Minimal Joint Entropy and Order-Preserving Couplings

Ma Yajing
Joint work with Wang Feng, Wu Xianyuan and Cai Kaiyuan

Capital Normal University

July 30, 2023

Contents

Research question

Suppose $\mathcal{X} = \mathcal{Y} = \{1, 2, \cdots n\}, n \geq 2$ and $\mathbf{p} = (p_1, p_2, \cdots, p_n)$, $\mathbf{q} = (q_1, q_2, \cdots, q_n)$ be two discrete probability distributions on $\mathcal{X}$. First we associate random variables X, Y in $\mathcal{X}$ to p and q in some way (see the following Definition 5) respectively, second we seek a minimum-entropy two-dimensional random vector (X, Y) in $\mathcal{X} \times \mathcal{X}$ with marginals $\mathbf{p}$ and $\mathbf{q}$ (see the optimization problem (8)).

- One strategy to solve the problem mentioned above is to calculate the exact value of the minimum entropy $H(X, Y)$. Since, for general case, the corresponding optimization problem is known to be NP-hard, people prefer to give it good estimates. Recently, F. Cicalese etc. [5] solved the problem almost perfectly in this respect. Actually, they obtained an efficient algorithm to find a joint distribution with entropy exceeding the minimum at most by 1.

Research question

- Another strategy to study the problem is to seek the unknown special structure of a minimum-entropy coupling (X, Y).
- What special structure in a coupling of (X, Y) (i.e. joint probablity distribution of X and Y) will determine the minimum entropy of the two-demensional random system?
- The main goal is to establish such a structure.

Shannon entropy

Definition 1 (Shannon entropy)

Denote by $\mathcal{P}_n$ the set of all discrete probability distributions on $\mathcal{X} = \{1, 2, \cdots, n\}$. The **Shannon entropy** of X (or $\mathbf{p}$) is defined by

$$H(X) = H(\mathbf{p}) := -\sum_{i=1}^{n} p_i \log p_i, \tag{1}$$

where X is a discrete random variable with probability mass $\mathbf{p} = \{p_1, p_2, \cdots, p_n\} \in \mathcal{P}_n$, log is the base-2 logarithm.

Clearly, $H(X)$ takes its minimum 0 when X is degenerated and takes its maximum $\log|\mathcal{X}|$ when X is uniformly distributed. In this sense, entropy is a measure of the uncertainty of a random element.

Proposition

Proposition 2 (uncertainty of univariate distribution)

Let X be a discrete random variable with probability mass $\mathbf{p} = \{p_1, p_2, \cdots, p_n\}$, then

$$0 \le H(X) = H(\mathbf{p}) \le \ln n. \tag{2}$$

- $H(X)$ *or* $H(\mathbf{p})$ *takes its minimum* 0 *only and if only* X *is degenerated, which means* $\exists i \in n$ *such that* $p_i = 1$ *(that is to say, the most ordered structure of* X *determines the minimum entropy)*
- $H(X)$ *or* $H(\mathbf{p})$ *takes its maximum* $\ln n$ *only and if only* X *is uniformly distributed with* $\mathbf{p} = \{\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\}$ *(that is to say, the most disordered structure of* X *determines the maximum entropy).*

Joint entropy

Definition 3 (Joint entropy)

Let (X, Y) be a two-dimensional discrete random vector in $\mathcal{X} \times \mathcal{X}$ with a joint distribution $\mathbf{P} = \{p_{ij} : i \in \mathcal{X}, j \in \mathcal{X}\}$, the **joint entropy** of (X, Y) (or $\mathbf{P}$) is defined by

$$H(X, Y) = H(\mathbf{P}) := -\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \log p_{ij}, \tag{3}$$

where $\mathbf{P}$ is a probability matrix with marginals $\mathbf{p}$ and $\mathbf{q}$ distributed on $\mathcal{X}$, we call $\mathbf{P}$ as a coupling of $\mathbf{p}$ and $\mathbf{q}$.

Mutual information

Another important concept on (X, Y) is the mutual information, which is a measure of the amount of information that one random variable contains about the other.

Definition 4 (Mutual information)

Let (X, Y) be a two-dimensional discrete random vector in $\mathcal{X} \times \mathcal{X}$ with a joint distribution $\mathbf{P} = \{p_{ij} : i \in \mathcal{X}, j \in \mathcal{X}\}$, the **mutual information** of (X, Y) (or $\mathbf{P}$) is defined by

$$I(X, Y) := \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{p_i p_j}, \tag{4}$$

where $\mathbf{P}$ is a probability matrix with marginals $\mathbf{p}$ and $\mathbf{q}$ distributed on $\mathcal{X}$, we call $\mathbf{P}$ as a coupling of $\mathbf{p}$ and $\mathbf{q}$.

By definition, one has

$$I(X, Y) = H(X) + H(Y) - H(X, Y). \tag{5}$$

Permutation

Denote by $\mathcal{P}_n$ the set of all discrete probability distributions on $\mathcal{X}$. For each $\mathbf{p} \in \mathcal{P}_n$, let $F_{\mathbf{p}}$ be the cumulative distribution function defined by

$$F_{\mathbf{p}} := \sum_{k=1}^{i} p_k, 1 \leq i \leq n. \tag{6}$$

Recall that a permutation σ is a bijective map from $\mathcal{X}$ into itself. For any given distribution $\mathbf{p} = (p_1, p_2, \cdots, p_n) \in \mathcal{P}_n$, define $\sigma\mathbf{p} := (p_{\sigma(1)}, p_{\sigma(2)}, \cdots, p_{\sigma(n)})$. By the definition of entropy, one has

$$H(\mathbf{p}) = H(\sigma\mathbf{p}), \tag{7}$$

holds for any permutation σ. Based on this fact, we identify all $\sigma\mathbf{p}'$ s with $\mathbf{p}$ as one distribution on $\mathcal{X}$.

Equivalence relation "$\sim$"

- To this end, we define an equivalence relation "$\sim$" in $\mathcal{P}_n$: for any $\mathbf{p}, \mathbf{p}' \in \mathcal{P}_n$, $\mathbf{p} \sim \mathbf{p}'$ if and only if for some permutation σ, $\mathbf{p}' = \sigma\mathbf{p}$.
- Denote by $\bar{\mathcal{P}}_n$ the subset of all $\mathbf{p} \in \mathcal{P}_n$ such that $p_1 \geq p_2 \geq \cdots \geq p_n$. Obviously, $\bar{\mathcal{P}}_n$ is an **isomorphism** of the quotient space $\mathcal{P}_n/\sim$, we should identify $\bar{\mathcal{P}}_n$ with $\mathcal{P}_n/\sim$ in case of necessity. For each $\mathbf{p} \in \bar{\mathcal{P}}_n$, we call it an isoentropy distribution

Isoentropy distributions

Definition 5

Given isoentropy distributions $\mathbf{p}, \mathbf{q} \in \bar{\mathcal{P}}_n$. Suppose X is a random variable in $\mathcal{X}$ and (X, Y) is a two-dimensional random vector in $\mathcal{X} \times \mathcal{X}$ with joint distribution matrix P.

1. Random variable X is distributed according to isoentropy distribution $\mathbf{p}$, if for some permutation σ, X is distributed according to $\sigma\mathbf{p}$.
2. Random vector (X, Y) (or its joint distribution P) is called having marginals $\mathbf{p}$ and $\mathbf{q}$, if for some permutations pair σ, σ', P has marginals $\sigma\mathbf{p}$ and $\sigma'\mathbf{q}$.

A coupling of $\mathbf{p}, \mathbf{q}$

Denote by $\bar{\mathcal{P}}_n$ the subset of all $\mathbf{p} \in \mathcal{P}_n$ such that $p_1 \geq p_2 \geq \cdots \geq p_n$. For any $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$, denote by $\mathcal{C}(\mathbf{p}, \mathbf{q})$ the set of all joint distributions with marginals $\mathbf{p}, \mathbf{q}$. For any $\mathbf{p}, \mathbf{q} \in \bar{\mathcal{P}}_n$, denote by $\mathcal{C}_e(\mathbf{p}, \mathbf{q})$ the set of all joint distribution matrix P with marginals $\mathbf{p}, \mathbf{q}$. For any $P \in \mathcal{C}_e(\mathbf{p}, \mathbf{q})$, with a little abuse of terminology, we call P a coupling of $\mathbf{p}, \mathbf{q}$.

Optimization problem

Now we turn to the following optimization problem: to find a $\hat{P} \in \mathcal{C}_e(\mathbf{p}, \mathbf{q})$, such that

$$H(\hat{P}) = \inf_{P \in \mathcal{C}_e(\mathbf{p},\mathbf{q})} H(P). \tag{8}$$

For any $\mathbf{p}, \mathbf{q} \in \bar{\mathcal{P}}_n \subset \mathcal{P}_n$, let $\mathbf{p} \wedge \mathbf{q}$ be the distribution with cumulative distribution function $F_{\mathbf{p} \wedge \mathbf{q}} = F_{\mathbf{p}} \wedge F_{\mathbf{q}}$. F. Cicalese etc. obtained the following relation in[5]

$$H(\mathbf{p} \wedge \mathbf{q}) \leq H(\hat{P}) = \inf_{P \in \mathcal{C}_e(\mathbf{p},\mathbf{q})} H(P) \leq H(\mathbf{p} \wedge \mathbf{q}) + 1. \tag{9}$$

In fact, to get the upper estimate, [5] constructed a $P \in \mathcal{C}_e(\mathbf{p}, \mathbf{q})$ from $\mathbf{p} \wedge \mathbf{q}$ such that $H(P) \leq H(\mathbf{p} \wedge \mathbf{q}) + 1$, but no special structure of that P is worthy of attention.

Optimization problem

Now, for any $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$, $\mathcal{C}(\mathbf{p}, \mathbf{q})$ is an isomorphism of the quotient space $\mathcal{C}_e(\bar{\mathbf{p}}, \bar{\mathbf{q}})/\sim$, where $\bar{\mathbf{p}}, \bar{\mathbf{q}} \in \bar{\mathcal{P}}_n$, $\bar{\mathbf{p}} \sim \mathbf{p}$, $\bar{\mathbf{q}} \sim \mathbf{q}$. On account of the fact that

$$\inf_{P \in \mathcal{C}(\mathbf{p}, \mathbf{q})} H(P) = \inf_{P \in \mathcal{C}_e(\bar{\mathbf{p}}, \bar{\mathbf{q}})} H(P).$$

the optimization problem (8) is equivalent to the following original one

$$\tilde{P} : H(\tilde{P}) = \inf_{P \in \mathcal{C}(\mathbf{p}, \mathbf{q})} H(P). \tag{10}$$

Order-preserving coupling

Definition 6 (order-preserving distribution)

For any $\mathbf{p}, \mathbf{q} \in \bar{\mathcal{P}}_n$, a coupling $P \in \mathcal{C}_e(\mathbf{p}, \mathbf{q})$ is called order-preserving, if P is upper triangular, i.e., for any $1 \leq j \leq i \leq n, p_{i,j} = 0$. In other words, if (X, Y) is distributed according to P, then

$$\mathbb{P}(X \leq Y) = 1. \tag{11}$$

Denote by $\mathcal{O}(\mathbf{p}, \mathbf{q})$ the set of all order-preserving couplings of $\mathbf{p}, \mathbf{q} \in \bar{\mathcal{P}}_n$.

Order-preserving coupling

Proposition 7

For any $n \geq 2$, *and for any* $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$, $\mathcal{O}(\mathbf{p}, \mathbf{q}) \neq \emptyset$

The main result

Now, we state our main result as the following.

The Main Theorem:

Suppose $n \geq 2$ *and* $\mathbf{p}, \mathbf{q} \in \bar{\mathcal{P}}_n$. *If* $\hat{P} \in \mathcal{C}_e(\mathbf{p}, \mathbf{q})$ *is a solution of the optimization problem 8, then* $\hat{P}$ *is order-preserving. In other words, for any* $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$, *suppose* $\mathbf{p} \sim \bar{\mathbf{p}}$, $\mathbf{q} \sim \bar{\mathbf{q}}$ *and* $\bar{\mathbf{p}}, \bar{\mathbf{q}} \in \bar{\mathcal{P}}_n$, *if* $\tilde{P} \in \mathcal{C}_e(\mathbf{p}, \mathbf{q})$ *is a solution of the optimization problem (10), then there exists* $\hat{P} \in \mathcal{O}(\mathbf{p}, \mathbf{q})$ *such that* $\tilde{P} \sim \hat{P}$.

By the theorem aboved, the optimization problem (8) can be simplified as the following

$$\hat{P} : H(\hat{P}) = \inf_{P \in \mathcal{O}(\mathbf{p}, \mathbf{q})} H(P). \tag{12}$$

With this simplification, firstly, the corresponding computational complexity is well reduced; secondly, the order-preserving structure may possibly help us to construct the concrete form of $\hat{P}$.

Local optimization lemmas

Lemma 8

For any second order positive square matrix $A = (a_{i,j})_{2\times 2}$. *Suppose that* $a_{1,1} \vee a_{2,2} \geq a_{1,2} \vee a_{2,1}$, *let* $b = a_{1,2} \wedge a_{2,1}$. *Let* $A' = (a'_{i,j})_{2\times 2}$ *such that* $a'_{i,i} = a_{i,i} + b, i = 1, 2$, $a'_{i,j} = a_{i,j} - b, i \neq j$. *Then* $H(A) \geq H(A')$. *Furthermore, if* $b > 0$, *then* $H(A) \geq H(A')$.

Local optimization lemmas

Lemma 9

For any second order positive square matrix $A = (a_{i,j})_{2\times 2}$. *Suppose that* $a_{1,1} + a_{1,2} \geq a_{2,1} + a_{2,2}$, $a_{1,1} + a_{2,1} \geq a_{1,2} + a_{2,2}$ *and* $a_{1,1} + a_{1,2} \geq a_{1,1} + a_{2,1}$. *Let* $b = a_{1,2} \wedge a_{2,1}$, *define* A' *as in Lemma 8, then* $H(A) \geq H(A')$.

Proof of the main thoerem I

$\forall n \geq 2$, $P = (p_{i,j})_{n\times n}$, $\sum_{j=1}^{n} p_{i,j} = \mathbf{p}_i$, $\sum_{i=1}^{n} p_{i,j} = \mathbf{q}_j$, Using Lemma8 to optimize P. Let A be any second-order square matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} p_{i_1 j_1} & p_{i_1 j_2} \\ p_{i_2 j_1} & p_{i_2 j_2} \end{pmatrix}. \tag{13}$$

If A can be optimized to A', by Lemma 8, $H(A) \geq H(A')$. Then $H(P) - H(P') = H(A) - H(A') \geq 0$. And proof by mathmatical induction column and column order can always be locally optimized and swapped. So that P eventually becomes the upper triangular matrix. When $n = 2$,

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}. \tag{14}$$

Let's assume p_{11} is the matrixmum and $\mathbf{p}_1 \geq \mathbf{q}_1$, then $p_{11} \vee p_{22} \geq p_{21} \vee p_{12}$, let $b = p_{21} \wedge p_{12}$, Without loss of generality, assume $p_{21} \leq p_{12}$,

$$P' = \begin{pmatrix} p_{11} + p_{21} & p_{12} - p_{21} \\ 0 & p_{22} + p_{21} \end{pmatrix} = \begin{pmatrix} \mathbf{q}_1 & \mathbf{p}_1 - \mathbf{q}_1 \\ 0 & \mathbf{p}_2 \end{pmatrix}. \tag{15}$$

Proof of the main thoerem II

Assume $n = k$ $(k \geq 2)$, the conclusion is tenable. When $n = k + 1$, $P = (p_{ij})_{(k+1)\times(k+1)}$. Let $p_{i_0,j_0} = \max_{1\leq i\leq k+1, 1\leq j\leq k+1} p_{ij}$. Assume $\mathbf{p}_{i_0} \geq \mathbf{q}_{j_0}$, exchange the 1st and i_0th row of matrix P, then swap the 1st and j_0th column of matrix P, such that p_{11} maximized , and $\sum_{j=1}^{k+1} p_{1j} \geq \sum_{i=1}^{k+1} p_{i1}$.

$$A = \begin{pmatrix} p_{11} & p_{1i} \\ p_{i1} & p_{ij} \end{pmatrix} \longrightarrow A' = \begin{pmatrix} p_{11} + p_{j1} & p_{1i} - p_{j1} \\ 0 & p_{ij} \end{pmatrix}. \qquad (16)$$

Repeate the exchange steps, we can obtain a new matrix $P : p_{j1} = 0, j \geq 2$, that is

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ 0 & * & * & * \\ \vdots & * & * & * \\ 0 & * & * & * \end{pmatrix}. \qquad (17)$$

Proof of the main thoerem III

Denote by C_k the sum of k^2 elements of P_k, let$\bar{P}_k = \frac{1}{C_k} P_k$, then $\bar{P}_k$ is a $k \times k$ probability matrix.

$$\begin{aligned} H(\bar{P}_k) &= -\sum_{i=2}^{k+1}\sum_{j=2}^{k+1} \frac{p_{ij}}{C_k} \ln \frac{p_{ij}}{C_k} \\ &= -\frac{1}{C_k}\sum_{i=2}^{k+1}\sum_{j=2}^{k+1} p_{ij}(\ln p_{ij} - \ln C_k) \\ &= \frac{1}{C_k} H(P_k) + \frac{1}{C_k}\sum_{i=2}^{k+1}\sum_{j=2}^{k+1} p_{ij} \ln C_k \\ &= \frac{1}{C_k} H(P_k) + \ln C_k \end{aligned} \tag{18}$$

When $\bar{P}_k$ is optimized by Lemma8, P_k is optimized accordingly. To make $\bar{P}_k$ an upper triangle. The order of the column vector also acts on the first row of P. The triangulation of P is realized by induction hypothesis.

References I

[1] Ludwig Boltzmann. Weitere studien über das wärmegleichgewicht unter gasmolekülen, sitzungs. akad. wiss. wein 66 (1872), 275–370; english: Further studies on the thermal equilibrium of gas molecules. *Kinetic Theory*, 2:88–174, 1872.

[2] Ludwig Boltzmann. *Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht*. Kk Hof-und Staatsdruckerei, 1877.

[3] Viktor Benes and Josef Stepán. *Distributions with given marginals and moment problems*. Springer Science & Business Media, 2012.

[4] Carlos María Cuadras, Josep Fortiana, and José A Rodriguez-Lallena. *Distributions with given marginals and statistical modelling*. Springer, 2002.

[5] Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. Minimum-entropy couplings and their applications. *IEEE Transactions on Information Theory*, 65(6):3436–3451, 2019.

References II

[6] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[7] Giorgio Dall Aglio, Giorgio Dall'Aglio, Samuel Kotz, and G Salinetti. *Advances in Probability Distributions with Given Marginals: Beyond the Copulas [; Lectures Presented at a" Symposium on Distributions with Given Marginals" Organized by the Dept. of Statistics of the University La Sapienza, Rome, Italy, Held in Rome in April 1990]*, volume 67. Springer Science & Business Media, 1991.

[8] Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, 3^ e serie, Sciences, Sect. A*, 14: 53–77, 1951.

[9] W Hoeffding. Masstabinvariante korrelationtheorie, schriften des mathematis chen instituts und des instituts für angewandte mathematik der universität berlin 5, 181# 233.(translated in fisher, ni and pk sen (1994). the collected works of wassily hoeffding, new york, 1940.

References III

[10] Gwo Dong Lin, Xiaoling Dou, Satoshi Kuriki, and Jin-Sheng Huang. Recent developments on the construction of bivariate distributions with fixed marginals. *Journal of Statistical Distributions and Applications*, 1(1):1–23, 2014.

[11] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[12] Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

Thank you for listening!