Lecture 3. Friday, April 1, 2022

In what follows, we consider the discret OT problem in Kantorovich's form.

Definition A vector $a = (a_i) \in \mathbb{R}^n$ is a probability vector if $a_i \geq 0$ $(i = 1, \cdots, m)$ and $\sum_{i=1}^n a_i = 1$.

Notation $\mathcal{P}_n = \{$ all probability vectors in $\mathbb{R}^n \}$.

Given $a \in \mathcal{P}_m$ and $b \in \mathcal{P}_n$,

$$C = [C_{ij}] \in \mathbb{R}^{m \times n}, \quad C \geq 0 \ (\text{i.e., } C_{ij} \geq 0 \ \forall i, j).$$

$$\mathcal{A}(a,b) \triangleq \Big\{ P = [P_{ij}] \in \mathbb{R}^{m \times n} : P \geq 0, \sum_{j=1}^n P_{ij} = a_i \ \forall i, \ \sum_{i=1}^m P_{ij} = b_j \ \forall j \Big\}.$$

The (discrete) OT problem:

$$\min_{P \in \mathcal{A}(a,b)} \sum_{i=1}^m \sum_{j=1}^n P_{ij} C_{ij}.$$

Today: ⊙ Basic properties

⊙ Wasserstein metric

Proposition The feasible set $\mathcal{A}(a,b)$ is a nonempty, convex, and compact subset of the vector space $\mathbb{R}^{m \times n}$.

Proof Let $P = [P_{ij}]$ with $P_{ij} = a_i b_j$.

Then $P \in \mathcal{A}(a,b)$. Clearly $\mathcal{A}(a,b)$ is convex.

If $P = [P_{ij}] \in \mathcal{A}(a,b)$, then $0 \leq P_{ij} \leq 1 \ \forall i, j$.

Hence $\mathcal{A}(a,b)$ is compact. QED

Now, let us try to understand the structure of $\mathcal{A}(a,b)$. First, reindex/relabel:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix} \xrightarrow{\text{relabel}} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \\ y_{n+1} & y_{n+2} & \cdots & y_{2n} \\ \vdots & \vdots & & \\ y_{(m-1)n+1} & y_{(m-1)n+2} & \cdots & y_{mn} \end{bmatrix}$$

Using the elementary row reduction to solve the system of equations

$$\left.\begin{aligned} \sum_{j=1}^{n} p_{ij} &= a_i, \quad i=1,\dots,m, \\ \sum_{i=1}^{n} p_{ij} &= b_j, \quad j=1,\dots,n. \end{aligned}\right\} \quad (*)$$

we obtain

$$y_1 = a_1 + b_1 - 1 + \sum_{i=2}^{m}\sum_{j=2}^{n} y_{(i-1)n+j}, \qquad (A)$$

$$y_j = b_j - \sum_{i=2}^{m} y_{(i-1)n+j} \quad (2 \le j \le n), \qquad (B)$$

$$y_{n+i} = a_i - \sum_{j=2}^{n} y_{(i-1)n+j} \quad (2 \le i \le m). \qquad (C)$$

Here, blue y's are free variables and red y's are dependent variables.

Denote $E_{ij} \in \mathbb{R}^{m\times n}$ to be the $m\times n$ matrix with the (i,j)-entry being 1 and all other entries being 0. Then the solution set for $(*)$ is given by

$$P = A + \sum_{i=2}^{m}\sum_{j=2}^{n} p_{ij} E_{ij}$$

where the matrix A is determined by (A)–(C).

Since $0 \le y_i \le 1$, we have

$$1-a_1-b_1 \le \sum_{i=2}^{m} \sum_{j=2}^{n} y_{(i-1)n+j} \le 2-a_1-b_1, \qquad (D)$$

$$(b_j-1 \le 0 \le) \sum_{i=2}^{m} y_{(i-1)n+j} \le b_j \quad (2 \le j \le n), \qquad (E)$$

$$(a_i-1 \le 0 \le) \sum_{j=2}^{n} y_{(i-1)n+j} \le a_i \quad (2 \le i \le m). \qquad (F)$$

In addition, all blue $y_k \ge 0$.

<u>Remarks</u> ① We can choose row i_0 and column j_0 for any i_0, j_0 instead of $i_0=1$, $j_0=1$.

② The equalities in (D), (E), (F) may not be reached.

Define the map $\mathcal{L}: \mathcal{A}(a,b) \longrightarrow \mathbb{R}^{(m-1)(n-1)}$, $\mathcal{L}(P)=x$, by the following:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix} \xrightarrow{\text{relabel}} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \\ y_{n+1} & y_{n+2} & \cdots & y_{2n} \\ \vdots & \vdots & & \\ y_{(m-1)n+1} & y_{(m-1)n+2} & \cdots & y_{mn} \end{bmatrix}$$

$$\downarrow \text{relabel/projection}$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_{(m-1)(n-1)} \end{bmatrix} \xleftarrow{\text{reformat}} \begin{bmatrix} x_1 & \cdots & x_{n-1} \\ x_n & \cdots & x_{2n-2} \\ & & \\ x_{(m-2)(n-1)+1} & \cdots & x_{(m-1)(n-1)} \end{bmatrix}$$

Clearly,

① $\mathcal{L}$ is a one-to-one linear map.

② $\mathcal{L}(\mathcal{A}(a,b))$ is a convex compact subset of the unit cube $[0,1]^{(m-1)(n-1)}$ of $\mathbb{R}^{(m-1)(n-1)}$.

Proposition Assume $m \geq 2$, $n \geq 2$, $a \in \mathcal{P}_m$, $a > 0$, $b \in \mathcal{P}_n$, $b > 0$. Then $\mathcal{L}(\mathcal{A}(a,b))$ is a convex and compact polyhydron in $\mathbb{R}^{(m-1)(n-1)}$ with nonzero interior. In particular, $\dim \mathcal{A}(a,b) = (m-1)(n-1)$. QED

Proposition (1) $\exists \hat{P} \in \mathcal{A}(a,b)$ s.t.

$$\sum_{i,j} \hat{P}_{ij} C_{ij} = \min_{P \in \mathcal{A}(a,b)} \sum_{i,j} P_{ij} C_{ij}. \quad (*)$$

(2) Let $\mathcal{M}_C$ denotes the subset of $\mathcal{A}(a,b)$ consisting of all $\hat{P}$ satisfying $(*)$. Then $\mathcal{M}_C = \mathcal{M}_C(a,b)$ is convex and compact.

(3) Each $P \in \mathcal{M}_C(a,b)$ has at most $m+n-1$ nonzero entries.

Proof. (1) and (2) are clear. (3): see some refs. QED

Definition A matrix $C = [C_{ij}] \in \mathbb{R}^{n \times n}$ is a metric matrix if $(i,j) \mapsto C_{ij}$ $(1 \leq i,j \leq n)$ defines a metric of $\{1, 2, \ldots, n\}$, i.e., it satisfies:

(1) $C_{ij} \geq 0$ $\forall i,j \in \{1, 2, \ldots, n\}$. $C_{ij} = 0 \iff i = j$

(2) $C_{ij} = C_{ji}$ $\forall i,j \in \{1, 2, \ldots, n\}$.

(3) $C_{ij} \leq C_{ik} + C_{kj}$ $\forall i,j,k \in \{1, 2, \ldots, n\}$.

Example Let $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$. Set

$C_{ij} = |x_i - x_j|$ $(1 \leq i,j \leq n)$. Then

$C = [C_{ij}] \in \mathbb{R}^{n \times n}$ is a metric matrix.

<u>Theorem</u> Let $C=[C_{ij}]\in\mathbb{R}^{n\times n}$ be a metric matrix. Define

$$W(a,b)=\min_{P\in\mathcal{A}(a,b)}\sum_{i=1}^{n}\sum_{j=1}^{n}P_{ij}C_{ij}\quad\forall a,b\in\mathcal{P}_n.$$

Then W is a metric on $\mathcal{P}_n$.

<u>Pf</u> (1) Clearly, $W(a,b)\geq 0$ $\forall a,b\in\mathcal{P}_n$. If $W(a,b)=0$, then for some $P\in\mathcal{A}(a,b)$, P is a minimizer, $\sum_i\sum_j P_{ij}C_{ij}=0$. Hence $P_{ij}C_{ij}=0$ $\forall i,j$. But $C_{ij}\neq 0$ if $i\neq j$. Thus, $P_{ij}=0$ if $i\neq j$, $P=\begin{bmatrix}p_{11} & & 0\\ & \ddots & \\ 0 & & p_{nn}\end{bmatrix}$

But $P\in\mathcal{A}(a,b)$. So, $P_{ii}=a_i=b_i$ $(i=1,2,\dots,n)$. Hence $a=b$.

(2) Suppose $a,b\in\mathcal{P}_n$. If $P\in\mathcal{A}(a,b)$ then $P^T\in\mathcal{A}(b,a)$, and since $C^T=C$,

$$\sum_{i,j}P_{ij}C_{ij}=\sum_{i,j}P_{ij}C_{ji}=\sum_{i,j}(P^T)_{ji}C_{ji}.$$

So, the matrix transpose defines a bijection between $\mathcal{A}(a,b)$ and $\mathcal{A}(b,a)$, which preserves the cost. Thus, $W(a,b)=W(b,a)$.

(3) Let $a,b,c\in\mathcal{P}_n$. We show the triangle inequality

$$W(a,c)\leq W(a,b)+W(b,c).$$

Let $R\in\mathcal{A}(a,b)$ and $S\in\mathcal{A}(b,c)$ be such that

$$W(a,b)=\sum_{i=1}^{m}\sum_{j=1}^{n}R_{ij}C_{ij},$$

$$W(b,c)=\sum_{i=1}^{m}\sum_{j=1}^{n}S_{ij}C_{ij}.$$

Define for any $i, j, k \in \{1, \cdots, n\}$:

$$\widetilde{Q}_{ijk} = \begin{cases} \frac{R_{ij} S_{jk}}{b_j} & \text{if } b_j \neq 0, \\ 0 & \text{if } b_j = 0. \end{cases}$$

We verify the following:

(1) $\widetilde{Q}_{ijk} \geq 0 \quad \forall i, j, k$.

(2) $\sum_{i,j,k} \widetilde{Q}_{ijk} = 1$.

In fact, since $R \in \mathcal{A}(a,b)$, $S \in \mathcal{A}(b,c)$,

$$\sum_{i,j,k} \widetilde{Q}_{ijk} = \sum_{j: b_j \neq 0} \underbrace{\left(\sum_i R_{ij}\right)}_{= b_j} \underbrace{\left(\sum_k S_{jk}\right)}_{= b_j} \cdot \frac{1}{b_j}$$

$$= \sum_{j: b_j \neq 0} b_j^2 \cdot \frac{1}{b_j} = \sum_{j: b_j \neq 0} b_j = \sum_{j=1}^{n} b_j = 1.$$

(3) $\forall i, j$: $\sum_{k=1}^{n} \widetilde{Q}_{ijk} = R_{ij}$.

In fact, if $b_j = 0$, then all $\widetilde{Q}_{ijk} = 0$, also $R_{ij} = 0$ since $\sum_{i=1}^{n} R_{ij} = b_j = 0$ and all $R_{ij} \geq 0$.

Suppose $b_j \neq 0$, then

$$\sum_k \widetilde{Q}_{ijk} = \sum_k R_{ij} S_{jk} \cdot \frac{1}{b_j}$$

$$= \frac{1}{b_j} R_{ij} \sum_k S_{jk} = \frac{1}{b_j} R_{ij} b_j = R_{ij}.$$

since $S \in \mathcal{A}(b,c)$

(4) $\forall j, k$: $\sum_{i=1}^{n} \widetilde{Q}_{ijk} = S_{jk}$. (similar).

Now, define

$$Q_{ik} = \sum_{j=1}^{n} \widetilde{Q}_{ijk} \quad \forall i, k.$$

Claim: $Q = [Q_{ik}] \in \mathcal{A}(a,c)$.

In fact, all $Q_{ik} \geq 0$.

$$\forall i: \quad \sum_k Q_{ik} = \sum_k \sum_j \tilde{Q}_{ijk} = \sum_j \sum_k \tilde{Q}_{ijk} \overset{(3)}{=} \sum_j R_{ij} \overset{R \in \mathcal{A}(a,b)}{=} a_i$$

$$\forall k: \quad \sum_i Q_{ik} = \sum_i \sum_j \tilde{Q}_{ijk} = \sum_j \sum_i \tilde{Q}_{ijk} \overset{(4)}{=} \sum_j S_{jk} \underset{S \in \mathcal{A}(b,c)}{=} c_k$$

(Hence, $Q \in \mathcal{A}(a,c)$.

Now, we have

$$W(a,c) \leq \sum_{i,k} Q_{ik} C_{ik} \overset{\text{def. of } Q_{ik}}{=} \sum_{i,k} \sum_j \tilde{Q}_{ijk} C_{ik}$$

$$\leq \sum_{i,j,k} \tilde{Q}_{ijk} (C_{ij} + C_{jk}) \quad \left[\begin{array}{l} C \text{ is a metric matrix} \\ \text{All } \tilde{Q}_{ijk} \geq 0 \end{array}\right]$$

$$= \sum_{i,j} C_{ij} \Big(\sum_k \tilde{Q}_{ijk} \Big) + \sum_{j,k} C_{jk} \Big(\sum_i \tilde{Q}_{ijk} \Big)$$

$$\overset{(3),(4)}{=} \sum_{i,j} C_{ij} R_{ij} + \sum_{j,k} C_{jk} S_{jk}$$

$$= W(a,b) + W(b,c).$$

QED