Optimal Transport

From Monge to Modern Applications

Avetik Arakelyan

Institute of Mathematics, NAS of Armenia

June 19, 2025

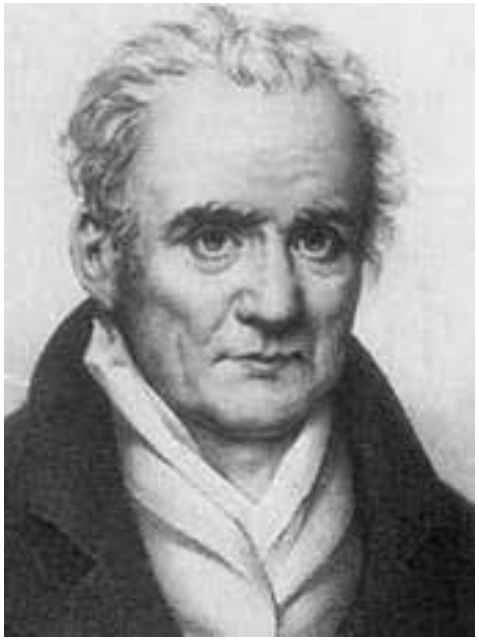
The Birth of Optimal Transport (1781)

Let's start at the beginning: what is Optimal transport (OT)?

The Birth of Optimal Transport (1781)

Let's start at the beginning: what is Optimal transport (OT)?

It all begins with Gaspard Monge, reading his mémoire in front of eminent scientists and engineers of the time.

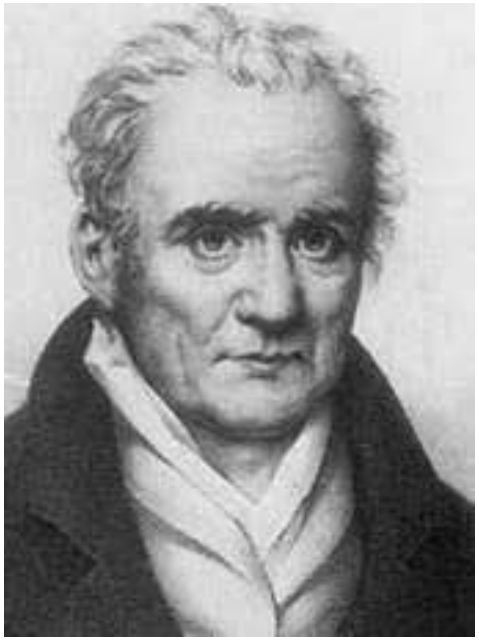


Gaspard Monge

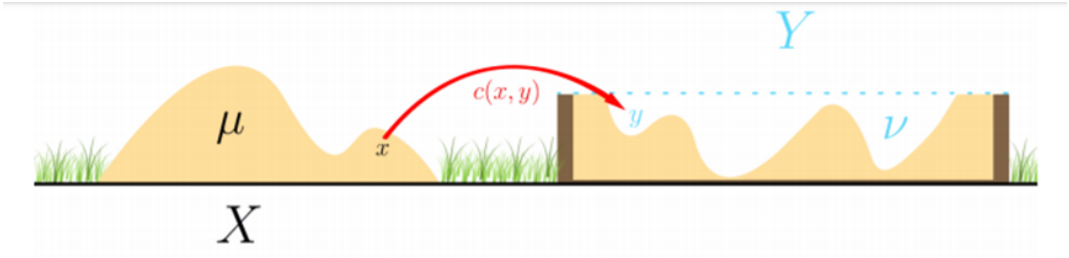
The Birth of Optimal Transport (1781)

Let's start at the beginning: what is Optimal transport (OT)?

It all begins with Gaspard Monge, reading his mémoire in front of eminent scientists and engineers of the time.



Gaspard Monge



Given two excavations D_1 and D_2, a cost of moving a unit of mass of dirt from any point of D_1 to any point of D_2. What is the cheapest way to move mass from D_1 to D_2?

Discrete Optimal Transport (Monge)

Each set (the above "excavations") can be represented as a histogram (or vector) $\mathbf{a}$ that belongs to the probability simplex — the vectors which components sum to 1:

$$\mathbf{a} \in S_N = \left\{ (z_1, \ldots, z_N) \in \mathbb{R}^N : \sum_{i=1}^{N} z_i = 1 \right\}$$

Discrete Optimal Transport (Monge)

Each set (the above "excavations") can be represented as a histogram (or vector) $\mathbf{a}$ that belongs to the probability simplex — the vectors which components sum to 1:

$$\mathbf{a} \in S_N = \left\{ (z_1, \ldots, z_N) \in \mathbb{R}^N : \sum_{i=1}^{N} z_i = 1 \right\}$$

We start with the simple case. Let the source and target sets are of the same total size.

Discrete Optimal Transport (Monge)

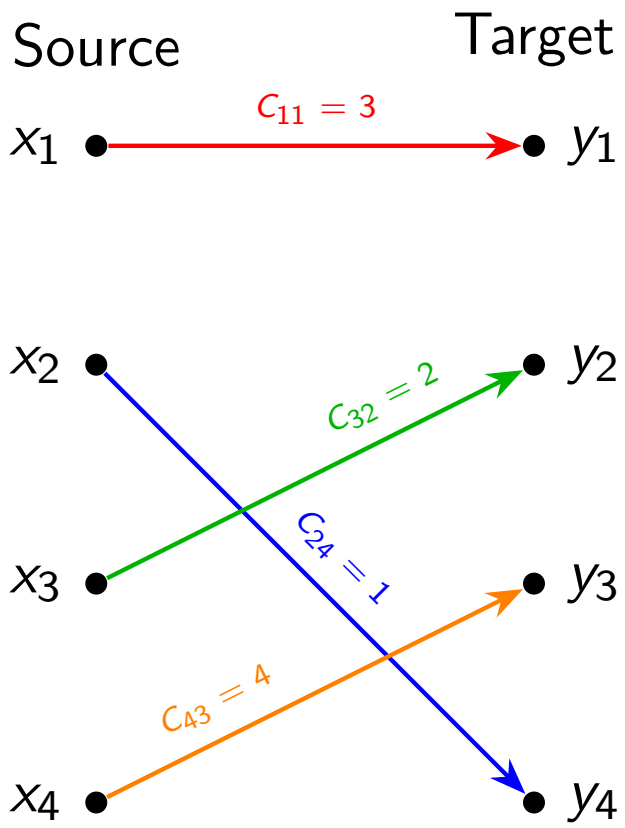
Each set (the above "excavations") can be represented as a histogram (or vector) **a** that belongs to the probability simplex — the vectors which components sum to 1:

$$\mathbf{a} \in S_N = \left\{ (z_1, \ldots, z_N) \in \mathbb{R}^N : \sum_{i=1}^{N} z_i = 1 \right\}$$

We start with the simple case. Let the source and target sets are of the same total size. If we write $\mathbf{C}_{i,j}$ the cost of moving an element from i to j, the quantity we want to minimize is $\sum_i \mathbf{C}_{i,\sigma(i)}$, where σ is a **permutation** of the set $\{1, \ldots, N\}$. This permutation represents an assignment of bin i of the first histogram to output positions j in the second histogram.

Visual Example of simple Monge Problem

Visual Example of simple Monge Problem



Total cost: $3 + 1 + 2 + 4 = 10$

$\sigma(1) = 1$, $\sigma(2) = 4$, $\sigma(3) = 2$, $\sigma(4) = 3$

Key: Each x_i sends mass to exactly one y_j (no splitting). Minimize $\sum_i C_{i,\sigma(i)}$.

Discrete Optimal Transport (Monge)

Lets assume, several points x_i can map to the same y_j and their weights are summed.

Discrete Optimal Transport (Monge)

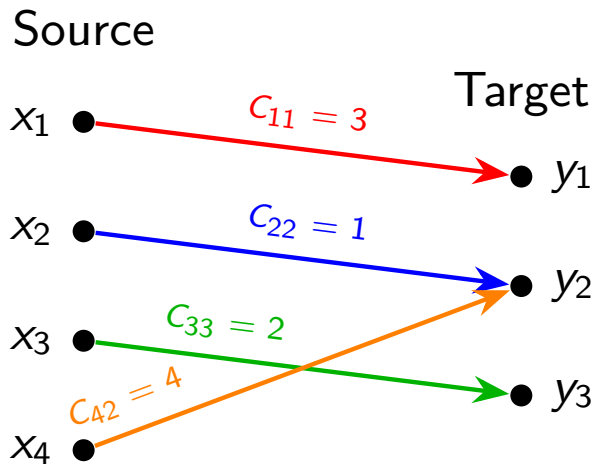
Lets assume, several points x_i can map to the same y_j and their weights are summed. In that case, the mapping between inputs and outputs is no longer a permutation, but a surjective map T. If points $x_1, \ldots, x_n$ have weights $\mathbf{a} = (a_1, \ldots, a_n)$ and points $y_1, \ldots, y_m$ have weights $\mathbf{b} = (b_1, \ldots, b_m)$, T must verify:

$$\forall j \in \{1, \ldots m\}, \quad b_j = \sum_{i : T(x_i) = y_j} a_i.$$

Discrete Optimal Transport (Monge)

Lets assume, several points x_i can map to the same y_j and their weights are summed. In that case, the mapping between inputs and outputs is no longer a permutation, but a surjective map T. If points $x_1, \ldots, x_n$ have weights $\mathbf{a} = (a_1, \ldots, a_n)$ and points $y_1, \ldots, y_m$ have weights $\mathbf{b} = (b_1, \ldots, b_m)$, T must verify:

$$\forall j \in \{1, \ldots m\}, \quad b_j = \sum_{i : T(x_i) = y_j} a_i.$$



Total cost: $3 + 1 + 2 + 4 = 10$

$T(x_1) = y_1$, $T(x_2) = y_2$, $T(x_3) = y_3$, $T(x_4) = y_2$

Key: Each x_i sends mass to exactly one y_j (no splitting). Minimize $\sum_i C_{i,T(i)}$.

Continuous Optimal Transport (Monge)

Continuous Optimal Transport (Monge)

Given two distributions of mass, how can we optimally reshape one into the other with respect to a given cost function?

Continuous Optimal Transport (Monge)

Given two distributions of mass, how can we optimally reshape one into the other with respect to a given cost function?

Let:

- $X, Y \subset \mathbb{R}^d$ be measurable spaces (typically compact domains)
- μ be a probability measure on X (source distribution)
- ν be a probability measure on Y (target distribution)
- $c : X \times Y \to \mathbb{R}^+$ be a cost function (typically $c(x, y) = \|x - y\|^p$)

Continuous Optimal Transport (Monge)

Given two distributions of mass, how can we optimally reshape one into the other with respect to a given cost function?

Let:

- $X, Y \subset \mathbb{R}^d$ be measurable spaces (typically compact domains)
- μ be a probability measure on X (source distribution)
- ν be a probability measure on Y (target distribution)
- $c : X \times Y \to \mathbb{R}^+$ be a cost function (typically $c(x, y) = \|x - y\|^p$)

Definition (Monge Problem)

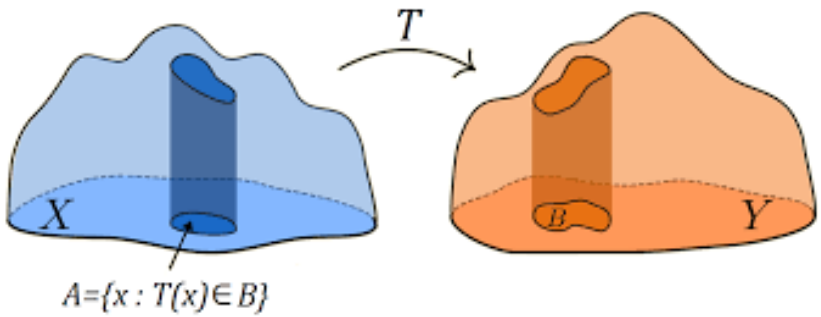
Find a transport map $T : X \to Y$ that minimizes the total transport cost:

$$\inf_T \left\{ \int_X c(x, T(x))\, d\mu(x) \;\middle|\; T_{\#}\mu = \nu \right\} \tag{1}$$

where $T_{\#}\mu = \nu$ denotes the *push-forward* condition:

$$\nu(B) = \mu(T^{-1}(B)) \quad \text{for all measurable } B \subset Y.$$

Continuous Optimal Transport (Monge)



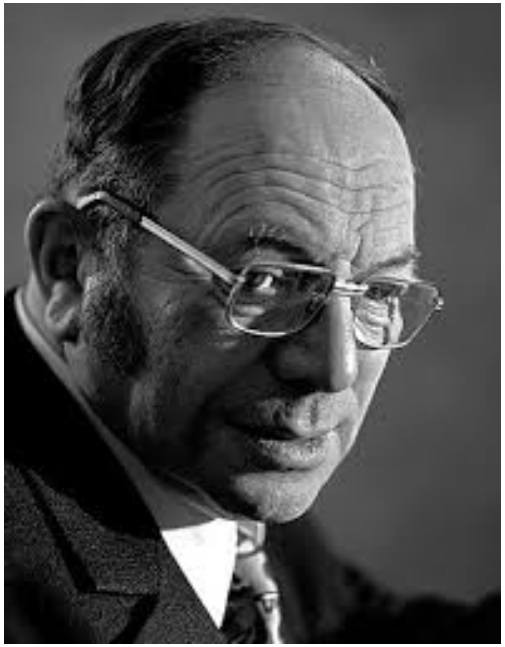
Monge's transport map

- **Push-forward constraint**: The condition $T_{\#}\mu = \nu$ ensures T properly redistributes μ to ν.
- **Deterministic transport**: Each point $x \in X$ is sent to exactly one destination $T(x) \in Y$.
- **Mass conservation**: Requires $\mu(X) = \nu(Y)$ (typically both normalized to 1).

Kantorovich Approach (Relaxation)

Kantorovich Approach (Relaxation)

Even with the above extension, this formulation of the optimal transport problem is still too constrained to be practically useful in many cases. In 1942, Leonid Kantorovitch introduced another key idea, which is to relax the deterministic portion of transportation. Source points x_i no longer have to map to a single target point and can be *fragmented* into smaller pieces, this is called **mass splitting**.



Leonid Kantorovich

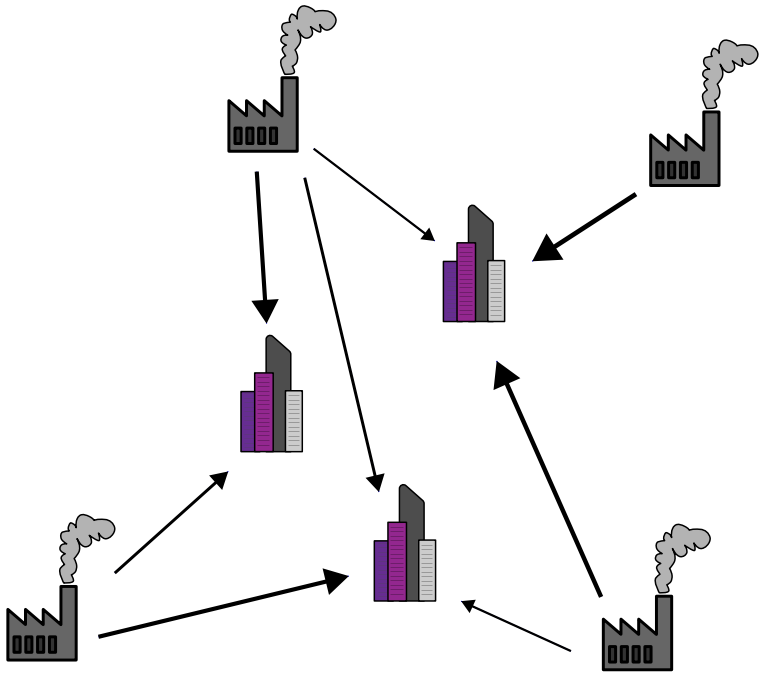
Nobel Prize in Economics

Kantorovich Approach (Relaxation)

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.

Kantorovich Approach (Relaxation)

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.



Factories with different supply capacities have to deliver goods to cities with various demands.

Kantorovich Linear Programming Problem

Notations

To reflect the above change, we modify our previous formulation slightly, by replacing the permutation σ by a coupling matrix $\mathbf{T} = \mathbf{T}_{ij} \in \mathbb{R}_+^{n \times m}$. In the example above, each $\mathbf{T}_{ij}$ would be the weight of the arrow from factory x_i to city y_j.

Kantorovich Linear Programming Problem

Notations

To reflect the above change, we modify our previous formulation slightly, by replacing the permutation σ by a coupling matrix $\mathbf{T} = \mathbf{T}_{ij} \in \mathbb{R}_+^{n \times m}$. In the example above, each $\mathbf{T}_{ij}$ would be the weight of the arrow from factory x_i to city y_j.

Suppose $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$. Then, $\mu \in Prob(X)$ can be written as a vector $v \in S_n$ and $\nu \in Prob(Y)$ can be written as a vector $w \in S_m$, where S_k denotes the k-dimensional probability simplex:

$$S_k := \left\{ v \in \mathbb{R}^k : v \geq 0 \text{ and } \sum_i v_i = 1 \right\}.$$

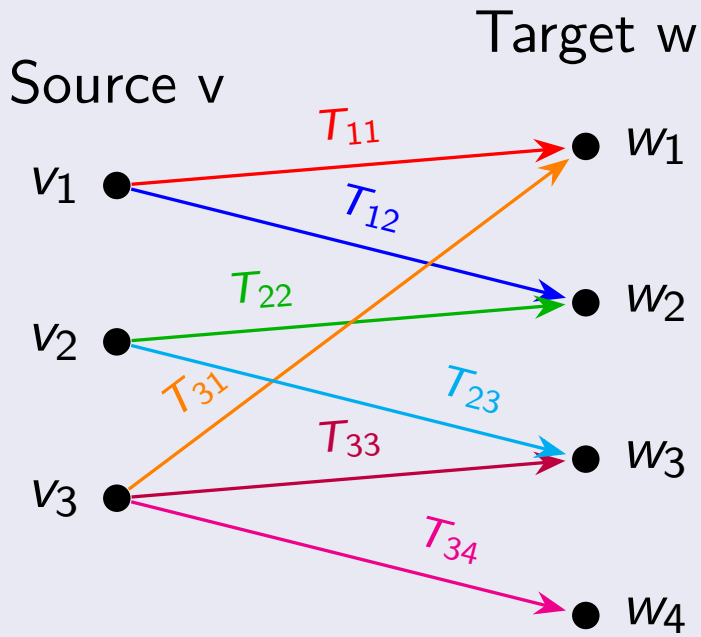
Kantorovich Linear Programming Problem

Visual Example

Kantorovich Linear Programming Problem

Visual Example

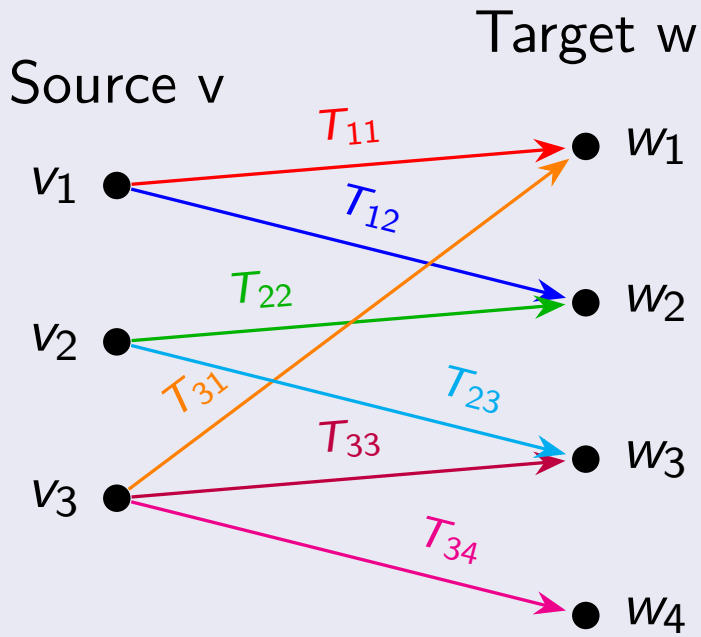
- $v \in S_3$, $w \in S_4$ (probability simplices)
- $C_{ij} = c(x_i, y_j)$ cost matrix
- T_{ij} = transport plan (couplings)



Kantorovich Linear Programming Problem

Visual Example

- $v \in S_3$, $w \in S_4$ (probability simplices)
- $C_{ij} = c(x_i, y_j)$ cost matrix
- T_{ij} = transport plan (couplings)



Cost Matrix C and Plan T

$$C = \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \end{pmatrix} \qquad T = \begin{pmatrix} T_{11} & T_{12} & 0 & 0 \\ 0 & T_{22} & T_{23} & 0 \\ T_{31} & 0 & T_{33} & T_{34} \end{pmatrix}$$

Kantorovich Linear Programming Problem

LP formulation

The cost function can be written as a matrix $C = (c_{ij})$. Thus, the transport problem between $v \in S_n$ and $w \in S_m$ given cost matrix C will be:

$$OT(v, w; C) = \begin{cases} \min\limits_{T \in \mathbb{R}_+^{n \times m}} & \sum_{ij} T_{ij} c_{ij} \\ \text{s.t.} & T \geq 0 \\ & \sum_j T_{ij} = v_i \ \forall i \in \{1, \ldots, n\} \\ & \sum_i T_{ij} = w_j \ \forall j \in \{1, \ldots, m\}. \end{cases}$$

Continuous Kantorovich Optimal Transport

Problem Setup

Given:

- Two probability measures μ on $Prob(X)$, ν on $Prob(Y)$ (e.g., $\mathbb{R}^d$)
- Cost function $c : X \times Y \to \mathbb{R}^+$ (e.g., $c(x, y) = \|x - y\|^p$)

Continuous Kantorovich Optimal Transport

Problem Setup

Given:

- Two probability measures μ on $Prob(X)$, ν on $Prob(Y)$ (e.g., $\mathbb{R}^d$)
- Cost function $c : X \times Y \to \mathbb{R}^+$ (e.g., $c(x, y) = \|x - y\|^p$)

Definition (Kantorovich Problem)

Find a **coupling** $\pi \in \mathcal{P}(X \times Y)$ that minimizes:

$$\inf_{\pi} \left\{ \int_{X \times Y} c(x, y) d\pi(x, y) \;\middle|\; \pi \in \Pi(\mu, \nu) \right\}$$

where $\Pi(\mu, \nu)$ is the set of **joint distributions** with marginals μ and ν.

Understanding Couplings $\Pi(\mu, \nu)$

Understanding Couplings $\Pi(\mu, \nu)$

Definition (Measure coupling)

A *measure coupling* $\pi \in Prob(X \times Y)$ is a probability measure on $X \times Y$ satisfying:

$$\pi(A \times Y) = \mu(A),$$
$$\pi(X \times B) = \nu(B),$$

for all measurable $A \subseteq X$ and $B \subseteq Y$. The set of measure couplings between μ and ν is denoted $\Pi(\mu, \nu)$.

Understanding Couplings $\Pi(\mu, \nu)$: Another Description

Key Components:

- Source $\mu(x)$: Factories
- Target $\nu(y)$: Cities to supply
- Coupling $\pi(x, y)$: Shipping plan

Rules:

1. Supply: $\int \pi(x, y)dy = \mu(x)$
   (All goods leave the factory x)
2. Demand: $\int \pi(x, y)dx = \nu(y)$
   (City y gets exact demand)

Understanding Couplings $\Pi(\mu, \nu)$: Another Description

Key Components:

- Source $\mu(x)$: Factories
- Target $\nu(y)$: Cities to supply
- Coupling $\pi(x, y)$: Shipping plan

Rules:

1. Supply: $\int \pi(x, y)dy = \mu(x)$
   (All goods leave the factory x)
2. Demand: $\int \pi(x, y)dx = \nu(y)$
   (City y gets exact demand)

Kantorovich Cost Minimization

Find π that minimizes:

$$OT(\mu, \nu; c) := \inf_{\pi \in \Pi(\mu,\nu)} \iint_{X \times Y} c(x, y)\pi(x, y)dxdy$$

Understanding Couplings $\Pi(\mu, \nu)$: Another Description

Key Components:

- Source $\mu(x)$: Factories
- Target $\nu(y)$: Cities to supply
- Coupling $\pi(x, y)$: Shipping plan

Rules:

1. Supply: $\int \pi(x, y)dy = \mu(x)$
   (All goods leave the factory x)
2. Demand: $\int \pi(x, y)dx = \nu(y)$
   (City y gets exact demand)

Kantorovich Cost Minimization

Find π that minimizes:

$$OT(\mu, \nu; c) := \inf_{\pi \in \Pi(\mu,\nu)} \iint_{X \times Y} c(x, y)\pi(x, y)dxdy$$

Key Features

- **Probabilistic**: π allows mass splitting (unlike Monge)
- **Existence**: Solutions always exist

Wasserstein Distance

Wasserstein Distance

For every $1 \leq p < +\infty$

Definition (**p-Wasserstein Distance Formula**)

$$W_p(\mu, \nu) \equiv \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{X \times Y} |x - y|^p d\pi(x, y) \right)^{1/p}$$

Wasserstein Distance

For every $1 \leq p < +\infty$

Definition (**p-Wasserstein Distance Formula**)

$$W_p(\mu, \nu) \equiv \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{X \times Y} |x - y|^p d\pi(x, y) \right)^{1/p}$$

Definition (**Discrete p-Wasserstein Distance Formula**)

Let $v \in S_n$ and $w \in S_m$ given cost matrix $C_{ij} = |x_i - y_j|^p$.

$$W_p^p(v, w) \equiv \begin{cases} \min\limits_{T \in \mathbb{R}_+^{n \times m}} & \sum_{ij} T_{ij} c_{ij} \\ \text{s.t.} & T \geq 0 \\ & \sum_j T_{ij} = v_i \ \ \forall i \in \{1, \ldots, n\} \\ & \sum_i T_{ij} = w_j \ \ \forall j \in \{1, \ldots, m\}. \end{cases}$$

Computational Optimal Transport

Computational Optimal Transport

Computing the discrete optimal transport (OT) or Wasserstein distance is quite challenging because it requires solving a Linear Programming Problem (LPP) numerically. However, well-known methods for solving LPPs are often slow in practice. For instance, the simplex method has exponential complexity, while interior-point methods scale as $\mathcal{O}(n^3 \log n)$.

Computational Optimal Transport

Computing the discrete optimal transport (OT) or Wasserstein distance is quite challenging because it requires solving a Linear Programming Problem (LPP) numerically. However, well-known methods for solving LPPs are often slow in practice. For instance, the simplex method has exponential complexity, while interior-point methods scale as $\mathcal{O}(n^3 \log n)$.

Regularizing optimal transport was proposed by Cuturi in 2013

- **Cuturi, M.**, ***Sinkhorn distances: Lightspeed computation of optimal transport***. **Advances in Neural Information Processing Systems, 26 (2013)**.

It is a method for approximating solutions to the optimal transport problem by adding a regularizing term to the objective function. This forces the solution to satisfy a number of constraints, making the problem easier to solve.

Thank You!

Questions?