On the Discontinuity of the Shannon Information Measures

Siu-Wai Ho and Raymond W. Yeung

Abstract

The Shannon information measures are well known to be continuous functions of the probability distribution for a given finite alphabet. In this paper, however, we show that these measures are discontinuous with respect to almost all commonly used "distance" measures when the alphabet is countably infinite. Such "distance" measures include the Kullback-Leibler divergence and the variational distance. Specifically, we show that all the Shannon information measures are in fact discontinuous at all probability distributions. The proofs are based on a probability distribution which can be realized by a discrete-time Markov chain with countably infinite number of states. Our findings reveal that the limiting probability distribution may not fully characterize the asymptotic behavior of a Markov chain. These results explain why certain existing information theoretical tools are restricted to finite alphabets, and provide hints on how these tools can be extended to countably infinite alphabet.

I. Introduction

The study of infinity is, as Georg Cantor realized, a form of soul's quest for God [1] and it has provided us many intellectual property including Gödel's Incompleteness Theorem [2]. In this paper, we will discuss an interesting concept brought by infinity and the concept can be illustrated through this statement: *We can be more and more sure that a particular event will happen as time goes, but at the same time, the uncertainty of the whole picture keeps on increasing.* If one finds the above statement counter-intuitive, he/she may have the preconception that entropy is

The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Adelaide, Sep., 2005.

S.-W. Ho is with Department of Electrical Engineering, Princeton University, NJ 08544, USA and he is now supported by The Croucher Foundation. He was with Department of Information Engineering, The Chinese University of Hong Kong, N.T., Hong Kong when part of this work was done. Email: siuho@princeton.edu

R. W. Yeung is with Department of Information Engineering, The Chinese University of Hong Kong, N.T., Hong Kong. Email: whyeung@ie.cuhk.edu.hk

continuous. But we will show that not only entropy but all the Shannon information measures are indeed discontinuous so that the above statement is possible. One implication in information theory, as we will see, is that we can now easily explain why certain theories or results are restricted to finite alphabets. Benefited from this observation, we can readily determine exactly what assumptions are necessary when we generalize these theories to countably infinite alphabets. In particular, we have already generalized the strong typicality of i.i.d. sequences and Fano's inequality to countably infinite alphabets in [3] and [4], respectively. Before getting into the details, we first review some basic facts about the continuity of entropy.

The Shannon information measures are functions mapping a probability distribution to a real value. If the input probability distribution is restricted to a given finite alphabet, then it is well-known that the Shannon information measures are continuous functions. Shannon, in fact, assumed the entropy function $H(p_1, p_2, ..., p_n)$ to be continuous in p_i when he introduced the definition of entropy in his seminal work [5]. Together with two other assumptions on entropy, he showed that the entropy function must take the form

$$H(\mathcal{P}) = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

for a probability distribution $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ with the base of the logarithm unspecified. The derivation of (1) is discussed in an exercise in [6, P.43] and the continuity of (1) has been discussed in [7]. Although the above definition is for probability distributions with finite alphabet, (1) is usually extended to

$$H(\mathcal{P}) = -\sum_{i=1}^{\infty} p_i \log p_i, \tag{2}$$

where H is applied to probability distributions with countably infinite alphabet. Since there is an ambiguity in (1) and (2) when $p_i = 0$ for some i, the convention (e.g., [6]) $0 \log 0 = 0$ is usually adopted due to the fact that $x \log x \to 0$ as $x \to 0$. McEliece [8, Problem 1.1] further shows that $H(X)$ being a continuous function for probability distributions with countably infinite alphabet is a necessary assumption for the convention $0 \log 0 = 0$. In this paper, however, we will adopt the definition [9]

$$H(\mathcal{P}) = -\sum_{i \in \mathcal{S}} p_i \log p_i,$$

where $\mathcal{S}$ is the support of $\mathcal{P}$. This definition avoids the ambiguity when $p_i = 0$ for some i and it is essentially the same as (1) and (2) with the convention $0 \log 0 = 0$. Note that all the Shannon information measures can be expressed as linear combinations of entropy. Therefore, the continuity of the Shannon information measures immediately follows from the continuity of entropy.

However when the alphabet is countably infinite, it is not clear whether the Shannon information measures are continuous or not. The situation can be even more complicated. Consider random variables X and Y where X takes value from a finite alphabet while Y takes value from a countably infinite alphabet. Then it is also not clear whether the mutual information $I(X;Y)$ is continuous or not. These would be trivial problems if entropy is always continuous. Unfortunately, there are certain results regarding the discontinuity of entropy on countably infinite alphabet. Consider a sequence of probability distributions $\mathcal{P}_n$ and a fixed probability distribution $\mathcal{Q}$. We want to know under what conditions

$$H(\mathcal{P}_n) \to H(\mathcal{Q}). \tag{3}$$

Suppose the Kullback-Leibler divergence [10] $D(\mathcal{P}_n||\mathcal{Q})$ is used. If $D(\mathcal{P}_n||\mathcal{Q}) \to 0$, then $H(\mathcal{P}_n) \to H(\mathcal{Q})$ if $\mathcal{Q}$ is power dominated but $H(\mathcal{P}_n)$ can tend to a value strictly greater than $H(\mathcal{Q})$ if $\mathcal{Q}$ is hyperbolic [11]. However, it is not clear whether entropy is continuous if $D(\mathcal{Q}||\mathcal{P}_n)$ instead is used to define convergence. Although entropy is discontinuous, note that it is lower-semi continuous [12], i.e., $\liminf_{n\to\infty} H(\mathcal{P}_n) \geq H(\mathcal{Q})$ if $\mathcal{P}_n$ pointwise converges to $\mathcal{Q}$.

The solutions for the above puzzles are organized in this paper as follows. In Section II, we will discuss some of the many well-known information divergence measures including the χ^2-divergence which will be used to define the continuity of a function of a probability distribution. If a function is discontinuous with respect to convergence in $\chi^2(\mathcal{Q}||\mathcal{P}_n)$, the function is also discontinuous with respect to convergence in most other common divergence measures, including the variational distance and the Kullback-Leibler divergence $D(\mathcal{Q}||\mathcal{P}_n)$. In Section III, we will give a sequence of probability distributions which converges to the deterministic distribution $\{1, 0, 0, ...\}$, while the entropy of the sequence may converge to any real number or tend to infinity. Then the result will be extended beyond the deterministic distribution to all probability distributions with finite entropy and countably infinite alphabet. After that, mutual information

will be shown to be discontinuous in Section IV by using two different approaches. Then all the Shannon information measures will be shown to be discontinuous at all probability distributions with countably infinite alphabet. In Section V, we will examine some constraints on the input distributions. For example, we will consider the set of joint distributions for random variables X and Y where X takes values in a finite alphabet but Y takes values in a countably infinity alphabet. Here, we will show that mutual information remains continuous, which can be viewed as an extension of the continuity of the Shannon information measures for finite alphabets. In Section VI, we will see that the discontinuity of entropy can provide satisfactory explanations to why certain theories or results, like strong typicality and Fano's inequality (see e.g., [6][9]), are restricted to finite alphabets. More importantly, this observation leads to some hints on how we can generalize these theoretical tools. In Section VII, we will show that the probability distribution which is used extensively in this paper to show the discontinuity of the Shannon information measures can be realized by a discrete-time Markov chain with a countably infinite number of states. This implies that the entropy of a Markov chain with infinite states may not tend to the entropy of its limiting probability distribution. As a result, the limiting probability distribution of a Markov chain may not fully characterize its asymptotic behavior. To conclude the paper, we give a mathematical proof of the statement in italics in the beginning paragraph of this introductory section.

II. Definitions

All the logarithms denoted by $\log$ in this paper are in the base 2. Consider probability distributions $\mathcal{P} = \{p_1, p_2, p_3, \ldots\}$ and $\mathcal{Q} = \{q_1, q_2, q_3, \ldots\}$ with countably infinite alphabet. Let $S_{\mathcal{P}}$ and $S_{\mathcal{Q}}$ be the support of $\mathcal{P}$ and $\mathcal{Q}$, respectively. The distance between $\mathcal{P}$ and $\mathcal{Q}$ can be measured by various divergence measures:

χ^2-Divergence

$$\chi^2(\mathcal{P}||\mathcal{Q}) = \sum_{i \in S_{\mathcal{P}}} \frac{p_i^2}{q_i} - 1,$$

where we adopt the convention $\chi^2(\mathcal{P}||\mathcal{Q}) = \infty$ if $q_i = 0$ but $p_i > 0$ for some i. Here, χ^2-divergence was due to Pearson [13] and it can be reexpressed as

$$\chi^2(\mathcal{P}||\mathcal{Q}) = \sum_{i \in S_{\mathcal{P}} \cup S_{\mathcal{Q}}} \frac{(p_i - q_i)^2}{q_i}. \tag{4}$$

Kullback-Leibler Divergence

$$D(\mathcal{P}||\mathcal{Q}) = \sum_{i \in S_{\mathcal{P}}} p_i \log \frac{p_i}{q_i},$$

where we adopt the convention $D(\mathcal{P}||\mathcal{Q}) = \infty$ if $q(x) = 0$ but $p(x) > 0$ for some x.

Variational Distance

$$V(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{\infty} |p_i - q_i|,$$

which is also known as the L_1-Norm. Note that

$$\chi^2(\mathcal{P}||\mathcal{Q}) \geq D(\mathcal{P}||\mathcal{Q}) \geq \frac{1}{2 \ln 2} V(\mathcal{P}, \mathcal{Q}), \tag{5}$$

where the first inequality can be verified by using $\ln x \leq x - 1$ and the second inequality follows from Pinsker's inequality [9].

In order to define the continuity of a function, a metric, like variational distance, should be used to define the convergence of a sequence of probability distributions. However, we will show that the χ^2-divergence alone can be used to obtain very general results.

Definition 1: Let $\mathcal{A}$ be a subset of all probability distributions and let $\mathcal{Q} \in \mathcal{A}$. Then a function $f : \mathcal{A} \to \mathcal{R}$ is **continuous** at $\mathcal{Q}$ if, given any $\varepsilon > 0$, there exists $\delta > 0$ such that if $\mathcal{P}$ is any distribution in $\mathcal{A}$ satisfying $\chi^2(\mathcal{Q}||\mathcal{P}) < \delta$, then $|f(\mathcal{P}) - f(\mathcal{Q})| < \varepsilon$.

If f fails to be continuous at $\mathcal{Q}$, then we say that f is **discontinuous** at $\mathcal{Q}$. The following definition, which is an alternative form of Definition 1, will be used to verify the discontinuity of a function.

Definition 2: Let $\mathcal{A}$ be a subset of all probability distributions and let $\mathcal{Q} \in \mathcal{A}$. Then a function $f : \mathcal{A} \to \mathcal{R}$ is **discontinuous** at $\mathcal{Q}$ if there exists a sequence $\mathcal{P}_n \in \mathcal{A}$ such that

$$\lim_{n \to \infty} \chi^2(\mathcal{Q}||\mathcal{P}_n) = 0,$$

but $f(\mathcal{P}_n)$ does not converge to $f(\mathcal{Q})$, i.e.,

$$\lim_{n \to \infty} f(\mathcal{P}_n) \neq f(\mathcal{Q}).$$

Two characteristics of χ^2-convergence are noteworthy: a) Due to (5), if $\mathcal{P}_n$ converges to $\mathcal{Q}$ with respect to the χ^2-divergence, then $\mathcal{P}_n$ also converges to $\mathcal{Q}$ with respect to the Kullback-Leibler divergence and variational distance. Therefore, if a function is discontinuous with respect to convergence in the χ^2-divergence, then it is also discontinuous with respect to convergence

in the Kullback-Leibler divergence or variational distance. Furthermore, the function is also discontinuous with respect to convergence in many other divergence measures as discussed in [14]. b) Following (5), $\chi^2(\mathcal{Q}||\mathcal{P}_n) \to 0$ implies $D(\mathcal{Q}||\mathcal{P}_n) \to 0$ instead of $D(\mathcal{P}_n||\mathcal{Q}) \to 0$. Our results are, therefore, different from the results obtained in [11].

III. The Discontinuity of Entropy

We first consider a probability distribution $\mathcal{P} = \{p_1, p_2, \ldots, p_L\}$ with L probability masses. For $0 < \alpha < 1$, let $\mathcal{Q} = \{\alpha p_1, \alpha p_2, \ldots, \alpha p_L, \frac{1-\alpha}{M}, \ldots, \frac{1-\alpha}{M}\}$ be a probability distribution with $L + M$ probability masses. Then

$$\chi^2(\mathcal{P}||\mathcal{Q}) = \sum_{i=1}^{L} \frac{p_i^2}{\alpha p_i} - 1 = \frac{1}{\alpha} - 1,$$

regardless of the value of M, and $\chi^2(\mathcal{P}||\mathcal{Q}) \to 0$ as $\alpha \to 1$. On the other hand, for any fixed $\alpha < 1$,

$$H(\mathcal{Q}) \geq \alpha H(\mathcal{P}) + (1-\alpha)\log M,$$

which tends to infinity as M tends to infinity. The above observation is summarized in the following proposition and a closer look in this case is given in [15].

Proposition 1: Suppose $\delta > 0$ and $\epsilon > 0$ are given. For any probability distribution $\mathcal{P}$ with L probability masses, there exists a sufficient large integer $M \geq L$ and a probability distribution $\mathcal{Q}$ with M probability masses such that $\chi^2(\mathcal{P}||\mathcal{Q}) < \epsilon$ but $H(\mathcal{Q}) - H(\mathcal{P}) > \delta$.

Now, we show that entropy is discontinuous at all probability distributions with countably infinite alphabet. In the following, let

$$\mathcal{D}_n = \left\{1 - \frac{\alpha}{\log n}, \frac{\alpha}{n\log n}, \frac{\alpha}{n\log n}, ..., 0, 0, ...\right\} \tag{6}$$

for $\alpha > 0$ and $n \geq 2^\alpha$. Let $\nu = \{1, 0, 0, \ldots\}$ be a deterministic distribution. Note that when $n \to \infty$, $\chi^2(\nu||\mathcal{D}_n) \to 0$ and

$$H(\mathcal{D}_n) \to \alpha. \tag{7}$$

Theorem 2: For any probability distribution $\mathcal{P}^0 = \{p_0, p_1, ...\}$ with countably infinite alphabet such that $H(\mathcal{P}^0) < \infty$ and any positive real number c including infinity, there exists a sequence of probability distributions $\mathcal{P}_n$ such that $\lim_{n\to\infty} \chi^2(\mathcal{P}^0||\mathcal{P}_n) = 0$ but $\lim_{n\to\infty} H(\mathcal{P}_n) = H(\mathcal{P}^0) + c$.

The proof of this theorem is given in Appendix A, where the sequence of probability distributions $\mathcal{P}_n$ is constructed explicitly. Note that we can still have the same conclusion if we restrict $\mathcal{P}_n$ to have a finite number of probability masses for all n.

IV. The Discontinuity of the Shannon Information Measures

We will first show the discontinuity of mutual information before the discontinuity of all the Shannon information measures will be discussed. Let $\tilde{\mathcal{P}}_{XY} = \{P_{XY}(x,y)\}$ be a joint probability distribution for random variables X and Y where $P_{XY}(x,y)$ is the probability that X equals x and Y equals y. The mutual information between X and Y with respect to $\tilde{\mathcal{P}}_{XY}$ is defined as

$$I_{X;Y}(\tilde{\mathcal{P}}_{XY}) = \sum_{xy:P_{XY}(x,y)>0} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}.$$

We can assume without loss of generality that $P_{XY}(0,0) > 0$, because if $P_{XY}(0,0) = 0$, we can always make $P_{XY}(0,0) > 0$ with an appropriate reindexing of the alphabets of X and Y. Let $q = 1 - P_{XY}(0,0)$ and

$$\tilde{\mathcal{P}}_{XY} = \begin{bmatrix} 0 & q^{-1}P_{XY}(1,0) & \cdots \\ q^{-1}P_{XY}(0,1) & q^{-1}P_{XY}(1,1) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

By letting $\tilde{\mathcal{D}}_n$ be a diagonal matrix with diagonal equals to the distribution

$$\mathcal{D}'_n = \left\{1 - \frac{1}{\sqrt{\log n}}, \frac{1}{n\sqrt{\log n}}, \frac{1}{n\sqrt{\log n}}, ..., 0, 0, ...\right\}, \tag{8}$$

that is

$$\tilde{\mathcal{D}}_n = \begin{bmatrix} 1 - \frac{1}{\sqrt{\log n}} & 0 & 0 & \cdots \\ 0 & \frac{1}{n\sqrt{\log n}} & 0 & \cdots \\ 0 & 0 & \frac{1}{n\sqrt{\log n}} & \\ \vdots & \vdots & & \ddots \end{bmatrix}.$$

Let Z be a binary random variable such that $\Pr\{Z = 0\} = q$ and $\Pr\{Z = 1\} = 1 - q$. Let $\tilde{\mathcal{Q}}^n_{XYZ}$ be the joint probability distribution of (X, Y, Z) for $n \geq 2$ such that

$$\tilde{\mathcal{Q}}^n_{XY|Z} = \begin{cases} \tilde{\mathcal{P}}_{XY} & \text{if } Z = 0 \\ \tilde{\mathcal{D}}_n & \text{if } Z = 1. \end{cases} \tag{9}$$

Therefore,

$$\tilde{\mathcal{Q}}_{XY}^n = \begin{bmatrix} 1-q-\frac{1-q}{\sqrt{\log n}} & P_{XY}(1,0) & \cdots \\ P_{XY}(0,1) & P_{XY}(1,1)+\frac{1-q}{n\sqrt{\log n}} & \\ \vdots & & \ddots \end{bmatrix}. \tag{10}$$

Theorem 3: Let $\tilde{\mathcal{P}}_{XY}$ be a joint probability distribution for random variables X and Y with countably infinite alphabet for both of the marginal probability distributions such that $I_{X;Y}(\tilde{\mathcal{P}}_{XY}) < \infty$. Then there exists a sequence of probability distributions $\tilde{\mathcal{P}}_{XY}^n$ such that $I_{X;Y}(\tilde{\mathcal{P}}_{XY}^n)$ is finite for $n \geq 2$ and $\lim_{n\to\infty} \chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{P}}_{XY}^n) = 0$ but $\lim_{n\to\infty} I_{X;Y}(\tilde{\mathcal{P}}_{XY}^n) = \infty$. Thus, $I_{X;Y}(\cdot)$ is discontinuous at $\tilde{\mathcal{P}}_{XY}$.

Proof: By the same argument used in the proof of Theorem 2, we have

$$\lim_{n\to 0} \chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{Q}}_{XY}^n) = 0.$$

On the other hand, it can readily be shown that

$$\begin{aligned} & I(X;Y) + I(X;Z|Y) \\ &= \Pr\{Z=0\} I(X;Y|Z=0) + \Pr\{Z=1\} I(X;Y|Z=1) + I(X;Z). \end{aligned} \tag{11}$$

For the joint probability distribution $\mathcal{Q}_{XYZ}^n$, the summations in $I(X;Z|Y)$, $I(X;Y|Z=0)$ and $I(X;Z)$ are bounded by

$$I(X;Z|Y) \leq H(Z) \leq \log 2,$$

$$I(X;Z) \geq 0,$$

$$\Pr\{Z=0\} I_{X;Y}(\tilde{\mathcal{Q}}_{XY|Z=0}^n) = q I_{X;Y}(\tilde{\mathcal{P}}_{XY}) \geq 0,$$

and

$$\begin{aligned} \Pr\{Z=1\} I_{X;Y}(\tilde{\mathcal{Q}}_{XY|Z=1}^n) &= (1-q) I_{X;Y}(\tilde{\mathcal{D}}_n) \\ &= (1-q) H(\mathcal{D}'_n), \end{aligned}$$

where $\tilde{\mathcal{Q}}_{XY|Z=0}^n = \tilde{\mathcal{P}}_{XY}$ and $\tilde{\mathcal{Q}}_{XY|Z=1}^n = \tilde{\mathcal{D}}_n$ follow from (9). Thus

$$\begin{aligned} I_{X;Y}(\tilde{\mathcal{Q}}_{XY}^n) &\geq 0 + (1-q) H(\mathcal{D}'_n) + 0 - \log 2 \\ &= (1-q) H(\mathcal{D}'_n) - \log 2. \end{aligned}$$

Therefore

$$\begin{aligned}\lim_{n\to\infty} I_{X;Y}(\tilde{\mathcal{Q}}_{XY}^n) &\geq \lim_{n\to\infty}(1-q)H(\mathcal{D}_n') - \log 2 \\ &= \infty.\end{aligned}$$

By Definition 2, the function $I_{X;Y}(\cdot)$ is discontinuous at the distribution $\tilde{\mathcal{P}}_{XY}$.

In Appendix B, we will show that $I_{X;Y}(\tilde{\mathcal{Q}}_{XY}^n)$ is finite for all integers n. ■

In the above theorem, we have constructed a sequence of probability distributions whose mutual information tends to infinity when $n \to \infty$ but the mutual information of each distribution in the sequence is finite. In the next theorem, we will resort to a different method to show that mutual information is discontinuous.

Theorem 4: Let $\tilde{\mathcal{P}}_{XY}$ be a joint probability distribution for random variables X and Y with countably infinite alphabet for both of the marginal probability distributions such that $I_{X;Y}(\tilde{\mathcal{P}}_{XY}) < \infty$. Then there exists a sequence of probability distributions $\tilde{\mathcal{P}}_{XY}^n$ such that $\lim_{n\to\infty} \chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{P}}_{XY}^n) = 0$ but $I_{X;Y}(\tilde{\mathcal{P}}_{XY}^n) = \infty$ for all integers n. Thus $I_{X;Y}(\cdot)$ is discontinuous at $\tilde{\mathcal{P}}_{XY}$.

Proof: Let $\tilde{\Phi}_{XY}$ be a joint probability distribution for random variables X and Y with $I_{X;Y}(\tilde{\Phi}_{XY}) = \infty$. An example is a diagonal matrix with elements equal to the distribution given in [9, Example 2.46].

Let S_n be a binary random variable such that $\Pr\{S_n = 0\} = 1 - \frac{1}{n}$ and $\Pr\{S_n = 1\} = \frac{1}{n}$. Let $\tilde{\mathcal{Q}}_{XYS_n}^n$ be the joint probability distribution of (X, Y, S_n) for $n \geq 2$ such that

$$\tilde{\mathcal{Q}}_{XY|S_n}^n = \begin{cases} \tilde{\mathcal{P}}_{XY} & \text{if } S_n = 0 \\ \tilde{\Phi}_{XY} & \text{if } S_n = 1. \end{cases} \tag{12}$$

Therefore,

$$\tilde{\mathcal{Q}}_{XY}^n = \left(1 - \frac{1}{n}\right)\tilde{\mathcal{P}}_{XY} + \frac{1}{n}\tilde{\Phi}_{XY}.$$

For any probability distributions $\mathcal{P} = \{p_1, p_2, \ldots\}$ and $\mathcal{Q} = \{q_1, q_2, \ldots\}$ with $q_i > 0$ for all i, we have

$$\begin{aligned}
&\lim_{n\to\infty} \chi^2\left(\mathcal{P}||\left(1-\frac{1}{n}\right)\mathcal{P}+\frac{1}{n}\mathcal{Q}\right)\\
&= \lim_{n\to\infty}\sum_{i=1}^{\infty}\frac{\left(p_i-(1-\frac{1}{n})p_i-\frac{1}{n}q_i\right)^2}{(1-\frac{1}{n})p_i+\frac{1}{n}q_i}\\
&= \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{\infty}\frac{(p_i-q_i)^2}{p_i-\frac{1}{n}(p_i-q_i)}\\
&= \lim_{n\to\infty}\frac{1}{n^2}\sum_{i:p_i>q_i}\frac{(p_i-q_i)^2}{p_i-\frac{1}{n}(p_i-q_i)}+\lim_{n\to\infty}\frac{1}{n^2}\sum_{i:p_i<q_i}\frac{(p_i-q_i)^2}{p_i-\frac{1}{n}(p_i-q_i)}\\
&\le \lim_{n\to\infty}\frac{1}{n^2}\sum_{i:p_i>q_i}\frac{(p_i-q_i)^2}{\left(1-\frac{1}{n}\right)(p_i-q_i)}+\lim_{n\to\infty}\frac{1}{n^2}\sum_{i:p_i<q_i}\frac{(p_i-q_i)^2}{\frac{1}{n}(q_i-p_i)}\\
&\le \lim_{n\to\infty}\frac{1}{n(n-1)}\sum_{i:p_i>q_i}|p_i-q_i|+\lim_{n\to\infty}\frac{1}{n}\sum_{i:p_i<q_i}|p_i-q_i|\\
&= 0,
\end{aligned}$$

where the first equality follows from (4). Therefore,

$$\lim_{n\to\infty}\chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{Q}}^n_{XY}) = 0.$$

On the other hand, it can be readily shown that

$$I(X;Y)+I(X;S_n|Y) = I(X;Y|S_n)+I(X;S_n).$$

For the joint probability distribution $\tilde{\mathcal{Q}}^n_{XYS_n}$, the summations in $I(X;S_n|Y)$ and $I(X;S_n)$ are bounded by

$$I(X;S_n|Y) \le H(S_n) \le \log 2,$$

and

$$I(X;S_n) \ge 0.$$

Follow from (12),

$$\Pr\{S_n=0\}I_{X;Y}(\tilde{\mathcal{Q}}^n_{XY|S_n=0}) = \Pr\{S_n=0\}I_{X;Y}(\tilde{\mathcal{P}}_{XY})$$

and

$$\Pr\{S_n=1\}I_{X;Y}(\tilde{\mathcal{Q}}^n_{XY|S_n=1}) = \Pr\{S_n=1\}I_{X;Y}(\tilde{\phi}_{XY}) = \infty.$$

Hence for the joint distribution $\tilde{\mathcal{Q}}^n_{XYS_n}$,

$$I(X;Y|S_n) = \infty.$$

Thus

$$I_{X;Y}(\tilde{\mathcal{Q}}^n_{XY}) \;\; \geq \;\; \infty + 0 - \log 2 = \infty, \tag{13}$$

for all integers n. Therefore,

$$\lim_{n\to\infty} I_{X;Y}(\tilde{\mathcal{Q}}^n_{XY}) = \infty \neq I_{X;Y}(\tilde{\mathcal{P}}_{XY}).$$

By Definition 2, the function $I_{X;Y}(\cdot)$ is discontinuous at the distribution $\tilde{\mathcal{P}}_{XY}$. ■

By letting $X = Y$ in Theorem 4, there exists a sequence of probability distributions $\mathcal{P}_n$ such that $\lim_{n\to\infty} \chi^2(\mathcal{P}_X||\mathcal{P}_n) = 0$ but $H(\mathcal{P}_n) = \infty$ for all integers n (cf. Theorem 2). Furthermore, the results in Theorem 3 and Theorem 4 can easily be extended to show that the conditional mutual information is discontinuous.

Theorem 5: For any Shannon information measure $\mathcal{H}(\mathcal{P})$ and any probability distribution $\mathcal{P}^0$ with $\mathcal{H}(\mathcal{P}^0) < \infty$, there exists a sequence of probability distributions $\mathcal{P}_n$ such that $\lim_{n\to\infty} \chi^2(\mathcal{P}^0||\mathcal{P}_n) = 0$ but $\lim_{n\to\infty} \mathcal{H}(\mathcal{P}_n) = \infty$. Thus, $\mathcal{H}(\mathcal{P})$ is discontinuous at $\mathcal{P}^0$.

Furthermore, if $\chi^2(\mathcal{P}^0||\mathcal{P}_n)$ is replaced by $V(\mathcal{P}^0, \mathcal{P}_n)$ or $D(\mathcal{P}^0||\mathcal{P}_n)$, the theorem still holds due to (5).

V. Finite Alphabets for some of the Marginal Distributions

In this section, we consider any Shannon information measure as a mapping from a set of probability distribution $\mathcal{A}$ to the set of extended real numbers. We have already shown the discontinuity of the Shannon information measures when $\mathcal{A}$ is the set of all probability distributions. In the following, we will consider different restrictions on $\mathcal{A}$. For any joint distribution $\tilde{\mathcal{P}}_{XY} = \{P_{XY}(x,y)\}$, define

$$H_{X|Y}(\tilde{\mathcal{P}}_{XY}) = \sum_{xy\in\mathcal{S}} P_{XY}(x,y) \log \frac{1}{P_{X|Y}(x|y)},$$

where $\mathcal{S} = \{xy : P_{XY}(x,y) > 0\}$.

Theorem 6: Let $\mathcal{A}$ be a set of joint distributions for random variables X and Y with countably infinite alphabet for X but finite alphabet for Y. Let $\tilde{\mathcal{P}}_{XY} \in \mathcal{A}$ with $H_{X|Y}(\tilde{\mathcal{P}}_{XY}) < \infty$. Then there exists a sequence of probability distributions $\tilde{\mathcal{Q}}^n_{XY} \in \mathcal{A}$ such that $\lim_{n\to\infty} \chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{Q}}^n_{XY}) = 0$ but $\lim_{n\to\infty} H_{X|Y}(\tilde{\mathcal{Q}}^n_{XY}) = H_{X|Y}(\tilde{\mathcal{P}}_{XY}) + c$ for any positive real number c including infinity. Thus $H_{X|Y}(\cdot)$ defined on $\mathcal{A}$ is discontinuous at $\tilde{\mathcal{P}}_{XY}$.

Proof: Suppose a joint distribution $\tilde{\mathcal{P}}_{XY} = \{P_{XY}(x,y)\}$ is given for a pair of random variables X and Y. For those y with $P_Y(y) > 0$, a sequence of conditional probability distributions $\tilde{\mathcal{Q}}^n_{X|Y=y} = \{Q^n_{X|Y=y}(x)\}$ such that $\lim_{n\to\infty} \chi^2(\tilde{\mathcal{P}}_{X|Y=y}||\tilde{\mathcal{Q}}^n_{X|Y=y}) = 0$ and $\lim_{n\to\infty} H(\tilde{\mathcal{Q}}^n_{X|Y=y}) = H(\tilde{\mathcal{P}}_{X|Y=y}) + c$ for any positive real number c including infinity can be found according to Theorem 2. Let $\tilde{\mathcal{Q}}^n_{XY} = \{Q^n_{XY}(x,y)\}$ with

$$Q^n_{XY}(x,y) = \begin{cases} Q^n_{X|Y=y}(x)P_Y(y) & \text{if } P_Y(y) > 0 \\ 0 & \text{if } P_Y(y) = 0. \end{cases}$$

Then

$$\begin{aligned} \lim_{n\to\infty} \chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{Q}}^n_{XY}) &= \lim_{n\to\infty} \sum_{xy} \frac{(P_{XY}(x,y))^2}{Q^n_{X|Y=y}(x)P_Y(y)} - 1 \\ &= \lim_{n\to\infty} \sum_{y} P_Y(y) \left(\sum_{x} \frac{\left(P_{X|Y=y}(x)\right)^2}{Q^n_{X|Y=y}(x)} - 1 \right) \\ &= \lim_{n\to\infty} \sum_{y} P_Y(y) \chi^2(\tilde{P}_{X|Y=y}||\tilde{Q}^n_{X|Y=y}) \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} \lim_{n\to\infty} H_{X|Y}(\tilde{\mathcal{Q}}^n_{XY}) &= \lim_{n\to\infty} \sum_{y} P_Y(y) H(\tilde{\mathcal{Q}}^n_{X|Y=y}) \\ &= H_{X|Y}(\tilde{\mathcal{P}}_{XY}) + c. \end{aligned}$$

Therefore, $H_{X|Y}(\cdot)$ defined on $\mathcal{A}$ is discontinuous at $\tilde{\mathcal{P}}_{XY}$. ■

So far, we have proved various cases for which the Shannon information measures are discontinuous. In the following, we will prove a case for which the Shannon information measures are continuous.

Theorem 7: Let $\mathcal{A}$ be a set of joint distributions for random variables X and Y with countably infinite alphabet for Y but alphabet with size M for X. For any $\tilde{\mathcal{P}}_{XY} \in \mathcal{A}$ and $\tilde{\mathcal{Q}}^n_{XY} \in \mathcal{A}$ such

that $\lim_{n\to\infty}\chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{Q}}^n_{XY})=0$, we have $\lim_{n\to\infty}H_{X|Y}(\tilde{\mathcal{Q}}^n_{XY})=H_{X|Y}(\tilde{\mathcal{P}}_{XY})$. Thus $H_{X|Y}(\cdot)$ defined on $\mathcal{A}$ is continuous at $\tilde{\mathcal{P}}_{XY}$ with respect to convergence in the χ^2-divergence.

Proof: Without lost of generality, we assume $P_Y(i)\geq P_Y(j)$ for $i<j$. Suppose an arbitrary $\varepsilon>0$ is given. Since $H_{X|Y}(\tilde{\mathcal{P}})$ is finite and P_Y is a probability distribution, we can find the smallest integer K such that

$$\sum_{y=K+1}^{\infty}P_Y(y)H(\tilde{\mathcal{P}}_{X|Y=y})<\frac{\varepsilon}{2}, \tag{14}$$

and

$$\sum_{y=K+1}^{\infty}P_Y(y)<\frac{\varepsilon}{4\log M}. \tag{15}$$

Hence, $P_Y(y)>0$ for $1\leq y\leq K$. In (14), we follow the convention that the summation is over all y with $P_Y(y)>0$. For any joint distributions $\tilde{\mathcal{P}}_{XY}=\{P_{XY}(x,y)\}$ and $\tilde{\mathcal{Q}}^n_{XY}=\{Q^n_{XY}(x,y)\}$ such that $\lim_{n\to\infty}\chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{Q}}^n_{XY})=0$,

$$\lim_{n\to\infty}V(\tilde{\mathcal{P}}_{XY},\tilde{\mathcal{Q}}^n_{XY})=0 \tag{16}$$

from (5), and hence

$$\lim_{n\to\infty}V(\tilde{\mathcal{P}}_{Y},\tilde{\mathcal{Q}}^n_{Y})=0. \tag{17}$$

Therefore, there exists an integer N_1 such that for $n\geq N_1$,

$$\mathcal{Q}^n_Y(y)>0 \tag{18}$$

for $1\leq y\leq k$ and

$$\left|\sum_{y=K+1}^{\infty}(P_Y(y)-Q^n_Y(y))\right|<\frac{\varepsilon}{4\log M}. \tag{19}$$

By combining (15) and (19), we have for $n\geq N_1$,

$$\sum_{y=K+1}^{\infty}Q^n_Y(y)<\frac{\varepsilon}{2\log M}. \tag{20}$$

On the other hand, follows from (16) and (18), for $1\leq y\leq K$,

$$\lim_{n\to\infty}V(\tilde{\mathcal{P}}_{X|Y=y},\tilde{\mathcal{Q}}^n_{X|Y=y})=0$$

and hence,

$$\lim_{n\to\infty} H(\tilde{\mathcal{Q}}^n_{X|Y=y}) = H(\tilde{\mathcal{P}}_{X|Y=y}) \tag{21}$$

since $\tilde{\mathcal{P}}_{X|Y=y}$ and $\tilde{\mathcal{Q}}^n_{X|Y=y}$ are in $\mathcal{A}$ and entropy is continuous for finite alphabet. Together with (17), there exists an integer $N \geq N_1$ such that for $n \geq N$,

$$\left| \sum_{y=1}^{K} P_Y(y) H(\tilde{\mathcal{P}}_{X|Y=y}) - \sum_{y=1}^{K} Q^n_Y(y) H(\tilde{\mathcal{Q}}^n_{X|Y=y}) \right| < \frac{\varepsilon}{2}. \tag{22}$$

Then for $n \geq N$,

$$\begin{aligned} |H_{X|Y}(\tilde{\mathcal{P}}_{XY}) - H_{X|Y}(\tilde{\mathcal{Q}}^n_{XY})| &\leq \left| \sum_{y=1}^{K} P_Y(y) H(\tilde{\mathcal{P}}_{X|Y=y}) - \sum_{y=1}^{K} Q^n_Y(y) H(\tilde{\mathcal{Q}}^n_{X|Y=y}) \right| + \\ &\quad \left| \sum_{y=K+1}^{\infty} P_Y(y) H(\tilde{\mathcal{P}}_{X|Y=y}) - \sum_{y=K+1}^{\infty} Q^n_Y(y) H(\tilde{\mathcal{Q}}^n_{X|Y=y}) \right| \\ &< \frac{\varepsilon}{2} + |J|, \end{aligned} \tag{23}$$

where

$$J = \sum_{y=K+1}^{\infty} P_Y(y) H(\tilde{\mathcal{P}}_{X|Y=y}) - \sum_{y=K+1}^{\infty} Q^n_Y(y) H(\tilde{\mathcal{Q}}^n_{X|Y=y}),$$

and the last inequality follows from (22). For the second term of J, we follow the convention that the summation is over all y with $Q^n_Y(y) > 0$. Due to (14),

$$J \leq \sum_{y=K+1}^{\infty} P_Y(y) H(\tilde{\mathcal{P}}_{X|Y=y}) < \frac{\varepsilon}{2}. \tag{24}$$

On the other hand, for $n \geq N$

$$\begin{aligned} J &\geq -\sum_{y=K+1}^{\infty} Q^n_Y(y) H(\tilde{\mathcal{Q}}^n_{X|Y=y}) \\ &\geq -\sum_{y=K+1}^{\infty} Q^n_Y(y) \log M \\ &> -\frac{\varepsilon}{2}, \end{aligned} \tag{25}$$

where the last inequality follows from (20). Therefore, $|J| < \frac{\varepsilon}{2}$. Together with (23), we have

$$|H_{X|Y}(\tilde{\mathcal{P}}_{XY}) - H_{X|Y}(\tilde{\mathcal{Q}}^n_{XY})| < \varepsilon$$

for $n \geq N$. Since $\varepsilon > 0$ is arbitrary, $|H_{X|Y}(\tilde{\mathcal{P}}_{XY}) - H_{X|Y}(\tilde{\mathcal{Q}}^n_{XY})| \to 0$ as $n \to \infty$. Hence, the theorem is proved. ■

Corollary 8: Let $\mathcal{A}$ be a set of joint distributions for random variables X and Y with finite alphabet for X but countably infinite alphabet for Y. For any $\tilde{\mathcal{P}}_{XY} \in \mathcal{A}$ and $\tilde{\mathcal{Q}}_{XY}^n \in \mathcal{A}$ such that $\lim_{n\to\infty} \chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{Q}}_{XY}^n) = 0$, we have $\lim_{n\to\infty} I_{X;Y}(\tilde{\mathcal{Q}}_{XY}^n) = I_{X;Y}(\tilde{\mathcal{P}}_{XY})$. Thus $I_{X;Y}(\cdot)$ defined on $\mathcal{A}$ is continuous at $\tilde{\mathcal{P}}_{XY}$ with respect to convergence in the χ^2-divergence.

Proof: Since X takes values in a finite alphabet, both $H(X)$ and $H(X|Y)$ are bounded. By observing that $I(X;Y) = H(X) - H(X|Y)$ which is the difference between two continuous and bounded functions, $I_{X;Y}(\cdot)$ defined on $\mathcal{A}$ is continuous. ■

Remark If $\lim_{n\to\infty} \chi^2(\tilde{\mathcal{P}}_{XY}||\tilde{\mathcal{Q}}_{XY}^n) = 0$ in Theorem 7 is replaced by pointwise convergence, the proof of Theorem 7 still goes through as (16) is still valid. Therefore, the continuity of mutual information in Corollary 8 can also be shown with respect to pointwise convergence. Furthermore, let $\mathcal{A}$ be a set of joint distributions for random variables X, Y and Z with finite alphabet for X but countably infinite alphabet for both Y and Z. An argument similar to the proof of Theorem 7 together with Corollary 8 can be used to show that the conditional mutual information $I(X;Y|Z)$ defined on $\mathcal{A}$ is continuous at any $\tilde{\mathcal{P}}_{XYZ} \in \mathcal{A}$.

VI. A Consequence of the Discontinuity of Entropy

The discontinuity of entropy on countably infinite alphabet can explain why certain information theoretical tools can only be applied for finite alphabet. Two important such tools are strong typicality for i.i.d. sequences and Fano's inequality to be discussed in this section. More importantly, this observation leads to some hints on how we can generalize these theoretical tools to countably infinite alphabet.

Strong typicality is more powerful than weak typicality as a tool for theorem-proving for finite alphabet memoryless systems. In fact, strong typicality is stronger than weak typicality. Specifically, for any sequence $\mathbf{x} \in \mathcal{X}^m$, where $\mathcal{X}$ is finite, if $\mathbf{x}$ is inside a strongly typical set, then $\mathbf{x}$ is also inside a weakly typical set with a suitable choice of parameters [9, p. 82]. Therefore, the strongly typical set is a subset of the weakly typical set. More importantly, with the notion of strong typicality, stronger coding results can often be proved. However, the strongly typical set is not necessarily a subset of the weakly typical set when $\mathcal{X}$ is countably infinite.

The definitions of weak typicality and strong typicality can be expressed in term of different divergence measures [3]. Specifically, let $\mathcal{P}$ be a probability distribution and let $\mathcal{Q} = \{Q_X(x)\}$

be the empirical distribution of the sequence **x**, where $Q_X(x) = m^{-1}N(x;\mathbf{x})$ and $N(x;\mathbf{x})$ is the number of occurrences of x in the sequence **x**. Then the weakly typical set is constructed by requiring the divergence measure

$$|D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})| \tag{26}$$

to be small while the strongly typical set is constructed by requiring the variational distance $V(\mathcal{Q},\mathcal{P})$ to be small. By the discontinuity of the Shannon entropy proved in Theorem 2, there exist probability distributions $\mathcal{P}$ and $\mathcal{Q}$ defined on a countably infinite alphabet such that $|H(\mathcal{Q})-H(\mathcal{P})|$ is large but $\chi^2(\mathcal{Q}||\mathcal{P})$ is small, and hence $D(\mathcal{Q}||\mathcal{P})$ and $V(\mathcal{Q},\mathcal{P})$ are also small. This implies that the value of (26) is large because

$$|D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})| \ge |D(\mathcal{Q}||\mathcal{P}) - |H(\mathcal{Q}) - H(\mathcal{P})||.$$

Let **x** be a sequence in $\mathcal{X}^m$ whose empirical distribution is $\mathcal{Q}$. Then **x** is strongly typical because $V(\mathcal{Q},\mathcal{P})$ is small, but **x** is not weakly typical because $|D(\mathcal{Q}||\mathcal{P})+H(\mathcal{Q})-H(\mathcal{P})|$ is large. Hence, strong typicality does not imply weak typicality when the alphabet is countably infinite. Instead of requiring $V(\mathcal{P},\mathcal{Q})$ to be small, if a set is defined by requiring $\chi^2(\mathcal{Q}||\mathcal{P})$ or $D(\mathcal{Q}||\mathcal{P})$ to be small, the fact that this set may not be a subset of the weakly typical set can also be seen from the above argument. So Theorem 2 shows that neither the χ^2-divergence nor the Kullback-Leibler divergence can be used to define a typicality stronger than weak typicality. Therefore, we need a new divergence measure for this purpose.

Definition 3: An asymmetric divergence measure E between probability distributions $\mathcal{P} = \{p_i\}$ and $\mathcal{Q} = \{q_i\}$ is defined by

$$E(\mathcal{Q}||\mathcal{P}) = D(\mathcal{Q}||\mathcal{P}) + |H(\mathcal{Q}) - H(\mathcal{P})|.$$

It is obvious that

$$\lim_{n\to\infty} E(\mathcal{Q}||\mathcal{P}_n) = 0 \Longleftrightarrow \lim_{n\to\infty} D(\mathcal{Q}||\mathcal{P}_n) = 0 \text{ and } \lim_{n\to\infty} H(\mathcal{P}_n) = H(\mathcal{Q}).$$

Furthermore,

$$E(\mathcal{Q}||\mathcal{P}) \ge |D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})|,$$

and

$$E(\mathcal{Q}||\mathcal{P}) \ge D(\mathcal{Q}||\mathcal{P}) \ge \frac{1}{2\ln 2} V(\mathcal{P},\mathcal{Q})$$

from Pinsker's inequality. Based on the divergence E, we have developed in [3] a unified typicality for finite or countably infinite alphabets which is stronger than both weak typicality and strong typicality while possessing asymptotic properties similar to strong typicality. The latter implies that many coding theorems proved by strong typicality for finite alphabet can readily be extended to countably infinite alphabet.

Note that E is an asymmetric divergence measure. It is interesting if a metric which has properties similar to E can be defined. In the following, a symmetric divergence measure with a simple form is presented and it will be shown to be a metric.

Definition 4: A symmetric divergence measure Ψ is defined by

$$\Psi(\mathcal{P},\mathcal{Q}) = H\left(\frac{1}{2}\mathcal{P}+\frac{1}{2}\mathcal{Q}\right) - \sqrt{H(\mathcal{P})H(\mathcal{Q})}.$$

Note that the definition of Ψ is similar to the definition of Jensen-Shannon Divergence (JSD) [16] given as

$$JSD(\mathcal{P},\mathcal{Q}) = H\left(\frac{1}{2}\mathcal{P}+\frac{1}{2}\mathcal{Q}\right) - \frac{1}{2}(H(\mathcal{P})+H(\mathcal{Q})).$$

The proofs of the following theorems are given in Appendix C.

Theorem 9: For any probability distributions $\mathcal{P}_n$ and $\mathcal{Q}$, the following two conditions are equivalent:

1) $\lim_{n\to\infty}\Psi(\mathcal{P}_n,\mathcal{Q}) = 0$.
2) $\lim_{n\to\infty} H(\mathcal{P}_n) = H(\mathcal{Q})$ and $\lim_{n\to\infty} V(\mathcal{P}_n,\mathcal{Q}) = 0$.

Theorem 10: The function $\sqrt{\Psi(\mathcal{P},\mathcal{Q})}$ is a metric.

Another well-know result restricted to finite alphabet is Fano's inequality which is crucial in proving converse coding theorems in information theory. Suppose Y is an estimate of X and both X and Y take values in the same alphabet $\mathcal{X}$. Let $P_e = \Pr\{X \neq Y\}$. Fano's inequality relates $H(X|Y)$ and P_e by

$$H(X|Y) \leq P_e \log(|\mathcal{X}|-1) + H(\{P_e, 1-P_e\}). \tag{27}$$

If $|\mathcal{X}| < \infty$, Fano's inequality says that $H(X|Y) \to 0$ as $P_e \to 0$. If $|\mathcal{X}| = \infty$, Fano's inequality becomes useless because the RHS of (27) equals infinity. In fact, $P_e \to 0$ does not imply

$H(X|Y) \to 0$, which can be shown by the discontinuity of entropy as follows. Suppose Y is a constant equal to 0 and the probability distribution of X is $\mathcal{D}_n$ in (6). Then

$$P_e = \Pr\{X \neq Y\} = \Pr\{X \neq 0\} = \frac{\alpha}{\log n},$$

which tends to zero as $n \to \infty$. However,

$$H(X|Y) = H(X,Y) - H(Y) = H(X) = H(\mathcal{D}_n),$$

which tends to a value depending on the choice of α as given in (7). Note that the distribution of X is changing while P_e is approaching to zero in the above example. Motivated by this observation, we have studied in [4] whether $H(X|Y) \to 0$ when X has a fixed distribution but Y has a varying distribution such that $P_e \to 0$. Toward this end, we have proved a generalized Fano's inequality which implies that for any X with a fixed probability distribution and finite $H(X)$, $H(X|Y) \to 0$ as $P_e \to 0$.

VII. Discontinuity of Entropy in Markov Chains

In the previous sections, we have used the probability distribution $\mathcal{D}_n$ to illustrate the discontinuity of the Shannon information measures. We will show in this section that the sequence of probability distributions $\mathcal{D}_n$ can be in fact realized by a discrete-time Markov chain with a countably infinite number of states. This means that $\mathcal{D}_n$ can possibly be observed from a physical system. Furthermore, this example shows that the entropy of a Markov chain with a countably infinite number of states may not tend to the entropy of its limiting probability distribution, so that the limiting probability distribution may not fully characterize the asymptotic behavior of a Markov chain.

Figure 1 shows the transition diagram of the following system. State 0 is an absorbing state on layer 0 and state 1 is the only state on layer 1. The state $(3, \gamma)$, for example, is referred to as the γ-th state on layer 3. The Markov chain at time 0 is in state 1 and it must make a transition to either one of the γ states on layer 2. When the system is in a state on layer l where $l > 1$, it may make a transition to state 0 with probability $1 - (\frac{l-1}{l})^\beta$ and then be absorbed, otherwise it can make a transition to any one of the γ states on layer $(l+1)$ with equal probability. It is clear that at time l, $l \geq 2$, the Markov chain is in either layer $(l+1)$ or state 0.

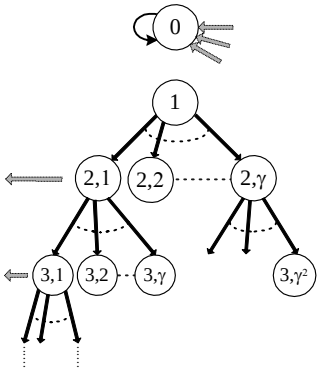

Fig. 1. A Markov chain realization of $\mathcal{D}_n$

Let X_m denote the system state at time m. Since the initial state X_0 is equal to 1, the probability distribution of X_0 is a deterministic distribution and

$$\begin{aligned}
&\Pr\{X_1 = (2,1)|X_0 = 1\} \\
&= \Pr\{X_1 = (2,2)|X_0 = 1\} \\
&= \cdots \\
&= \Pr\{X_1 = (2,\gamma)|X_0 = 1\} \\
&= \frac{1}{\gamma}.
\end{aligned}$$

For a state (l,i) belonging to layer $l > 1$, we have

$$1 \leq i \leq \gamma^{(l-1)}$$

and transition probabilities are

$$\Pr\{X_{m+1} = 0|X_m = (l,i)\} = 1 - \left(\frac{l-1}{l}\right)^{\beta} \tag{28}$$

and

$$\begin{aligned}
&\Pr\{X_{m+1} = (l+1, (i-1)\gamma + 1) | X_m = (l,i)\} \\
&= \Pr\{X_{m+1} = (l+1, (i-1)\gamma + 2) | X_m = (l,i)\} \\
&= \cdots \\
&= \Pr\{X_{m+1} = (l+1, (i-1)\gamma + \gamma) | X_m = (l,i)\} \\
&= \frac{1}{\gamma}\left(\frac{l-1}{l}\right)^{\beta}, \qquad (29)
\end{aligned}$$

for all $m \geq 1$. At time $m-1 \geq 0$, the system must be in either state 0 or on layer m. By (28),

$$\begin{aligned}
\Pr\{X_m \neq 0 | X_{m-1} \neq 0\} &= 1 - \Pr\{X_m = 0 | X_{m-1} \neq 0\} \\
&= 1 - \left(1 - \left(\frac{m-1}{m}\right)^{\beta}\right) \\
&= \left(\frac{m-1}{m}\right)^{\beta} \qquad (30)
\end{aligned}$$

for $m \geq 2$. Since $X_m \neq 0$ means that the system has never been in state 0, we have

$$\begin{aligned}
&\Pr\{X_m \neq 0\} \\
&= \Pr\{X_m \neq 0, X_{m-1} \neq 0\} \\
&= \Pr\{X_m \neq 0 | X_{m-1} \neq 0\} \Pr\{X_{m-1} \neq 0\} \\
&= \Pr\{X_m \neq 0 | X_{m-1} \neq 0\} \Pr\{X_{m-1} \neq 0 | X_{m-2} \neq 0\} \cdots \\
&\quad \Pr\{X_2 \neq 0 | X_1 \neq 0\} \Pr\{X_1 \neq 0\} \\
&= \left(\frac{m-1}{m}\right)^{\beta} \left(\frac{m-2}{m-1}\right)^{\beta} \left(\frac{m-3}{m-2}\right)^{\beta} \cdots \left(\frac{1}{2}\right)^{\beta} \cdot 1 \\
&= \left(\frac{1}{m}\right)^{\beta}, \qquad (31)
\end{aligned}$$

and

$$\Pr\{X_m = 0\} = 1 - \left(\frac{1}{m}\right)^{\beta}. \qquad (32)$$

Then

$$\lim_{m \to \infty} \Pr\{X_m = 0\} = 1,$$

which means that the limiting probability distribution exists and equals the deterministic distribution. If $X_m \neq 0$, the system must be in one of the γ^m states on layer $(m+1)$ with the same probability due to symmetry. Therefore, we can explicitly express the probability distribution of X_m. This probability distribution, however, contains a lot of zeros when m is large. Since we are only interested in the value of $H(X_m)$, we can neglect all the zeros and let

$$\mathcal{F}_m^{(\gamma,\beta)} = \left\{1 - \frac{1}{m^\beta}, \frac{1}{\gamma^m m^\beta}, ..., \frac{1}{\gamma^m m^\beta}\right\},$$

where the first element is $1 - \frac{1}{m^\beta}$ and the other γ^m elements are $\frac{1}{\gamma^m m^\beta}$. Here, the probability mass $1 - \frac{1}{m^\beta}$ denotes the probability that X_m is in state 0, and each other probability mass $\frac{1}{\gamma^m m^\beta}$ denotes the probability that X_m is in any particular state on layer $m+1$. In fact, by letting $\alpha = \log\gamma$ and $n = \gamma^m$ in (6), $\mathcal{D}_n$ is equivalent to $\mathcal{F}_m^{(\gamma,1)}$. On the other hand, by letting $n = 2^m$ in (8), $\mathcal{D}_n$ is equivalent to $\mathcal{F}_m^{(2,0.5)}$. It can readily be verified that

$$\begin{aligned}\lim_{m\to\infty} H(X_m) &= \lim_{m\to\infty} H(\mathcal{F}_m^{(\gamma,\beta)}) \\ &= \begin{cases} 0 & \beta > 1, \\ \log\gamma & \beta = 1, \\ \infty & 0 < \beta < 1. \end{cases} \end{aligned} \tag{33}$$

Therefore, we see that $\lim_{m\to\infty} H(X_m)$ and $H(\lim_{m\to\infty} X_m)$ are not equal if $0 < \beta \leq 1$. When $0 < \beta < 1$, $\Pr\{X_m = 0\}$ is increasing to 1 and $H(X_m)$ is increasing without bound. This is a mathematical proof of the seemingly counter-intuitive statement in the beginning paragraph in Section I.

The Markov chain in Fig. 1 consists of one absorbing state and an infinite number of transient states. In particular, all the states except for state 0 can be visited only once. For arbitrary state i and state j where $0 < i < j$, they may not communicate with each other. In the following, we present another example of a Markov chain in which all the states except for the absorbing state communicate with each other. This Markov chain is closely related to the gambler's ruin problem [17, P.184].

Consider the random walk as shown in Fig. 2. Let Z_m be the state of the system in Fig. 2 at time m. Here, State 0 is an absorbing state and for a state $i > 0$, we have

$$\Pr\left\{Z_{m+1} = \left\lfloor \frac{i}{2} \right\rfloor \middle| Z_m = i\right\} = \frac{1}{2}$$

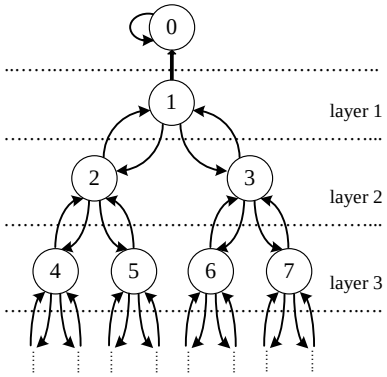

Fig. 2. A random walk on a tree structure

and

$$\Pr\{Z_{m+1} = 2i | Z_m = i\} = \Pr\{Z_{m+1} = 2i+1 | Z_m = i\} = \frac{1}{4}.$$

If we only consider the layers but not the states, the system is simply a one-dimensional symmetric random walk within the layers together with an absorbing state at the top, and is equivalent to the gambler's ruin problem. Therefore, starting on layer $l \geq 0$, the system will eventually be absorbed in state 0 with probability 1. This implies that $\lim_{m\to\infty} \Pr\{Z_m = 0\} = 1$.

By numerically computing the first $50,000$ values of $H(Z_m)$, we obtain the plot in Fig. 3. At the end of the computation, $H(Z_{50000}) \approx 1.0595$ and $\Pr\{Z_{50000} = 0\} \approx 0.9964$. It appears that as $m \to \infty$, $H(Z_m)$ does not end to 0, the entropy of the limiting distribution.

For a discrete-time Markov chain with countably infinite states, we have seen in two examples that the entropy of a Markov chain may not tend to the entropy of its limiting probability distribution. Therefore, the limiting probability distribution may not fully characterize the asymptotic behavior of a Markov chain. It is an interesting problem for future research to obtain the conditions for the limiting entropy of a Markov chain to be equal to the entropy of the limiting probability distribution of the Markov chain.

VIII. Conclusion

We have demonstrated different behaviors of the Shannon information measures on different alphabet sizes. With respect to the convergence in χ^2-divergence, all the Shannon information

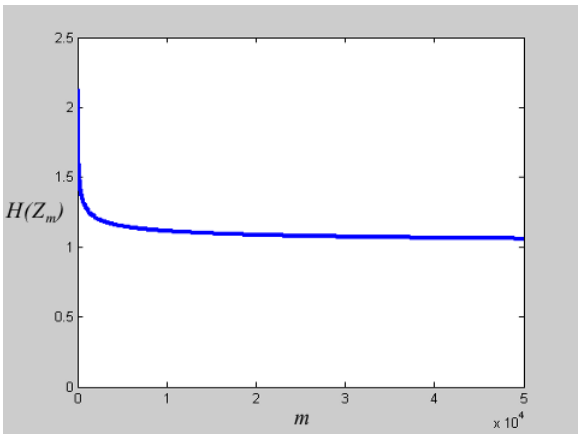

Fig. 3. Plotting of $H(Z_m)$ verses m

measures are discontinuous at every probability distribution when the alphabet is countably infinite. Also, these measures are discontinuous with respect to the convergence in many other divergence measures including variational distance and Kullback-Leibler divergence. These findings explain why the applications of certain information theoretical tools are restricted to finite alphabets. Perhaps more importantly, these discontinuity results have provided hints on how the aforementioned tools can be extended to countably infinite alphabets. In particular, we have formulated the notion of unified typicality of sequences in a related work [3], and Fano's inequality has been extended in [4]. Therefore, these information theoretical tools do not fail in the general setting due to the discontinuity of entropy, but they have to assume a more refined form. We expect that more information theoretical tools can be generalized along the same direction. Furthermore, we have demonstrated in two examples that the discontinuity of entropy can be exhibited in a discrete-time Markov chain with one of them being closely related to the gambler's ruin problem. As a whole, the results have enriched the understanding of uncertainty.

APPENDIX A

Proof of Theorem 2 We first assume c is finite. By letting $\mathcal{P}_n = \mathcal{D}_n$, the theorem is proved for $\mathcal{P}^0 = \nu$, so we assume $\mathcal{P}^0 \neq \nu$. Then without loss of generality, we can assume that $0 < p_0 < 1$. Let $q = 1 - p_0$ and

$$\mathcal{P}^1 = \{0, q^{-1}p_1, q^{-1}p_2, q^{-1}p_3, ...\}, \tag{34}$$

which is seen to be a probability distribution. Let V and W_n be random variables with probability distributions $\mathcal{P}^1$ and $\mathcal{D}_n$, respectively. Let the probability distribution of an independent binary random variable Z be such that $\Pr\{Z=0\}=q$ and $\Pr\{Z=1\}=1-q$. By letting

$$X_n = \begin{cases} V & \text{if } Z=0 \\ W_n & \text{if } Z=1, \end{cases} \tag{35}$$

the probability distribution of X_n for $n \geq 2$ is given by

$$\begin{aligned} \mathcal{P}_n^2 &= \Big\{(1-q)-\frac{(1-q)\alpha}{\log n}, p_1+\frac{(1-q)\alpha}{n\log n}, \\ &\quad p_2+\frac{(1-q)\alpha}{n\log n}, ..., p_n+\frac{(1-q)\alpha}{n\log n}, p_{n+1}, p_{n+2}, ...\Big\} \\ &= \Big\{p_0-\frac{\alpha p_0}{\log n}, p_1+\frac{\alpha p_0}{n\log n}, p_2+\frac{\alpha p_0}{n\log n}, \ldots, \\ &\quad p_n+\frac{\alpha p_0}{n\log n}, p_{n+1}, p_{n+2}, \ldots\Big\}. \end{aligned} \tag{36}$$

Then follows from (4),

$$\begin{aligned} \lim_{n\to\infty}\chi^2(\mathcal{P}^0||\mathcal{P}_n^2) &= \lim_{n\to\infty}\left(\frac{\alpha p_0}{\log n}\right)^2\left(p_0-\frac{\alpha p_0}{\log n}\right)^{-1}+\lim_{n\to\infty}\sum_{i=1}^{n}\left(\frac{\alpha p_0}{n\log n}\right)^2\left(p_i+\frac{\alpha p_0}{n\log n}\right)^{-1} \\ &= \lim_{n\to\infty}\sum_{i=1}^{n}\left(\frac{\alpha p_0}{n\log n}\right)^2\left(p_i+\frac{\alpha p_0}{n\log n}\right)^{-1} \\ &\leq \lim_{n\to\infty}\sum_{i=1}^{n}\left(\frac{\alpha p_0}{n\log n}\right)^2\left(\frac{\alpha p_0}{n\log n}\right)^{-1} \\ &= 0. \end{aligned}$$

Since $\chi^2(\mathcal{P}^0||\mathcal{P}_n^2)$ is nonnegative, we have proved that $\lim_{n\to\infty}\chi^2(\mathcal{P}^0||\mathcal{P}_n^2)=0$. Consider

$$\begin{aligned} H(\mathcal{P}_n^2) &= H(X_n) \\ &= H(X_n|Z)+I(X_n;Z) \\ &= qH(X_n|Z=0)+(1-q)H(X_n|Z=1)+I(X_n;Z) \\ &= qH(\mathcal{P}^1)+(1-q)H(\mathcal{D}_n)+I(X_n;Z). \end{aligned}$$

Note that

$$\begin{aligned} qH(\mathcal{P}^1) &= -q\sum_{i=1}^{\infty}(q^{-1}p_i)\log(q^{-1}p_i) \\ &= -\sum_{i=1}^{\infty}p_i\log q^{-1} - \sum_{i=1}^{\infty}p_i\log p_i \\ &= q\log q + H(\mathcal{P}^0) + p_0\log p_0 \\ &= H(\mathcal{P}^0) - H(\{p_0, 1-p_0\}). \end{aligned}$$

Hence,

$$\begin{aligned} H(\mathcal{P}_n^2) &= H(\mathcal{P}^0) - H(\{p_0, 1-p_0\}) + p_0 H(\mathcal{D}_n^{(\alpha,1)}) + I(X_n; Z) \\ &= H(\mathcal{P}^0) - H(Z) + p_0 H(\mathcal{D}_n) + I(X_n; Z) \\ &= H(\mathcal{P}^0) + p_0 H(\mathcal{D}_n) - H(Z|X_n). \end{aligned} \tag{37}$$

Toward finding $\lim_{n\to\infty} H(Z|X_n)$, let

$$f(X_n) = \begin{cases} 0 & \text{if } X_n = 0 \\ 1 & \text{if } X_n > 0, \end{cases}$$

and consider

$$\begin{aligned} \lim_{n\to\infty} H(Z|X_n) &= \lim_{n\to\infty} H(Z|f(X_n), X_n) \\ &\le \lim_{n\to\infty} H(Z|f(X_n)) \\ &= \lim_{n\to\infty} \left[\left(p_0 - \frac{p_0\alpha}{\log n}\right) H(Z|f(X_n)=0) \right. \\ &\qquad \left. + \left(1 - p_0 + \frac{p_0\alpha}{\log n}\right) H(Z|f(X_n)=1)\right] \\ &= \lim_{n\to\infty} \left(1 - p_0 + \frac{p_0\alpha}{\log n}\right) H(Z|f(X_n)=1) \\ &= (1-p_0) \lim_{n\to\infty} H(Z|f(X_n)=1), \end{aligned}$$

where the inequality follows from conditioning does not increase entropy and $H(Z|f(X_n) = 0) = 0$ because $f(X_n) = 0$ implies $Z = 1$. Note that

$$\begin{aligned} \Pr\{Z=0, f(X_n)=1\} &= \Pr\{f(X_n)=1|Z=0\}\Pr\{Z=0\} \\ &= 1-p_0, \end{aligned}$$

and

$$\begin{aligned}\Pr\{Z=1, f(X_n)=1\} &= \Pr\{f(X_n)=1|Z=1\}\Pr\{Z=1\} \\ &= \frac{p_0\alpha}{\log n}.\end{aligned}$$

Therefore

$$\lim_{n\to\infty} H(Z|f(X_n)=1)=0.$$

Together with

$$H(Z|X_n)\geq 0,$$

we have

$$\lim_{n\to\infty} H(Z|X_n)=0.$$

By taking $n\to\infty$ on the both sides of (37) and using (7), we have

$$\lim_{n\to\infty} H(\mathcal{P}_n^2)=H(\mathcal{P}^0)+p_0\alpha.$$

Finally, for any $c>0$, by letting

$$\alpha=\frac{c}{p_0},$$

the theorem is proved for the case that C is a positive real number. Now, consider the distribution of W is $\mathcal{D}'_n$ in (8. We can repeat the above argument and show the theorem for $C=\infty$. Therefore, the theorem is proved. ■

APPENDIX B

Proof of $I_{X;Y}(\tilde{\mathcal{Q}}_{XY}^n)$ is finite.

Note that $I(X;Z|Y)\geq 0$. Together with (11), we have

$$I(X;Y)\leq I(X;Y)+I(X;Z|Y) = qI(X;Y|Z=0)+(1-q)I(X;Y|Z=1)+I(X;Z).$$

Thus

$$\begin{aligned}I_{X;Y}(\tilde{\mathcal{Q}}_{XY}^n) &\leq qI_{X;Y}(\tilde{\mathcal{P}}'_{XY})+(1-q)H(\mathcal{D}_n)+I(X;Z) \\ &\leq qI_{X;Y}(\tilde{\mathcal{P}}'_{XY})+(1-q)\log(n+1)+\log 2. \qquad (38)\end{aligned}$$

The summation in $I_{X;Y}(\tilde{\mathcal{P}}'_{XY})$ can be broken into three parts which will be considered in the following. Consider

$$\begin{aligned}
&\sum_{x=1}^{\infty}\sum_{y=1}^{\infty} q^{-1}P_{XY}(x,y)\log\frac{q^{-1}P_{XY}(x,y)}{q^{-1}P_X(x)q^{-1}P_Y(y)} \\
&= \sum_{x=1}^{\infty}\sum_{y=1}^{\infty} q^{-1}P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} + \sum_{x=1}^{\infty}\sum_{y=1}^{\infty} q^{-1}P_{XY}(x,y)\log q \\
&\leq q^{-1}\sum_{x=1}^{\infty}\sum_{y=1}^{\infty} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}. \qquad (39)
\end{aligned}$$

On the other hand, let $P'_Y(0) = \sum_{x=1}^{\infty} q^{-1}P_{XY}(x,0)$ and it is readily checked that $P'_Y(0) \leq P_Y(0) = \sum_{x=0}^{\infty} P_{XY}(x,0)$. We temporary assume $P'_Y(0) > 0$ which will not be required at the end of this proof. We have

$$\begin{aligned}
&\sum_{x=1}^{\infty} q^{-1}P_{XY}(x,0)\log\frac{q^{-1}P_{XY}(x,0)}{q^{-1}P_X(x)P'_Y(0)} \\
&= q^{-1}\sum_{x=1}^{\infty} P_{XY}(x,0)\log\frac{P_{XY}(x,0)}{P_X(x)P_Y(0)} + q^{-1}\sum_{x=1}^{\infty} P_{XY}(x,0)\log\frac{P_Y(0)}{P'_Y(0)} \\
&\leq q^{-1}\sum_{x=1}^{\infty} P_{XY}(x,0)\log\frac{P_{XY}(x,0)}{P_X(x)P_Y(0)} + P_Y(0)\log\frac{P_Y(0)}{P'_Y(0)}. \qquad (40)
\end{aligned}$$

By a similar argument, we can see that

$$\begin{aligned}
&\sum_{y=1}^{\infty} q^{-1}P_{XY}(0,y)\log\frac{q^{-1}P_{XY}(0,y)}{P'_X(0)q^{-1}P_Y(y)} \\
&\leq q^{-1}\sum_{y=1}^{\infty} P_{XY}(0,y)\log\frac{P_{XY}(0,y)}{P_X(0)P_Y(y)} + P_X(0)\log\frac{P_X(0)}{P'_X(0)}, \qquad (41)
\end{aligned}$$

where $P'_X(0) = \sum_{y=1}^{\infty} q^{-1}P_{XY}(0,y)$ and it is readily checked that $P'_X(0) \leq P_X(0) = \sum_{y=0}^{\infty} P_{XY}(0,y)$. In the meantime, we have assumed $P'_X(0) > 0$ which will not be required at the end of this proof. By summing (39), (40) and (41) together, we get

$$\begin{aligned}
&I_{X;Y}(\tilde{\mathcal{P}}'_{XY}) \\
&\leq q^{-1}I_{X;Y}(\tilde{\mathcal{P}}^0) - q^{-1}P_{XY}(0,0)\log\frac{P_{XY}(0,0)}{P_X(0)P_Y(0)} + P_Y(0)\log\frac{P_Y(0)}{P'_Y(0)} + P_X(0)\log\frac{P_X(0)}{P'_X(0)},
\end{aligned}$$

which is finite. Together with (38), we have shown that $I_{X;Y}(\tilde{\mathcal{Q}}^n_{XY})$ is finite for all finite n.

In the above, we have assumed $P'_X(0) > 0$ and $P'_Y(0) > 0$. If $P'_Y(0) = 0$, then $P_{XY}(x,0) = 0$ for $x \geq 1$ and the L.H.S. of (40) does not appear in $I_{X;Y}(\tilde{\mathcal{P}}'_{XY})$. So the upper bound on

$I_{X;Y}(\tilde{\mathcal{P}}'_{XY})$ is given by the summation of (39) and (41) and we can still show that $I_{X;Y}(\tilde{\mathcal{P}}'_{XY})$ is finite and hence $I_{X;Y}(\tilde{\mathcal{Q}}^n_{XY})$ is finite for all finite n. The same conclusion holds if $P'_X(0) = 0$ or $P'_X(0) = P'_Y(0) = 0$. Therefore, the claim has been proved. ■

APPENDIX C

Proof of Theorem 9 Assume 2) holds. Since the proof of [16, Theorem 3] is valid for probability distributions with countably infinite support, we have

$$\lim_{n\to\infty} 2JSD(\mathcal{P}_n, \mathcal{Q}) \leq \lim_{n\to\infty} V(\mathcal{P}_n, \mathcal{Q}) = 0.$$

Note that

$$\Psi(\mathcal{P}_n, \mathcal{Q}) \;=\; JSD(\mathcal{P}_n, \mathcal{Q}) + \frac{1}{2}\left(\sqrt{H(\mathcal{P}_n)} - \sqrt{H(\mathcal{Q})}\right)^2. \tag{42}$$

Therefore, 2) implies 1).

Now, assume 1) holds. Follows from (42),

$$\lim_{n\to\infty} H(\mathcal{P}_n) = H(\mathcal{Q})$$

and

$$0 = \lim_{n\to\infty} JSD(\mathcal{P}_n, \mathcal{Q}) = \frac{1}{2}D\left(\mathcal{P}_n \middle\| \frac{1}{2}\mathcal{P}_n + \frac{1}{2}\mathcal{Q}\right) + \frac{1}{2}D\left(\mathcal{Q} \middle\| \frac{1}{2}\mathcal{P}_n + \frac{1}{2}\mathcal{Q}\right).$$

Therefore,

$$\begin{aligned}
\lim_{n\to\infty} V(\mathcal{P}_n, \mathcal{Q}) &= \lim_{n\to\infty} 2V\left(\mathcal{P}_n, \frac{1}{2}\mathcal{P}_n + \frac{1}{2}\mathcal{Q}\right) \\
&\leq \lim_{n\to\infty} 2\sqrt{(2\ln 2)D\left(\mathcal{P}_n \middle\| \frac{1}{2}\mathcal{P}_n + \frac{1}{2}\mathcal{Q}\right)} \\
&= 0,
\end{aligned}$$

where the inequality follows from Pinsker's inequality. The theorem is proved. ■

Proof of Theorem 10 It is easy to show that Ψ has the properties of a metric except for the proof of the triangle inequality which is shown here. We put

$$\xi_1 = \sqrt{JSD(\mathcal{P}, \mathcal{S})},$$

$$\xi_2 = \frac{1}{\sqrt{2}}(\sqrt{H(\mathcal{P})} - \sqrt{H(\mathcal{S})}),$$
$$\eta_1 = \sqrt{JSD(\mathcal{S},\mathcal{Q})},$$
$$\eta_2 = \frac{1}{\sqrt{2}}(\sqrt{H(\mathcal{S})} - \sqrt{H(\mathcal{Q})})$$

and

$$\xi_k = \eta_k = 0$$

for $k \geq 3$ into the Minkowski inequality

$$\left(\sum_{j=1}^{\infty} |\xi_j + \eta_j|^p\right)^{\frac{1}{p}} \leq \left(\sum_{k=1}^{\infty} |\xi_k|^p\right)^{\frac{1}{p}} + \left(\sum_{m=1}^{\infty} |\eta_m|^p\right)^{\frac{1}{p}}, \tag{43}$$

and take $p = 2$, the right hand side of (43) becomes

$$\sqrt{\Psi(\mathcal{P},\mathcal{S})} + \sqrt{\Psi(\mathcal{S},\mathcal{Q})}$$

and consideration of the L.H.S. gives

$$\begin{aligned}
&\left(\left|\sqrt{JSD(\mathcal{P},\mathcal{S})} + \sqrt{JSD(\mathcal{S},\mathcal{Q})}\right|^2\right. \\
&\left. + \left|\frac{1}{\sqrt{2}}(\sqrt{H(\mathcal{P})} - \sqrt{H(\mathcal{S})}) + \frac{1}{\sqrt{2}}(\sqrt{H(\mathcal{S})} - \sqrt{H(\mathcal{Q})})\right|^2\right)^{\frac{1}{2}} \\
= &\left(\left|\sqrt{JSD(\mathcal{P},\mathcal{S})} + \sqrt{JSD(\mathcal{S},\mathcal{Q})}\right|^2 + \frac{1}{2}\left|\sqrt{H(\mathcal{P})} - \sqrt{H(\mathcal{Q})}\right|^2\right)^{\frac{1}{2}} \\
\geq &\left(\left|\sqrt{JSD(\mathcal{P},\mathcal{Q})}\right|^2 + \frac{1}{2}\left|\sqrt{H(\mathcal{P})} - \sqrt{H(\mathcal{Q})}\right|^2\right)^{\frac{1}{2}} \\
= &\sqrt{\Psi(\mathcal{P},\mathcal{Q})}
\end{aligned}$$

where the first inequality is due to the metric properties of $\sqrt{JSD}$ [18]. Thus we have proved the triangle inequality

$$\sqrt{\Psi(\mathcal{P},\mathcal{Q})} \leq \sqrt{\Psi(\mathcal{P},\mathcal{S})} + \sqrt{\Psi(\mathcal{S},\mathcal{Q})}.$$

■

ACKNOWLEDGMENT

The author would like to thank Sergio Verdú for his valuable comments.

REFERENCES

[1] R. Rucker, *Infinity and the Mind: The Science and Philosophy of the Infinite*, Princeton University Press, 2005.

[2] J. D. Barrow, *The Infinite Book: A Short Guide to the Boundless, Timeless and Endless* , Pantheon, 2005.

[3] S.-W. Ho and R. W. Yeung, "On Information Divergence Measures and a Unified Typicality," in *Proc. 2006 IEEE Int. Symposium Inform. Theory (ISIT 2006)*, Seattle, United States, July 9-14, 2006.

[4] S.-W. Ho, "Generalizing the Fano Inequality under a Fixed Marginal Distribution," submitted to *ISIT2008*, 2008.

[5] C. E. Shannon, The Mathematical Theory of Communication, *Bell Tech. J.*, V. 27, pp.379-423, July 1948.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley-Interscience, 1991.

[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* Academic Press, New York, 1981.

[8] R. J. McEliece. *The Theory of Information and Coding*. Cambridge University Press, 2nd edition, 2002.

[9] R. W. Yeung, *A First Course in Information Theory*, Kluwer Academic/Plenum Publishers, 2002.

[10] S. Kullback and R. Leibler. "On information and sufficiency," *Ann. Math. Stat.*, 22:79-86, 1951.

[11] P. Harremoës, "Information topologies with applications," Special volume of Bolyi Series. Springer, 2004. To appear.

[12] F. Topsøe. "Basic Concepts, Identities and Inequalities – the Toolkit of Information Theory," *Entropy*, 3:162-190, Sept. 2001.

[13] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling," *Philosophy Magazine Series (5)*, 50, 157-172, 1900.

[14] S.-W. Ho, "On the Discontinuity of the Shannon Information Measures and Typical Sequences," *PhD Thesis*, The Chinese University of Hong Kong, 2006.

[15] S.-W. Ho and R. W. Yeung, "The Interplay between Entropy and Variational Distance," in *Proc. 2007 IEEE Int. Symposium Inform. Theory (ISIT 2007)*, Nice, France, June 24-29, 2007.

[16] J. Lin and S. K. M. Wong. A new directed divergence measure and its characterization. *Int. J. Gen. Syst.,* 17:73-81, 1990.

[17] S. M. Ross. *Introduction to Probability Models*. Academic Press, 6th edition, 1997.

[18] D. M. Endres and J. E. Schindelin. A New Metric for Probability Distributions. *IEEE Trans. Inform. Theory,* 49:1858-1860, July 2003.