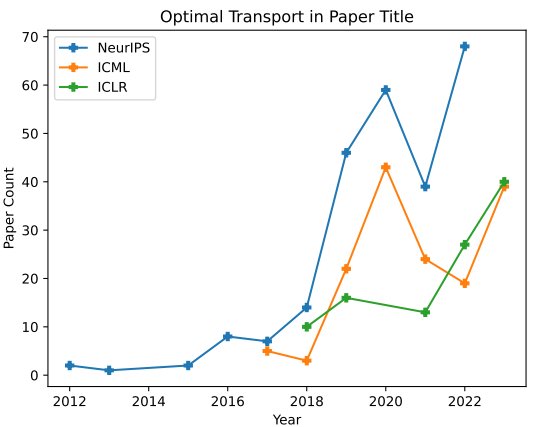
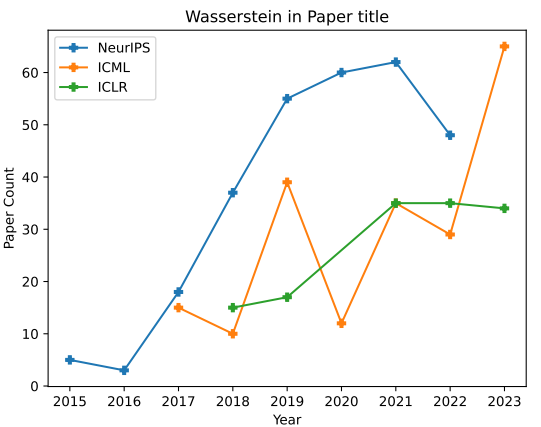
Introduction to Optimal Transport

ML Reading Group

Tareq Si Salem
Northeastern University

Why You Should Care About Optimal Transport?

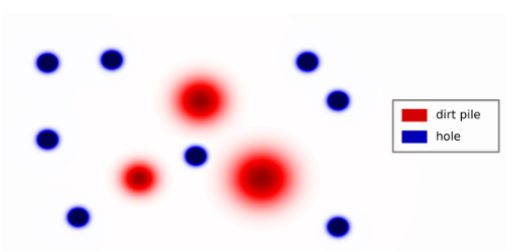




Monge Soil-transport Problem

Figure. Gaspard Monge (Optimal Transport, 1781)

In 1781, Gaspard Monge (founder of ENS and École Polytechnique and participant in the French Revolution) tackled a seemingly simple task: efficiently moving dirt for embankments. This birthed the field of optimal transport, with applications far beyond just dirt!



Monge Formulation

Consider the probem involving efficient distribution of muffins from bakeries to coffee shops. For simplicity, assume six bakeries and coffee shops represented in Figure 1 (red points - bakeries, blue points - coffee shops). We aim to minimize the total travel time. Denote by $C_{i,j} \in \mathbb{R}_{\geq 0}$ the travel time between bakery $i \in [6] \triangleq \{1, 2, 3, 4, 5, 6\}$ and coffee shop $j \in [6]$ ($C_{3,4} = 10$ implies a ten-minute commute). Each café is connected to one and only one bakery. We will note the permutation

$$\sigma : i \in [6] \longrightarrow j \in [6] \tag{1}$$

such a choice of connections.

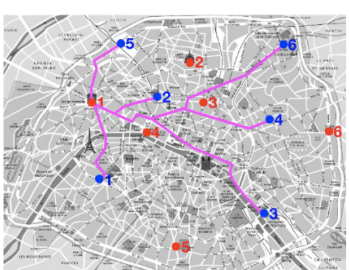
Monge Formulation

Consider the probem involving efficient distribution of muffins from bakeries to coffee shops. For simplicity, assume six bakeries and coffee shops represented in Figure 1 (red points - bakeries, blue points - coffee shops). We aim to minimize the total travel time. Denote by $C_{i,j} \in \mathbb{R}_{\geq 0}$ the travel time between bakery $i \in [6] \triangleq \{1,2,3,4,5,6\}$ and coffee shop $j \in [6]$ ($C_{3,4} = 10$ implies a ten-minute commute). Each café is connected to one and only one bakery. We will note the permutation

$$\sigma : i \in [6] \longrightarrow j \in [6] \tag{1}$$

such a choice of connections.

c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15



c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15

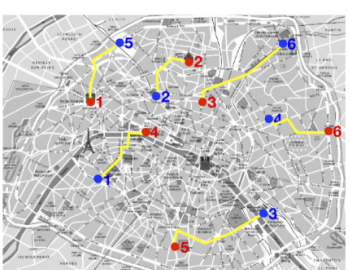


Figure. Cost matrix and associated connections. Left: a row of the cost matrix. Right: a particular example of permutation ($\sigma(1) = 5, \sigma(2) = 2, \sigma(3) = 6, \sigma(4) = 1, \sigma(5) = 3, \sigma(6) = 4$).

Monge Formulation

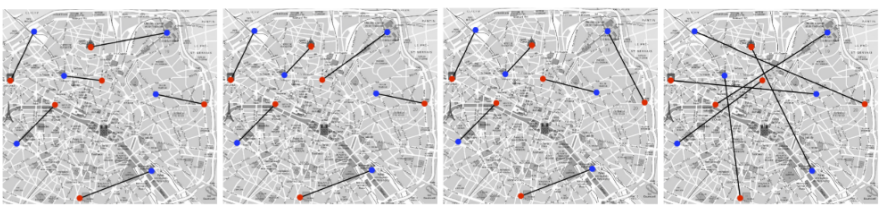


Figure. Examples of permutations with different costs.

The transport cost associated with a choice σ is the sum of the costs $C_{i,\sigma(i)}$ selected by σ:

$$\text{Cost}(\sigma) \triangleq C_{1,\sigma(1)} + C_{2,\sigma(2)} + C_{3,\sigma(3)} + C_{4,\sigma(4)} + C_{5,\sigma(5)} + C_{6,\sigma(6)}. \tag{2}$$

For example, for the permutation σ depicted in the figure, the cost is:

$$C_{1,5} + C_{2,2} + C_{3,6} + C_{4,1} + C_{5,3} + C_{6,4} = 10 + 7 + 15 + 10 + 14 + 9 = 65.$$

Monge Formulation

Monge's problem seeks the permutation σ that minimizes the cost, formulated as the optimization problem:

$$\min_{\sigma \in \Sigma_6} \text{Cost}(\sigma), \tag{3}$$

where Σ_6 denotes the set of permutations of set $[6]$.

- Solving this problem becomes computationally intractable due to its *combinatorial* nature. An exhaustive search for the optimal transport plan necessitates checking all permutations, leading to $|\Sigma_6| = 6! = 720$ possibilities in our small example.
- However, this approach quickly becomes infeasible for larger problems. For instance, with $n = 100$, the number of possibilities explodes to 10^{100}, exceeding the estimated numbers of neurons in the human brain (10^{11}) and atoms in the universe (10^{79}). This highlights the need for efficient alternative optimization methods for larger instances.

Monge Formulation — Permutation Matrices

Note that each permutation $\sigma \in \Sigma_n$ can be represented by a permutation matrix $\mathbf{P}$. This matrix is binary (containing only 0s and 1s) and has dimensions $n \times n$. Specifically, $P_{i,j} = 0$ unless $j = \sigma(i)$, in which case $P_{i,\sigma(i)} = 1$. As an example, consider $n = 3$ points. Permutations $(1, 2, 3) \to (1, 2, 3)$, $(1, 2, 3) \to (3, 2, 1)$ are represented by their respective 3×3 matrices.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{4}$$

Optimal Transport of Kantorovich

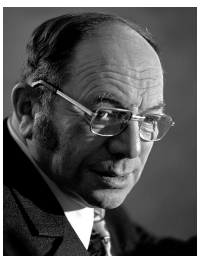


Figure. Leonid Kantorovich (Kantorovich's formulation of OT, 1942)

In the 1940s, Kantorovich's key insight hinges on replacing Monge's set of permutations with a larger, more tractable set.

$$\mathcal{P}_n = \left\{ \mathbf{P} \in \{0,1\}^{n\times n} : \forall i, \sum_j P_{i,j} = 1, \forall j, \sum_i P_{i,j} = 1 \right\} \Longrightarrow \mathcal{B}_n \triangleq \operatorname{conv}(\mathcal{P}_n) \in [0,1]^{n\times n}. \tag{5}$$

The Kantorovitch problem aims to solve

$$\min_{P\in\mathcal{B}_n} \sum_{i,j} P_{i,j} C_{i,j} \tag{6}$$

over the relaxed set of *doubly stochastic* matrices.

Optimal Transport of Kantorovich — Simplex Algorithm

Figure. George Dantzig (Simplex Algorithm, 1947)

George Dantzig's groundbreaking work that introduced the simplex algorithm to solve linear programs, marks a second revolution in this field. (One could recognize that the OT problem is a *linear* program.)

Why Relax?

This is good news because we have moved from the *computationally difficult* realm of combinatorics to the comfortable world of convex optimization. Optimization over a matrix space might sound difficult but it is usually much simpler than searching among possible assignments. (The simplex algorithm runs in $\mathcal{O}\left(n^3\right)$ iterations under this problem.)

The Weighted Case

Need for the Weighted Case

Monge formulation of OT assumes an equal number of coffee shops and bakeries. However, this creates an impossible scenario because sets with different sizes cannot be perfectly matched (not a bijection mapping).
Instead of considering the number of bakeries and coffee shops, a more relevant approach employs their production and consumption distributions. These are denoted by $(a_1, \dots, a_n)$ for bakeries and $(b_1, \dots, b_m)$ for coffee shops, respectively. Each a_i represents the daily production of the ith bakery, while each b_j represents the daily consumption of the jth coffee shop. For instance, $a_1 = 45$ indicates that the first bakery produces 45 croissants daily, and $b_3 = 34$ signifies that the third coffee shop consumes 34 croissants daily. Naturally, these quantities must be positive, and satisfy

$$a_1 + \cdots + a_n = b_1 + \cdots + b_m. \tag{7}$$

The Weighted Case

Need for the Weighted Case

Monge formulation of OT assumes an equal number of coffee shops and bakeries. However, this creates an impossible scenario because sets with different sizes cannot be perfectly matched (not a bijection mapping).
Instead of considering the number of bakeries and coffee shops, a more relevant approach employs their production and consumption distributions. These are denoted by $(a_1, \dots, a_n)$ for bakeries and $(b_1, \dots, b_m)$ for coffee shops, respectively. Each a_i represents the daily production of the ith bakery, while each b_j represents the daily consumption of the jth coffee shop. For instance, $a_1 = 45$ indicates that the first bakery produces 45 croissants daily, and $b_3 = 34$ signifies that the third coffee shop consumes 34 croissants daily. Naturally, these quantities must be positive, and satisfy

$$a_1 + \cdots + a_n = b_1 + \cdots + b_m. \tag{7}$$

Coupling Matrices and Distributions

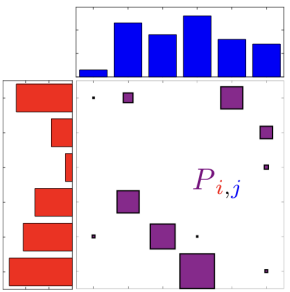
Kantorovich's construction naturally adapts to the case of general distributions, replacing the doubly stochastic matrices by matrices of *coupling* which satisfies the mass conservation constraint:

$$\mathcal{B}(a, b) \triangleq \left\{ \mathbf{P} \in [0,1]^{n \times m}; \forall i, \sum_j P_{i,j} = a_i, \forall j, \sum_i P_{i,j} = b_j \right\}, \tag{8}$$

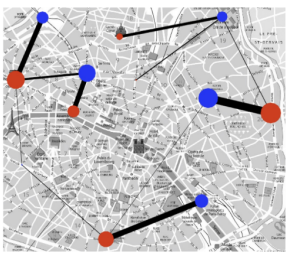
Alternative Ways of Representing Couplings

	3	23	18	26	16	14
24	1	7	0	0	16	0
9	0	0	0	0	0	9
3	0	0	0	0	0	3
16	0	16	0	0	0	0
21	2	0	18	1	0	0
27	0	0	0	25	0	2

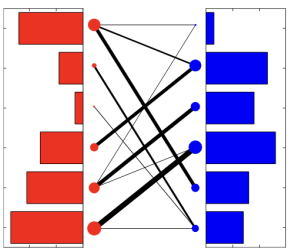
(a) matrix



(b) histograms



(c) segments

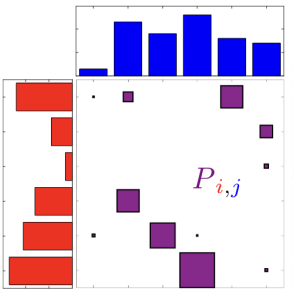


(d) bipartite graph

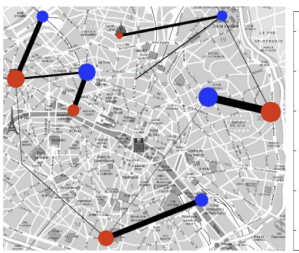
Alternative Ways of Representing Couplings

	3	23	18	26	16	14
24	1	7	0	0	16	0
9	0	0	0	0	0	9
3	0	0	0	0	0	3
16	0	16	0	0	0	0
21	2	0	18	1	0	0
27	0	0	0	25	0	2

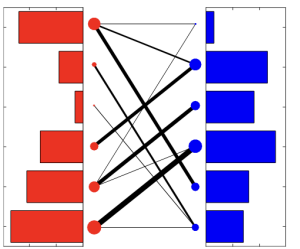
(a) matrix



(b) histograms



(c) segments



(d) bipartite graph

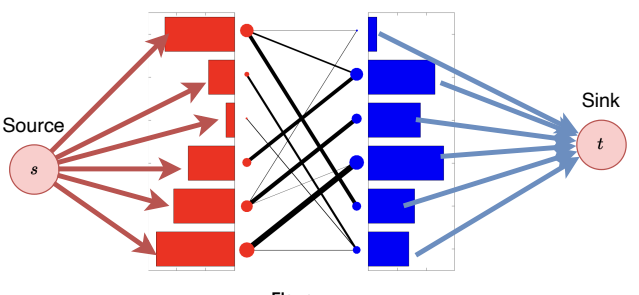
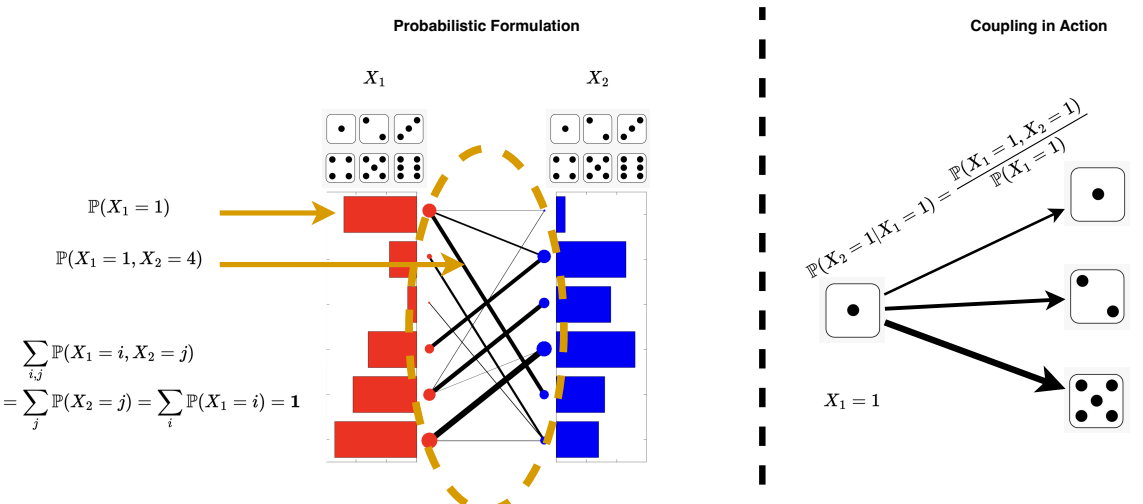


Figure. Subfig (d) depicts the connection of OT and min-cost flow problems.

Coupling Matrices: Demystifying the Name



Matrices $\mathbf{P} \in \mathcal{B}(a, b)$ are indeed coupling matrices. Random state X_2 can be fully determined by coupling matrix (transport plan) $\mathbf{P}$ and random state X_1. The random states X_1 and X_2 are then said to be *coupled*!

Computational Optimal Transport: Regularization

In 2013, Marco Cuturi developed computational speedup techniques for the optimal transport problem. This trick involves penalizing transport plans that lack diversity, as measured by the Shannon entropy. To achieve this, we tweak the optimization problem in this way:

$$\min_{P \in \mathcal{B}_n(\mathbf{a},\mathbf{b})} \sum_{i,j} P_{i,j} C_{i,j} - \epsilon \cdot H(\mathbf{P}), \tag{9}$$

where $H(\mathbf{P}) = -\sum_{i,j} P_{i,j} \log(P_{i,j})$ and $\epsilon \geq 0$ is a regularization parameter.

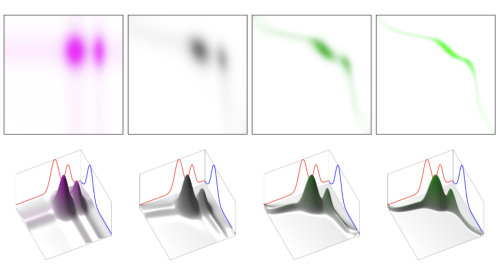
Computational Optimal Transport: Regularization

In 2013, Marco Cuturi developed computational speedup techniques for the optimal transport problem. This trick involves penalizing transport plans that lack diversity, as measured by the Shannon entropy. To achieve this, we tweak the optimization problem in this way:

$$\min_{P \in \mathcal{B}_n(\mathbf{a},\mathbf{b})} \sum_{i,j} P_{i,j} C_{i,j} - \epsilon \cdot H(\mathbf{P}), \tag{9}$$

where $H(\mathbf{P}) = -\sum_{i,j} P_{i,j} \log(P_{i,j})$ and $\epsilon \geq 0$ is a regularization parameter.

An Example



When $\epsilon \to \infty$, the solution is $\mathbf{P}^\star = (a_i b_j)_{i,j}$ (product distribution) and when $\epsilon \to 0$ the original formulation is recovered.

Computational Optimal Transport: Regularization

In 2013, Marco Cuturi developed computational speedup techniques for the optimal transport problem. This trick involves penalizing transport plans that lack diversity, as measured by the Shannon entropy. To achieve this, we tweak the optimization problem in this way:

$$\min_{P \in \mathcal{B}_n(\mathbf{a},\mathbf{b})} \sum_{i,j} P_{i,j} C_{i,j} - \epsilon \cdot H(\mathbf{P}), \tag{10}$$

where $H(\mathbf{P}) = -\sum_{i,j} P_{i,j} \log(P_{i,j})$ and $\epsilon \geq 0$ is a regularization parameter.

Why Regularize?

Marco Cuturi (2013) only demonstrated speedups empirically via Sinkhorn's Algorithm. It was later proved (Altschuler et al., 2019; Dvurechensky et al., 2019) that we achieve nearly linear time convergence after adding the entropy regularization to an approximator (the quality of the solution depends on ϵ).

Optimal Transport Theory — Wasserstein distance or Kantorovich–Rubinstein metric

Motivation

We wish to compare the following distributions. A natural metric to compare probability distributions $\mathbf{p}$ and $\mathbf{q}$ is the KL divergence $D_{\mathrm{KL}}(\mathbf{p}\,||\,\mathbf{q}) \triangleq \sum_i p_i \log(p_i/q_i)$.

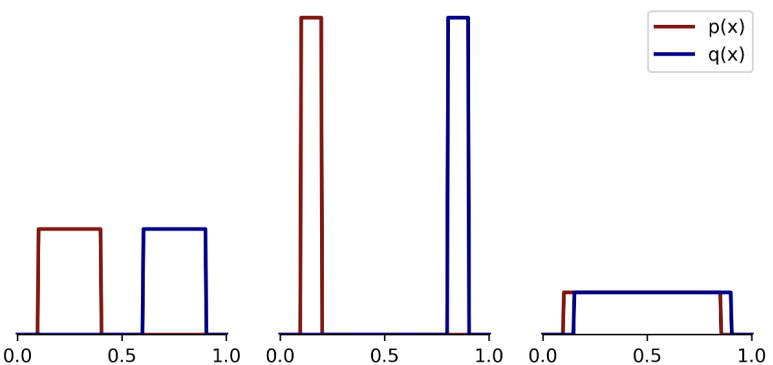


Figure. Three examples with infinite KL divergence. These distributions are infinitely far apart according to KL divergence.

Optimal Transport Theory — Wasserstein distance or Kantorovich–Rubinstein metric

A Robust Metric

The name "Wasserstein distance" was coined by R. L. Dobrushin in 1970 from the work of Leonid Vaserštein on Markov processes. However, the metric was first defined by Kantorovich in 1939. (Some scholars encourage the use of the terms "Kantorovich metric" and "Kantorovich distance".)

$$\mathcal{W}_p(\mathbf{p},\mathbf{q}) \triangleq \min_{\mathbf{P}\in\mathcal{B}(\mathbf{p},\mathbf{q})} \left(\mathbb{E}_{(X_1,X_2)\sim\mathbf{P}}\left[d(X_1,X_2)^p\right]\right)^{\frac{1}{p}}. \tag{11}$$

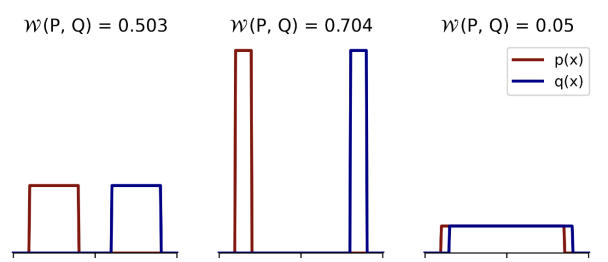


Figure. Unlike KL divergence, the Wasserstein distances in these examples are finite and intuitive.

Continuous Formulation (Infinitesimal Mass)

Monge's Formulation

Given probability measures μ on X and ν on Y, Monge's formulation of the optimal transportation problem is to find a transport map $T : X \to Y$ that realizes the infimum

$$\inf \left\{ \int_X c(x, T(x)) \mathrm{d}\mu(x) \mid T \sharp \mu = \nu \right\},$$

where $T \sharp \mu$ denotes the push forward of μ by T. A map T that attains this infimum is called an "optimal transport map".

Kantorovich's Formulation

Kantorovich's formulation of the optimal transportation problem is to find a probability measure γ on $X \times Y$ that attains the infimum

$$\inf \left\{ \int_{X \times Y} c(x, y) \mathrm{d}\gamma(x, y) \mid \gamma \in \Gamma(\mu, \nu) \right\},$$

where $\Gamma(\mu, \nu)$ denotes the collection of all probability measures on $X \times Y$ with marginals μ on X and ν on Y.

Applications of the Continuous Framework

Yann Brenier established the equivalence between Monge's and Kantorovich's formulations under specific conditions. By bridging the transport problem with partial differential equations, it paved the way for groundbreaking discoveries, including Fields Medals awarded to Cédric Villani (2010) and Alessio Figalli (2018).

(a) Cédric Villani (Fields Medal, 2010)

(b) Alessio Figalli (Fields Medal for his contributions to the theory of optimal transport, 2018)

Conclusion

Thank you for your attention!