

On Information Divergence Measures and a Unified Typicality

Siu-Wai Ho and Raymond W. Yeung

Abstract

Strong typicality, which is more powerful for theorem proving than weak typicality, can be applied to finite alphabet only, while weak typicality can be applied to both finite and countably infinite alphabets. In this paper, the relation between typicality and information divergence measures is discussed. The new definition of information divergence measure in this paper leads to the definition of a unified typicality for finite or countably infinite alphabets which is stronger than both weak typicality and strong typicality. Unified typicality retains the asymptotic equipartition property and the structural properties of strong typicality.

I. INTRODUCTION

Weak typicality was first introduced by Shannon [1] to establish the source coding theorem, while strong typicality was first used by Wolfowitz [2] and then by Berger [3]. The concept of typicality was elaborated by Wolfowitz in the book [2] which used some ideas of the method of types. Together with the others works (more history can be found in [4]), the method of types was systematically developed in [5]. Both versions of typicality are widely used in information theory and their details can be found in standard textbooks [6][7]. Strong typicality can be used only for random variables with finite alphabet but it processes stronger properties compared with weak typicality [6]. The aim of this paper is to define a unified typicality for finite or countably infinite alphabet which is stronger than both weak and strong typicalities.

The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Seattle, July, 2006.

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

In the next section, we express the definitions of weak typicality and strong typicality in terms of information divergence measures. In Section III, we introduce a unified typicality and show that it shares the same asymptotic equipartition property with both weak and strong typicalities. Finally, joint typicality is discussed in Section IV before we conclude our paper in Section V. In this paper, the base of logarithm denoted by \log is 2. The natural logarithm is denoted by \ln and the natural number to the power x is denoted by $\exp(x)$.

II. WEAK TYPICALITY AND STRONG TYPICALITY

The main observation in this section is that the definitions of weak typicality and strong typicality can be expressed in terms of entropy and information divergence measures. Consider an information source $\{X_k, k \geq 1\}$ where X_k are i.i.d. with distribution $\mathcal{P} = \{p(x)\}$ on an alphabet \mathcal{X} which can be finite or countably infinite. We use X to denote the generic random variable and $H(X)$ to denote the common entropy for all X_k , where $H(X) < \infty$. Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$. For a sequence $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$, we call $\mathcal{Q} = \{q(x; \mathbf{x})\}$ the *empirical distribution* of the sequence \mathbf{x} , where $q(x; \mathbf{x}) = n^{-1}N(x; \mathbf{x})$ and $N(x; \mathbf{x})$ is the number of occurrences of x in the sequence \mathbf{x} . The empirical distribution of the sequence \mathbf{x} is also called the *type* of \mathbf{x} [5]. Then the probability of observing a sequence \mathbf{x} from the source $\{X_k\}$ is

$$p(\mathbf{x}) = \prod_{x \in \mathcal{X}} p(x)^{N(x; \mathbf{x})} = \prod_{x \in \mathcal{X}} p(x)^{nq(x; \mathbf{x})},$$

so that the *empirical entropy* can be written as

$$\begin{aligned} -\frac{1}{n} \log p(\mathbf{x}) &= -\frac{1}{n} \sum_x nq(x; \mathbf{x}) \log p(x) \\ &= \sum_x q(x; \mathbf{x}) \log \frac{q(x; \mathbf{x})}{p(x)} - \sum_x q(x; \mathbf{x}) \log q(x; \mathbf{x}) \\ &= D(\mathcal{Q} \parallel \mathcal{P}) + H(\mathcal{Q}), \end{aligned} \tag{1}$$

where $D(\mathcal{Q} \parallel \mathcal{P})$ is the Kullback-Leibler distance between the empirical distribution of the sequence \mathbf{x} and the probability distribution of X . Thus the definition of weak typicality [6][7] can be rewritten as follows.

Definition 1 (Weak typicality): For any $\epsilon > 0$, the weakly typical set $W_{[X]\epsilon}^n$ with respect to $p(x)$ is the set of sequences $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ such that

$$|D(\mathcal{Q} \parallel \mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})| \leq \epsilon. \tag{2}$$

Strong typicality has been defined in slightly different forms in [3][5][6], but these definitions are essentially the same when the alphabet is finite. Here we adopt the definition in [6] which is the simplest and also the most convenient for our discussion. By using the same notation except that \mathcal{X} is assumed to be finite, the definition of strong typicality in [6] can be rewritten as follows.

Definition 2 (Strong typicality): For any $\delta > 0$, the strongly typical set $T_{[X]\delta}^n$ with respect to $p(x)$ is the set of sequences $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ such that $q(x; \mathbf{x}) = 0$ for $p(x) = 0$ and

$$V(\mathcal{Q}, \mathcal{P}) \leq \delta, \quad (3)$$

where

$$V(\mathcal{Q}, \mathcal{P}) = \sum_x |q(x; \mathbf{x}) - p(x)|$$

is the variational distance between the empirical distribution of the sequence \mathbf{x} and the probability distribution of X .

Weak typicality has significant implications due to the *weak Asymptotic Equipartition Property* (weak AEP) [6][7].

Theorem 1 (Weak AEP): For any $\epsilon > 0$:

1) If $\mathbf{x} \in W_{[X]\epsilon}^n$, then

$$2^{-n(H(X)+\epsilon)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)}.$$

2) For sufficiently large n ,

$$\Pr\{\mathbf{X} \in W_{[X]\epsilon}^n\} > 1 - \epsilon.$$

3) For sufficiently large n ,

$$(1 - \epsilon)2^{n(H(X)-\epsilon)} \leq |W_{[X]\epsilon}^n| \leq 2^{n(H(X)+\epsilon)}.$$

Strong typicality applying to finite alphabet shares similar properties with weak typicality, namely the *strong AEP* [6], which will not be repeated here. Although strong typicality applying to countably infinite alphabet does not have properties similar to Property 1 and Property 3 in Theorem 1, a property similar to Property 2 still holds which is shown in the following lemma.

Lemma 2: Let $T_{[X]\delta}^n$ be defined according to Definition 2 except that \mathcal{X} is possibly countably infinite. Then for any $\delta > 0$ and sufficiently large n ,

$$\Pr\{\mathbf{X} \in T_{[X]\delta}^n\} > 1 - (2^M - 2) \exp\left(\frac{-n\delta^2}{18}\right),$$

where M is an integer such that

$$\sum_{i=M}^{\infty} p(x) \leq \frac{\delta}{3}.$$

This lemma says that the variational distance between the true distribution and the empirical distribution converges to zero. This is in some sense a generalization of [8, eq. (8)] which says that if the true distribution \mathcal{P} has a finite number of probability masses, say L , then

$$\Pr\{V(\mathcal{P}, \mathcal{Q}) \leq \delta\} \geq 1 - (2^L - 2) \exp\left(-\frac{n\delta^2}{2}\right). \quad (4)$$

This result will be used in the following proof.

Proof: Let M be an integer such that

$$\sum_{i=M}^{\infty} p(x) \leq \frac{\delta}{3}. \quad (5)$$

Let

$$\mathcal{P}' = \left\{ p(1), p(2), \dots, p(M-1), \sum_{x=M}^{\infty} p(x) \right\}$$

and

$$\mathcal{Q}' = \left\{ q(1; \mathbf{x}), q(2; \mathbf{x}), \dots, q(M-1; \mathbf{x}), \sum_{i=M}^{\infty} q(i; \mathbf{x}) \right\},$$

where \mathcal{P}' and \mathcal{Q}' both have M probability masses. Assume $V(\mathcal{P}', \mathcal{Q}') \leq \frac{\delta}{3}$, i.e.

$$V(\mathcal{P}', \mathcal{Q}') = \sum_{x=1}^{M-1} |p(x) - q(x; \mathbf{x})| + \left| \sum_{x=M}^{\infty} p(x) - \sum_{x=M}^{\infty} q(x; \mathbf{x}) \right| \leq \frac{\delta}{3}. \quad (6)$$

Let

$$\gamma_1 = \sum_{x=1}^{M-1} |p(x) - q(x; \mathbf{x})| \quad (7)$$

and

$$\gamma_2 = \left| \sum_{x=M}^{\infty} p(x) - \sum_{x=M}^{\infty} q(x; \mathbf{x}) \right|, \quad (8)$$

so that $\gamma_1 + \gamma_2 \leq \frac{\delta}{3}$. Consider

$$\begin{aligned}
\sum_{x=M}^{\infty} |p(x) - q(x; \mathbf{x})| &\leq \sum_{x=M}^{\infty} q(x; \mathbf{x}) + \sum_{x=M}^{\infty} p(x) \\
&= \sum_{x=M}^{\infty} q(x; \mathbf{x}) - \sum_{x=M}^{\infty} p(x) + 2 \sum_{x=M}^{\infty} p(x) \\
&\leq \sum_{x=M}^{\infty} q(x; \mathbf{x}) - \sum_{x=M}^{\infty} p(x) + \frac{2\delta}{3} \\
&\leq \gamma_2 + \frac{2\delta}{3},
\end{aligned}$$

where the second inequality follows from (5) and the last inequality follows from (8). Then by (7), we get

$$\sum_{x=1}^{\infty} |p(x) - q(x; \mathbf{x})| \leq \gamma_1 + \gamma_2 + \frac{2\delta}{3} \leq \delta. \quad (9)$$

Since (6) implies (9), we have

$$\begin{aligned}
\Pr \left\{ \sum_{x=1}^{\infty} |p(x) - q(x; \mathbf{x})| \leq \delta \right\} &\geq \Pr \left\{ V(\mathcal{P}', \mathcal{Q}') \leq \frac{\delta}{3} \right\} \\
&\geq 1 - (2^M - 2) \exp \left(\frac{-n\delta^2}{18} \right), \quad (10)
\end{aligned}$$

where the last inequality follows from (4). Therefore,

$$\Pr \left\{ \sum_{x=1}^{\infty} |p(x) - q(x; \mathbf{x})| < \delta \right\} > 1 - \delta$$

for sufficiently large n . ■

Strong typicality is more powerful than weak typicality as a tool for theorem proving for memoryless problems. For finite alphabet, strong typicality is in fact stronger than weak typicality. Specifically, for any $\mathbf{x} \in \mathcal{X}^n$, if $\mathbf{x} \in T_{[X]\delta}^n$, then $\mathbf{x} \in W_{[X]\epsilon}^n$ where $\epsilon = -\delta \log(\min_x p(x))$ [6, p. 82]. However, this is not true when \mathcal{X} is countably infinite. By the discontinuity of the Shannon entropy [9], there exist probability distributions \mathcal{P} and \mathcal{Q} defined on a countably infinite alphabet such that both $D(\mathcal{Q}||\mathcal{P})$ and $V(\mathcal{Q}, \mathcal{P})$ are small but $|H(\mathcal{Q}) - H(\mathcal{P})|$ is large. This means that \mathcal{P} and \mathcal{Q} satisfy the condition in (3) but not the condition in (2) since

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

In short, strong typicality does not imply weak typicality when the alphabet is countably infinite.

III. UNIFIED TYPICALITY

We have seen in the last section that weak typicality and strong typicality can be defined in terms of properly chosen information divergence measures. In this section, we introduce a new information divergence measure and discuss the properties of the typicality it induces. We will show that this new typicality unifies both weak typicality and strong typicality.

We again consider an information source $\{X_k, k \geq 1\}$ where X_k are i.i.d. with distribution $\mathcal{P} = \{p(x)\}$ on an alphabet set \mathcal{X} which can be finite or countably infinite, with $H(\mathcal{P}) < \infty$.

Definition 3 (Unified typicality): For any $\eta > 0$, the unified typical set $U_{[X]\eta}^n$ with respect to $p(x)$ is the set of sequences $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ such that

$$D(\mathcal{Q}||\mathcal{P}) + |H(\mathcal{Q}) - H(\mathcal{P})| \leq \eta. \quad (11)$$

Unified typicality shares a similar AEP with weak and strong typicalities to be proved in the following theorem. The proof can also illustrate the relationship among weak typicality, strong typicality and unified typicality.

Theorem 3 (Unified AEP): For any $\eta > 0$:

1) If $\mathbf{x} \in U_{[X]\eta}^n$, then

$$2^{-n(H(X)+\eta)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}.$$

2) For sufficiently large n ,

$$\Pr\{\mathbf{X} \in U_{[X]\eta}^n\} > 1 - \eta.$$

3) For sufficiently large n ,

$$(1 - \eta)2^{n(H(X)-\eta)} \leq |U_{[X]\eta}^n| \leq 2^{n(H(X)+\eta)}.$$

The following Lemma 4 [10, Theorem 8] will be used in the proof.

Definition 4: For an unnormalized distribution $\tilde{\mathcal{P}} = (\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_L)$ which can be normalized by a positive constant $\alpha \leq 1$ so that $(\alpha^{-1}\tilde{p}_1, \alpha^{-1}\tilde{p}_2, \dots, \alpha^{-1}\tilde{p}_L)$ is a probability distribution with L probability masses, let

$$H(\tilde{\mathcal{P}}) = - \sum_{i=1}^L \tilde{p}_i \log \tilde{p}_i.$$

Lemma 4: Let $\tilde{\mathcal{P}} = (\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_L)$ and $\tilde{\mathcal{Q}} = (\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_M)$ be two unnormalized distributions which can be normalized by two positive constants $\alpha \leq 1$ and $\beta \leq 1$ so that $(\alpha^{-1}\tilde{p}_1, \alpha^{-1}\tilde{p}_2, \dots, \alpha^{-1}\tilde{p}_L)$ and $(\beta^{-1}\tilde{q}_1, \beta^{-1}\tilde{q}_2, \dots, \beta^{-1}\tilde{q}_M)$ with $M \geq L$ are two probability distributions. If

$$V(\tilde{\mathcal{P}}, \tilde{\mathcal{Q}}) \leq \epsilon,$$

then

$$|H(\tilde{\mathcal{Q}}) - H(\tilde{\mathcal{P}})| \leq \begin{cases} -\epsilon \log \epsilon + \epsilon \log M & \epsilon < 1 \\ \log M & \epsilon \geq 1. \end{cases}$$

Proof of Theorem 3: To prove Property 1, we have for any $\mathbf{x} \in U_{[X]\eta}^n$,

$$\begin{aligned} \eta &\geq D(\mathcal{Q}||\mathcal{P}) + |H(\mathcal{Q}) - H(\mathcal{P})| \\ &\geq |D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})|. \end{aligned}$$

Thus $\mathbf{x} \in W_{[X]\eta}^n$. By Property 1 in Theorem 1,

$$2^{-n(H(X)+\eta)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}.$$

This proves Property 1.

To prove Property 2, assume that a random sequence \mathbf{X} is generated from the information source $\{X_k, k \geq 1\}$. Fix $\eta > 0$ and let

$$\epsilon = \frac{\eta}{3}. \quad (12)$$

For any probability distribution $\mathcal{P} = \{p(x)\}$ such that $H(\mathcal{P}) < \infty$, we can find an integer N such that

$$-\sum_{x=N+1}^{\infty} p(x) \log p(x) \leq \frac{\epsilon}{2}. \quad (13)$$

Let $0 < \delta < \min\{\epsilon, \frac{1}{2}\}$ be a positive real number satisfying

$$-\delta \log \delta + \delta \log N \leq \frac{\epsilon}{2}. \quad (14)$$

Such a δ exists because the L.H.S. of (14) tends to 0 as $\delta \rightarrow 0$.

We now show that

$$W_{[X]\epsilon}^n \cap T_{[X]\delta}^n \subset U_{[X]\eta}^n. \quad (15)$$

Consider any $\mathbf{x} \in W_{[X]\epsilon}^n \cap T_{[X]\delta}^n$. Then \mathbf{x} satisfies

$$|D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})| \leq \epsilon \quad (16)$$

and

$$V(\mathcal{Q}, \mathcal{P}) \leq \delta. \quad (17)$$

By (16),

$$\epsilon \geq D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P}) \geq H(\mathcal{Q}) - H(\mathcal{P}), \quad (18)$$

and by (17),

$$\delta \geq V(\mathcal{Q}, \mathcal{P}) = \sum_{x=1}^{\infty} |q(x) - p(x)| \geq \sum_{x=1}^N |q(x) - p(x)|. \quad (19)$$

If the two finite distributions

$$\mathcal{P}' = \{p(1), p(2), \dots, p(N)\}$$

and

$$\mathcal{Q}' = \{q(1), q(2), \dots, q(N)\}$$

were normalized, then we have $|H(\mathcal{Q}') - H(\mathcal{P}')| \rightarrow 0$ as $\delta \rightarrow 0$ by (19) and the continuity of the entropy function for finite alphabet. Although here \mathcal{P}' and \mathcal{Q}' are not normalized, Lemma 4 shows that $|H(\mathcal{Q}') - H(\mathcal{P}')|$ is upper bounded by the L.H.S. of (14). It then follows from (14) that

$$-\frac{\epsilon}{2} \leq -\sum_{x=1}^N q(x) \log q(x) + \sum_{x=1}^N p(x) \log p(x) \leq \frac{\epsilon}{2}.$$

So

$$\begin{aligned} H(\mathcal{Q}) &\geq -\sum_{x=1}^N q(x) \log q(x) \\ &\geq -\sum_{x=1}^N p(x) \log p(x) - \frac{\epsilon}{2} \\ &= H(\mathcal{P}) + \sum_{x=N+1}^{\infty} p(x) \log p(x) - \frac{\epsilon}{2} \\ &\geq H(\mathcal{P}) - \epsilon, \end{aligned}$$

where the last inequality follows from (13). Together with (18), we obtain

$$|H(\mathcal{Q}) - H(\mathcal{P})| \leq \epsilon. \quad (20)$$

An upper bound on $D(\mathcal{Q}||\mathcal{P})$ can also be obtained as

$$\begin{aligned} D(\mathcal{Q}||\mathcal{P}) &\leq D(\mathcal{Q}||\mathcal{P}) - |H(\mathcal{Q}) - H(\mathcal{P})| + \epsilon \\ &\leq |D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})| + \epsilon \\ &\leq 2\epsilon, \end{aligned}$$

where the last inequality follows from (16). Together with (20), this gives

$$D(\mathcal{Q}||\mathcal{P}) + |H(\mathcal{Q}) - H(\mathcal{P})| \leq 3\epsilon = \eta \quad (21)$$

(cf. (12)), i.e., $\mathbf{x} \in U_{[X]\eta}^n$, proving (15). It then follows that

$$\begin{aligned} &\Pr\{\mathbf{X} \in U_{[X]\eta}^n\} \\ &\geq \Pr\{\mathbf{X} \in W_{[X]\epsilon}^n \cap T_{[X]\delta}^n\} \\ &= 1 - \Pr\{\mathbf{X} \in (W_{[X]\epsilon}^n)^c \cup (T_{[X]\delta}^n)^c\} \\ &\geq 1 - (\Pr\{\mathbf{X} \in (W_{[X]\epsilon}^n)^c\} + \Pr\{\mathbf{X} \in (T_{[X]\delta}^n)^c\}) \\ &= \Pr\{\mathbf{X} \in W_{[X]\epsilon}^n\} + \Pr\{\mathbf{X} \in T_{[X]\delta}^n\} - 1. \end{aligned} \quad (22)$$

From Property 2 in Theorem 1 and Lemma 2, we have

$$\Pr\{\mathbf{X} \in (W_{[X]\epsilon}^n)\} > 1 - \epsilon,$$

and

$$\Pr\{\mathbf{X} \in (T_{[X]\delta}^n)\} > 1 - \delta > 1 - \epsilon,$$

for sufficiently large n , where $\delta \leq \epsilon$ can be seen from the choice of δ in (14). By using $\epsilon = \frac{\eta}{3}$ from (12), we obtain

$$\Pr\{\mathbf{X} \in U_{[X]\eta}^n\} > 1 - 2\epsilon > 1 - \eta,$$

proving Property 2.

To prove Property 3, we use the lower bound on $p(\mathbf{x})$ for $\mathbf{x} \in U_{[X]\eta}^n$ obtained in Property 1 and consider

$$|U_{[X]\eta}^n| 2^{-n(H(X)+\eta)} \leq \Pr\{\mathbf{X} \in U_{[X]\eta}^n\} \leq 1,$$

which implies

$$|U_{[X]\eta}^n| \leq 2^{n(H(X)+\eta)}.$$

On the other hand, by using the upper bound on $p(\mathbf{x})$ for $\mathbf{x} \in U_{[X]\eta}^n$ obtained in Property 1 and the lower bound on $\Pr\{\mathbf{X} \in U_{[X]\eta}^n\}$ obtained in Property 2 for sufficiently large n , we have

$$|U_{[X]\eta}^n| 2^{-n(H(X)-\eta)} \geq \Pr\{\mathbf{X} \in U_{[X]\eta}^n\} \geq 1 - \eta,$$

which implies

$$|U_{[X]\eta}^n| \geq (1 - \eta) 2^{n(H(X)-\eta)}.$$

Thus

$$(1 - \eta) 2^{n(H(X)-\eta)} \leq |U_{[X]\eta}^n| \leq 2^{n(H(X)+\eta)}.$$

The theorem is proved. ■

Remark In the above proof, for a countably infinite alphabet \mathcal{X} and any $\eta > 0$, we have shown that for sufficiently large n , i) $|H(\mathcal{Q}) - H(\mathcal{P})| < \eta$, i.e., the *entropy of the empirical distribution* of the sequence \mathbf{x} is close to the true entropy $H(X)$ [11][12] and ii) $D(\mathcal{Q}||\mathcal{P}) < \eta$, with very high probability. Note that the entropy of the empirical distribution, $H(\mathcal{Q})$, is different from the empirical entropy in (1). Furthermore, the bound $|U_{[X]\eta}^n| \leq 2^{n(H(X)+\eta)}$ in Point 3 is always true even for small n .

This result is closely related to the work of Kullback-Leibler distance estimation. The application of our results on Kullback-Leibler distance estimation will be shown after the following theorem. This theorem, giving a bound on the probability of obtaining a non-typical sequence, enhances Property 2 of the Unified AEP.

Theorem 5: For any probability distribution $\{p(x)\}$ with finite entropy and finite $E[-\log p(X)^2]$, we have

$$\Pr\{\mathbf{X} \in (U_{[X]\eta}^n)^c\} = O(n^{-1}).$$

Proof: From Chebyshev's inequality, we have

$$\begin{aligned} \Pr\{\mathbf{X} \in (W_{[X]\epsilon}^n)^c\} &= \Pr\left\{\left|-\frac{1}{n} \sum_{k=1}^n \log p(X_k) - H(X)\right| \geq \epsilon\right\} \\ &\leq \frac{\sigma^2}{n\epsilon^2}, \end{aligned}$$

where

$$\sigma^2 = \sum_x p(X=x)(-\log p(X=x))^2 - (H(X))^2.$$

is the the entropy “variance”. On the other hand, we have

$$\begin{aligned} \Pr\{\mathbf{X} \in (T_{[X]\delta}^n)^c\} &= 1 - \Pr\{\mathbf{X} \in T_{[X]\delta}^n\} \\ &< (2^M - 2) \exp\left(\frac{-n\delta^2}{18}\right), \end{aligned}$$

from (10) where M depends on $\{p(x)\}$ and δ . By (22), we have

$$\begin{aligned} \Pr\{\mathbf{X} \in (U_{[X]\eta}^n)^c\} &= 1 - \Pr\{\mathbf{X} \in U_{[X]\eta}^n\} \\ &\leq \Pr\{\mathbf{X} \in (W_{[X]\epsilon}^n)^c\} + \Pr\{\mathbf{X} \in (T_{[X]\delta}^n)^c\} \\ &< \frac{\sigma^2}{n\epsilon^2} + (2^M - 2) \exp\left(\frac{-n\delta^2}{18}\right) \\ &= O(n^{-1}). \end{aligned}$$

Hence we have proved the theorem. ■

We have seen that the Kullback-Leibler distance between the empirical distribution of the sequence, \mathcal{Q} , and the true distribution, \mathcal{P} , is tending to zero with a linear rate in the last theorem, i.e.,

$$\Pr\{D(\mathcal{Q}||\mathcal{P}) > \eta\} = O(n^{-1}).$$

This result gives a new understanding on the Kullback-Leibler distance estimation on countably infinite alphabet.

The fact that strong typicality is stronger than weak typicality for finite alphabet has been discussed in Section II. In the following theorem, we will prove that unified typicality is stronger than both weak typicality and strong typicality on finite or countably infinite alphabets.

Theorem 6: For any $\mathbf{x} \in \mathcal{X}^n$, if $\mathbf{x} \in U_{[X]\eta}^n$, then $\mathbf{x} \in W_{[X]\eta}^n$ and $\mathbf{x} \in T_{[X]\delta}^n$, where $\delta = \sqrt{\eta \cdot 2 \ln 2}$.

Proof: For any $\mathbf{x} \in U_{[X]\eta}^n$, we have proved that $\mathbf{x} \in W_{[X]\eta}^n$ in the proof of Property 1 in Theorem 3. Moreover,

$$\begin{aligned} \eta &\geq D(\mathcal{Q}||\mathcal{P}) + |H(\mathcal{Q}) - H(\mathcal{P})| \\ &\geq D(\mathcal{Q}||\mathcal{P}) \\ &\geq \frac{1}{2 \ln 2} V(\mathcal{Q}, \mathcal{P})^2 \end{aligned}$$

by Pinsker's inequality (see e.g. [6]). Thus

$$V(\mathcal{Q}, \mathcal{P}) \leq \sqrt{\eta \cdot 2 \ln 2}.$$

At the same time, $q(x; \mathbf{x}) = 0$ for those x such that $p(x) = 0$ because $D(\mathcal{Q}||\mathcal{P})$ is bounded. Therefore $\mathbf{x} \in T_{[X]\delta}^n$, where $\delta = \sqrt{\eta \cdot 2 \ln 2}$. ■

Remark It can be seen from the above proof that the definition of strong typicality in Definition 2 can be strengthened by replacing the variational distance by the Kullback-Leibler distance, while preserving the AEP.

IV. UNIFIED JOINT TYPICALITY

In this section, we discuss unified joint typicality with respect to a bivariate distribution. Generalization to a multivariate distribution is straightforward. Consider a bivariate information source $\{(X_k, Y_k), k \geq 1\}$ where (X_k, Y_k) are i.i.d. with distribution $\mathcal{P} = p(xy)$. We use (X, Y) to denote the pair of generic random variables.

Definition 5: The unified jointly typical set $U_{[XY]\eta}^n$ with respect to $p(xy)$ is the set of sequences $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that

$$\begin{aligned} & D(\mathcal{Q}||\mathcal{P}) + |H_{XY}(\mathcal{Q}) - H_{XY}(\mathcal{P})| + \\ & |H_X(\mathcal{Q}) - H_X(\mathcal{P})| + |H_Y(\mathcal{Q}) - H_Y(\mathcal{P})| \leq \eta, \end{aligned} \quad (23)$$

where $H_{XY}(\mathcal{P})$, $H_X(\mathcal{P})$, and $H_Y(\mathcal{P})$ denote the entropies of the joint distribution $p(xy)$ and the marginal distributions $p(x)$ and $p(y)$, respectively, while $H_{XY}(\mathcal{Q})$, $H_X(\mathcal{Q})$, and $H_Y(\mathcal{Q})$ are the corresponding entropies of the empirical distributions of the pair of sequence (\mathbf{x}, \mathbf{y}) , i.e., $\mathcal{Q} = \{q(x, y; \mathbf{x}, \mathbf{y})\}$.

Unified typicality preserves the consistency property [6] of strong typicality as below.

Theorem 7 (Consistency): If $(\mathbf{x}, \mathbf{y}) \in U_{[XY]\eta}^n$, then $\mathbf{x} \in U_{[X]\eta}^n$ and $\mathbf{y} \in U_{[Y]\eta}^n$.

Proof: By the log-sum inequality (see e.g. [6]), we have

$$\begin{aligned} \sum_{xy} q(xy) \log \frac{q(xy)}{p(xy)} & \geq \sum_x \left(\sum_y q(xy) \right) \log \frac{\sum_y q(xy)}{\sum_y p(xy)} \\ & \geq \sum_x q(x) \log \frac{q(x)}{p(x)}. \end{aligned} \quad (24)$$

Therefore,

$$\begin{aligned}
\eta &\geq D(\mathcal{Q}||\mathcal{P}) + |H_{XY}(\mathcal{Q}) - H_{XY}(\mathcal{P})| + \\
&\quad |H_X(\mathcal{Q}) - H_X(\mathcal{P})| + |H_Y(\mathcal{Q}) - H_Y(\mathcal{P})| \\
&\geq D(\mathcal{Q}||\mathcal{P}) + |H_X(\mathcal{Q}) - H_X(\mathcal{P})| \\
&\geq D_X(\mathcal{Q}||\mathcal{P}) + |H_X(\mathcal{Q}) - H_X(\mathcal{P})|,
\end{aligned}$$

where $D_X(\mathcal{Q}||\mathcal{P})$ denotes the R.H.S. of (24). Therefore $\mathbf{x} \in U_{[X]_\eta}^n$. By symmetry, it is readily seen that $\mathbf{y} \in U_{[Y]_\eta}^n$. ■

The unified joint asymptotic equipartition property (Unified JAEP) is shown in the following theorem.

Theorem 8 (Unified JAEP): Let

$$(\mathbf{X}, \mathbf{Y}) = ((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)),$$

where (X_i, Y_i) are i.i.d. with generic pair of random variables (X, Y) . The following hold for any $\eta > 0$.

1) If $(\mathbf{x}, \mathbf{y}) \in U_{[XY]_\eta}^n$, then

$$2^{-n(H(X,Y)+\eta)} \leq p(\mathbf{x}, \mathbf{y}) \leq 2^{-n(H(X,Y)-\eta)}.$$

2) For sufficiently large n ,

$$\Pr\{(\mathbf{X}, \mathbf{Y}) \in U_{[XY]_\eta}^n\} > 1 - \eta.$$

3) For sufficiently large n ,

$$(1 - \eta)2^{n(H(X,Y)-\eta)} \leq |U_{[XY]_\eta}^n| \leq 2^{n(H(X,Y)+\eta)}.$$

Proof: We will first prove Property 2 by letting $\delta = \frac{\eta}{3}$. By an argument similar to the proof of Property 2 of Theorem 3, we can prove that

$$D(\mathcal{Q}||\mathcal{P}) + |H_{XY}(\mathcal{Q}) - H_{XY}(\mathcal{P})| \leq \delta \tag{25}$$

is true with probability greater than $1 - \delta$ for sufficiently large n . By applying Property 2 of Theorem 3 to the information source $\{X_k, k \geq 1\}$, we have

$$D_X(\mathcal{Q}||\mathcal{P}) + |H_X(\mathcal{Q}) - H_X(\mathcal{P})| \leq \delta \tag{26}$$

is true with probability greater than $1 - \delta$ for sufficiently large n . Since (26) implies

$$|H_X(\mathcal{Q}) - H_X(\mathcal{P})| \leq \delta, \quad (27)$$

(27) is true with probability greater than $1 - \delta$. Similarly for the information source $\{Y_k, k \geq 1\}$, we have

$$|H_Y(\mathcal{Q}) - H_Y(\mathcal{P})| \leq \delta, \quad (28)$$

which is true with probability greater than $1 - \delta$ for sufficiently large n . Note that if (25), (27) and (28) are true, then (23) is true because $\delta = \frac{\eta}{3}$. By the union bound, we have

$$\Pr\{(\mathbf{X}, \mathbf{Y}) \in U_{[XY]^\eta}^n\} > 1 - 3\delta = 1 - \eta,$$

for sufficiently large n , proving Property 2.

Finally, the proofs of Property 1 and Property 3 follow the same arguments as in Theorem 3, so they are omitted. ■

It is natural to suspect that the condition in (23) can be simplified by omitting one of the terms on the L.H.S. However, this is not possible due to the discontinuity of entropy. This will be illustrated by the following probability distribution. For a fixed real number γ and an integer n , where $\gamma > 0$ and $n > 2^\gamma$, let \mathcal{D}_n^γ be a probability distribution such that one of the elements is $1 - \frac{\gamma}{\log n}$, n of them are $\frac{\gamma}{n \log n}$ and the rest are all 0, i.e.,

$$\mathcal{D}_n^\gamma = \left\{ 1 - \frac{\gamma}{\log n}, \frac{\gamma}{n \log n}, \frac{\gamma}{n \log n}, \dots, 0, 0, \dots \right\}. \quad (29)$$

The above distribution is a special case of the distribution $\mathcal{D}_n^{\alpha, \beta}$ in [9] with $\log \alpha = \gamma$ and $\beta = 1$. Then it can readily be checked (see (3) in [9]) that

$$\lim_{n \rightarrow \infty} H(\mathcal{D}_n^\gamma) = \gamma. \quad (30)$$

Moreover, for any $\alpha > 0$ and $\beta > 0$, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} D(\mathcal{D}_n^\alpha || \mathcal{D}_n^\beta) &= \lim_{n \rightarrow \infty} \left(1 - \frac{\alpha}{\log n} \right) \log \frac{1 - \frac{\alpha}{\log n}}{1 - \frac{\beta}{\log n}} + \\ &\quad \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{\alpha}{n \log n} \log \frac{\frac{\alpha}{n \log n}}{\frac{\beta}{n \log n}} \\ &= 0 + \lim_{n \rightarrow \infty} n \cdot \frac{\alpha}{n \log n} \log \frac{\alpha}{\beta} \\ &= 0. \end{aligned} \quad (31)$$

Thus we can find an integer m such that $D(\mathcal{D}_m^1||\mathcal{D}_m^2)$, $D(\mathcal{D}_m^3||\mathcal{D}_m^2)$, $|H(\mathcal{D}_m^1) - 1|$, $|H(\mathcal{D}_m^2) - 2|$ and $|H(\mathcal{D}_m^3) - 3|$ are all less than ϵ . Let the distributions of independent random variables Φ_Q^X , Φ_Q^C , Φ_Q^Y , Φ_P^X , Φ_P^C , and Φ_P^Y be \mathcal{D}_m^1 , \mathcal{D}_m^3 , \mathcal{D}_m^1 , \mathcal{D}_m^2 , \mathcal{D}_m^2 and \mathcal{D}_m^2 respectively. Now, we construct the probability distributions $\{q(xy)\}$ and $\{p(xy)\}$ as prescribed by Fig. 1(a). Specifically, the probability distribution $\{q(xy)\}$ is defined by letting $X = (\Phi_Q^X, \Phi_Q^C)$ and $Y = (\Phi_Q^Y, \Phi_Q^C)$. On the other hand, the distribution of $\{p(xy)\}$ is defined by letting $X = (\Phi_P^X, \Phi_P^C)$ and $Y = (\Phi_P^Y, \Phi_P^C)$. The information diagrams of $\{q(xy)\}$ and $\{p(xy)\}$ are shown in Fig. 1(b) where the approximate values shown in the diagrams have error range within ϵ . Then it can readily be checked that

$$\begin{aligned} D(\mathcal{Q}||\mathcal{P}) &= D(\mathcal{D}_m^1||\mathcal{D}_m^2) + D(\mathcal{D}_m^3||\mathcal{D}_m^2) + D(\mathcal{D}_m^1||\mathcal{D}_m^2) \\ &< 3\epsilon. \end{aligned}$$

Moreover,

$$\begin{aligned} |H_X(\mathcal{Q}) - H_X(\mathcal{P})| &= |H(\mathcal{D}_m^1) + H(\mathcal{D}_m^3) - H(\mathcal{D}_m^2) - H(\mathcal{D}_m^2)| \\ &\leq 4\epsilon, \end{aligned}$$

and similarly,

$$|H_Y(\mathcal{Q}) - H_Y(\mathcal{P})| \leq 4\epsilon.$$

However,

$$\begin{aligned} |H_{XY}(\mathcal{Q}) - H_{XY}(\mathcal{P})| &= |H(\mathcal{D}_m^1) + H(\mathcal{D}_m^3) + H(\mathcal{D}_m^1) - \\ &\quad H(\mathcal{D}_m^2) - H(\mathcal{D}_m^2) - H(\mathcal{D}_m^2)| \\ &\geq 1 - 6\epsilon. \end{aligned}$$

Therefore, the example in Fig. 1(b) shows that if $|H_{XY}(\mathcal{Q}) - H_{XY}(\mathcal{P})|$ is dropped from (23), then the meaning of Definition 5 is changed and Theorem 8 may not be proved.

On the other hand, even if only $|H_Y(\mathcal{Q}) - H_Y(\mathcal{P})|$ is dropped from (23), Theorem 7 cannot be proved which can be seen from the information diagram in Fig. 1(c). By repeating the setup used in Fig. 1(b) except that we replace the distribution of Φ_Q^Y by \mathcal{D}_m^2 , we have

$$|H_{XY}(\mathcal{Q}) - H_{XY}(\mathcal{P})| \leq 6\epsilon,$$

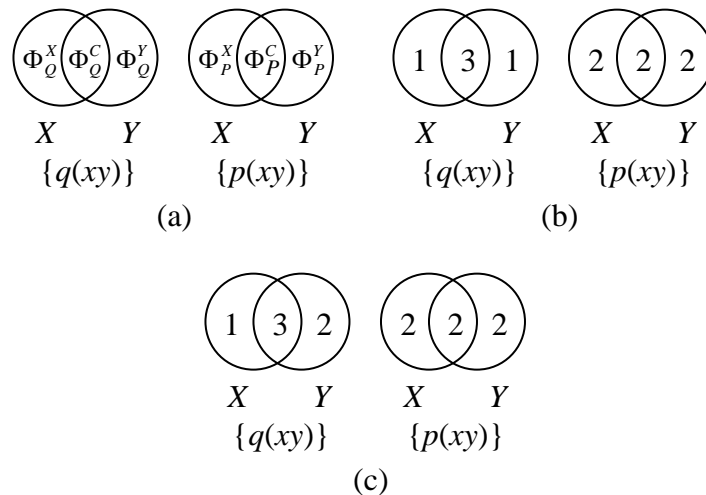


Fig. 1. (a) To illustrate how to construct $q(xy)$ and $p(xy)$. (b)-(c) Two cases illustrating that (23) cannot be simplified.

and

$$|H_X(\mathcal{Q}) - H_X(\mathcal{P})| \leq 4\epsilon,$$

but

$$|H_Y(\mathcal{Q}) - H_Y(\mathcal{P})| \geq 1 - 4\epsilon.$$

Thus we conclude that (23) cannot be simplified.

For weak typicality and for a typical \mathbf{x} , the number of \mathbf{y} such that (\mathbf{x}, \mathbf{y}) is jointly typical is approximately $2^{nH(Y|X)}$ on the average. For strong typicality, this is not only true on the average, but it is also true for every typical \mathbf{x} as long as there exists at least a \mathbf{y} such that (\mathbf{x}, \mathbf{y}) is jointly typical [6]. This result can be extended to countably infinite alphabet by using the unified JAEP, as to be proved in Theorem 10.

Definition 6: For any $\mathbf{x} \in U_{[X]\eta}^n$, the conditional typical set is defined as

$$U_{[Y|X]\eta}^n(\mathbf{x}) = \{\mathbf{y} \in U_{[Y]\eta}^n : (\mathbf{x}, \mathbf{y}) \in U_{[XY]\eta}^n\}.$$

Lemma 9: For any $\mathbf{x} \in U_{[X]\eta}^n$,

$$|U_{[Y|X]\eta}^n(\mathbf{x})| \leq 2^{n(H(Y|X)+2\eta)}$$

Proof: Since $\mathbf{x} \in U_{[X]\eta}^n$, by the unified AEP (Theorem 3), we have

$$\begin{aligned}
2^{-n(H(X)-\eta)} &\geq p(\mathbf{x}) \\
&= \sum_{\mathbf{y} \in \mathcal{Y}^n} p(\mathbf{x}, \mathbf{y}) \\
&\geq \sum_{\mathbf{y} \in U_{[Y|X]\eta}^n(\mathbf{x})} p(\mathbf{x}, \mathbf{y}) \\
&\geq \sum_{\mathbf{y} \in U_{[Y|X]\eta}^n(\mathbf{x})} 2^{-n(H(XY)+\eta)} \\
&= |U_{[Y|X]\eta}^n(\mathbf{x})| \cdot 2^{-n(H(XY)+\eta)},
\end{aligned}$$

so that

$$|U_{[Y|X]\eta}^n(\mathbf{x})| \leq 2^{n(H(Y|X)+2\eta)}.$$

■

Theorem 10: For any $\mathbf{x} \in U_{[X]\eta}^n$, if

$$|U_{[Y|X]\eta}^n(\mathbf{x})| \geq 1,$$

then

$$2^{n(H(Y|X)-\nu)} \leq |U_{[Y|X]\eta}^n(\mathbf{x})| \leq 2^{n(H(Y|X)+\nu)}$$

where $\nu \rightarrow 0$ as $\eta \rightarrow 0$ and then $n \rightarrow \infty$.

The following lemma [6, Lemma 5.10] will be used in the proof.

Lemma 11: For any $n > 0$,

$$n \ln n - n \leq \ln n! \leq (n+1) \ln(n+1) - n.$$

Proof of Theorem 10: In the following, we adopt the notations $H_e(XY)$ and $H_e(X)$ to represent the entropies of $\{p(xy)\}$ and $\{p(x)\}$ in the unit of nat, respectively. Without loss of generality, we assume $\eta < 1$ and let

$$A = B = \left\lfloor \frac{1}{\sqrt{\eta}} \right\rfloor.$$

For any probability distribution $\mathcal{P} = \{p(xy)\}$, find the smallest ν' such that

$$\left| -\sum_{x=1}^A \sum_{y=1}^B p(xy) \ln p(xy) - H_e(XY) \right| \leq \nu' \quad (32)$$

and

$$\left| -\sum_{x=1}^A \left(\sum_{y=1}^B p(xy) \right) \ln \left(\sum_{y=1}^B p(xy) \right) - H_e(X) \right| \leq \nu'. \quad (33)$$

Here $\nu' \rightarrow 0$ as $A \rightarrow \infty$ and $B \rightarrow \infty$. Assume that $|U_{[Y|X]\eta}^n(\mathbf{x})| \geq 1$. Then there exists a $\mathbf{y} \in U_{[Y|X]\eta}^n(\mathbf{x})$ such that

$$D(n^{-1}N(x, y; \mathbf{x}, \mathbf{y}) || p(xy)) \leq \eta$$

so that

$$V(n^{-1}N(x, y; \mathbf{x}, \mathbf{y}), p(xy)) \leq \sqrt{2\eta \ln 2} \quad (34)$$

by Pinsker's inequality. Now let

$$K(x, y) = N(x, y; \mathbf{x}, \mathbf{y}), \quad (35)$$

$$K_X(x) = \sum_{y=1}^{\infty} K(x, y) = N(x; \mathbf{x}),$$

and

$$K'_X(x) = \sum_{y=1}^B K(x, y)$$

for $1 \leq x \leq A$. Straightforward combinatorics reveals that the number of \mathbf{y} satisfying the constraint in (35) is equal to

$$M(K) = \prod_{x=1}^{\infty} \frac{K_X(x)!}{\prod_{y=1}^{\infty} K(x, y)!}.$$

Note that for any such \mathbf{y} , the empirical joint distribution \mathcal{Q} is the same. Let

$$M'(K) = \prod_{x=1}^A \frac{K'_X(x)!}{\prod_{y=1}^B K(x, y)!},$$

which is obviously less than or equal to $M(K)$.

Consider

$$\begin{aligned}
& n^{-1} \ln M'(K) \\
&= n^{-1} \ln \prod_{x=1}^A \frac{K'_X(x)!}{\prod_{y=1}^B K(x, y)!} \\
&= n^{-1} \sum_{x=1}^A \left\{ \ln(K'_X(x)!) - \sum_{y=1}^B \ln(K(x, y)!) \right\} \\
&\stackrel{a)}{\geq} n^{-1} \sum_{x=1}^A \left\{ K'_X(x) \ln K'_X(x) - K'_X(x) \right. \\
&\quad \left. - \sum_{y=1}^B [(K(x, y) + 1) \ln(K(x, y) + 1) - K(x, y)] \right\} \\
&= n^{-1} \sum_{x=1}^A \left\{ K'_X(x) \ln K'_X(x) \right. \\
&\quad \left. - \sum_{y=1}^B (K(x, y) + 1) \ln(K(x, y) + 1) \right\} \\
&= \sum_{x=1}^A \left\{ \frac{K'_X(x)}{n} \left(\ln \frac{K'_X(x)}{n} + \ln n \right) \right. \\
&\quad \left. - \sum_{y=1}^B \frac{K(x, y) + 1}{n} \left(\ln \frac{K(x, y) + 1}{n} + \ln n \right) \right\}.
\end{aligned}$$

In the above, (a) follows from Lemma 11. Since

$$\begin{aligned}
& \frac{K'_X(x) \ln n}{n} - \sum_{y=1}^B \left(\frac{K(x, y) + 1}{n} \right) \ln n \\
&= \frac{K'_X(x) \ln n}{n} - \sum_{y=1}^B \left(\frac{K(x, y) \ln n}{n} \right) - \sum_{y=1}^B \frac{\ln n}{n} \\
&= -\frac{B \ln n}{n},
\end{aligned}$$

we have

$$\begin{aligned}
& n^{-1} \ln M'(K) \\
& \geq \sum_{x=1}^A \left(\frac{K'_X(x)}{n} \right) \ln \left(\frac{K'_X(x)}{n} \right) \\
& \quad - \sum_{x=1}^A \sum_{y=1}^B \left(\frac{K(x, y) + 1}{n} \right) \ln \left(\frac{K(x, y) + 1}{n} \right) \\
& \quad - \frac{AB \ln n}{n}.
\end{aligned} \tag{36}$$

The proof can be completed if we can relate the R.H.S. of (36) to the entropies $H_e(X)$ and $H_e(XY)$. By (34), we have

$$\begin{aligned}
\sqrt{2\eta \ln 2} & \geq \sum_{x=1}^{\infty} \sum_{y=1}^{\infty} |n^{-1}K(x, y) - p(xy)| \\
& \geq \sum_{x=1}^A \sum_{y=1}^B |n^{-1}K(x, y) - p(xy)| \\
& \geq \sum_{x=1}^A \left| \sum_{y=1}^B n^{-1}K(x, y) - \sum_{y=1}^B p(xy) \right| \\
& = \sum_{x=1}^A \left| n^{-1}K'_X(x) - \sum_{y=1}^B p(xy) \right| \\
& = V \left(n^{-1}K'_X(x), \sum_{y=1}^B p(xy) \right).
\end{aligned} \tag{37}$$

Then letting ϵ and M in Lemma 4 be $\sqrt{2\eta \ln 2}$ and A , respectively, we can obtain the upper bound

$$\begin{aligned}
& - \sum_{x=1}^A \left(\frac{K'_X(x)}{n} \right) \ln \left(\frac{K'_X(x)}{n} \right) + \sum_{x=1}^A \left(\sum_{y=1}^B p(xy) \right) \ln \left(\sum_{y=1}^B p(xy) \right) \\
& \leq \phi(A, \sqrt{2\eta \ln 2}),
\end{aligned}$$

where

$$\phi(M, \epsilon) = -\epsilon \log \epsilon + \epsilon \log M$$

for $M > 1$ and $0 < \epsilon < 1$. Therefore,

$$\begin{aligned}
& \sum_{x=1}^A \left(\frac{K'_X(x)}{n} \right) \ln \left(\frac{K'_X(x)}{n} \right) \\
& \geq \sum_{x=1}^A \left(\sum_{y=1}^B p(xy) \right) \ln \left(\sum_{y=1}^B p(xy) \right) - \phi(A, \sqrt{2\eta \ln 2}) \\
& \geq -H_\epsilon(X) - \nu' - \phi(A, \sqrt{2\eta \ln 2}),
\end{aligned} \tag{38}$$

from (33). Now, we consider the second summation in (36) and let $e = \exp(1)$. Since $-x \ln x$ is an increasing function for $0 < x \leq e^{-1}$, we have

$$-\frac{K(x, y) + 1}{n} \ln \frac{K(x, y) + 1}{n} \geq -\frac{K(x, y)}{n} \ln \frac{K(x, y)}{n}$$

for $\frac{K(x, y) + 1}{n} \leq e^{-1}$. Let C be the number of (x, y) such that $\frac{K(x, y) + 1}{n} > e^{-1}$. Then

$$\begin{aligned}
1 + \frac{AB}{n} & \geq \sum_{x=1}^A \sum_{y=1}^B \frac{K(x, y)}{n} + \frac{AB}{n} \\
& = \sum_{x=1}^A \sum_{y=1}^B \frac{K(x, y) + 1}{n} \\
& \geq Ce^{-1}.
\end{aligned}$$

Therefore,

$$C \leq e \left(1 + \frac{AB}{n} \right). \tag{39}$$

Since $-x \ln x$ is a strictly concave function, it is easily checked that if $\frac{K(x, y) + 1}{n} > e^{-1}$, then

$$-\frac{K(x, y) + 1}{n} \ln \frac{K(x, y) + 1}{n} + \frac{K(x, y)}{n} \ln \frac{K(x, y)}{n} \geq -\frac{n-1+1}{n} \ln \frac{n-1+1}{n} + \frac{n-1}{n} \ln \frac{n-1}{n}.$$

That is

$$-\frac{K(x, y) + 1}{n} \ln \frac{K(x, y) + 1}{n} \geq -\frac{K(x, y)}{n} \ln \frac{K(x, y)}{n} - \frac{n-1}{n} \ln \frac{n}{n-1}.$$

Together with (39), we have

$$\begin{aligned}
& -\sum_{x=1}^A \sum_{y=1}^B \left(\frac{K(x, y) + 1}{n} \right) \ln \left(\frac{K(x, y) + 1}{n} \right) \\
& \geq -\sum_{x=1}^A \sum_{y=1}^B \left(\frac{K(x, y)}{n} \right) \ln \left(\frac{K(x, y)}{n} \right) - C \frac{n-1}{n} \ln \frac{n}{n-1} \\
& \geq -\sum_{x=1}^A \sum_{y=1}^B \left(\frac{K(x, y)}{n} \right) \ln \left(\frac{K(x, y)}{n} \right) - e \left(1 + \frac{AB}{n} \right) \frac{n-1}{n} \ln \frac{n}{n-1}.
\end{aligned}$$

By considering (37), we obtain

$$\sum_{x=1}^A \sum_{y=1}^B \left| \frac{K(x,y)}{n} - p(xy) \right| \leq \sqrt{2\eta \ln 2}.$$

Again by using Lemma 4 and applying the same argument leading to (38), we conclude that

$$\begin{aligned} & - \sum_{x=1}^A \sum_{y=1}^B \left(\frac{K(x,y) + 1}{n} \right) \ln \left(\frac{K(x,y) + 1}{n} \right) \\ & \geq - \sum_{x=1}^A \sum_{y=1}^B p(xy) \ln p(xy) - \phi \left(AB, \sqrt{2\eta \ln 2} \right) - e \left(1 + \frac{AB}{n} \right) \frac{n-1}{n} \ln \frac{n}{n-1} \\ & \geq H_e(XY) - \nu' - \phi \left(AB, \sqrt{2\eta \ln 2} \right) - e \left(1 + \frac{AB}{n} \right) \frac{n-1}{n} \ln \frac{n}{n-1}, \end{aligned} \quad (40)$$

from (32). By substituting (38) and (40) into (36), we have

$$\begin{aligned} & n^{-1} \ln M'(K) \\ & \geq -H_e(X) - \nu' - \phi(A, \sqrt{2\eta \ln 2}) + \\ & \quad H_e(XY) - \nu' - \phi \left(AB, \sqrt{2\eta \ln 2} \right) - e \left(1 + \frac{AB}{n} \right) \frac{n-1}{n} \ln \frac{n}{n-1} - \frac{AB \ln n}{n} \\ & = H_e(XY) - H_e(X) - \nu'' \ln 2, \end{aligned}$$

where

$$\begin{aligned} & \nu'' \ln 2 \quad (41) \\ & = 2\nu' + \phi(A, \sqrt{2\eta \ln 2}) + \phi \left(AB, \sqrt{2\eta \ln 2} \right) + e \left(1 + \frac{AB}{n} \right) \frac{n-1}{n} \ln \frac{n}{n-1} + \frac{AB \ln n}{n} \\ & \leq 2\nu' + \phi \left(\frac{1}{\sqrt{\eta}}, \sqrt{2\eta \ln 2} \right) + \phi \left(\frac{1}{\eta}, \sqrt{2\eta \ln 2} \right) + e \left(1 + \frac{1}{\eta n} \right) \frac{n-1}{n} \ln \frac{n}{n-1} + \frac{\ln n}{\eta n} \quad (42) \end{aligned}$$

By changing the base of the logarithm to 2, we have

$$n^{-1} \log M(K) \geq n^{-1} \log M'(K) \geq H(Y|X) - \nu''.$$

Hence we have

$$|U_{[Y|X]_\eta}^n(\mathbf{x})| \geq M(K) \geq 2^{n(H(Y|X) - \nu'')}.$$

We now check that $\nu'' \rightarrow 0$ as $\eta \rightarrow 0$ and then $n \rightarrow \infty$. When $\eta \rightarrow 0$, A and B tend to infinity so that ν' tends to zero. Moreover,

$$\begin{aligned} 0 &\leq \phi\left(\frac{1}{\sqrt{\eta}}, \sqrt{2\eta \ln 2}\right) \\ &\leq \phi\left(\frac{1}{\eta}, \sqrt{2\eta \ln 2}\right) \\ &= -\left(\sqrt{2\eta \ln 2}\right) \log\left(\sqrt{2\eta \ln 2}\right) + 2 \cdot \left(\sqrt{2\eta \ln 2}\right) \log \frac{1}{\sqrt{\eta}} \rightarrow 0. \end{aligned}$$

At the same time, let $\sqrt{n} > \frac{1}{\eta}$ and let n tend to infinity. Then

$$0 \leq e\left(1 + \frac{1}{\eta n}\right) \frac{n-1}{n} \ln \frac{n}{n-1} \leq e\left(1 + \frac{1}{\sqrt{n}}\right) \left(\frac{n-1}{n} \ln \frac{n}{n-1}\right) \rightarrow 0$$

and

$$0 \leq \lim_{\eta \rightarrow 0} \lim_{n \rightarrow \infty} \frac{\ln n}{\eta n} \leq \lim_{\eta \rightarrow 0} \lim_{n \rightarrow \infty} \frac{\ln n}{\sqrt{n}} = 0$$

Therefore, ν'' as defined in (42) tends to zero. This proves the lower bound on $|U_{[Y|X]\eta}^n(\mathbf{x})|$.

The upper bound $|U_{[Y|X]\eta}^n(\mathbf{x})| \leq 2^{n(H(Y|X)+2\eta)}$ has been obtained in Lemma 9. In summary, by letting $\nu = \max\{2\eta, \nu''\}$, we have

$$2^{n(H(Y|X)-\nu)} \leq |U_{[Y|X]\eta}^n(\mathbf{x})| \leq 2^{n(H(Y|X)+\nu)},$$

where $\nu \rightarrow 0$ as $\eta \rightarrow 0$ and then $n \rightarrow \infty$. ■

In the above theorem, we see that the set containing all \mathbf{x} such that $|U_{[Y|X]\eta}^n(\mathbf{x})| \geq 1$ exhibits a nice property. Moreover, this set has essentially the same property as the set $U_{[X]\eta}^n$ that is summarized as in the next theorem.

Definition 7: The set $S_{[X]\eta}^n$ is defined as the set of all sequences $\mathbf{x} \in U_{[X]\eta}^n$ such that $U_{[Y|X]\eta}^n(\mathbf{x})$ is nonempty, i.e.,

$$S_{[X]\eta}^n = \{\mathbf{x} \in U_{[X]\eta}^n : |U_{[Y|X]\eta}^n(\mathbf{x})| > 0\}.$$

Theorem 12: For any $\eta > 0$:

1) If $\mathbf{x} \in S_{[X]\eta}^n$, then

$$2^{-n(H(X)+\eta)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}.$$

2) For sufficiently large n ,

$$\Pr\{\mathbf{X} \in S_{[X]\eta}^n\} > 1 - \eta.$$

3) For sufficiently large n ,

$$(1 - \eta)2^{n(H(X) - \eta)} \leq |S_{[X]\eta}^n| \leq 2^{n(H(X) + \eta)}.$$

Proof: Since $S_{[X]\eta}^n \subset U_{[X]\eta}^n$, Property 1 inherits from Theorem 3. To prove Property 2, we consider

$$1 - \eta \leq \Pr\{(\mathbf{X}, \mathbf{Y}) \in U_{[XY]\eta}^n\} \leq \Pr\{\mathbf{X} \in S_{[X]\eta}^n\},$$

where the first inequality follows from Theorem 8 and the second inequality follows because

$$(\mathbf{X}, \mathbf{Y}) \in U_{[XY]\eta}^n \Rightarrow \mathbf{X} \in S_{[X]\eta}^n.$$

Finally, the proof of Property 3 follows from the same argument as in Theorem 3, so it is omitted here. ■

Another nice property regarding the typical set $S_{[X]\eta}^n$ is presented in the next theorem. With this property, the proof of the achievability of the rate-distortion function in [6, Section 9.5] can readily be extended to countably infinite alphabet.

Theorem 13: For any $\epsilon > 0$, let

$$M = 2^{n(I(X;Y) + \epsilon)}.$$

Define a set of sequences $\Omega = \{\mathbf{y}_i \in U_{[Y]\eta}^n : 1 \leq i \leq M\}$ which is independently and randomly picked from $U_{[Y]\eta}^n$. If $\mathbf{x} \in S_{[X]\eta}^n$, then

$$\Pr\{|U_{[Y|X]\eta}^n(\mathbf{x}) \cap \Omega| > 0\} \geq 1 - \gamma,$$

where $\gamma \rightarrow 0$ as $\eta \rightarrow 0$ and then $n \rightarrow \infty$.

Proof:

We have proved that $|U_{[Y]\eta}^n| \leq 2^{nH(Y) + \eta}$ in Theorem 3. At the same time, if $\mathbf{x} \in S_{[X]\eta}^n$, we have shown that $|U_{[Y|X]\eta}^n(\mathbf{x})| \geq 2^{n(H(Y|X) - \nu)}$ in Theorem 10. Since the M sequences in Ω are randomly and independently picked from $U_{[Y]\eta}^n$, we have

$$\begin{aligned} \Pr\{|U_{[Y|X]\eta}^n(\mathbf{x}) \cap \Omega| = 0\} &= \left(1 - \frac{|U_{[Y|X]\eta}^n(\mathbf{x})|}{|U_{[Y]\eta}^n|}\right)^M \\ &\leq \left(1 - \frac{2^{n(H(Y|X) - \nu)}}{2^{n(H(Y) + \eta)}}\right)^M \\ &= \left(1 - 2^{-n(I(X;Y) + \nu + \eta)}\right)^M. \end{aligned}$$

Then

$$\begin{aligned}
\ln \Pr\{|U_{[Y|X]\eta}^n(\mathbf{x}) \cap \Omega| = 0\} &\leq M \ln(1 - 2^{-n(I(X;Y)+\nu+\eta)}) \\
&\leq -M2^{-n(I(X;Y)+\nu+\eta)} \\
&= -2^{n(\epsilon-\nu-\eta)},
\end{aligned} \tag{43}$$

where (43) follows from $\ln a \leq a - 1$ for $a > 0$. Therefore, we have

$$\begin{aligned}
\Pr\{|U_{[Y|X]\eta}^n(\mathbf{x}) \cap \Omega| > 0\} &= 1 - \Pr\{|U_{[Y|X]\eta}^n(\mathbf{x}) \cap \Omega| = 0\} \\
&\geq 1 - \gamma,
\end{aligned}$$

where

$$\gamma = \exp(-2^{n(\epsilon-\nu-\eta)}).$$

According to Theorem 10, there exist $\eta' < \frac{\epsilon}{3}$ and $n' > \frac{1}{\eta'^2}$ such that $\nu < \frac{\epsilon}{3}$ for $\eta < \eta'$ and $n > n'$. Let $\eta < \frac{\epsilon}{3}$ and $\nu < \frac{\epsilon}{3}$, so that $\epsilon - \nu - \eta > \frac{\epsilon}{3} > 0$. Therefore, $\gamma \rightarrow 0$ as $\eta \rightarrow 0$ and then $n \rightarrow \infty$. ■

V. CONCLUSION

We have introduced a unified typicality for both finite and countably infinite alphabets which is stronger than weak typicality and strong typicality. In the proofs, we have also shown the convergence of the Kullback-Leibler distance between the empirical distribution and the true distribution on a countably infinite alphabet. If the Kullback-Leibler distance is replaced by the variational distance, an exponential convergence rate is shown. The results in this work are closely related to the discontinuity of entropy on countably infinite alphabet with respect to commonly used information divergence measures previously reported in [9]. Unified typicality is potentially useful for proving unified coding theorems that apply to both finite and infinite alphabets.

REFERENCES

- [1] C. E. Shannon, The Mathematical Theory of Communication, *Bell Tech. J.*, V. 27, pp.379-423, July 1948.
- [2] J. Wolfowitz, *Coding Theorems of Information Theory*, Springer, Berlin-Heidelberg, 2nd ed., 1964, 3rd ed., 1978.
- [3] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications*, G. Longo, Ed., Springer-Verlag, New York, 1978.

- [4] I. Csiszár, “The Method of Types,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505-2523, Oct 1998.
- [5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.
- [6] R. W. Yeung, *A First Course in Information Theory*, Kluwer Academic/Plenum Publishers, 2002.
- [7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley-Interscience, 1991.
- [8] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú and M.J. Weiberger, “Inequalities for the L_1 Deviation of the Empirical Distribution,” HP Laboratories Palo Alto, HPL-2003-97 (R.1), Jun. 2003.
- [9] S.-W. Ho and R. W. Yeung, “On the Discontinuity of the Shannon Information Measures”, in *Proc. 2005 IEEE Int. Symposium Inform. Theory (ISIT 2005)*, Adelaide, Australia, Sept. 4-9, 2005.
- [10] 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.
- [11] A. Antos and I. Kontogiannis, “Convergence Properties of Functional Estimates of Discrete Distributions,” *Random Structures and Algorithms*, 2002.
- [12] A. J. Wyner and D. Foster, “On the lower limits of entropy estimation,” *IEEE Trans. Inform. Theory*, submitted for publication.