One problem on relationship between mutual information and variational distance *

Viacheslav V. Prelov

Institute for Information Transmission Problems
of the Russian Academy of Sciences
19 Bol'shoi Karetnyi, 127994 Moscow, Russia.
Email: prelov@iitp.ru

Abstract – *A generalization of one Pinsker's problem on estimation of mutual information via variation is considered. We obtain some upper and lower bounds for the maximum of the absolute value of the difference between the mutual information of several random variables via variational distance between the probability distributions of these random variables. In some cases, these bounds are optimal.*

1 Introduction

In [1], Pinsker considered the following problem. Let X and Y be two discrete random variables with a joint distribution P_{XY} and marginal distributions P_X and P_Y, respectively. The problem is to upper estimate the maximum of the mutual information $I(X;Y)$ under the condition that the distribution P_X is given and the variational distance $V(P_{XY}, P_X \times P_Y)$ between the joint distribution P_{XY} and the product of marginal distributions P_X and P_Y does not exceed a fixed $\tau \geq 0$. In [1], an upper bound for this maximum in terms of P_X and τ was obtained, which in most cases is better than the Csiszár –Körner estimate [2,3]

$$I(X;Y) \leq \tau \ln \frac{N}{\tau},$$

where N is cardinality of range of the random variable X. Note that a lower estimate (known as *Pinsker's inequality*)

$$I(X;Y) \geq \frac{1}{2}\tau^2$$

was obtained earlier [2,4]. A lower estimate (which is optimal or asymptotically optimal in some special cases) for the maximum of $I(X;Y)$ in terms of P_X and τ was obtained in [5].

The Pinsker problem admits generalizations in several directions (see, e.g., [6-8]). Here we consider the most general problem of estimating the maximum of the absolute value of the difference $|I(X_1;...;X_n;Y) - I(X_1';...;X_n';Y')|$ via the variational distance $V(P_{X_1...X_nY}, P_{X_1'...X_n'Y'})$ and the distribution $P_{X_1...X_n}$ of the random variables $X_1,...,X_n$ (here $I(U_1;...;U_n)$ denotes the mutual information of the random variables $U_1,...,U_n$ (see [2] and Section 2)). This problem is reduced to the Pinsker problem above in the special case $n=1$ and under the additional assumptions that $P_{X'} = P_X$, $P_{Y'} = P_Y$, and X' and Y' are independent.

*This work was carried out with the partial support of the Russian Foundation for Basic Research (project no. 09-01-00536) .

2 Main definitions and results

Let $X_1, ..., X_n$ be discrete random variables ranging in finite sets $\mathcal{I}_i = \{1, 2, ..., N_i\}$, $i = 1, ..., n$, respectively, such that $|\mathcal{I}_i| = N_i$, where N_i are given integers and $|\mathcal{I}|$ denotes the cardinality of a set $\mathcal{I}$. Denote by

$$I(X_1; ...; X_n) = D\left(P_{X_1...X_n} \,||\, P_{X_1} \times \cdot \times P_{X_n}\right)$$

the information divergence which is usually called the *mutual information* of $X_1, ..., X_n$ (see, e.g., [2]). In the special case $n = 2$, the quantity $I(X_1; ...; X_n)$ coincides with the standard mutual information $I(X_1; X_2)$ of two random variables. In what follows, we denote vectors by boldfaceletters, e.g., $\mathbf{X} = (X_1, ..., X_n)$, $\mathbf{N} = (N_1, ..., N_n)$, etc.

Given a random vector $\mathbf{X}$ and a positive τ, define

$$J_\tau(\mathbf{X}) = \sup_{\mathbf{X}', Y, Y'} |I(X_1; \ldots; X_n; Y) - I(X'_1; \ldots; X'_n; Y')|,$$

where the supremum is over all random variables $\mathbf{X}' = (X'_1, \ldots, X'_n)$, Y, and Y' such that

$$V(P_{\mathbf{X}Y}, P_{\mathbf{X}'Y'}) = V(P_{X_1...X_nY}, P_{X'_1...X'_nY'}) \leq \tau.$$

Here we assume that the random variables X_k and X'_k take values in the set $\mathcal{I}_k$, $|\mathcal{I}_k| = N_k$, $k = 1, \ldots, n$, and the random variables Y and Y' range in a finite or countable set $\mathcal{J}$.

Denote also

$$J_\tau^{(\mathbf{N})} = \sup_{\mathbf{X}} J_\tau(\mathbf{X}),$$

where $\mathbf{N} = (N_1, \ldots, N_n)$, and the supremum is over all random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ whose components X_i take values in $\mathcal{I}_i$, $|\mathcal{I}_i| = N_i$, $i = 1, \ldots, n$. Finally, define the quantities

$$\begin{aligned} \tau^0_{\mathbf{X}} &= \left\{\inf \tau \,:\, J_\tau(\mathbf{X}) = \max_\nu J_\nu(\mathbf{X}) \right. \\ &= \left. \max\left\{\sum_{i=1}^n H(X_i), \ln N - I(X_1; ...; X_n)\right\}\right\} \end{aligned}$$

and

$$\tau^0 = \tau^0(\mathbf{N}) = \left\{\inf \tau \,:\, J_\tau^{(N)} = \ln N\right\},$$

where $N = \prod\limits_{k=1}^n N_k$.

We obtain some upper and lower bounds for $J_\tau(\mathbf{X})$, $J_\tau^{(\mathbf{N})}$, $\tau^0_{\mathbf{X}}$, and τ^0 which are contained in the theorems formulated below. In some special cases, these upper and lower bounds coincide or are rather close.

To state our main results, we introduce the following definitions. For any ordered collection of $L \geq 1$ nonnegative numbers $\xi = \{\xi_i\}$ such that $1 \geq \xi_1 \geq \xi_2 \geq \ldots \geq \xi_L \geq 0$, $\sum\limits_{i=1}^L \xi_i \leq 1$, and for some integers $K \geq 1$ and $m \geq 1$, put

$$\lambda(\xi, K) = \begin{cases} 2\left[1 - \dfrac{t}{K} - A(\xi, K)\right] & \text{if } \sum\limits_{k=1}^L \dfrac{1}{\xi_k} \geq K \\ 2\left[1 - \dfrac{L}{K}\right] & \text{if } \sum\limits_{k=1}^L \dfrac{1}{\xi_k} < K, \end{cases}$$

where

$$A(\xi, K) = \xi_{t+1} \left(1 - \frac{1}{K} \sum_{k=1}^{t} \frac{1}{\xi_k} \right)$$

and the integer t is defined by the conditions

$$\sum_{k=1}^{t} \frac{1}{\xi_k} \leq K, \quad \sum_{k=1}^{t+1} \frac{1}{\xi_k} > K,$$

and

$$\mu(\xi, m) = \begin{cases} 2 \left[1 - \sum\limits_{k=1}^{s} \xi_k - B(\xi, m) \right] & \text{if } \sum\limits_{k=1}^{L} \sqrt[m]{\xi_k} \geq 1 \\ 2 \left[1 - \sum\limits_{k=1}^{L} \xi_k \right] & \text{if } \sum\limits_{k=1}^{L} \sqrt[m]{\xi_k} < 1, \end{cases}$$

where

$$B(\xi, m) = \left(1 - \sum_{k=1}^{s} \sqrt[m]{\xi_k} \right)^m$$

and s is an integer satisfying inequalities

$$\sum_{k=1}^{s} \sqrt[m]{\xi_k} \leq 1, \quad \sum_{k=1}^{s+1} \sqrt[m]{\xi_k} > 1.$$

In what follows, we always assume that the components $p(\mathbf{i}) = \Pr\{\mathbf{X} = \mathbf{i}\}$ of the probability distribution $p = \{p(\mathbf{i}),\ \mathbf{i} \in \mathcal{I} = \prod\limits_{k=1}^{n} \mathcal{I}_k\}$ of the random vector $\mathbf{X}$ and the numbers $N_k = |\mathcal{I}_k|,\ k = 1, \ldots, n$, are ordered in such a way that $p_{\max} = p(\mathbf{i}_1) \geq p(\mathbf{i}_2) \geq \ldots \geq p(\mathbf{i}_N) = p_{\min}$, where $N = \prod\limits_{k=1}^{n} N_k$, and $N_1 \geq N_2 \geq \ldots \geq N_n$.

An ordered collection of probabilities $\widehat{p} = \{p(\mathbf{j}_1) \geq p(\mathbf{j}_2) \geq \ldots \geq p(\mathbf{j}_{N_n})\}$ of several components of the probability distribution $p = \{p(\mathbf{i})\}$ is called *admissible* if every column of the matrix $||j_{kl}||,\ k = 1, \ldots, N_n,\ l = 1, \ldots, n$, constructed from components of the vectors $\mathbf{j}_k = (j_{k1}, \ldots, j_{kn}),\ k = 1, \ldots, N_n$, consists of different elements of the corresponding sets $\mathcal{I}_l,\ l = 1, \ldots, n$.

Theorem 1: The following statements are valid:

- If $\ln N - I(X_1; \ldots; X_n) > \sum\limits_{k=1}^{n} H(X_k)$, then

$$\lambda(p, N_1) \leq \tau^0_{\mathbf{X}} \leq \lambda(p, N) \tag{1}$$

and

$$\tau^0_{\mathbf{X}} \leq \lambda(\widehat{p}, N_1). \tag{2}$$

- If $\ln N - I(X_1; \ldots; X_n) < \sum\limits_{k=1}^{n} H(X_k)$, then

$$\mu(p, 2) \leq \tau^0_{\mathbf{X}} \leq 2(1 - p_{\max}) \tag{3}$$

and

$$\tau^0_{\mathbf{X}} \leq \mu(\widehat{p}, n+1). \tag{4}$$

In the upper estimates (2) and (4), $\widehat{p} = \{p(\mathbf{j}_k),\ k = 1, \ldots, N_n\}$ denotes any admissible collection of components of the probability distribution of the random vector $\mathbf{X}$.

Remark 1: If $n = 1$ and $H(X) < \ln N$, then the upper and lower estimates in (1) coincide and we have $\tau_X^0 = \lambda(p, N)$. If $H(X) = \ln N$, i.e., if X is uniformly distributed, then it is easy to show that $\tau_X^0 = \mu(p, 2)$.

Example: Let all components of a random vector $\mathbf{X} = (X_1, \dots, X_n)$ be the same with probability 1, i.e., $X_i = X_1,\ i = 2, \dots, n$. It is easy to see that $\tau_{\mathbf{X}}^0 = \lambda(p, N_1)$ if $H(X_1) < \dfrac{n \ln N_1}{2n-1}$. Indeed, this fact immediately follows from estimates (1) and (2) since we can put $\widehat{p} = p$. On the other hand, if $H(X_1) > \dfrac{n \ln N_1}{2n-1}$, then one can show that $\tau_{\mathbf{X}}^0 = \mu(p, n+1)$.

Theorem 2: The following estimates for $\tau^0 = \tau^0(\mathbf{N})$ are valid:

$$\tau^0 \geq 2\left[1 - \frac{\lfloor\sqrt{N_1}\rfloor}{N_1} - \left(1 - \frac{\lfloor\sqrt{N_1}\rfloor}{\sqrt{N_1}}\right)^2\right] \tag{5}$$

and

$$\tau^0 \leq \begin{cases} 2\left[1 - \dfrac{\lfloor\sqrt[n+1]{N_1}\rfloor}{N_1} - \alpha(N_1)\right] & \text{if } N_n > \sqrt[n+1]{N_1}, \\ 2\left[1 - \dfrac{N_n}{N_1} - \beta(N_1, N_n)\right] & \text{if } N_n \leq \sqrt[n+1]{N_1} \text{ and } \left(\sqrt[n+1]{N_1} - N_n\right)^\nu < \dfrac{1}{N_n}, \\ 2\left[1 - \dfrac{N_n}{N_1} - \dfrac{1}{N_1 N_n}\right] & \text{if } N_n \leq \sqrt[n+1]{N_1} \text{ and } \left(\sqrt[n+1]{N_1} - N_n\right)^\nu \geq \dfrac{1}{N_n}, \end{cases} \tag{6}$$

where

$$\alpha(N_1) = \left(1 - \frac{\lfloor\sqrt[n+1]{N_1}\rfloor}{\sqrt[n+1]{N_1}}\right)^{n+1}, \quad \beta(N_1, N_n) = \frac{\left(\sqrt[n+1]{N_1} - N_n\right)^\nu}{N_1}$$

and an integer $\nu \geq 2$ is defined by the relations $N_{\nu-1} > N_\nu = N_n$.

Remark 2: Note that for $n = 1$, the upper and lower estimates (5) and (6) coincide. We conjecture that the upper estimate (6) is tight for any $n \geq 2$, i.e., it gives an exact expression for τ^0 if $N_n > \sqrt[n+1]{N_1}$.

In the definitions of $J_\tau(\mathbf{X})$ and $J_\tau^{\mathbf{N}}$, besides the main condition $V(P_{\mathbf{X}Y}, P_{\mathbf{X}'Y'}) \leq \tau$, it is sometimes introduced some additional conditions. In particular, if we additionally assume that $\mathbf{X}' \sim \mathbf{X}$ (the notation $U' \sim U$ means that $P_{U'} = P_U$) or $\mathbf{X}' \sim \mathbf{X}$ and $Y' \sim Y$ simultaneously, then it is easy to see that

$$\max_\tau J_\tau(\mathbf{X}) = \sum_{k=1}^n H(X_k) - I(X_1; \dots; X_n).$$

On the other hand, if we assume that only $Y' \sim Y$, then

$$\max_\tau J_\tau(\mathbf{X}) = \max\left\{\sum_{k=1}^n H(X_k),\ \ln N - I(X_1; \dots; X_n)\right\}.$$

In the next theorem, we give explicit expressions for $\tau_{\mathbf{X}}^0$ and τ^0 under the additional conditions on $\mathbf{X}'$, Y, and Y' given above.

Theorem 3: The following statements are valid:

- Let $\mathbf{X} \sim \mathbf{X}'$. Then

$$\tau_{\mathbf{X}}^0 = 2(1 - p_{\max}) \quad \text{and} \quad \tau^0 = 2\left(1 - \frac{1}{N}\right); \tag{7}$$

- Let $\mathbf{X} \sim \mathbf{X}'$ and $Y \sim Y'$. Then

$$\tau^0_{\mathbf{X}} = 2\left(1 - \sum_{\mathbf{i}\in\mathcal{I}} p^2(\mathbf{i})\right) \quad \text{and} \quad \tau^0 = 2\left(1 - \frac{1}{N}\right); \tag{8}$$

- Let $Y \sim Y'$. Then

$$\tau^0 = 2\left(1 - \frac{1}{N_1}\right). \tag{9}$$

Moreover, if $\sum\limits_{k=1}^{n} H(X_k) > \ln N - I(X_1;\ldots;X_n)$, then

$$\tau^0_{\mathbf{X}} = 2\left(1 - p_{\max}\right), \tag{10}$$

and if $\sum\limits_{k=1}^{n} H(X_k) < \ln N - I(X_1;\ldots;X_n)$, then

$$\tau^0_{\mathbf{X}} \geq 2\left(1 - \frac{1}{N_1}\sum_{k=1}^{N_1} p(\mathbf{i}_k)\right), \tag{11}$$

and

$$\tau^0_{\mathbf{X}} \leq \min\left\{2\left(1 - \frac{1}{N_n}\sum_{k=1}^{N_n} p(\mathbf{j}_k)\right),\ 2\left(1 - \frac{1}{N}\sum_{k=1}^{N} p(\mathbf{i}_k)\right)\right\}, \tag{12}$$

where $\{p(\mathbf{j}_k),\ k = 1,\ldots,N_n\}$ is any admissible collection of components of the probability distribution p.

In the next two theorems, some upper and lower bounds for $J_\tau(\mathbf{X})$ and $J_\tau^{\mathbf{N}}$ are given.

Theorem 4: The quantity $J_\tau(\mathbf{X})$ satisfies the following inequality:

$$J_\tau(\mathbf{X}) \geq \frac{\widehat{\tau}}{2}\ln(N-1) + h\left(\frac{\widehat{\tau}}{2}\right), \quad 0 \leq \tau < 2\left(1 - \frac{1}{N}\right), \tag{13}$$

where

$$\widehat{\tau} = \widehat{\tau}(\mathbf{X}) = \begin{cases} \tau & \text{if} \quad p_{\min} \geq \dfrac{\tau}{2(N-1)}, \\ 2(N-1)p_{\min} & \text{if} \quad p_{\min} < \dfrac{\tau}{2(N-1)}, \end{cases} \tag{14}$$

and $h(x) = -x\ln x - (1-x)\ln(1-x)$ is the binary entropy function. Moreover, for all $\tau,\ \gamma \leq \tau \leq \tau^*$, we also have

$$J_\tau(\mathbf{X}) \geq \sum_{k=1}^{n} H(X_k) - \frac{\tau^* - \tau}{\tau^* - \gamma} H(\mathbf{X}), \tag{15}$$

where

$$\gamma = \gamma(\mathbf{X}) = V(P_{X_1\ldots X_n}, P_{X_1} \times \ldots \times P_{X_n}),$$
$$\tau^* = \tau^*(\mathbf{X}) = 2\Big(1 - \sum_{\mathbf{i}=(i_1\ldots i_n)} p(\mathbf{i})p^{(1)}(i_1)\ldots p^{(n)}(i_n)\Big)$$

and

$$p(\mathbf{i}) = \Pr\{\mathbf{X} = \mathbf{i}\} = \Pr\{X_1 = i_1,\ldots,X_n = i_n\},$$
$$p^{(k)}(i_k) = \Pr\{X_k = i_k\},\ i_k \in \mathcal{I}_k,\ k = 1,\ldots,n.$$

Theorem 5: The quantity $J_\tau^{(\mathbf{N})}$, $0 \leq \tau < \tau^0$ satisfies the following inequalities:

- $$J_\tau^{(\mathbf{N})} \geq \frac{\tau}{2} \ln \left[\prod_{k=1}^{n} (N_k - 1) \right] + nh\left(\frac{\tau}{2}\right) \tag{16}$$

  and

  $$J_\tau^{(\mathbf{N})} \leq \frac{\tau}{2} \ln \left[N \prod_{k=1}^{n} (N_k - 1) \right] + (n+1)h\left(\frac{\tau}{2}\right); \tag{17}$$

- If $\mathbf{X} \sim \mathbf{X}'$, then we have

  $$\frac{\tau}{2} \ln(N-1) + h\left(\frac{\tau}{2}\right) \leq J_\tau^{(\mathbf{N})} \leq \frac{\tau}{2} \ln N + h\left(\frac{\tau}{2}\right); \tag{18}$$

- If $\mathbf{X} \sim \mathbf{X}'$ and $Y \sim Y'$, then

  $$J_\tau^{(\mathbf{N})} = \frac{\tau}{2} \ln(N-1) + h\left(\frac{\tau}{2}\right); \tag{19}$$

- If $Y \sim Y'$, then

  $$J_\tau^{(\mathbf{N})} \geq \frac{\tau}{2} \ln \left[\prod_{k=1}^{n} (N_k - 1) \right] + nh\left(\frac{\tau}{2}\right) \tag{20}$$

  and

  $$J_\tau^{(\mathbf{N})} \leq \frac{\tau}{2} \ln \left[(N-1) \prod_{k=1}^{n} (N_k - 1) \right] + (n+1)h\left(\frac{\tau}{2}\right). \tag{21}$$

The proofs of theorems formulated above can be found in [9].

References

[1] M. S. Pinsker, “On estimation of information via variation,” *Problemi Peredachi Informatsii,* vol. 41, no. 2, pp. 3–18, 2005.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Budapest: Academiai Kiado,1981.

[3] I. Csiszár, “Almost independence and secrecy capacity,”*Problemi Peredachi Informatsii,* vol. 32, no. 1, pp. 48–57, 1996.

[4] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes.* San Francisco: Holden-Day, 1964.

[5] V. V. Prelov, “On inequalities between mutual information and variation,” *Problemi Peredachi Informatsii,* vol. 43 , no. 1, pp. 15–27, 2007.

[6] V. V. Prelov, “Mutual information of several random variables and its estimation via variation ,” *Problemi Peredachi Informatsii,* vol. 45 , no. 4, pp. 3–17, 2009.

[7] Z. Zhang, “Estimating mutual information via Kolmogorov distance,” *IEEE Trans. Inform. Theory,* vol. 53, no. 9, pp. 3280–3282, 2007.

[8] V. V. Prelov and E. C. van der Meulen, “Mutual information, variation , and Fano’s inequality,” *Problemi Peredachi Informatsii,* vol. 44 , no. 3, pp. 19–32, 2008.

[9] V. V. Prelov, “Generalization of a Pinsker problem,” accepted for publication in *Problemi Peredachi Informatsii.*