Optimal Transport

Theory, Computation and Applications

Wenchong Huang

School of Mathematical Sciences,
Zhejiang University.

Dec. 30th, 2024

Overview

Principal concern: the distance between two probability measures.
First introduced in 1781 by Monge.
Relative subjects: probability theory, geometry, graph theory, machine learning...
Applications:

- Image registration and warping;
- Reflector design;
- Retrieving information from shadowgraphy and proton radiography;
- Seismic tomography and reflection seismology.

Some well-known researchers:

- Gasoard Monge (France);
- Leonid Kantorovich (Russia);
- Yann Brenier (France);
- Xianfeng Gu (顾险峰, China);

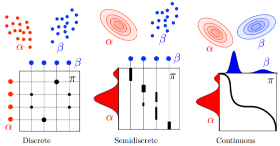


Fig. 1. Three main scenarios for Kantorovich OT

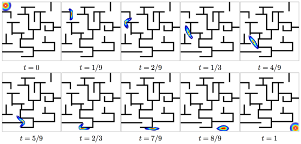


Fig. 2. Solving maze with OT

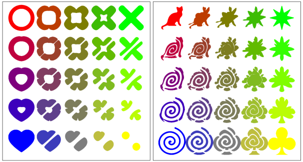


Fig. 3. 2D shape interpolation with OT

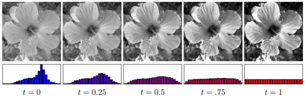


Fig. 4. Histogram equalization with OT

1. Theory
2. Computation

The sand-moving problem

A child wants to make a pile of sand in the shape of a castle.
Cost: 1 kcal per shovel and per meter horizontally.
Target: Minimize the total cost.

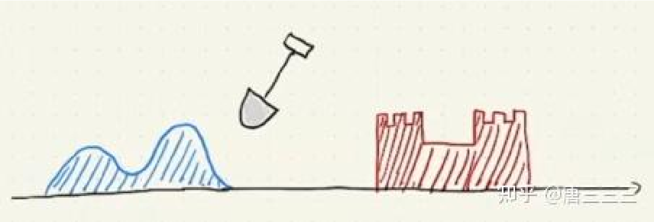


Fig. 5. The sand-moving problem.

The sand-moving problem

A child wants to make a pile of sand in the shape of a castle.
Cost: 1 kcal per shovel and per meter horizontally.
Target: Minimize the total cost.

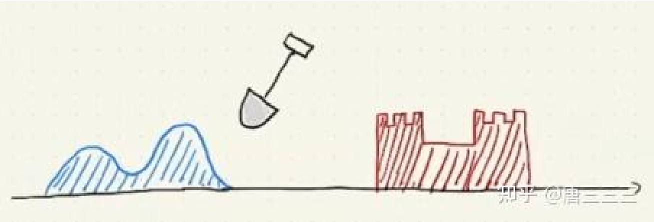


Fig. 5. The sand-moving problem.

Let's denote the source shape by $f(x)$ and the target by $g(x)$. The sand-moving problem cound be formulated as: find a **transport mapping** $T : \mathbb{R} \to \mathbb{R}$ to minimize

$$\int_{\mathbb{R}} |T(x) - x| f(x)\ dx, \tag{1}$$

which satisfies

$$\int_{T(U)} g(x)\ dx = \int_{U} f(x)\ dx \text{ for all open interval } U \subset \mathbb{R}. \tag{2}$$

The allocation problem

There are some steel coils to be transported from warehouses to factories. The transport cost is \$1 per coil and per kilometer. How to minimize the total cost?

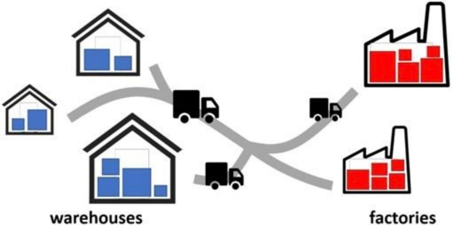


Fig. 6. The allocation problem.

The allocation problem

There are some steel coils to be transported from warehouses to factories. The transport cost is $1 per coil and per kilometer. How to minimize the total cost?

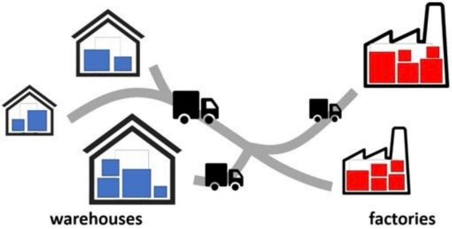


Fig. 6. The allocation problem.

Assume the i-th warehouse has a_i coils and the j-th factory needs b_i coils. And assume the distance between the i-th warehouse and the j-th factory is d_{ij}. The allocation problem could be formulated as: find a **transport matrix** v_{ij} to minimize

$$\sum_{i,j} d_{ij} v_{ij} \tag{3}$$

which satisfies

$$a_i = \sum_j v_{ij}, \quad \forall i, \qquad \text{and} \qquad b_j = \sum_i v_{ij}, \quad \forall j. \tag{4}$$

The Monge formulation

Denote $\mathcal{M}_+^1(\mathcal{X})$ the set of probability measures on $\mathcal{X}$.

Definition (push-forward)

Suppose $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and a map $T : \mathcal{X} \to \mathcal{Y}$. Say $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ is the push-forward of μ by T if

$$\int_{\mathcal{Y}} h(y)\, d\nu(y) = \int_{\mathcal{X}} h(T(x))\, d\mu(x), \quad \forall h \in \mathcal{C}(\mathcal{Y}). \tag{5}$$

Write $T_\# \mu := \nu$.

Example (push-forward of a discrete measure)

Suppose α is a discrete measure

$$\alpha = \sum_{i=1}^n a_i \delta_{x_i}.$$

Then the push-forward of α by T is

$$T_\# \alpha = \sum_{i=1}^n a_i \delta_{T(x_i)}.$$

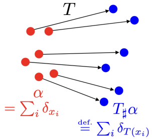


Fig. 7. push-forward of a discrete measure

[1] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences* (1781).

The Monge formulation

Denote $\mathcal{M}_+^1(\mathcal{X})$ the set of probability measures on $\mathcal{X}$.

Definition (push-forward)

Suppose $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and a map $T : \mathcal{X} \to \mathcal{Y}$. Say $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ is the push-forward of μ by T if

$$\int_{\mathcal{Y}} h(y)\, d\nu(y) = \int_{\mathcal{X}} h(T(x))\, d\mu(x), \quad \forall h \in \mathcal{C}(\mathcal{Y}). \tag{5}$$

Write $T_\# \mu := \nu$.

Example (push-forward of a discrete measure)

Suppose α is a discrete measure

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}.$$

Then the push-forward of α by T is

$$T_\# \alpha = \sum_{i=1}^{n} a_i \delta_{T(x_i)}.$$

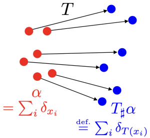


Fig. 7. push-forward of a discrete measure

Given two probability measures μ on $\mathcal{X}$ and ν on $\mathcal{Y}$, and a cost function $c(x, y)$. Optimal transport could be generally formulated as the Monge problem:

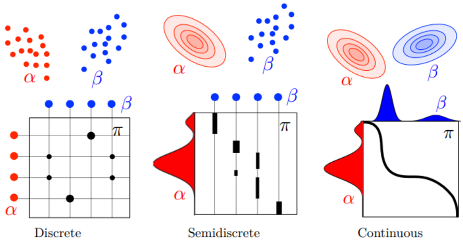
$$\min_{T} \left\{ \int_{\mathcal{X}} c(x, T(x))\, d\mu(x) : T_\# \mu = \nu \right\} \tag{6}$$

The Monge problem between discrete measures is introduced by Monge[1].

[1] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences* (1781).

The Kantorovich formulation

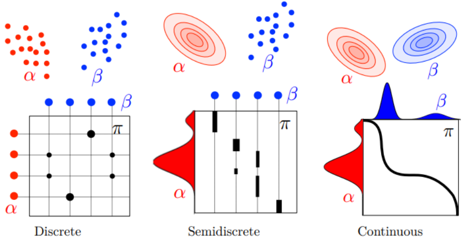
Here's another general formulation of OT, we first recall the three main scenarios for OT.



[2] Leonid Kantorovich. "On the transfer of masses". In: *Doklady Akademii Nauk* 37.2 (1942).

The Kantorovich formulation

Here's another general formulation of OT, we first recall the three main scenarios for OT.



Given two probability measures μ on $\mathcal{X}$ and ν on $\mathcal{Y}$, and a cost function $c(x, y)$. Optimal transport could be generally formulated as the Kantorovich problem[2]:

$$\mathcal{L}_c(\mu, \nu) = \min_{\pi} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\ d\pi(x, y), \tag{7}$$

where π is a measure on $\mathcal{X} \times \mathcal{Y}$, whose marginals are μ and ν, that is,

$$\mu = \int_{\mathcal{Y}} \pi(\cdot, y)\ dy, \qquad \nu = \int_{\mathcal{X}} \pi(x, \cdot)\ dx. \tag{8}$$

[2] Leonid Kantorovich. "On the transfer of masses". In: *Doklady Akademii Nauk* 37.2 (1942).

Wasserstein disrtance

Here we suppose $\mathcal{X} = \mathcal{Y}$ and $c(x, y) = d(x, y)^p$ $(p > 1)$, where d is a distance on $\mathcal{X}$.

[3] Cédric Villani. *Optimal Transport: Old and New*. Vol. 338. Springer Verlag, 2009.

Wasserstein disrtance

Here we suppose $\mathcal{X} = \mathcal{Y}$ and $c(x, y) = d(x, y)^p$ $(p > 1)$, where d is a distance on $\mathcal{X}$.

Theorem (Wasserstein distance)

Under the above assumptions, $\mathcal{L}_c(\mu, \nu)^{1/p}$ is a distance on $\mathcal{M}_+^1(\mathcal{X})$.

The distance $\mathcal{W}_p(\mu, \nu) := \mathcal{L}_c(\mu, \nu)^{1/p}$ is called p-Wasserstein distance.

[3] Cédric Villani. *Optimal Transport: Old and New*. Vol. 338. Springer Verlag, 2009.

Wasserstein disrtance

Here we suppose $\mathcal{X} = \mathcal{Y}$ and $c(x, y) = d(x, y)^p$ $(p > 1)$, where d is a distance on $\mathcal{X}$.

Theorem (Wasserstein distance)

Under the above assumptions, $\mathcal{L}_c(\mu, \nu)^{1/p}$ is a distance on $\mathcal{M}_+^1(\mathcal{X})$.

The distance $\mathcal{W}_p(\mu, \nu) := \mathcal{L}_c(\mu, \nu)^{1/p}$ is called p-Wasserstein distance.

Definition (weak convergence)

Suppose $\mathcal{X}$ is compact. Say $(\mu_k)_{k\geq 1} \subset \mathcal{M}_+^1(\mathcal{X})$ converges weakly to $\mu \in \mathcal{M}_+^1(\mathcal{X})$ if

$$\int_{\mathcal{X}} g \, d\mu_k \to \int_{\mathcal{X}} g \, d\mu, \quad \forall g \in \mathcal{C}(\mathcal{X}). \tag{9}$$

Theorem (Wasserstein distance and weak convergence[3])

On a compact domain $\mathcal{X}$, $(\mu_k)_{k\geq 1} \subset \mathcal{M}_+^1(\mathcal{X})$ converges weakly to $\mu \in \mathcal{M}_+^1(\mathcal{X})$ if and only if $\mathcal{W}_p(\mu_k, \nu) \to 0$.

[3] Cédric Villani. *Optimal Transport: Old and New.* Vol. 338. Springer Verlag, 2009.

Equivalence between the Kantorovich and Monge problems

Theorem (Kantorovich dual problem)

The Kantorovich problem can be solved in the dual space by

$$\mathcal{L}_c(\mu, \nu) = \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\ d\mu(x) + \int_{\mathcal{Y}} g(y)\ d\nu(y), \tag{10}$$

where the set of admissible dual potential is

$$\mathcal{R}(c) := \{(f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) : \forall (x, y), f(x) + g(y) \leq c(x, y)\}. \tag{11}$$

The pair (f, g) is called Kantorovich potentials.

[4] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics* 44.4 (1991).

Equivalence between the Kantorovich and Monge problems

Theorem (Kantorovich dual problem)

The Kantorovich problem can be solved in the dual space by

$$\mathcal{L}_c(\mu,\nu) = \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\ d\mu(x) + \int_{\mathcal{Y}} g(y)\ d\nu(y), \tag{10}$$

where the set of admissible dual potential is

$$\mathcal{R}(c) := \{(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) : \forall (x,y), f(x) + g(y) \le c(x,y)\}. \tag{11}$$

The pair (f, g) is called Kantorovich potentials.

Theorem (Brenier[4])

In the case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|_2^2$, if at least one of the two input measures (denoted μ) has a density ρ_μ with respect to the Lebesgue measure, then the optimal π in the Kantorovich formulation is unique and is supported on the graph $(x, T(x))$ of a Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$. This means that $\pi = (\mathrm{Id}, T)_{\#}\mu$, i.e.

$$\int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\ d\pi(x,y) = \int_{\mathcal{X}} h(x, T(x))\ d\mu(x), \quad \forall h \in \mathcal{C}(\mathcal{X}\times\mathcal{Y}). \tag{12}$$

Furthermore, this map T is uniquely defined as the gradient of a convex function φ, $T(x) = \nabla\varphi(x)$, where φ is the unique (up to an additive constant) convex function such that $(\nabla\varphi)_{\#}\mu = \nu$. This convex function is related to the dual potential f solving (10) as

$$\varphi(x) = \frac{\|x\|_2^2}{2} - f(x). \tag{13}$$

[4] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics* 44.4 (1991).

1-D discrete case

Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. Suppose $\alpha = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\beta = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$ where $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$.

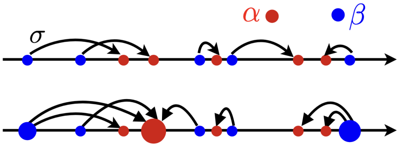


Fig. 8. 1-D optimal transport in discrete case

Then the p-Wasserstein distance can be simply computed by

$$\mathcal{W}_p(\alpha, \beta)^p = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|^p. \tag{14}$$

It's in fact a greedy algorithm.

1-D continuous case

If μ, ν are 1-D measures with densities. Suppose their cummulative distribution functions are $\mathcal{C}_\mu$ and $\mathcal{C}_\nu$, respectively. Then the $\mathcal{W}_1$ distance could be computed by

$$\mathcal{W}_1(\mu, \nu) = \int_{\mathbb{R}} |\mathcal{C}_\mu(x) - \mathcal{C}_\nu(x)| \, dx = \int_{\mathbb{R}} \left| \int_{-\infty}^{x} d(\mu - \nu) \right| \, dx. \tag{15}$$

And the Monge map is then defined by

$$T = \mathcal{C}_\nu^{-1} \circ \mathcal{C}_\mu. \tag{16}$$

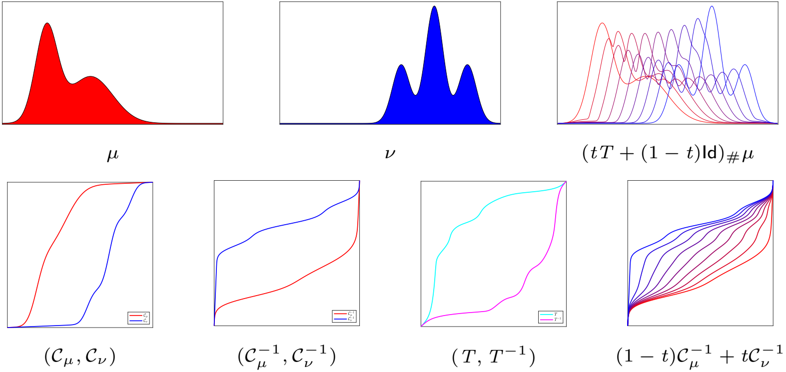


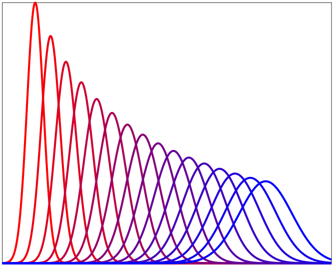
Fig. 9. Computation of OT and displacement interpolation between two 1-D measures.

1-D Gaussian

If $\mu = \mathcal{N}(m_1, \sigma_1^2), \nu = \mathcal{N}(m_2, \sigma_2^2)$ are 1-D Gaussians. Then the $\mathcal{W}_2$ distance can be directly computed by

$$\mathcal{W}_2(\mu, \nu) = \sqrt{|m_1 - m_2|^2 + |\sigma_1 - \sigma_2|^2}, \tag{17}$$

which is thus the Euclidean distance on the 2-D plane plotting the mean and the standard deviation of a Gaussian $\mathcal{N}(m, \sigma)$.



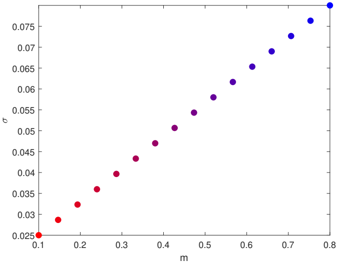


Fig. 10. Computation of displacement interpolation between two 1-D Gaussians.

Learn more in [Takatsu, 2011][5].

[5] Asuka Takatsu. "Wasserstein geometry of Gaussian measures". In: *Osaka Journal of Mathematics* 48.4 (2011).

Discretization

Suppose μ is a measure with density ρ, supported on $[0, 1]$. Let

$$\tilde{\mu} = \sum_{i=0}^{N} u_i \delta_{x_i}, \tag{18}$$

where

$$u_i = \frac{\rho(x_i)}{N+1}, \quad x_i = \frac{i}{N}, \quad i = 0, ..., N. \tag{19}$$

We call $\tilde{\mu}$ the *discretization* of μ.

Discretization

Suppose μ is a measure with density ρ, supported on $[0,1]$. Let

$$\tilde{\mu} = \sum_{i=0}^{N} u_i \delta_{x_i}, \tag{18}$$

where

$$u_i = \frac{\rho(x_i)}{N+1}, \quad x_i = \frac{i}{N}, \quad i = 0, ..., N. \tag{19}$$

We call $\tilde{\mu}$ the *discretization* of μ.

Let $\tilde{\nu} = \sum_{i=0}^{M} v_i \delta_{y_i}$ and $(\boldsymbol{C})_{ij}$ be the cost matrix. The Kantorovich problem then becomes

$$L_{\boldsymbol{C}}(\mathbf{u}, \mathbf{v}) := \min_{\boldsymbol{P} \in U(\mathbf{u}, \mathbf{v})} \langle \boldsymbol{P}, \boldsymbol{C} \rangle := \min_{\boldsymbol{P} \in U(\mathbf{u}, \mathbf{v})} \sum_{i,j} \boldsymbol{P}_{ij} \boldsymbol{C}_{ij}, \tag{20}$$

where

$$U(\mathbf{u}, \mathbf{v}) := \left\{ \boldsymbol{P} \middle| \sum_j \boldsymbol{P}_{ij} = u_i, \forall i, \quad \text{and} \quad \sum_i \boldsymbol{P}_{ij} = v_j, \forall j \right\}. \tag{21}$$

Entropy regularization

Define the entropy

$$H(\boldsymbol{P}) := -\sum_{i,j} \boldsymbol{P}_{ij}(\log(\boldsymbol{P}_{ij}) - 1). \tag{22}$$

Then the regularized Kantorovich problem[6] is defined by

$$L^{\varepsilon}_{\boldsymbol{C}}(\boldsymbol{u}, \boldsymbol{v}) := \min_{\boldsymbol{P} \in U(\boldsymbol{u}, \boldsymbol{v})} \langle \boldsymbol{P}, \boldsymbol{C} \rangle - \varepsilon H(\boldsymbol{P}). \tag{23}$$

It can be shown that $L^{\varepsilon}_{\boldsymbol{C}}(\boldsymbol{u}, \boldsymbol{v}) = L_{\boldsymbol{C}}(\boldsymbol{u}, \boldsymbol{v}) + O(\varepsilon)$.

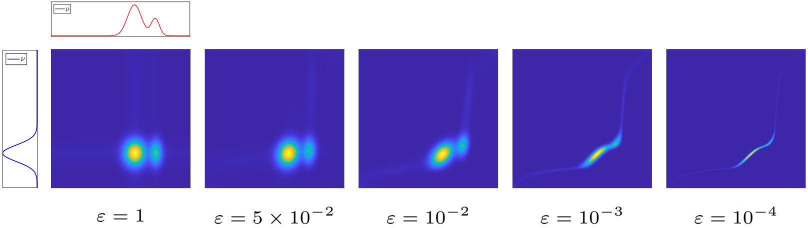


Fig. 11. Graphs of optimal $\boldsymbol{P}$s when choose different ε. Set $\boldsymbol{C}_{ij} = |x_i - x_j|^2$.

[6] Alan G. Wilson. "The use of entropy maximizing models, in the theory of trip distribution, mode split and route split". In: *Journal of Transport Economics and Policy* (1969), pp. 108–126.

Sinkhorn iteration

Let $\boldsymbol{K}_{ij} = e^{-\frac{\boldsymbol{C}_{ij}}{\varepsilon}}$. Sinkhorn iteration writes

$$\mathbf{a}^{(l+1)} \leftarrow \frac{\boldsymbol{u}}{\boldsymbol{K}\boldsymbol{b}^{(l)}}, \quad \text{and} \quad \boldsymbol{b}^{(l+1)} \leftarrow \frac{\mathbf{v}}{\boldsymbol{K}^T \mathbf{a}^{(l+1)}}, \quad \text{for } l = 0, 1, \dots \tag{24}$$

which starts with an arbitary $\boldsymbol{b}^{(0)}$. The transport matrix $\boldsymbol{P}$ can be rebuilt by

$$\boldsymbol{P}^{(l)} = \operatorname{diag}\left(\boldsymbol{b}^{(l)}\right) \cdot \boldsymbol{K} \cdot \operatorname{diag}\left(\mathbf{a}^{(l)}\right). \tag{25}$$

The convergence is proved by Sinkhorn[7]. And Altschuler et al[8] give an analysis of the computational complexity.

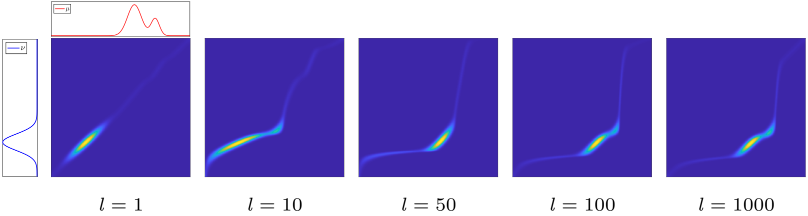


Fig. 12. Graphs of $\boldsymbol{P}^{(l)}$. Set $\boldsymbol{C}_{ij} = |x_i - x_j|^2$ and $\varepsilon = 10^{-3}$.

[7] Richard Sinkhorn. "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *Annals of Mathematical Statistics* 35 (1964).

[8] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems* (2017).

Thank You