Optimal Transport

Makoto Yamada (Kyoto University)

Notation

Matrix operation

- scalar: $x \in \mathbb{R}$
- vector: $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)^\top \in \mathbb{R}^D$
- matrix: $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)^\top \in \mathbb{R}^{D \times N}$
- transpose: $^\top$
- ℓ_2 norm: $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{D} x_i^2}$ (Euclid norm)

Probability

- random variable: X
- probability density: $p(\cdot)$

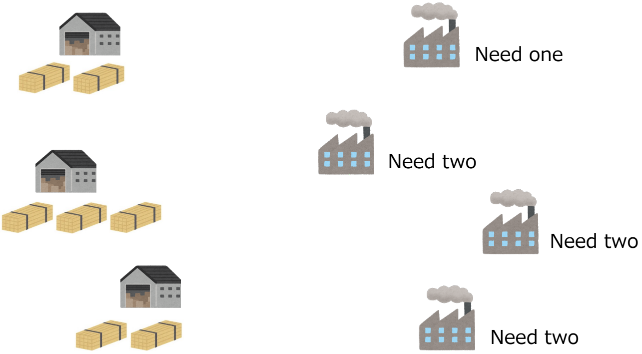
Transportation problem

The optimal transport has been studied since the 17th century (many Nobel Prizes and Fields Medals recipients). In particular, in the last few years, a great deal of research has been reported in the field of machine learning.

- Wasserstein GAN (Machine learning)
- Earth Mover's Distance (Computer vision)
- Word Mover's Distance (NLP)

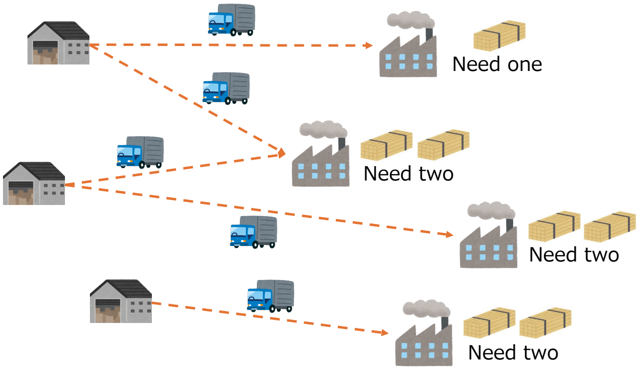
輸送問題

The problem of transporting goods from one location (warehouse) to another (factory). How is the best way to transport the goods?



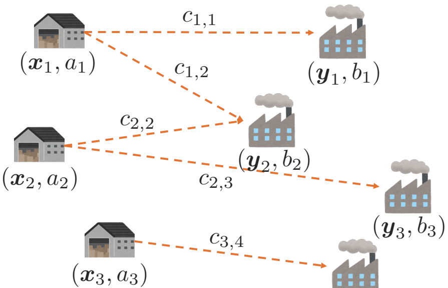
Transportation problem

Use trucks to transport goods (**costly**). Consider transportation that is less costly and satisfies the constraints.



Transportation problem

The percentage of the quantity of goods in the warehouse is $\boldsymbol{a} = (a_1, a_2, \ldots, a_N)^\top$, the percentage of consumption in the factory is $\boldsymbol{b} = (b_1, b_2, \ldots, b_N)^\top$, $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^D$ are the locations of the warehouses and factories. transportation costs is $c_{i,j} = c(\boldsymbol{x}_i, \boldsymbol{y}_j)$. The quantity to be transported (by truck) is $[\boldsymbol{P}]_{ij} = p_{i,j}$.



Optimal transport (OT)

Optimal transport problem

$$\min_{\boldsymbol{P} \in U(\mu,\nu)} \sum_{i,j} p_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j)$$

or

$$\min_{\boldsymbol{P} \in \mathbb{R}_+^{N \times M}} \sum_{i,j} p_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j) \ \text{ s.t. } \ \boldsymbol{P}\mathbf{1}_M = \boldsymbol{a}, \boldsymbol{P}^\top \mathbf{1}_N = \boldsymbol{b}$$

- $\mu = \sum_{i=1}^N a_i \delta_{\boldsymbol{x}_i}$, $\nu = \sum_{j=1}^M b_j \delta_{\boldsymbol{y}_j}$ are discrete measures.
- $U(\mu, \nu) = \{\boldsymbol{P} \in \mathbb{R}_+^{N \times M} : \boldsymbol{P}\mathbf{1}_M = \boldsymbol{a}, \boldsymbol{P}^\top \mathbf{1}_M = \boldsymbol{b}\}$
- $\boldsymbol{a}^\top \mathbf{1}_N = 1$ (probability)
- $\boldsymbol{b}^\top \mathbf{1}_M = 1$ (probability)
- $\boldsymbol{P}\mathbf{1}_M = \boldsymbol{a}, \boldsymbol{P}^\top \mathbf{1}_N = \boldsymbol{b}$.

Wasserstein distance

p-Wasserstein distance

$$W_p(\mu,\nu) = \left(\min_{\boldsymbol{P} \in U(\mu,\nu)} \sum_{i,j} p_{ij} d(\boldsymbol{x}_i, \boldsymbol{y}_j)^p \right)^{1/p}$$

$d(\boldsymbol{x},\boldsymbol{y})$ is a distance. (e.g., Euclid distance $d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2$)

1-Wasserstein distance

$$W_1(\mu,\nu) = \min_{\boldsymbol{P} \in U(\mu,\nu)} \sum_{i,j} p_{ij} d(\boldsymbol{x}_i, \boldsymbol{y}_j)$$

2-Wasserstein distance

$$W_2(\mu,\nu) = \left(\min_{\boldsymbol{P} \in U(\mu,\nu)} \sum_{i,j} p_{ij} d(\boldsymbol{x}_i, \boldsymbol{y}_j)^2 \right)^{1/2}$$

Wasserstein distance

2-Wasserstein distance

$$W_2(\mu, \nu) = \left(\min_{\boldsymbol{P} \in U(\mu,\nu)} \sum_{i,j} p_{ij} d(\boldsymbol{x}_i, \boldsymbol{y}_j)^2 \right)^{1/2}$$

2-Wasserstein distance If $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$ is a Euclidean distance, we have

$$W_2(\mu, \nu) = \left(\min_{\boldsymbol{P} \in U(\mu,\nu)} \sum_{i,j} p_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 \right)^{1/2}$$

$\min_{\boldsymbol{P} \in U(\mu,\nu)} \sum_{i,j} p_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 = (W_2(\mu, \nu))^2$ is not a distance!

Entropic regularized OT

Optimization problem

$$\min_{\boldsymbol{P}\in U(\mu,\nu)} \sum_{i,j} p_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j) + \lambda \sum_{i,j} p_{ij}(\log(p_{ij}) - 1)$$

$\lambda \geq 0$ is a regularization parameter.

- $c(\boldsymbol{x}, \boldsymbol{y})$ does **not need to be a distance**.
- When $\lambda = 0$, it is Earth Mover's Distance (EMD).
- **Strongly convex function** ($0 < p_k \leq 1$).

$$\frac{\partial}{\partial p_k \partial p_{k'}} \sum_i p_i(\log(p_i) - 1) = \frac{\partial}{\partial p_{k'}} \log(p_k) = \begin{cases} \frac{1}{p_k} & k = k' \\ 0 & k \neq k' \end{cases}$$

For $p \in [0\ 1]$, $p \log(p)$ is convex & The Hessian is positive definite $\rightarrow$ Strongly convex.

Sinkhorn Algorithm

Optimization problem

$$\min_{\boldsymbol{P}} \quad \sum_{i,j} p_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j) + \lambda \sum_{i,j} p_{ij}(\log(p_{ij}) - 1)$$
$$\text{s.t.} \quad \boldsymbol{P}\mathbf{1}_M = \boldsymbol{a}, \boldsymbol{P}^\top \mathbf{1}_N = \boldsymbol{b}$$

Using Lagrange multipliers $\widetilde{\boldsymbol{u}} \in \mathbb{R}^N, \widetilde{\boldsymbol{v}} \in \mathbb{R}^M$

$$\begin{aligned} J(\boldsymbol{P}, \boldsymbol{u}, \boldsymbol{v}) = &\sum_{i,j} p_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j) + \lambda \sum_{i,j} p_{ij}(\log(p_{ij}) - 1) \\ &+ \widetilde{\boldsymbol{u}}^\top (\boldsymbol{a} - \boldsymbol{P}\mathbf{1}_M) + \widetilde{\boldsymbol{v}}^\top (\boldsymbol{b} - \boldsymbol{P}^\top \mathbf{1}_N) \end{aligned}$$

Sinkhorn Algorithm

Taking derivative w.r.t. p_{kl}, and solving for p_{kl}.

$$\frac{\partial}{\partial p_{kl}} J(\boldsymbol{P}, \boldsymbol{u}, \boldsymbol{v}) = c(\boldsymbol{x}_k, \boldsymbol{y}_l) + \lambda \log(p_{kl}) - u_k - v_l = 0$$

The optimal solution of p_{kl} is given as

$$\begin{aligned}
\log(p_{kl}) &= -\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l) + \frac{1}{\lambda}\widetilde{u}_k + \frac{1}{\lambda}\widetilde{v}_l \\
p_{kl} &= \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right) \exp\left(\frac{1}{\lambda}\widetilde{u}_k\right) \exp\left(\frac{1}{\lambda}\widetilde{v}_l\right) \\
p_{kl} &= \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right) \exp\left(\frac{1}{\lambda}\widetilde{u}_k\right) \exp\left(\frac{1}{\lambda}\widetilde{v}_l\right) \\
&= \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right) u_k v_l
\end{aligned}$$

Sinkhorn Algorithm

Updating v_l:

$$v_l \sum_{k=1}^{N} \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right) u_k = \sum_{k=1}^{N} p_{kl}$$

$$v_l = \frac{b_l}{\sum_{k=1}^{N} \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right) u_k}$$

Matrix version:

$$\boldsymbol{v}^{(\ell+1)} = \boldsymbol{b} / (\boldsymbol{K}^\top \boldsymbol{u}^{(\ell)})$$

Note: $\boldsymbol{K} \in \mathbb{R}^{n \times m}$, $[\boldsymbol{K}]_{ij} = \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right)$.

Sinkhorn Algorithm

Updating u_k:

$$v_k \sum_{l=1}^{N} \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right) v_l = \sum_{l=1}^{M} p_{kl}$$

$$u_k = \frac{a_k}{\sum_{l=1}^{M} \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right) v_l}$$

Matrix version:

$$\boldsymbol{u}^{(\ell+1)} = \boldsymbol{a} / (\boldsymbol{K} \boldsymbol{v}^{(\ell)})$$

Note: $\boldsymbol{K} \in \mathbb{R}^{n \times m}$, $[\boldsymbol{K}]_{ij} = \exp\left(-\frac{1}{\lambda} c(\boldsymbol{x}_k, \boldsymbol{y}_l)\right)$.

Dual problem

Optimization problem

$$\min_{\boldsymbol{P}} \quad \sum_{i,j} p_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j) \text{ s.t. } \quad \boldsymbol{P}\mathbf{1}_M = \boldsymbol{a}, \boldsymbol{P}^\top \mathbf{1}_N = \boldsymbol{b}$$

Using the Lagrange multipliers $\boldsymbol{u} \in \mathbb{R}^N, \boldsymbol{v} \in \mathbb{R}^M$ We denote $\langle \boldsymbol{P}, \boldsymbol{C} \rangle = \sum_{i,j} p_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j)$, then we have

$$\min_{\boldsymbol{P} \geq 0} \max_{(\boldsymbol{u},\boldsymbol{v}) \in \mathbb{R}^N \times \mathbb{R}^M} \langle \boldsymbol{P}, \boldsymbol{C} \rangle + \boldsymbol{u}^\top (\boldsymbol{a} - \boldsymbol{P}\mathbf{1}_M) + \boldsymbol{v}^\top (\boldsymbol{b} - \boldsymbol{P}^\top \mathbf{1}_N)$$

For a linear problem, min and max can be exchangeable

$$\max_{(\boldsymbol{u},\boldsymbol{v}) \in \mathbb{R}^N \times \mathbb{R}^M} \langle \boldsymbol{u}, \boldsymbol{a} \rangle + \langle \boldsymbol{v}, \boldsymbol{b} \rangle + \min_{\boldsymbol{P} \geq 0} \langle \boldsymbol{P}, \boldsymbol{C} \rangle - \boldsymbol{u}^\top \boldsymbol{P}\mathbf{1}_M - \boldsymbol{v}^\top \boldsymbol{P}^\top \mathbf{1}_N$$

Dual problem

We use $\boldsymbol{u}^\top \boldsymbol{P}\mathbf{1}_M = \mathrm{tr}(\boldsymbol{u}^\top \boldsymbol{P}\mathbf{1}_M) = \mathrm{tr}(\boldsymbol{P}\mathbf{1}_M \boldsymbol{u}^\top) = \mathrm{tr}(\boldsymbol{u}\mathbf{1}_M^\top \boldsymbol{P}^\top) = \langle \boldsymbol{P}, \boldsymbol{u}\mathbf{1}_M^\top \rangle$ and $\boldsymbol{v}^\top \boldsymbol{P}^\top \mathbf{1}_N = \langle \boldsymbol{P}, \mathbf{1}_N \boldsymbol{v}^\top \rangle$.

$$\max_{(\boldsymbol{u},\boldsymbol{v})\in\mathbb{R}^N\times\mathbb{R}^M} \langle \boldsymbol{u}, \boldsymbol{a}\rangle + \langle \boldsymbol{v}, \boldsymbol{b}\rangle + \min_{\boldsymbol{P}\geq 0} \langle \boldsymbol{P}, \boldsymbol{C}\rangle - \langle \boldsymbol{P}, \boldsymbol{u}\mathbf{1}_M^\top\rangle - \langle \boldsymbol{P}, \mathbf{1}_N\boldsymbol{v}^\top\rangle$$

$$\max_{(\boldsymbol{u},\boldsymbol{v})\in\mathbb{R}^N\times\mathbb{R}^M} \langle \boldsymbol{u}, \boldsymbol{a}\rangle + \langle \boldsymbol{v}, \boldsymbol{b}\rangle + \min_{\boldsymbol{P}\geq 0} \langle \boldsymbol{P}, \boldsymbol{C} - \boldsymbol{u}\mathbf{1}_M^\top - \mathbf{1}_N\boldsymbol{v}^\top\rangle$$

$$\min_{\boldsymbol{P}\geq 0} \langle \boldsymbol{P}, \boldsymbol{C} - \boldsymbol{u}\mathbf{1}_M^\top - \mathbf{1}_N\boldsymbol{v}^\top\rangle = \begin{cases} 0 & \boldsymbol{C} - \boldsymbol{u}\oplus\boldsymbol{v} \geq 0 \\ -\infty & \text{Otherwise} \end{cases}$$

$\boldsymbol{u}\oplus\boldsymbol{v} = \boldsymbol{u}\mathbf{1}_M^\top - \mathbf{1}_N\boldsymbol{v}^\top$. If the following condition satisfied, the second term is zero.

$$\boldsymbol{u}\oplus\boldsymbol{v} \leq \boldsymbol{C}$$

Dual problem

Dual problem

$$\max_{(\boldsymbol{u},\boldsymbol{v})\in R(\boldsymbol{C})} \langle \boldsymbol{u}, \boldsymbol{a} \rangle + \langle \boldsymbol{v}, \boldsymbol{b} \rangle$$

$$R(\boldsymbol{C}) = \{(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^N \times \mathbb{R}^M : \forall (i,j) \in [\![N]\!] \times [\![M]\!], \boldsymbol{u} \oplus \boldsymbol{v} \leq \boldsymbol{C}\}.$$

$$[\![N]\!] = \{1, 2, \ldots N\}.$$

The relationship between primal and dual problems

$$\sum_{i,j} p_{ij} c_{ij} \geq \sum_{i,j} p_{ij}(u_i + v_j) = \sum_{i=1}^{N} a_i u_i + \sum_{j=1}^{M} b_i v_j = \langle \boldsymbol{u}, \boldsymbol{a} \rangle + \langle \boldsymbol{v}, \boldsymbol{b} \rangle$$

Other OT based distances

- Sliced Wasserstein distance
- Tree-Wasserstein distance
- Gromov Wasserstein distance

Summary

- Optimal transport (OT)
- Wasserstein distance
- Entropic regularized Optimal transport
- Dual problem