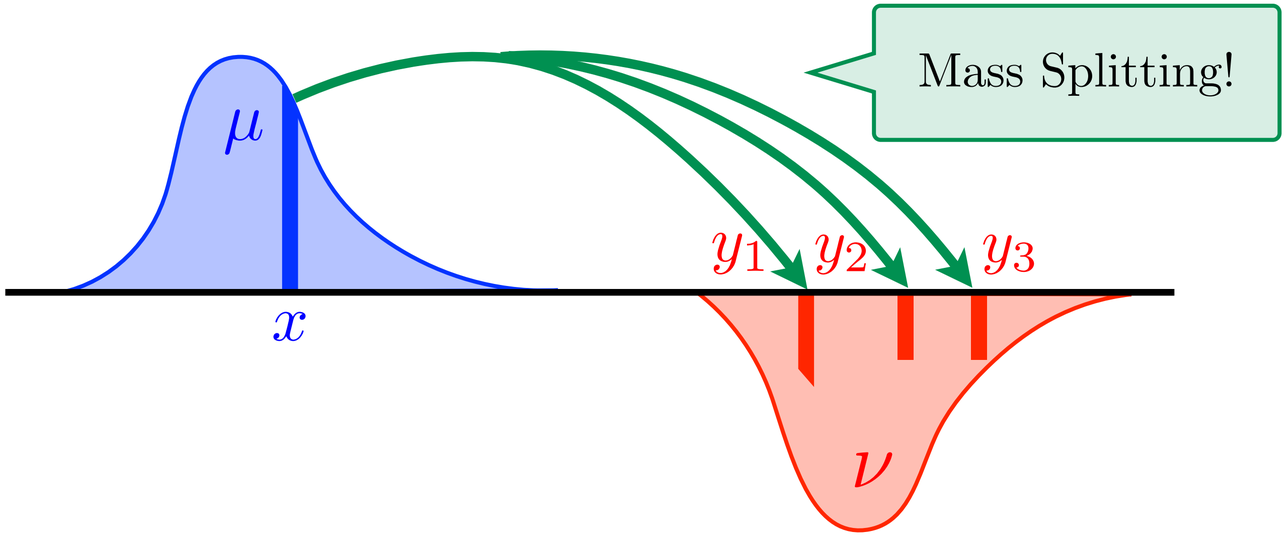
MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.

Par M. MONGE.

LORSQU'ON doit tranſporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit tranſporter, & le nom de *Remblai* à l'eſpace qu'elles doivent occuper après le tranſport.

Le prix du tranſport d'une molécule étant, toutes choſes d'ailleurs égales, proportionnel à ſon poids & à l'eſpace qu'on lui fait parcourir, & par conſéquent le prix du tranſport total devant être proportionnel à la ſomme des produits des molécules multipliées chacune par l'eſpace parcouru, il s'enſuit que le déblai & le remblai étant donnés de figure & de poſition, il n'eſt pas indifférent que telle molécule du déblai ſoit tranſportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine diſtribution à faire des molécules du premier dans le ſecond, d'après laquelle la ſomme de ces produits ſera la moindre poſſible, & le prix du tranſport total ſera un *minimum*.

What is Optimal Transport?



$$\inf_{\pi \in \mathcal{M}(\mathbb{X}, \mathbb{Y})} \quad \mathbb{E}_\pi \left[c(x, y) \right]$$

$$\text{s.t.} \quad \pi \in \Pi(\mu, \nu)$$

Infinite Dimensional LP

THE DISTRIBUTION OF A PRODUCT FROM SEVERAL SOURCES TO NUMEROUS LOCALITIES **1941**

By Frank L. Hitchcock

1. Statement of the problem. When several factories supply a product to a number of cities we desire the least costly manner of distribution. Due to freight rates and other matters the cost of a ton of product to a particular city will vary according to which factory supplies it, and will also vary from city to city.

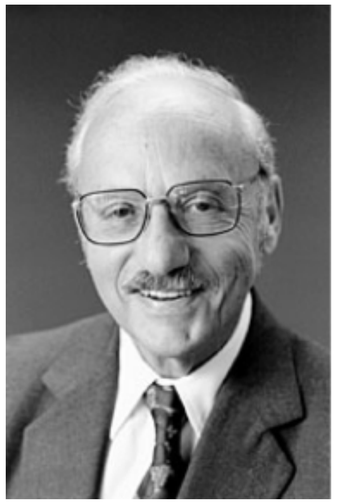
OPTIMUM UTILIZATION OF THE TRANSPORTATION SYSTEM* **1949**

by Tjalling C. Koopmans

Professor of Economics, The University of Chicago, and Research Associate, Cowles Commission for Research in Economics

The purpose of this paper is to give an application of the theory of optimum allocation of resources to one particular industry. I shall, therefore, not speak on that theory in general. I shall use one of its

Discrete OP Problems

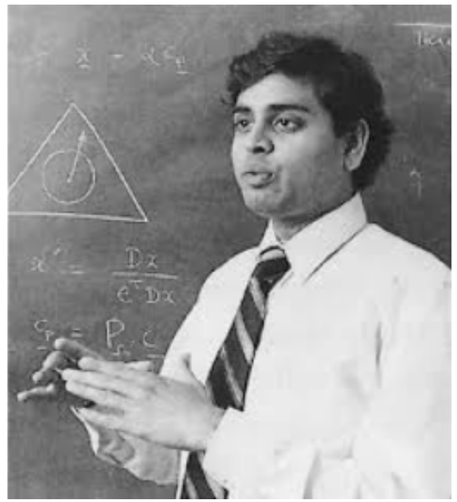


Dantzig

$\mathcal{O}(2^n)$

Khachiyan

$\mathcal{O}(n^4)$



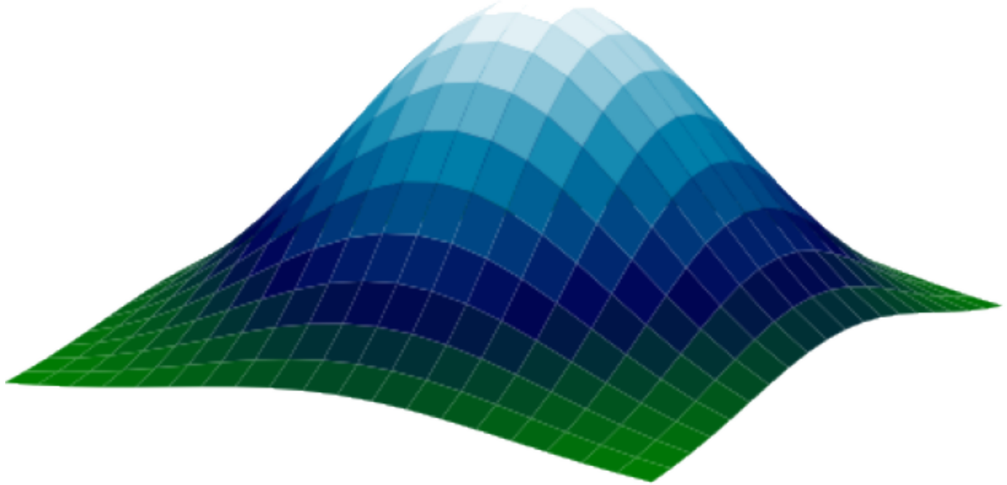
Karmarkar

$\mathcal{O}(n^{3.5})$

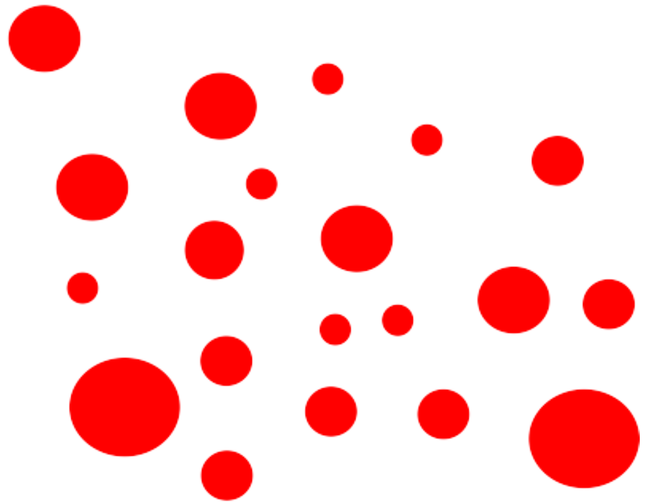
Sinkhorn

$\mathcal{O}(n^2)$

Semi-Discrete OT Problems



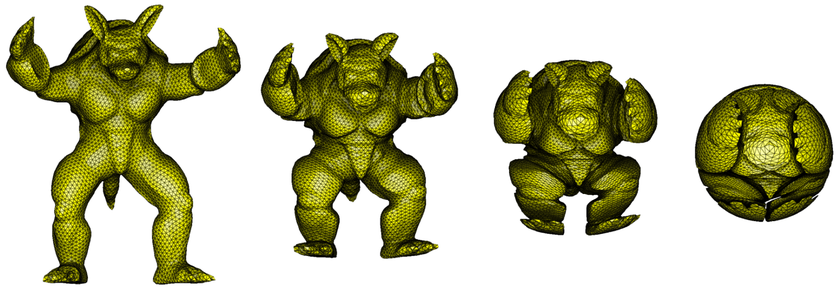
μ



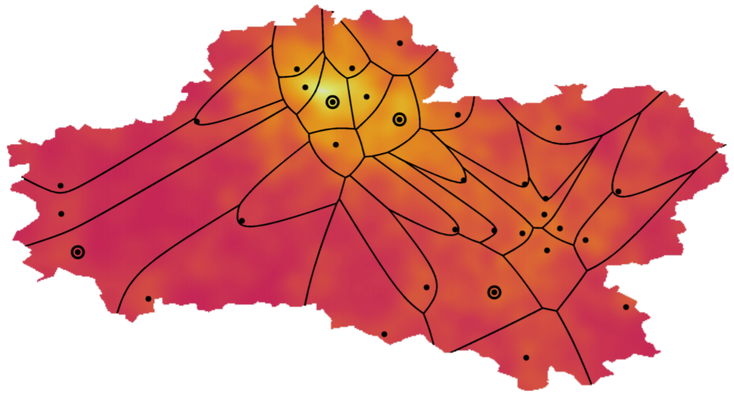
$$\nu = \sum_{i=1}^{n} \nu_i \delta_{y_i}$$

Applications

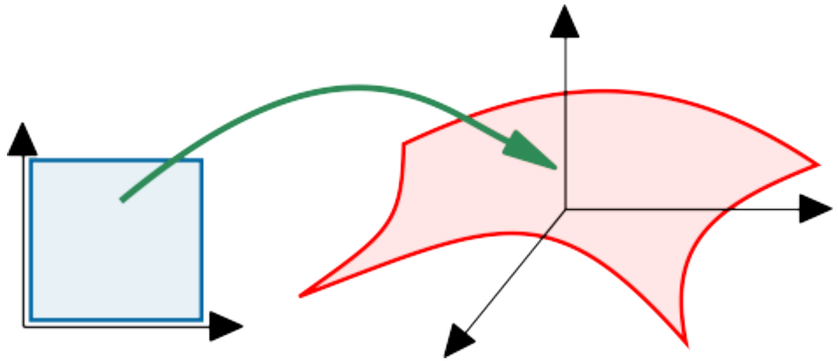
3D morphing [L14]



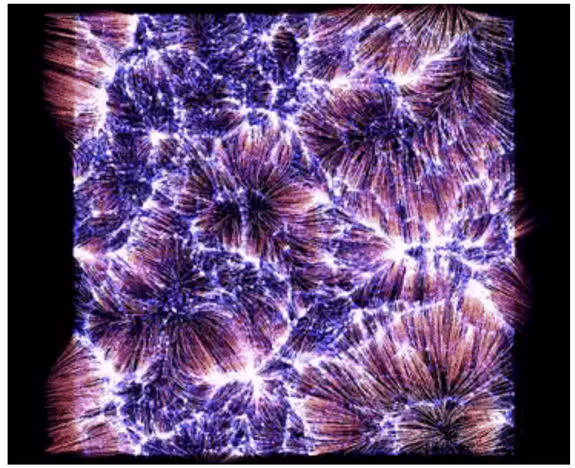
Resource allocation [HS20]



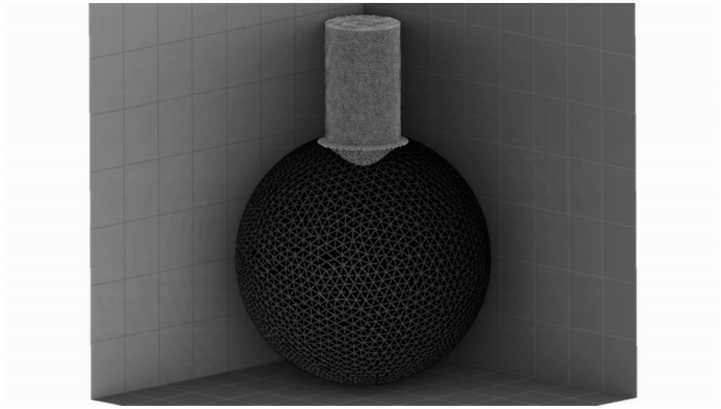
Generative models [HS20]



Reconstruction of early universe [LMHN21]

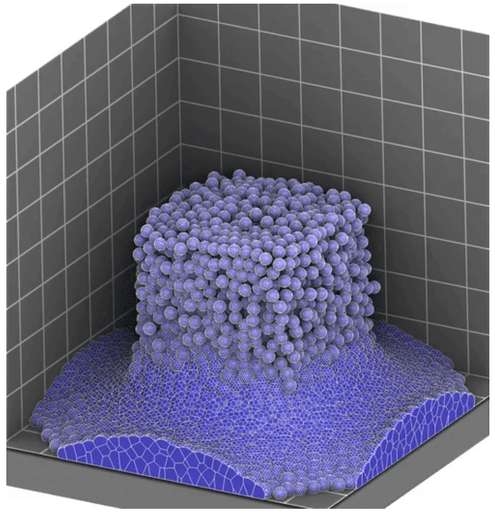


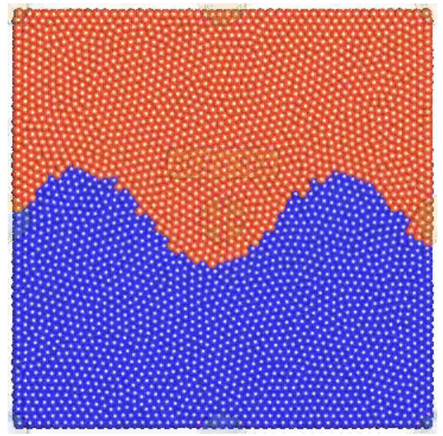
Applications



Simulation of an incompressible bi-phasic flow in a bottle [GWH18]

Free-surface fluid simulation with Gallouet-Merigot scheme [GWH18]





Taylor-Rayleigh instability using the Gallouet-Merigot scheme [GWH18]

Agenda

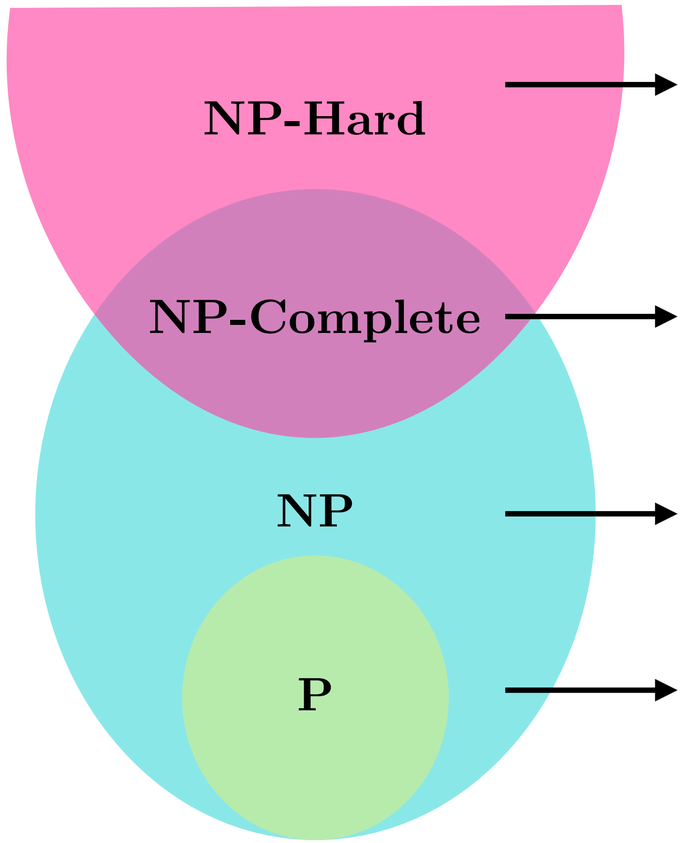
1- Complexity of OT?

2- Smooth OT?

3- Algorithms for OT?

Computational Complexity

Yes **No**



Harder than NP-Complete!

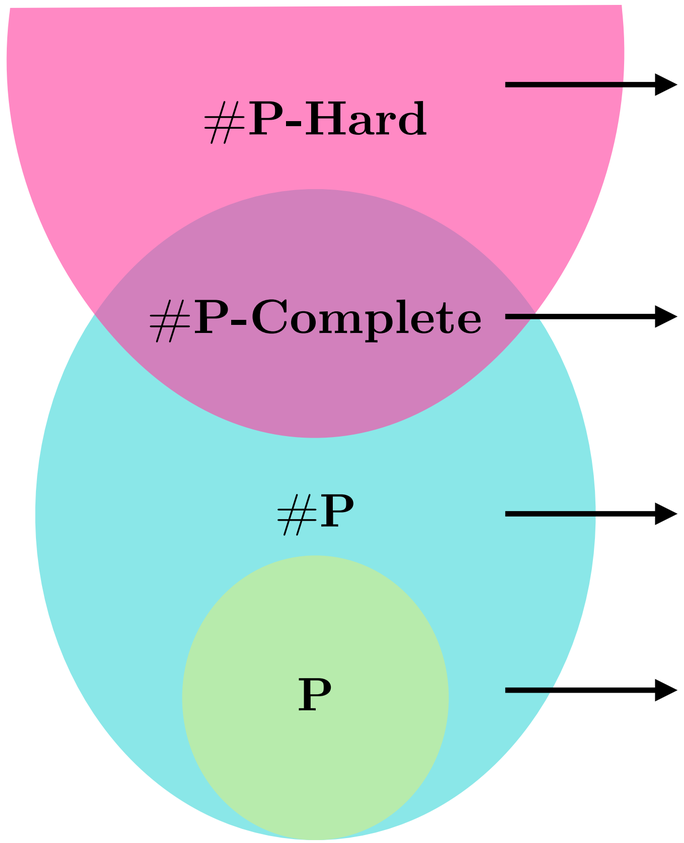
Hardest problem in NP!

Solvable in polynomial time with NTM!

Solvable in polynomial time with DTM!

Computational Complexity

How many yes?



Harder than #P-Complete!

Hardest problem in #P!

Solvable in polynomial time with NTM!

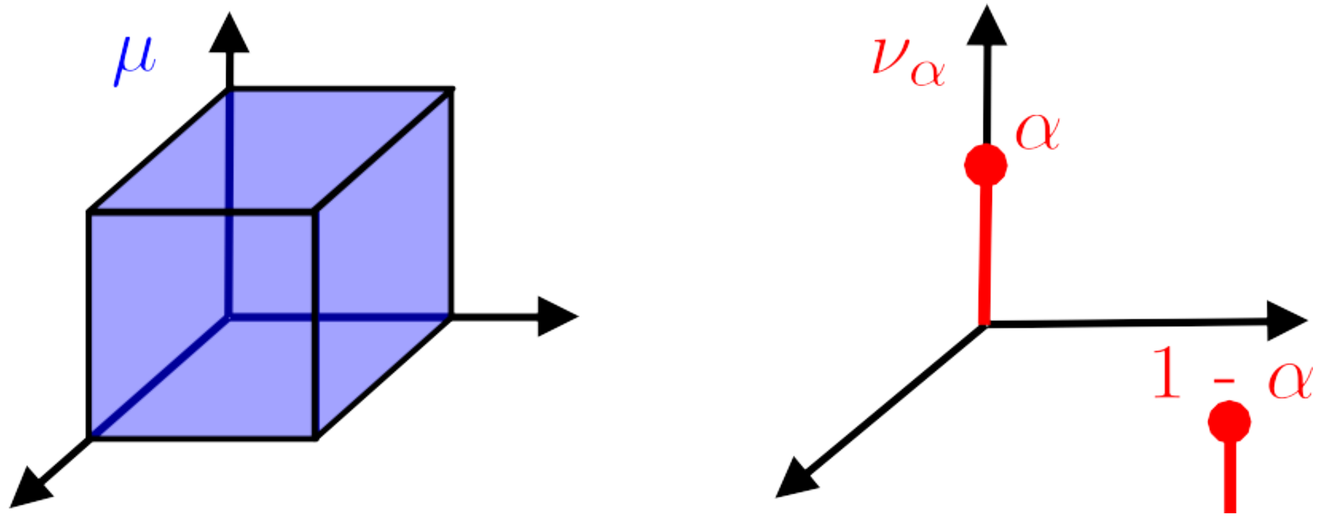
Solvable in polynomial time with DTM!

Computational Complexity of OT

Theorem 1. Computing the OT cost is #P-hard!

Computational Complexity of OT

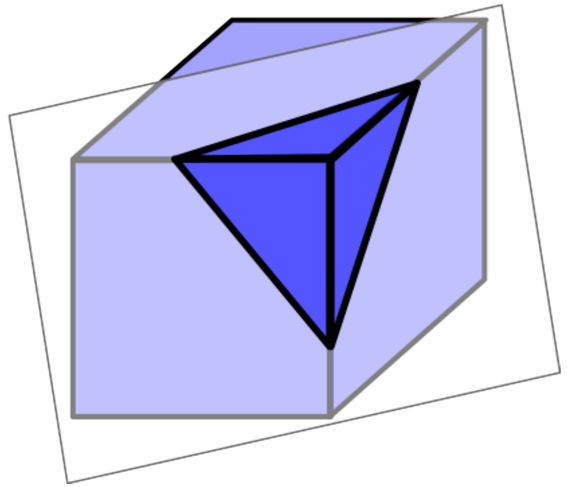
Theorem 1. Computing the OT cost is #P-hard!



$$\min_{\alpha \in [0,1]} W_c(\mu, \nu_\alpha)$$

Computational Complexity of OT

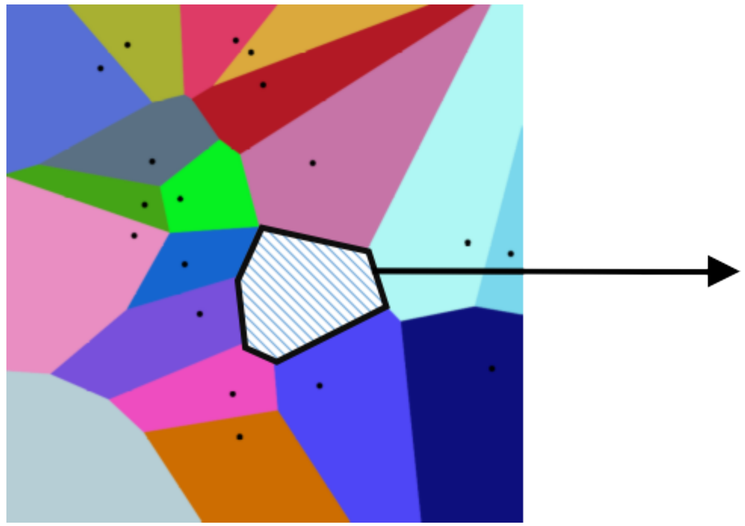
Theorem 1. Computing the OT cost is #P-hard!



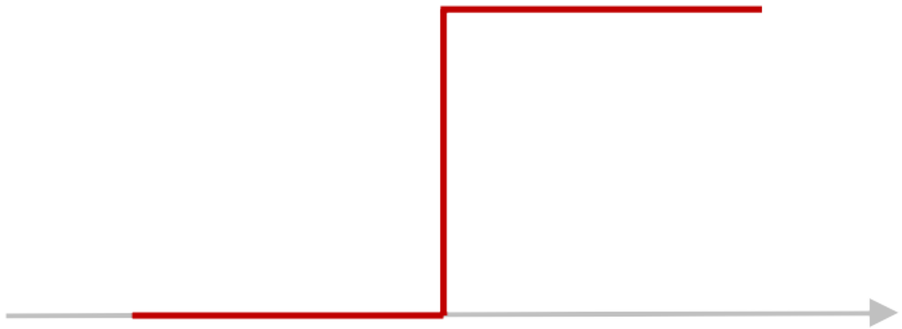
$$\operatorname*{argmin}_{\alpha \in [0,1]} W_c(\mu, \nu_\alpha) = \text{volume of Knapsack polytope}$$

Why is #P-hard?

$$W_c(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi \left[c(x, y) \right]$$

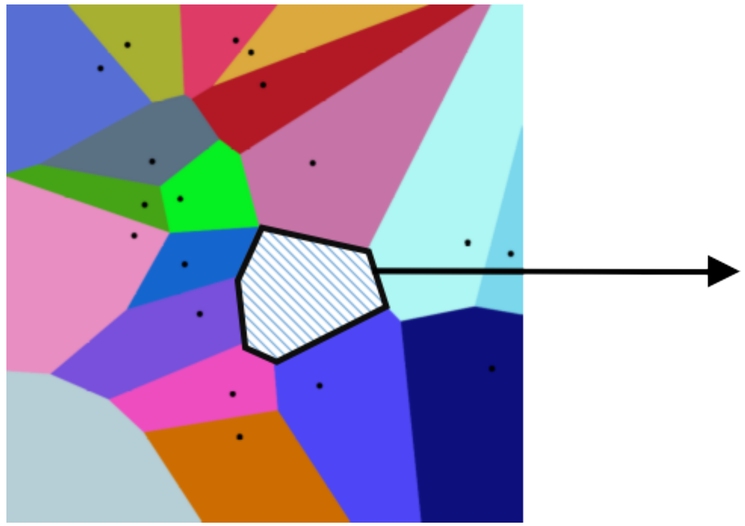


Non-smooth Integrand!

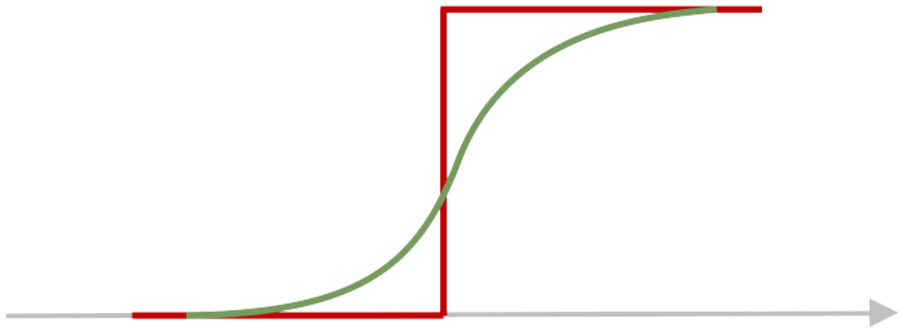


Why is #P-hard?

$$W_c(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi \left[c(x, y) \right]$$



Smooth approximation!



What is #P-hard?

Input: 1- two distributions representable by n bits
2- Accuracy parameter ϵ

Output: OT distance between two distributions

No algorithm can return the OT cost to within accuracy ϵ in time $\mathcal{O}(\text{poly}(n)\log(1/\epsilon))$

What is #P-hard?

Input: 1- two distributions representable by n bits
2- Accuracy parameter ϵ

Output: OT distance between two distributions

~~No~~ algorithm can return the OT cost to within accuracy ϵ in time $\mathcal{O}(\text{poly}(n)\,\text{~~log}(1/\epsilon))$
$\text{poly}(1/\varepsilon)$

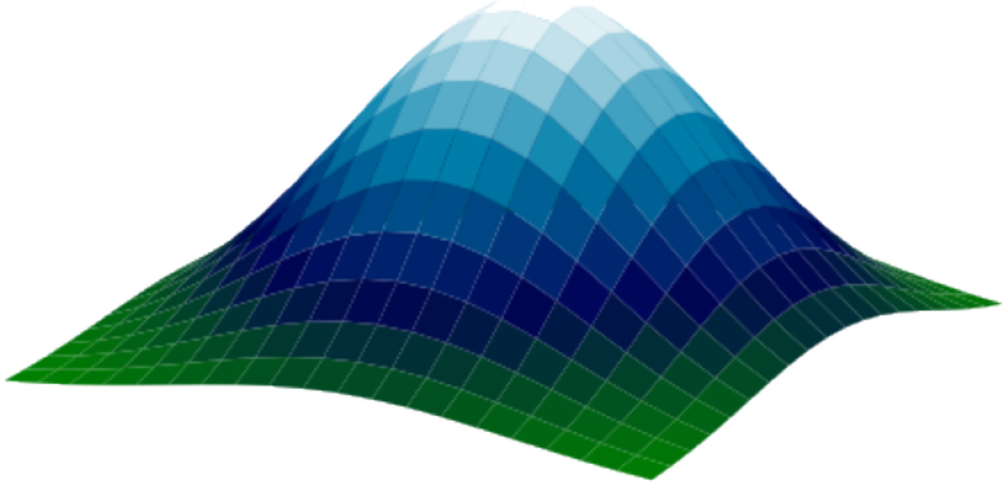
Agenda

1- Complexity of OT? #P-Hard

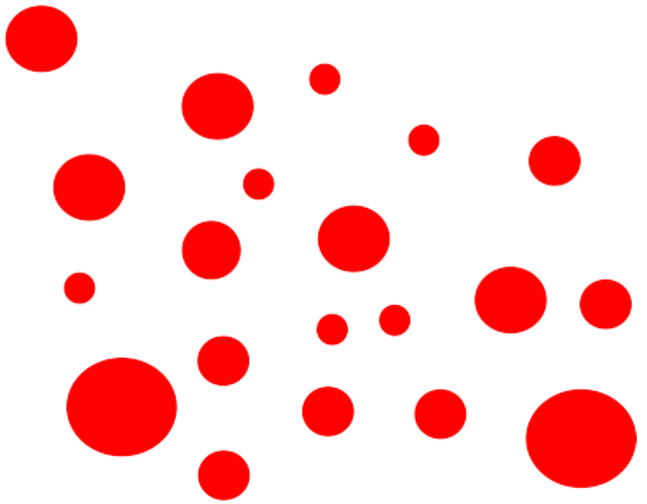
2- Smooth OT?

3- Algorithms for OT?

Semi-discrete case



μ

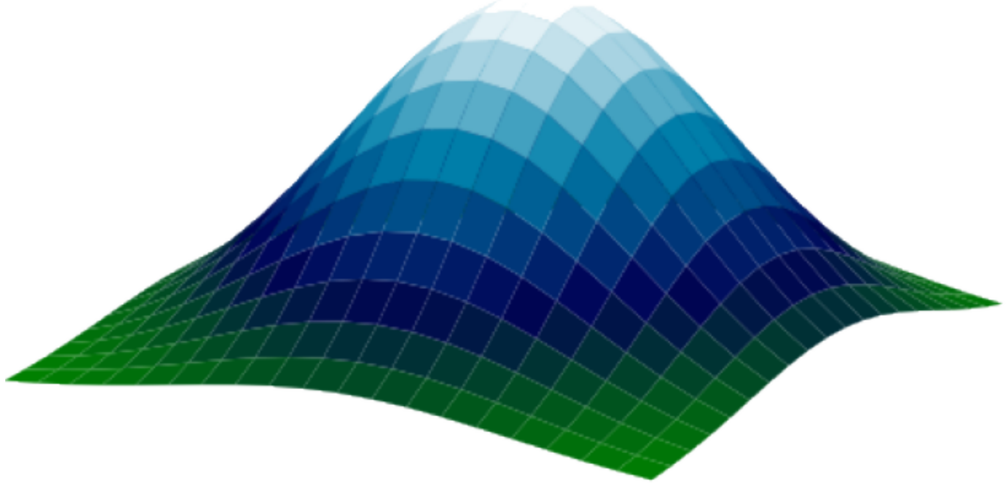


$$\nu = \sum_{i=1}^{n} \nu_i \delta_{y_i}$$

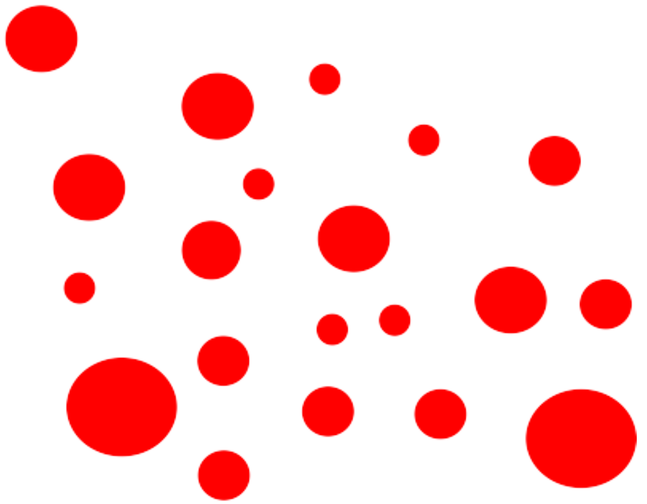
$$W_c(\mu, \nu) = \sup_{\phi \in \mathbb{R}^n} \sum_{i=1}^{n} \nu_i \phi_i - \mathbb{E}_\mu \left[\max_{i \in [n]} \phi_i - c(x, y_i) \right]$$

c-transform: $\psi_c(\phi, x)$

Semi-discrete case



μ

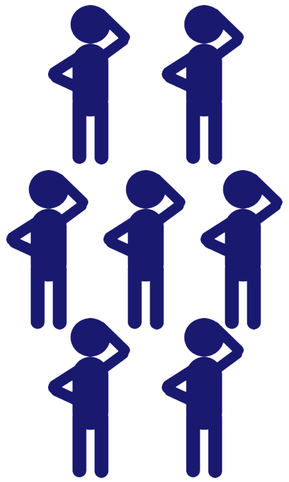


$$\nu = \sum_{i=1}^{n} \nu_i \delta_{y_i}$$

$$W_c(\mu, \nu) = \sup_{\phi \in \mathbb{R}^n} \sum_{i=1}^{n} \nu_i \phi_i - \mathbb{E}_\mu \left[\max_{i \in [n]} \phi_i - c(x, y_i) \right]$$

smooth c-transform: $\overline{\psi}_c(\phi, x) = \log\left(\sum_{i \in [n]} \exp(\phi_i - c(x, y_j)) \right)$

Multinomial Logit Models

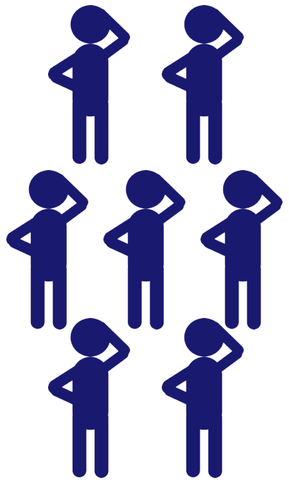


$$\mathbb{E}_{z\sim\omega}[\max_{i\in[n]} u_i + z_i] = \log(\sum_{i=1}^{n} \exp(u_i))$$

Gumbel distribution

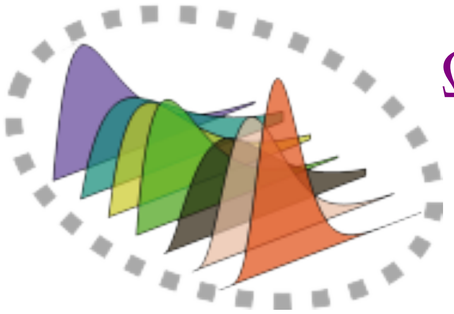
[McF73, B85, ASA88, VB98]

Ambiguous Discrete Choice Models



$$\sup_{\omega \in \Omega} \; \mathbb{E}_{z \sim \omega} \left[\max_{i \in [n]} u_i + z_i \right]$$

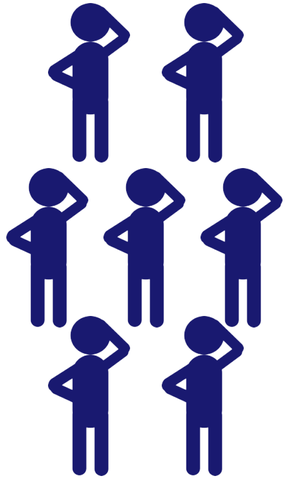
Fréchet Ambiguity Set



$$\Omega = \{\omega \in \mathcal{M}(\mathbb{R}^n) : \omega(z_i \leq s) = F_i(s)\}$$

[NST09, MNPTL14, FLW17]

Ambiguous Discrete Choice Models



$$\sup_{\omega \in \Omega} \ \mathbb{E}_{z \sim \omega} \left[\max_{i \in [n]} \ u_i + z_i \right]$$

$$= \max_{p \in \Delta^n} \sum_{i \in [n]} u_i p_i + \int_{1-p_i}^{1} F_i^{-1}(s) \mathrm{d}s$$

[NST09, MNPTL14, FLW17]

Smooth OT

- Smooth c-transform

$$\psi_c(\phi, x) = \max_{i \in [n]} \phi_i - c(x, y_i) \longrightarrow \overline{\psi}_c(\phi, x) = \sup_{\omega \in \Omega} \mathbb{E}_{z \sim \omega} \left[\max_{i \in [N]} \phi_i - c(x, y_i) + z_i \right]$$

- Smooth OT

$$\overline{W}_c(\mu, \nu) = \sup_{\phi \in \mathbb{R}^n} \sum_{i=1}^{n} \nu_i \phi_i - \mathbb{E}_\mu \left[\overline{\psi}_c(\phi, x) \right]$$

Regularized OT = Smooth OT

Theorem 2. If $\Omega = \{\omega \in \mathcal{M}(\mathbb{R}^n) : \omega(z_i \leq s) = F_i(s)\}$ with

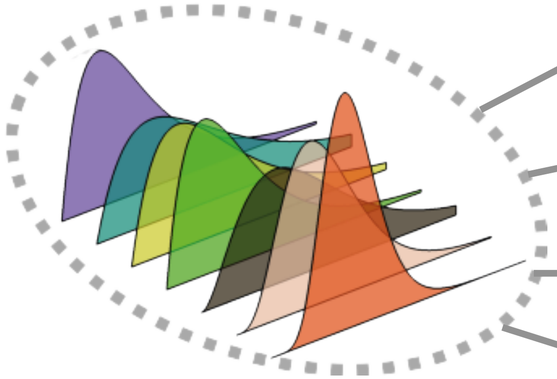
$$F_i(s) = \min\{1, \max\{0, 1 - \nu_i F(-s)\}\}$$

Let $f(s) = \int_0^s F^{-1}(t)\mathrm{d}t$. Then we have

$$\overline{W}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi \left[c\left(x, y\right)\right] + D_f(\pi \parallel \mu \otimes \nu)$$

$$D_f(\tau \parallel \rho) = \int_{\mathcal{Z}} f(\mathrm{d}\tau/\mathrm{d}\rho(z))\rho(\mathrm{d}z)$$

Examples



Exponential distributions $\Longrightarrow$ KL divergence

Uniform distributions $\Longrightarrow$ χ^2-divergence

Pareto distributions $\Longrightarrow$ Tsallis divergence

$t-$distributions $\Longrightarrow$ Tikhonov regularization

Agenda

1- Complexity of OT? #P-Hard

2- Smooth OT = Regularized OT

3- Algorithms for OT?

Averaged Stochastic Gradient Descent

$$x \sim \mu \implies \nabla_\phi H(\phi, x)$$

$$\min_{\phi \in \mathbb{R}^n} \left\{ h(\phi) = \mathbb{E}_\mu \left[H(\phi, x) \right] \right\}$$

Input: $\phi_0 \in \mathbb{R}^n, \eta \in [0, 1]$

for $t = 0, \ldots, T-1$

$$\phi_{t+1} \leftarrow \phi_t - \eta \nabla_\phi H(\phi_t, x_t)$$

endfor

Output: average$(\phi_0, \ldots, \phi_T)$

Definitions

- **R-Lipschitz:** $|h(\phi) - h(\phi')| \leq R\|\phi - \phi'\|$

- **L-Smooth:** $\|\nabla h(\phi) - \nabla h(\phi')\| \leq L\,\|\phi - \phi'\|$

- **U-Strongly Convex:** $(\nabla h(\phi) - \nabla h(\phi'))^\top(\phi - \phi') \geq U\,\|\phi - \phi'\|^2$

- **M-(Generalized) Self-Concordant:** $v(s) = h(\phi + s(\phi' - \phi))$

$$\left|\frac{\mathrm{d}^3 v(s)}{\mathrm{d}s^3}\right| \leq M\,\|\phi - \phi'\|\,\frac{\mathrm{d}^2 v(s)}{\mathrm{d}s^2}$$

Averaged Stochastic Gradient Descent

$$\min_{\phi \in \mathbb{R}^n} \left\{ h(\phi) = \mathbb{E}_\mu \left[H(\phi, x) \right] \right\}$$

$$x \sim \mu \implies \nabla_\phi H(\phi, x)$$

Input: $\phi_0 \in \mathbb{R}^n, \eta \in [0, 1]$

for $t = 0, \ldots, T-1$

$$\phi_{t+1} \leftarrow \phi_t - \eta \nabla_\phi H(\phi_t, x_t)$$

endfor

Output: average$(\phi_0, \ldots, \phi_T)$

→ Lipschitz: $\mathcal{O}(1/\sqrt{T})$

→ Smooth: $\mathcal{O}(1/\sqrt{T})$

→ Strongly Convex: $\mathcal{O}(1/T)$

→ Self-Concordant: $\mathcal{O}(1/T)$

[Tsy03, NV08, NJLS09, SST10, Lan12, Bac13, BM14]

Averaged Stochastic Gradient Descent for OT

$$\min_{\phi \in \mathbb{R}^n} \left\{ h(\phi) = \mathbb{E}_\mu \left[H(\phi, x) \right] \right\}$$

$$H(\phi, x) = \overline{\psi}_c(\phi, x) - \sum_{i \in [n]} \nu_i \phi_i$$

$$x \sim \mu \implies \nabla_\phi H(\phi, x) = \nabla_\phi \overline{\psi}_c(\phi, x) - \nu$$

Structural Results

Theorem 3. If $\Omega = \{\omega \in \mathcal{M}(\mathbb{R}^n) : \omega(z_i \leq s) = F_i(s)\}$, then $\overline{\psi}_c$ is

- **1-Lipschitz:** F_i is continuous
- **L-Smooth:** F_i is L-Lipschitz
- **M-(Generalized) Self-Concordant:**

$$\sup_{s \in F_i^{-1}(0,1)} \frac{|\mathrm{d}^2 F_i(s)/\mathrm{d}s^2|}{\mathrm{d}F_i(s)/\mathrm{d}s} \leq M$$

Inexact Averaged Stochastic Gradient Descent

$$\min_{\phi \in \mathbb{R}^n} \left\{ h(\phi) = \mathbb{E}_{\mu} \left[H(\phi, x) \right] \right\}$$

$$H(\phi, x) = \overline{\psi}_c(\phi, x) - \sum_{i \in [n]} \nu_i \phi_i$$

$$x \sim \mu \implies \nabla_\phi H(\phi, x) = \nabla_\phi \overline{\psi}_c(\phi, x) - \nu$$

- computing gradients = solving regularized LPs
- numerical errors

Inexact Averaged SGD

Theorem 4. Set $\varepsilon_t \leq \mathcal{O}(1/\sqrt{t})$ and $\eta = \mathcal{O}(1/\sqrt{T})$

- **Lipschitz:** $\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right] - h(\boldsymbol{\phi}^\star) \leq \mathcal{O}(1/\sqrt{T})$
- **Smooth:** $\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t}\right)\right] - h(\boldsymbol{\phi}^\star) \leq \mathcal{O}(1/T) + \mathcal{O}(1/\sqrt{T})$
- **Self-Concordant:** $\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right] - h(\boldsymbol{\phi}^\star) \leq \mathcal{O}(1/T)$

Convergence Behavior