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 alphabets only, while weak typicality can be applied to countable 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, and it can potentially be used to generalize those theorems which are previously established by strong typicality to countable alphabets. The applications in rate-distortion theory and multi-source network coding problems are discussed.

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] for proving channel coding theorems and then by Berger [3] for proving the rate-distortion theorem and various results in multi-terminal source coding. The concept of typicality was elaborated by Wolfowitz in the book [2]. Together with others works (more history can be found in [4]), the ideas in [2] were systematically developed into the method of types by Csiszár and Körner [5]. Both strong typicality and weak typicality are widely used in information theory, and their details can be found in standard

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 the Institute for Telecommunications Research, University of South Australia, Adelaide SA 5095, Australia. He was with Department of Information Engineering, The Chinese University of Hong Kong, N.T., Hong Kong and Department of Electrical Engineering, Princeton University, NJ 08544, USA when part of this work was done. Email: siuwai.ho@unisa.edu.au

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

textbooks [6][7]. Strong typicality possesses stronger properties compared with weak typicality [6], and hence it is instrumental to proving stronger results. The additional power afford by strong typicality is particularly useful in universal coding, rate-distortion theory and large deviation theory [7, P. 357]. For example, the rate-distortion theorem established by strong typicality is stronger than the one established by weak typicality [7, Sec. 13.6]. Strong typicality is also used in proving results in source coding with side information [7, P. 579], rate distortion with side information [7, P. 585] and relay channel [8]. The remark on [8, Sec. II] asserts that strong typicality is crucial in the proof of [8, Theorem 6].

Unfortunately, strong typicality can only be used for random variables defined on finite alphabets, and hence those theories proved by it suffer the same limitation. When it is important for a theorem, e.g., the source coding theorem, to be independent of the alphabet size [9], we can only use weak typicality in the proof because weak typicality can be applied to countable alphabets[1]. Therefore, weak typicality and strong typicality are used in different problems in information theory. In other words, a notion of typicality that can fully characterize the asymptotic behavior of a memoryless source is lacking. One of the aims in this paper is to define a new typicality which can be applied to countable alphabets while retaining the structural properties of strong typicality. Then those theories that have been established by strong typicality can readily be extended to countable alphabets. Furthermore, researchers can apply this new typicality in place of strong typicality to avoid the assumption of finite alphabet and to prove some results which cannot be proved by weak typicality.

This paper also serves to characterize the asymptotic behavior of a memoryless source. New results on estimating the source distribution and entropy which are independent of the alphabet size are obtained. These results can improve some existing results for those source distribution that have a finite but long tail. They are also instrumental in proving the main results in this paper, and they may have further applications in different problems.

The most important observation in this paper is the one-to-one correspondence between different definitions of typicality and divergence measures. We first express the definitions of weak typicality and strong typicality in terms of information divergence measures in Section II. After that, we define in Section III a new divergence measure which induces a new typicality for

[1] Countable alphabet means an alphabet which can be finite or countably infinite

univariate distributions. Then in Section IV, we extend the results to bivariate distributions. The new typicality, called unified typicality, shares the same asymptotic equipartition property with both weak and strong typicalities. Moreover, it satisfies a conditional asymptotic equipartition property that is satisfied by strong typicality but not weak typicality. After that, the applications of unified typicality to rate-distortion theory and multi-source network coding are discussed before we conclude our paper in Section VI. In this paper, the base of the logarithm is 2.

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 a countable alphabet $\mathcal{X}$. 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 $\mathbf{X} = (X_1, X_2, ..., X_n)$. For a sequence $\mathbf{x} = (x_1, x_2, ..., x_n) \in \mathcal{X}^n$, we call $\mathcal{Q} = \{q(x;\mathbf{x})\}$[2] 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

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

where $D(\mathcal{Q}||\mathcal{P})$ is the Kullback-Leibler divergence 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.

[2] When there is no ambiguity, $q(x;\mathbf{x})$ is simplified as $q(x)$.

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

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

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 $T^n_{[X]\delta}$ with respect to $p(x)$ is the set of sequences $\mathbf{x} = (x_1, x_2, ..., x_n) \in \mathcal{X}^n$ such that $q(x; \mathbf{x}) = 0$ for $p(x) = 0$ and

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

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 $\mathbf{x} \in W^n_{[X]\epsilon}$, then

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

2) For sufficiently large n,

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

3) For sufficiently large n,

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

Strong typicality applying to finite alphabet shares similar properties with weak typicality, namely the *strong AEP* [6], which will not be repeated here. As we will see later, strong typicality can show properties which cannot be shown by weak typicality. That is why strong typicality is more powerful comparing with weak typicality. However, strong typicality cannot be applied to countable infinite alphabets. Since $|\mathcal{X}|$ is involved in the proofs of strong AEP [3][7][6], the proofs become invalid when $|\mathcal{X}| = \infty$. Therefore, the theorems proved by using strong typicality cannot be applied to any distribution with countable infinite alphabet. Although strong typicality applying to countable 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 stated in the following lemma.

Lemma 2: For any $\delta > 0$,

$$\Pr\{\mathbf{X} \in T^n_{[X]\delta}\} = \Pr\{V(\mathcal{P}, \mathcal{Q}) \leq \delta\} \geq 1 - (2^M - 2) \exp\left(\frac{-n\delta^2}{18}\right), \tag{4}$$

where M is the smallest integer satisfying

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

Proof: In this proof, we use [10, Prop. 1] which says that if the true distribution $\mathcal{P}$ has a finite number of probability masses, say L, then

$$\Pr\{V(\mathcal{P}, \mathcal{Q}) \leq \delta\} \geq 1 - (2^L - 2) \exp\left(-\frac{n\delta^2}{2}\right). \tag{6}$$

Let M be the integer as prescribed in the lemma. Let

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

and

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

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

$$V(\mathcal{P}', \mathcal{Q}') = \sum_{x=1}^{M-1} |p(x) - q(x;\mathbf{x})| + \left|\sum_{x=M}^{\infty} p(x) - \sum_{x=M}^{\infty} q(x;\mathbf{x})\right| \leq \frac{\delta}{3}. \tag{7}$$

Let

$$\gamma_1 = \sum_{x=1}^{M-1} |p(x) - q(x;\mathbf{x})| \tag{8}$$

and

$$\gamma_2 = \left|\sum_{x=M}^{\infty} p(x) - \sum_{x=M}^{\infty} q(x;\mathbf{x})\right|, \tag{9}$$

[3] A similar trick can be found in [11].

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 (9). Then by (8), we get

$$\sum_{x=1}^{\infty} |p(x) - q(x;\mathbf{x})| \leq \gamma_1 + \gamma_2 + \frac{2\delta}{3} \leq \delta. \tag{10}$$

Since (7) implies (10), we have

$$\begin{aligned}\Pr\left\{\sum_{x=1}^{\infty} |p(x) - q(x;\mathbf{x})| \leq \delta\right\} &\geq \Pr\left\{V(\mathcal{P}', \mathcal{Q}') \leq \frac{\delta}{3}\right\} \\ &\geq 1 - (2^M - 2)\exp\left(\frac{-n\delta^2}{18}\right), \qquad (11)\end{aligned}$$

where the last inequality follows from (6). ■

Remarks:

i) This lemma says that the variational distance between the true distribution and the empirical distribution converges to 0 in probability as long as the alphabet is countable. This is in some sense an enhancement of [10, Prop. 1].

ii) If the true distribution has a long tail, this result can give a bound tighter than that in [10, Prop. 1].

iii) Neither finite alphabet nor finite variance is assumed in this lemma. This is important in the later parts of the paper.

When we consider a finite alphabet $\mathcal{X}$, strong typicality is more powerful than weak typicality as a tool for theorem proving for memoryless problems. In this case, strong typicality is in fact stronger than weak typicality because for any $\mathbf{x} \in \mathcal{X}^n$, if $\mathbf{x} \in T^n_{[X]\delta}$, then $\mathbf{x} \in W^n_{[X]\epsilon}$ where $\epsilon = -\delta \log(\min_x p(x))$ [6, p. 82]. However, this does not hold when we consider a countable alphabet $\mathcal{X}$. We now argue that strongly typical set may not be a subset of weakly typical set no matter how small δ is. By the discontinuity of the Shannon entropy [12], there exist probability distributions $\mathcal{P}$ and $\mathcal{Q}$ defined on a countable 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 countable. Moreover, the strong AEP does not necessarily hold for countably infinite alphabet. We refer the reader to Problem 3 in Chapter 6 in [6].

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 $\mathcal{P}=\{p(x)\}$ defined on a countable alphabet $\mathcal{X}$ and $H(\mathcal{P})<\infty$.

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

$$D(\mathcal{Q}||\mathcal{P})+|H(\mathcal{Q})-H(\mathcal{P})|\leq\eta. \tag{12}$$

Note that in the above definition, we do not specify any constraint on $|\mathcal{X}|$. The support of the distribution $\mathcal{P}$ of the generic random variable X can be either finite or countably infinite. The former case can be regarded as equivalent to requiring $|\mathcal{X}|$ to be finite as in the definition of strong typicality.

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 $\mathbf{x} \in U^n_{[X]\eta}$, then

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

2) For sufficiently large n,

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

3) For sufficiently large n,

$$(1-\eta)2^{n(H(X)-\eta)} \le |U^n_{[X]\eta}| \le 2^{n(H(X)+\eta)}.$$

The following Lemma 4 [13, Theorem 8] will be used in the proof.

Definition 4: For an unnormalized distribution $\tilde{\mathcal{P}} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_L)$ which can be normalized by a positive constant $\alpha \le 1$ so that $(\alpha^{-1}\tilde{p}_1, \alpha^{-1}\tilde{p}_2, \ldots, \alpha^{-1}\tilde{p}_L)$ 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 $\tilde{\mathcal{P}} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_N)$ and $\tilde{\mathcal{Q}} = (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_N)$ be two unnormalized distributions which can be normalized by two positive constants $\alpha \le 1$ and $\beta \le 1$ so that $(\alpha^{-1}\tilde{p}_1, \alpha^{-1}\tilde{p}_2, \ldots, \alpha^{-1}\tilde{p}_N)$ and $(\beta^{-1}\tilde{q}_1, \beta^{-1}\tilde{q}_2, \ldots, \beta^{-1}\tilde{q}_N)$ are two probability distributions. If[4]

$$V(\tilde{\mathcal{P}}, \tilde{\mathcal{Q}}) = \sum_{i=1}^{N} |\tilde{p}_i - \tilde{q}_i| \le \epsilon,$$

then

$$|H(\tilde{\mathcal{Q}}) - H(\tilde{\mathcal{P}})| \le \begin{cases} -\epsilon \log \epsilon + \epsilon \log N & \epsilon < 1 \\ \log N & \epsilon \ge 1. \end{cases}$$

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

$$\begin{aligned} \eta &\ge D(\mathcal{Q}||\mathcal{P}) + |H(\mathcal{Q}) - H(\mathcal{P})| \\ &\ge |D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})|. \end{aligned} \tag{13}$$

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

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

[4] Here, the definition of variational distance is extended to accept unnormalized distributions as arguments.

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

$$\epsilon = \frac{\eta}{3}. \tag{14}$$

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

$$-\sum_{x=N+1}^{\infty} p(x) \log p(x) \leq \frac{\epsilon}{2}. \tag{15}$$

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

$$-\delta \log \delta + \delta \log N \leq \frac{\epsilon}{2}. \tag{16}$$

Such a δ exists because the L.H.S. of (16) tends to 0 as $\delta \to 0$.

We now show that

$$W^n_{[X]\epsilon} \cap T^n_{[X]\delta} \subseteq U^n_{[X]\eta}. \tag{17}$$

Consider any $\mathbf{x} \in W^n_{[X]\epsilon} \cap T^n_{[X]\delta}$. Then $\mathbf{x}$ satisfies

$$|D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P})| \leq \epsilon \tag{18}$$

and

$$V(\mathcal{Q}, \mathcal{P}) \leq \delta. \tag{19}$$

By (18),

$$\epsilon \geq D(\mathcal{Q}||\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{P}) \geq H(\mathcal{Q}) - H(\mathcal{P}), \tag{20}$$

and by (19),

$$\delta \geq V(\mathcal{Q}, \mathcal{P}) = \sum_{x=1}^{\infty} |q(x) - p(x)| \geq \sum_{x=1}^{N} |q(x) - p(x)|. \tag{21}$$

If the two finite distributions $\mathcal{P}' = \{p(1), p(2), ..., p(N)\}$ and $\mathcal{Q}' = \{q(1), q(2), ..., q(N)\}$ were normalized, then we have $|H(\mathcal{Q}') - H(\mathcal{P}')| \to 0$ as $\delta \to 0$ by (21) 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 (16). It then follows from (16) that

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

So

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

where the last inequality follows from (15). Together with (20), we obtain

$$|H(\mathcal{Q}) - H(\mathcal{P})| \le \epsilon. \tag{22}$$

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

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

where the last inequality follows from (18). Together with (22), this gives

$$D(\mathcal{Q}||\mathcal{P}) + |H(\mathcal{Q}) - H(\mathcal{P})| \le 3\epsilon = \eta \tag{23}$$

(cf. (14)), i.e., $\mathbf{x} \in U^n_{[X]\eta}$, proving (17). It then follows that

$$\begin{aligned} &\Pr\{\mathbf{X} \in U^n_{[X]\eta}\} \\ &\ge \Pr\{\mathbf{X} \in W^n_{[X]\epsilon} \cap T^n_{[X]\delta}\} \\ &= 1 - \Pr\{\mathbf{X} \in (W^n_{[X]\epsilon})^c \cup (T^n_{[X]\delta})^c\} \\ &\ge 1 - (\Pr\{\mathbf{X} \in (W^n_{[X]\epsilon})^c\} + \Pr\{\mathbf{X} \in (T^n_{[X]\delta})^c\}) \qquad (24) \\ &= \Pr\{\mathbf{X} \in W^n_{[X]\epsilon}\} + \Pr\{\mathbf{X} \in T^n_{[X]\delta}\} - 1. \end{aligned}$$

From Property 2 in Theorem 1 and Lemma 2, we have $\Pr\{\mathbf{X} \in W^n_{[X]\epsilon}\} > 1 - \epsilon$, and $\Pr\{\mathbf{X} \in T^n_{[X]\delta}\} > 1 - \delta > 1 - \epsilon$, for sufficiently large n, where $\delta \leq \epsilon$ can be seen from the choice of δ in (16). By using $\epsilon = \frac{\eta}{3}$ from (14), we obtain $\Pr\{\mathbf{X} \in U^n_{[X]\eta}\} > 1 - 2\epsilon > 1 - \eta$, proving Property 2.

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

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

which implies $|U^n_{[X]\eta}| \leq 2^{n(H(X)+\eta)}$. On the other hand, by using the upper bound on $p(\mathbf{x})$ for $\mathbf{x} \in U^n_{[X]\eta}$ obtained in Property 1 and the lower bound on $\Pr\{\mathbf{X} \in U^n_{[X]\eta}\}$ obtained in Property 2 for sufficiently large n, we have

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

which implies

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

Thus

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

The theorem is proved. ■

Remarks:

i) Since the weak law of large numbers for i.i.d. random variables requires only finite mean [14], the assumption $H(\mathcal{P}) < \infty$ implies $\lim_{n\to\infty} \Pr\{\mathbf{X} \in W^n_{[X]\epsilon}\} = 1$. Together with Lemma 2, here in Theorem 3 we only require that $\mathcal{X}$ is countable and $H(\mathcal{P}) < \infty$.

ii) Theorem 3 shows that for a countable alphabet $\mathcal{X}$, a) $H(\mathcal{Q})$ converges in probability to $H(\mathcal{P})$, i.e., the *entropy of the empirical distribution* of the sequence $\mathbf{x}$ converges in probability to the true entropy $H(X)$ and b) $\mathcal{Q}$ converges in probability to $\mathcal{P}$ with respect to Kullback-Leibler divergence, i.e., $\lim_{n\to\infty} \Pr\{D(\mathcal{Q}||\mathcal{P}) > \eta\} = 0$ for any $\eta > 0$. Note that the entropy of the empirical distribution, $H(\mathcal{Q})$, is different from the empirical entropy in (1).

iii) The bound $|U^n_{[X]\eta}| \leq 2^{n(H(X)+\eta)}$ in 3) holds even for small n.

The proof of Theorem 3 implies the following corollary.

Corollary 5: Let ϵ and δ be as specified in (14) and (16), respectively. Then

$$W^n_{[X]\epsilon} \cap T^n_{[X]\delta} \subseteq U^n_{[X]\eta}. \tag{25}$$

Together with the following theorem, the relationship among weakly typical set, strongly typical set and unified typical set is illustrated. In the following theorem, we will prove that unified typicality is stronger than both weak typicality and strong typicality for countable alphabet. This is analogous to the fact that strong typicality is stronger than weak typicality for finite alphabet as discussed in Section II.

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

Proof: For any $\mathbf{x} \in U^n_{[X]\eta}$, $\mathbf{x} \in W^n_{[X]\eta}$ due to (13). Moreover,

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

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

Remarks:

i) Theorem 6 can readily be extended to multivariate distributions. The proof is omitted.

ii) 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 divergence, while preserving the AEP.

iii) The unified AEP in Theorem 3 implies that for any countable alphabet $\mathcal{X}$, both $D(\mathcal{Q}||\mathcal{P})$ and $|H(\mathcal{Q})-H(\mathcal{P})|$ vanish in probability as $n \to \infty$. Since $D(\mathcal{Q}||\mathcal{P}) \to 0$ implies $V(\mathcal{P},\mathcal{Q}) \to 0$, where the latter is in fact the strong AEP when $\mathcal{X}$ is finite. We conclude that the unified AEP implies the strong AEP. However, the converse is not true.

The following theorem giving a bound on the probability of obtaining a non-typical sequence, enhances Property 2 of unified AEP.

Theorem 7: For any probability distribution $\{p(x)\}$ with finite entropy and finite $E[-\log p(X)^2]$, we have

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

Proof: From Chebyshev's inequality, we have

$$\Pr\{\mathbf{X} \in (W^n_{[X]\epsilon})^c\} \quad = \quad \Pr\left\{\left|-\frac{1}{n}\sum_{k=1}^{n} \log p(X_i) - H(X)\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2},$$

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

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

from (11) where M depends on $\{p(x)\}$ and δ. By (24), we have

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

Hence we have proved the theorem. ■

Theorem 7 can be applied to the Kullback-Leibler divergence estimation and entropy estimation [15][16], which we now discuss. If $|\mathcal{X}| < \infty$,

$$\Pr\{D(\mathcal{Q}||\mathcal{P}) > \eta\} \leq 2^{-n\left(\eta - |\mathcal{X}|\frac{\log(n+1)}{n}\right)} \tag{26}$$

is well known (see e.g., [7, Theorem 12.2.1]. By the definition of unified typicality, Theorem 7 implies that

$$\Pr\{D(\mathcal{Q}||\mathcal{P}) > \eta\} = O(n^{-1}), \tag{27}$$

and

$$\Pr\{|H(\mathcal{P}) - H(\mathcal{Q})| > \eta\} = O(n^{-1}). \tag{28}$$

These results give new understanding on the Kullback-Leibler divergence estimation and entropy estimation on countable alphabet.

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 $\mathcal{P}_{XY} = \{p(xy)\}$ and $H(\mathcal{P}_{XY}) < \infty$. We use (X, Y) to denote the pair of generic random variables.

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

$$\begin{aligned} &D(\mathcal{Q}_{XY}||\mathcal{P}_{XY}) + |H(\mathcal{Q}_{XY}) - H(\mathcal{P}_{XY})| + \\ &|H(\mathcal{Q}_X) - H(\mathcal{P}_X)| + |H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)| \leq \eta, \end{aligned} \tag{29}$$

where $\mathcal{P}_X$ and $\mathcal{P}_Y$ denote the marginal distributions $\{p(x)\}$ and $\{p(y)\}$, respectively, while $\mathcal{Q}_{XY}$, $\mathcal{Q}_X$, and $\mathcal{Q}_Y$ are the corresponding empirical distributions of the pair of sequence $(\mathbf{x}, \mathbf{y})$, i.e., $\mathcal{Q}_{XY} = \{q(x, y; \mathbf{x}, \mathbf{y})\}$[5]. Here, $q(x, y; \mathbf{x}, \mathbf{y}) = n^{-1}N(x, y; \mathbf{x}, \mathbf{y})$ and $N(x, y; \mathbf{x}, \mathbf{y})$ is the number of occurrences of (x, y) in the pair of sequences $(\mathbf{x}, \mathbf{y})$.

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

Theorem 8 (Consistency): If $(\mathbf{x}, \mathbf{y}) \in U^n_{[XY]\eta}$, then $\mathbf{x} \in U^n_{[X]\eta}$ and $\mathbf{y} \in U^n_{[Y]\eta}$.

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

$$\begin{aligned} \sum_{xy} q(xy) \log \frac{q(xy)}{p(xy)} &\geq \sum_x \left(\sum_y q(xy) \right) \log \frac{\sum_y q(xy)}{\sum_y p(xy)} \\ &\geq \sum_x q(x) \log \frac{q(x)}{p(x)}. \end{aligned} \tag{30}$$

Therefore,

$$\begin{aligned} \eta &\geq D(\mathcal{Q}_{XY}||\mathcal{P}_{XY}) + |H(\mathcal{Q}_{XY}) - H(\mathcal{P}_{XY})| + \\ &\quad |H(\mathcal{Q}_X) - H(\mathcal{P}_X)| + |H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)| \\ &\geq D(\mathcal{Q}_{XY}||\mathcal{P}_{XY}) + |H(\mathcal{Q}_X) - H(\mathcal{P}_X)| \\ &\geq D(\mathcal{Q}_X||\mathcal{P}_X) + |H(\mathcal{Q}_X) - H(\mathcal{P}_X)|, \end{aligned}$$

[5] When there is no ambiguity, $q(x, y; \mathbf{x}, \mathbf{y})$ is simplified as $q(xy)$.

where $D(\mathcal{Q}_X||\mathcal{P}_X)$ denotes the R.H.S. of (30). Therefore $\mathbf{x} \in U^n_{[X]\eta}$. By symmetry, it is readily seen that $\mathbf{y} \in U^n_{[Y]\eta}$. ■

Theorem 9 (Preservation): For any function $f : \mathcal{X} \to \mathcal{Y}$, denote $(f(x_1), f(x_2), \ldots, f(x_n))$ as $f(\mathbf{x})$. If $\mathbf{x} \in U^n_{[X]\eta}$, then $f(\mathbf{x}) \in U^n_{[f(X)]\gamma}$ with $\gamma \to 0$ as $\eta \to 0$.

Proof: Let $Y = f(X)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ where $y_i = f(x_i)$ for $1 \leq i \leq n$. Then

$$q(xy) = q(x)\mathbf{1}\{y = f(x)\} \tag{31}$$

and

$$p(xy) = p(x)\mathbf{1}\{y = f(x)\}, \tag{32}$$

where $\mathbf{1}\{y = f(x)\} = 1$ if and only if $y = f(x)$. Since $\mathbf{x} \in U^n_{[X]\eta}$,

$$D(\mathcal{Q}_X||\mathcal{P}_X) + |H(\mathcal{Q}_X) - H(\mathcal{P}_X)| \leq \eta. \tag{33}$$

Therefore,

$$\begin{aligned} D(\mathcal{Q}_{XY}||\mathcal{P}_{XY}) &= \sum_{xy} q(x)\mathbf{1}\{y = f(x)\} \log \frac{q(x)\mathbf{1}\{y = f(x)\}}{p(x)\mathbf{1}\{y = f(x)\}} && (34) \\ &= D(\mathcal{Q}_X||\mathcal{P}_X) && (35) \\ &\leq \eta. && (36) \end{aligned}$$

Then

$$D(\mathcal{Q}_Y||\mathcal{P}_Y) \leq D(\mathcal{Q}_{XY}||\mathcal{P}_{XY}) \leq \eta, \tag{37}$$

where the first inequality follows from the log-sum inequality (see for example [6]). So we have $D(\mathcal{Q}_Y||\mathcal{P}_Y) + |H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)| \leq \eta + |H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)|$. By letting $\gamma = \eta + |H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)|$, $f(\mathbf{x}) \in U^n_{[f(X)]\gamma}$ is shown.

The proof is completed if we can show that $\gamma \to 0$ as $\eta \to 0$. It is equivalent to showing

$$\lim_{\eta \to 0} |H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)| = 0. \tag{38}$$

By (37), $D(\mathcal{Q}_Y||\mathcal{P}_Y) \to 0$ and hence $\mathcal{Q}_Y \to \mathcal{P}_Y$ pointwise as $\eta \to 0$. Consider any $\epsilon > 0$. Since entropy is lower-semicontinuous with respect to pointwise convergence [17], (37) shows that there exists ζ such that for $0 < \eta \leq \zeta$

$$H(\mathcal{Q}_Y) \geq H(\mathcal{P}_Y) - \epsilon. \tag{39}$$

Now, we find an upper bound on $H(\mathcal{Q}_Y)$ as follows. For any $\epsilon > 0$, there exists sufficient large L and M such that

$$H(\mathcal{P}_{X|Y}) \leq \sum_{y=1}^{M} p(y)\tilde{H}(\mathcal{P}_{X|Y=y}) + \epsilon, \tag{40}$$

where

$$\tilde{H}(\mathcal{P}_{X|Y=y}) = -\sum_{x=1}^{L} p(x|y)\log p(x|y). \tag{41}$$

On the other hand,

$$\begin{aligned} H(\mathcal{Q}_{X|Y}) &\geq \sum_{y=1}^{M} q(y)H(\mathcal{Q}_{X|Y=y}) & (42)\\ &\geq \sum_{y=1}^{M} q(y)\tilde{H}(\mathcal{Q}_{X|Y=y}), & (43)\end{aligned}$$

where the RHS of (43) is a continuous function in $\{q(xy) : 1 \leq x \leq L \text{ and } 1 \leq y \leq M\}$. From (36), as $\eta \to 0$, we have $D(\mathcal{Q}_{XY}||\mathcal{P}_{XY}) \to 0$, so that $q(x,y) \to p(x,y)$ for all $1 \leq x \leq L$ and $1 \leq y \leq M$. Following (43), by replacing q by p and $\mathcal{Q}$ by $\mathcal{P}$ on the RHS, for any $\epsilon > 0$, there exists ζ' such that for $0 < \eta \leq \min\{\zeta, \zeta'\}$,

$$\begin{aligned} H(\mathcal{Q}_{X|Y}) &\geq \sum_{y=1}^{M} p(y)\tilde{H}(\mathcal{P}_{X|Y=y}) - \epsilon & (44)\\ &\geq H(\mathcal{P}_{X|Y}) - 2\epsilon, & (45)\end{aligned}$$

where the last inequality follows from (40). Therefore,

$$\begin{aligned} H(\mathcal{Q}_Y) &= H(\mathcal{Q}_{XY}) - H(\mathcal{Q}_{X|Y}) & (46)\\ &= H(\mathcal{Q}_X) - H(\mathcal{Q}_{X|Y}) & (47)\\ &\leq H(\mathcal{P}_X) + \eta - H(\mathcal{Q}_{X|Y}) & (48)\\ &\leq H(\mathcal{P}_X) + \eta - H(\mathcal{P}_{X|Y}) + 2\epsilon & (49)\\ &= H(\mathcal{P}_{XY}) + \eta - H(\mathcal{P}_{X|Y}) + 2\epsilon & (50)\\ &= H(\mathcal{P}_Y) + \eta + 2\epsilon, & (51)\end{aligned}$$

where (47) follows from $\mathbf{y} = f(\mathbf{x})$, (48) follows from (33), (49) follows from (45) and (50) follows from $Y = f(X)$. Together with (39),

$$|H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)| \leq \eta + 2\epsilon. \tag{52}$$

By choosing $\epsilon > 0$ to be arbitrarily small, (38) is shown and the proof is completed. ■

Note that Theorem 9 is somewhat weaker than [6, Theorem 6.8]. Nevertheless, since $\gamma \to 0$ as $\eta \to 0$, Theorem 9 still preserves the essential property of [6, Theorem 6.8] and it is good enough for the purpose in Section VI. In the following theorem, the unified joint asymptotic equipartition property (unified JAEP) is proved.

Theorem 10 (Unified JAEP): Let

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

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 $(\mathbf{x}, \mathbf{y}) \in U^n_{[XY]\eta}$, 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,

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

3) For sufficiently large n,

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

Proof: We will first prove Property 2 by letting $\delta = \frac{\eta}{3}$. By applying Property 2 of Theorem 3 to the information source $\{(X_k, Y_K), k \geq 1\}$, we have

$$D(\mathcal{Q}_{XY}||\mathcal{P}_{XY}) + |H(\mathcal{Q}_{XY}) - H(\mathcal{P}_{XY})| \leq \delta \tag{53}$$

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

$$D(\mathcal{Q}_X||\mathcal{P}_X) + |H(\mathcal{Q}_X) - H(\mathcal{P}_X)| \leq \delta \tag{54}$$

is true with probability greater than $1 - \delta$ for sufficiently large n. Since (54) implies

$$|H(\mathcal{Q}_X) - H(\mathcal{P}_X)| \leq \delta, \tag{55}$$

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

$$|H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)| \leq \delta, \tag{56}$$

which is true with probability greater than $1-\delta$ for sufficiently large n. Note that if (53), (55) and (56) are true, then (29) is true because $\delta=\frac{\eta}{3}$. By the union bound, we have

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

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. ■

In Definition 5, $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be countable. If they are also finite, then a joint typicality can be defined in a way simpler than (29). Since entropy is continuous when the alphabet is finite, a small $D(\mathcal{Q}_{XY}||\mathcal{P}_{XY})$ implies that the L.H.S. of (29) is small. In this case, it is sufficient to require $D(\mathcal{Q}_{XY}||\mathcal{P}_{XY})\leq\eta$ in order to define joint typicality. In the general case that the alphabets are countable, the following example shows that omitting any term on the L.H.S. of (29) will lead to a different definition. This will be illustrated by the following probability distribution which has been used in [12] to show the discontinuity of entropy. For a fixed real number γ and an integer n, where $\gamma>0$ and $n>2^\gamma$, let $\mathcal{D}^\gamma_n$ 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.,

$$\mathcal{D}^\gamma_n=\left\{1-\frac{\gamma}{\log n},\frac{\gamma}{n\log n},\frac{\gamma}{n\log n},...,0,0,..\right\}. \tag{57}$$

The above distribution is a special case of the distribution $\mathcal{D}^{\alpha,\beta}_n$ in [12] with $\log\alpha=\gamma$ and $\beta=1$. Then it can readily be checked (see (3) in [12]) that

$$\lim_{n\to\infty}H(\mathcal{D}^\gamma_n)=\gamma. \tag{58}$$

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

$$\begin{aligned}\lim_{n\to\infty}D(\mathcal{D}^\alpha_n||\mathcal{D}^\beta_n) &= \lim_{n\to\infty}\left(1-\frac{\alpha}{\log n}\right)\log\frac{1-\frac{\alpha}{\log n}}{1-\frac{\beta}{\log n}}+\\ &\quad \lim_{n\to\infty}\sum_{i=1}^n\frac{\alpha}{n\log n}\log\frac{\frac{\alpha}{n\log n}}{\frac{\beta}{n\log n}}\\ &= 0+\lim_{n\to\infty}n\cdot\frac{\alpha}{n\log n}\log\frac{\alpha}{\beta}\\ &= 0. \end{aligned}\tag{59}$$

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, 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 probability distributions $\{q(xy)\}$ and $\{p(xy)\}$ as prescribed by Fig. 1(a) and the information diagrams [6] 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

$$D(\mathcal{Q}_{XY}||\mathcal{P}_{XY}) = D(\mathcal{D}_m^1||\mathcal{D}_m^2) + D(\mathcal{D}_m^3||\mathcal{D}_m^2) + D(\mathcal{D}_m^1||\mathcal{D}_m^2) < 3\epsilon.$$

Moreover,

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

and similarly, $|H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)| \leq 4\epsilon$. However,

$$|H(\mathcal{Q}_{XY}) - H(\mathcal{P}_{XY})| = |H(\mathcal{D}_m^1) + H(\mathcal{D}_m^3) + H(\mathcal{D}_m^1) - H(\mathcal{D}_m^2) - H(\mathcal{D}_m^2) - H(\mathcal{D}_m^2)| \geq 1 - 6\epsilon.$$

Therefore, the example in Fig. 1(b) shows that if $|H(\mathcal{Q}_{XY}) - H(\mathcal{P}_{XY})|$ is dropped from (29), then the meaning of Definition 5 is changed and Theorem 10 may not be proved.

On the other hand, even if only $|H(\mathcal{Q}_Y)-H(\mathcal{P}_Y)|$ is dropped from (29), Theorem 8 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 $|H(\mathcal{Q}_{XY})-H(\mathcal{P}_{XY})| \leq 6\epsilon$, and $|H(\mathcal{Q}_X) - H(\mathcal{P}_X)| \leq 4\epsilon$, but $|H(\mathcal{Q}_Y) - H(\mathcal{P}_Y)| \geq 1 - 4\epsilon$. Thus we conclude that (29) 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 is useful in the proof of a version of rate-distortion theorem (mentioned in Exercise 10.16 in [7]) and it can be generalized to countable alphabet by using the unified JAEP, as to be proved in Theorem 12.

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

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

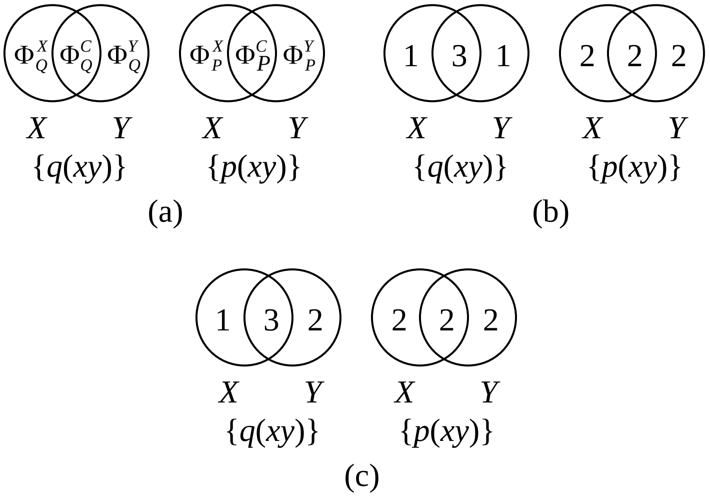

Fig. 1. (a) To illustrate how to construct $q(xy)$ and $p(xy)$. (b)-(c) Two cases illustrating that (29) cannot be simplified.

Lemma 11: For any $\mathbf{x} \in U^n_{[X]\eta}$,

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

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

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

so that

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

■

Theorem 12 (Conditional AEP): For any $\mathbf{x} \in U^n_{[X]\eta}$, if

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

then

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

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

Proof: 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

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

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

$$\left| -\sum_{x=1}^{A}\sum_{y=1}^{A} p(xy)\ln p(xy) - H_e(XY) \right| \leq \nu' \tag{60}$$

and

$$\left| -\sum_{x=1}^{A} \left(\sum_{y=1}^{A} p(xy) \right) \ln \left(\sum_{y=1}^{A} p(xy) \right) - H_e(X) \right| \leq \nu'. \tag{61}$$

Here $\nu' \to 0$ as $A \to \infty$. Consider any $\mathbf{x}$ such that $|U^n_{[Y|X]\eta}(\mathbf{x})| \geq 1$. Then there exists a $\mathbf{y} \in U^n_{[Y|X]\eta}(\mathbf{x})$ such that

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

so that

$$V(n^{-1}N(x,y;\mathbf{x},\mathbf{y}), p(xy)) \leq \sqrt{2\eta \ln 2} \tag{62}$$

by Pinsker's inequality. Now let

$$K(x,y) = N(x,y;\mathbf{x},\mathbf{y}), \tag{63}$$

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

and

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

for $1 \leq x \leq A$. Straightforward combinatorics reveals that the number of $\mathbf{y}$ satisfying the constraint in (63) 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

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

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

By [6, Lemma 6.11], it can easily be verified that

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

Since

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

we have

$$\begin{aligned} & n^{-1} \ln M'(K) \\ \geq & \sum_{x=1}^{A} \left(\frac{K'_X(x)}{n} \right) \ln \left(\frac{K'_X(x)}{n} \right) \\ & - \sum_{x=1}^{A} \sum_{y=1}^{A} \left(\frac{K(x,y)+1}{n} \right) \ln \left(\frac{K(x,y)+1}{n} \right) \\ & - \frac{A^2 \ln n}{n}. \end{aligned} \tag{64}$$

The proof can be completed if we can relate the R.H.S. of (64) to the entropies $H_e(X)$ and $H_e(XY)$. By (62), we have

$$\begin{aligned}
\sqrt{2\eta \ln 2} &\geq \sum_{x=1}^{\infty}\sum_{y=1}^{\infty} |n^{-1}K(x,y) - p(xy)| \\
&\geq \sum_{x=1}^{A}\sum_{y=1}^{A} |n^{-1}K(x,y) - p(xy)| \qquad (65)\\
&\geq \sum_{x=1}^{A} \left| \sum_{y=1}^{A} n^{-1}K(x,y) - \sum_{y=1}^{A} p(xy) \right| \\
&= \sum_{x=1}^{A} \left| n^{-1}K'_X(x) - \sum_{y=1}^{A} p(xy) \right| \\
&= V\left(n^{-1}K'_X(x), \sum_{y=1}^{A} p(xy) \right).
\end{aligned}$$

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

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

where

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

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

$$\begin{aligned}
&\sum_{x=1}^{A}\left(\frac{K'_X(x)}{n}\right)\ln\left(\frac{K'_X(x)}{n}\right) \\
&\quad \geq \sum_{x=1}^{A}\left(\sum_{y=1}^{A} p(xy)\right)\ln\left(\sum_{y=1}^{A} p(xy)\right) - \phi(A, \sqrt{2\eta \ln 2}) \\
&\quad \geq -H_e(X) - \nu' - \phi(A, \sqrt{2\eta \ln 2}), \qquad (66)
\end{aligned}$$

from (61). Now, we consider the second summation in (64) 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

$$\begin{aligned} 1+\frac{A^2}{n} &\geq \sum_{x=1}^{A}\sum_{y=1}^{B}\frac{K(x,y)}{n}+\frac{A^2}{n} \\ &= \sum_{x=1}^{A}\sum_{y=1}^{B}\frac{K(x,y)+1}{n} \\ &\geq Ce^{-1}. \end{aligned}$$

Therefore,

$$C \leq e\left(1+\frac{A^2}{n}\right). \tag{67}$$

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 (67), we have

$$\begin{aligned} &-\sum_{x=1}^{A}\sum_{y=1}^{A}\left(\frac{K(x,y)+1}{n}\right)\ln\left(\frac{K(x,y)+1}{n}\right) \\ \geq\ & -\sum_{x=1}^{A}\sum_{y=1}^{A}\left(\frac{K(x,y)}{n}\right)\ln\left(\frac{K(x,y)}{n}\right)-C\frac{n-1}{n}\ln\frac{n}{n-1} \\ \geq\ & -\sum_{x=1}^{A}\sum_{y=1}^{A}\left(\frac{K(x,y)}{n}\right)\ln\left(\frac{K(x,y)}{n}\right)-e\left(1+\frac{A^2}{n}\right)\frac{n-1}{n}\ln\frac{n}{n-1}. \end{aligned}$$

By considering (65), we obtain

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

By an argument similar to the one leading to (66), we can show that

$$\begin{aligned} &-\sum_{x=1}^{A}\sum_{y=1}^{A}\left(\frac{K(x,y)+1}{n}\right)\ln\left(\frac{K(x,y)+1}{n}\right) \\ \geq\ & -\sum_{x=1}^{A}\sum_{y=1}^{A}p(xy)\ln p(xy)-\phi\left(A^2,\sqrt{2\eta\ln 2}\right)-e\left(1+\frac{A^2}{n}\right)\frac{n-1}{n}\ln\frac{n}{n-1} \\ \geq\ & H_e(XY)-\nu'-\phi\left(A^2,\sqrt{2\eta\ln 2}\right)-e\left(1+\frac{A^2}{n}\right)\frac{n-1}{n}\ln\frac{n}{n-1}, \end{aligned} \tag{68}$$

from (60). By substituting (66) and (68) into (64), we have

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

where

$$\begin{aligned} & \nu''\ln 2 \qquad (69) \\ =\ & 2\nu' + \phi(A, \sqrt{2\eta\ln 2}) + \phi\left(A^2, \sqrt{2\eta\ln 2}\right) + e\left(1 + \frac{A^2}{n}\right)\frac{n-1}{n}\ln\frac{n}{n-1} + \frac{A^2\ln n}{n} \\ \leq\ & 2\nu' + \phi\left(\frac{1}{\sqrt{\eta}}, \sqrt{2\eta\ln 2}\right) + \phi\left(\frac{1}{\eta}, \sqrt{2\eta\ln 2}\right) + e\left(1 + \frac{1}{\eta n}\right)\frac{n-1}{n}\ln\frac{n}{n-1} + \frac{\ln n}{\eta n} \qquad (70) \end{aligned}$$

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

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

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

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

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

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

and

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

Therefore, ν'' as defined in (70) tends to zero. This proves the lower bound on $|U^n_{[Y|X]\eta}(\mathbf{x})|$.

The upper bound $|U^n_{[Y|X]\eta}(\mathbf{x})| \leq 2^{n(H(Y|X)+2\eta)}$ has been obtained in Lemma 11. In summary, by letting $\nu = \max\{2\eta, \nu''\}$, we have

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

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

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

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

$$S^n_{[X]\eta} = \{\mathbf{x} \in U^n_{[X]\eta} : |U^n_{[Y|X]\eta}(\mathbf{x})| > 0\}.$$

Theorem 13: For any $\eta > 0$:

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

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

2) For sufficiently large n,

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

3) For sufficiently large n,

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

Proof: Since $S^n_{[X]\eta} \subset U^n_{[X]\eta}$, Property 1 follows Theorem 3. To prove Property 2, we consider

$$1 - \eta \leq \Pr\{(\mathbf{X}, \mathbf{Y}) \in U^n_{[XY]\eta}\} \leq \Pr\{\mathbf{X} \in S^n_{[X]\eta}\},$$

where the first inequality follows from Theorem 10 and the second inequality follows because

$$(\mathbf{X}, \mathbf{Y}) \in U^n_{[XY]\eta} \Rightarrow \mathbf{X} \in S^n_{[X]\eta}.$$

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 $S^n_{[X]\eta}$ is presented in the next theorem. This property is used in the proof of the achievability of the rate-distortion function in [6, Section 9.5].

Theorem 14: For any $\epsilon > 0$, let

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

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

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

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

Proof:

We have proved that $|U^n_{[Y]\eta}| \leq 2^{nH(Y)+\eta}$ in Theorem 3. At the same time, if $\mathbf{x} \in S^n_{[X]\eta}$, we have shown that $|U^n_{[Y|X]\eta}(\mathbf{x})| \geq 2^{n(H(Y|X)-\nu)}$ in Theorem 12. Since the M sequences in Ω are randomly and independently picked from $U^n_{[Y]\eta}$, we have

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

Then

$$\begin{aligned} \ln \Pr\{|U^n_{[Y|X]\eta}(\mathbf{x}) \cap \Omega| = 0\} &\leq M \ln\left(1 - 2^{-n(I(X;Y)+\nu+\eta)}\right) \\ &\leq -M 2^{-n(I(X;Y)+\nu+\eta)} \qquad (71) \\ &= -2^{n(\epsilon-\nu-\eta)}, \end{aligned}$$

where (71) follows from $\ln a \leq a - 1$ for $a > 0$. Therefore, we have

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

where $\gamma = \exp(-2^{n(\epsilon-\nu-\eta)})$. According to Theorem 12, 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, $\gamma \to 0$ as $\eta \to 0$ and then $n \to \infty$. ■

V. Applications

In this section, we show how the results we have obtained can be used for enhancing results in information theory problems. Since this essentially involves nothing but replacing strong typicality by unified typicality in the original proofs, instead of presenting a tedious complete proof of the result, we will only point out the critical arguments.

A. Multi-Source Network Coding

In [6], a complete characterization of the information rate region is given in Theorem 21.5. This characterization is in terms of $\Gamma_{\mathcal{N}}^{**}$, the set of all entropy functions defined on finite alphabets. In this section, we will show that the information rate region so characterized is unchanged if $\Gamma_{\mathcal{N}}^{**}$ is replaced by $\Gamma_{\mathcal{N}}^{*}$, the set of all entropy functions (possibly defined on countable alphabets).

For the converse proof of Theorem 21.5 given in Section 21.6, since $\Gamma_{\mathcal{N}}^{**} \subset \Gamma_{\mathcal{N}}^{*}$, exactly the same proof can be used without modification. So we only need to show that the achievability proof of Theorem 21.5 given in Section 21.7, where strong typicality is used, continues to work if unified typicality is used instead. Toward this end, we point out that in this proof, the essential properties of strong typicality that are invoked are the joint AEP [6, Theorem 6.9], conditional AEP [6, Theorem 6.10], consistency [6, Theorem 6.7] and preservation [6, Theorem 6.8]. Note that the preservation property is instrumental in proving Lemma 21.9.

In Section IV of the current paper, we have obtained direct generalizations of the first three properties for unified typicality. We also have obtained a somewhat weaker generalization of the preservation property for unified typicality. In this regard, we need the following lemma which is a modification of Lemma 21.9. In the following lemma, Ξ denotes the unified typical set.

Lemma 15: Let

$$X_S = x_S$$

$$\mathbf{Y}_S(x_S) = \mathbf{y}_S \in \Xi^{\hat{n}}_{[Y_S]\delta},$$

and for $e \in E$, let C_e take the value c_e, which by the code construction is a function of x_S and $\mathbf{y}_S$. Then

$$\mathbf{U}_{\mathrm{In}(t)}(c_{\mathrm{In}(t)}) = \tilde{u}_{\mathrm{In}(t)}(\mathbf{y}_S).$$

and

$$(\mathbf{y}_S, \mathbf{U}_{\mathbf{In}(t)}(c_{\mathbf{In}(t)})) \in \Xi^{\hat{n}}_{[Y_S U_{\mathbf{In}(t)}]\xi}$$

for all $t \in \Xi$, where $\xi \to 0$ as $\delta \to 0$.

Then this lemma can be proved by essentially the same proof for Lemma 21.9. With this lemma, the proof of the achievability of the information rate region can be completed accordingly.

The significance of characterizing the information rate region in terms of $\Gamma^*_{\mathcal{N}}$ instead of $\Gamma^{**}_{\mathcal{N}}$ is as follows. The set $\Gamma^*_{\mathcal{N}}$, introduced in [18], has been studied extensive in the literature [19][20][21][22]. On the other hand, relatively little about the set $\Gamma^{**}_{\mathcal{N}}$ is known except that $\overline{\Gamma}^{**}_{\mathcal{N}}$, the closure of $\Gamma^{**}_{\mathcal{N}}$, is equal to $\Gamma^*_{\mathcal{N}}$ [6, Appendix 2.A].

B. Rate-Distortion Theory

The use of unified typicality can readily generalize some existing coding theorems on finite alphabet to countably infinite alphabet. The version of the rate-distortion theorem for finite alphabet in [7, Ch. 10.6] and [6, Theorem 8.17] is one of the examples, where the proof of the achievability of the rate-distortion function $R_I(D)$ is established by using strong typicality. By the same proof with strong typicality replaced by unified typicality, this version of the rate-distortion can immediately be extended to countably infinite alphabet.

The same result has been obtained in [23, Prop. 2b]. In this work, they first prove a weaker version of the rate-distortion theorem by using weak typicality and then strengthen the result by constructing a supercode. Thus unified typicality gives an alternative direct proof. This demonstrates the potential of unified typicality for generalizing coding theorems to countably infinite alphabet which have previously been proved by strong typicality for finite alphabets.

VI. Conclusion

We have introduced a new notion of typical sequences, called *unified typicality*, which works for countable alphabets. This notion of typicality is stronger than both strong typicality and weak typicality that were previously defined in the literature. Fundamental properties of unified typicality, including the asymptotic equipartition properties, have been proved. We have shown how unified typicality can be used for obtaining a new characterization of the information rate region for multi-source network coding and for giving a direct proof of a version of the rate-distortion theorem for countably infinite alphabet. In summary, the notion of unified typicality provides a more complete understanding of the asymptotic behavior of a discrete memoryless source.

ACKNOWLEDGMENT

The author would like to thank the anonymous reviewers for their valuable comments which benefit this paper a lot.

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, *Information Theory and Network Coding*, Springer, 2008.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Ed.*, New York: Wiley-Interscience, 2006.

[8] T. M. Cover and A. El Gamal, “Capacity Theorems for the Relay Channel,” *IEEE Trans. Inform. Theory,* vol. 25, pp. 572–584, Sep. 1979.

[9] A. Orlitsky and N. P. Santhanam, “Speaking of Infinity,” *IEEE Trans. Inform. Theory,* vol. 50, pp. 2215–2230, Oct. 2004.

[10] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú and M. Weinberger, “Universal Discrete Denoising: Known Channel,” *IEEE Trans. Information Theory*, vol. 51, no. 1, pp. 5-28, Jan. 2005.

[11] R. G. Gallager, *Information Theory and Reliability Communication*, John Wiley & Sons, 1968.

[12] 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.

[13] 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.

[14] R. M. Dudley, *Real Analysis and Probability*, Second edition, Cambridge University Press, 2003.

[15] A. Antos and I. Kontoiannis, “Convergence Properties of Functional Estimates of Discrete Distributions,” *Random Structures and Algorithms*, 2002.

[16] A. J. Wyner and D. Foster, “On the lower limits of entropy estimation,” *IEEE Trans. Inform. Theory,* submitted for publication.

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

[18] R. W. Yeung, “A framework for linear information inequalities,” *IEEE Trans. Info. Theory*, IT-43: 1924-1934, 1997.

[19] Z. Zhang and R. W. Yeung, “A non-Shannon-type conditional inequality of information quantities,” *IEEE Trans. Info. Theory*, IT-43: 1982-1986, 1997.

[20] Z. Zhang and R. W. Yeung, “On characterization of entropy function via information inequalities,” *IEEE Trans. Info. Theory*, IT-44: 1440-1452, 1998. ıYeung, R.W. ıZhang, Z.

[21] R. Dougherty, C. Freiling, and K. Zeger, “Six new non-Shannon information inequalities,” 2006 IEEE International Symposium on Information Theory, Seattle, WA, Jul. 9-14, 2006.

[22] F. Matúš, “Infinitely many information inequalities,” 2007 IEEE International Symposium on Information Theory, Nice, France, Jun. 24-29, 2007.

[23] J. C. Kieffer, “Sample Converses in Source Coding Theory,” *IEEE Trans. Inform. Theory,* vol. 37, pp. 263–268, Mar. 1991.