

# On the Interplay Between Conditional Entropy and Error Probability

Siu-Wai Ho, *Member, IEEE*, and Sergio Verdú

**Abstract**—Fano’s inequality relates the error probability of guessing a finitely-valued random variable  $X$  given another random variable  $Y$  and the conditional entropy of  $X$  given  $Y$ . It is not necessarily tight when the marginal distribution of  $X$  is fixed. This paper gives a tight upper bound on the conditional entropy of  $X$  given  $Y$  in terms of the error probability and the marginal distribution of  $X$ . A new lower bound on the conditional entropy for countably infinite alphabets is also found. The relationship between the reliability criteria of vanishing error probability and vanishing conditional entropy is also discussed. A strengthened form of the Schur-concavity of entropy which holds for finite or countably infinite random variables is given.

**Index Terms**—Entropy, equivocation, Fano’s inequality, majorization theory, Schur-concavity, Shannon theory.

## I. INTRODUCTION

IN Shannon theory, coding theorems show reliability in the sense of vanishing decoding error. Shannon [1] showed the converse of the channel coding theorem in the sense that operating above capacity cannot lead to vanishing equivocation (conditional entropy of the message given the decoder output). Fano’s inequality [2] serves to show that reliability in the sense of vanishing error probability implies reliability in the sense of vanishing equivocation. The fact that vanishing equivocation implies vanishing error probability follows from the lower bound on conditional entropy found independently in [3], [4] and [5]. However, both Fano’s inequality and the lower bound in [3]–[5] assume finite alphabets. The results in this paper enable to extend the above results to the general case of possibly countably infinite alphabets. The relationship among the reliability criteria follows from various relationships we find in this paper between the conditional entropy and the minimal error probability. In particular, we obtain the tightest upper bound on

$H(X|Y)$  for a given marginal  $P_X$  and a given minimal error probability  $\min_f \mathbb{P}[X \neq f(Y)]$ .

For discrete random variables  $X$  and  $Y$  taking values on the same alphabet  $\mathcal{X} = \{1, 2, \dots\}$ , we denote for brevity

$$\varepsilon = \mathbb{P}[X \neq Y] = 1 - \sum_{w \in \mathcal{X}} P_{XY}(w, w). \quad (1)$$

If  $\mathcal{X}$  is a finite set, Fano’s inequality relates the conditional entropy of  $X$  given  $Y$  and the error probability  $\varepsilon$  by<sup>1</sup>

$$H(X|Y) \leq \varepsilon \log(|\mathcal{X}| - 1) + h(\varepsilon) \quad (2)$$

where

$$h(x) = x \log \frac{1}{x} + (1 - x) \log \frac{1}{1 - x} \quad (3)$$

for  $0 < x < 1$  and  $h(0) = h(1) = 0$ . Fano’s inequality is tight in the sense that there exist joint distributions for which (2) holds with equality. However, as we will see, there are distributions  $P_X$  such that

$$\max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) < \varepsilon \log(|\mathcal{X}| - 1) + h(\varepsilon). \quad (4)$$

This motivates the search for strengthened versions of Fano’s inequality for which the left side of (4) is achieved with equality.

If  $\mathcal{X}$  is countably infinite, (2) no longer gives an upper bound on  $H(X|Y)$ . In fact, it is possible that in that case  $H(X|Y)$  does not tend to 0 as  $\varepsilon \rightarrow 0$ . This can be explained by the discontinuity of entropy [6] (see also examples in [7] and [8, Ex. 2.49]). Therefore, it is interesting to explore a generalized form of Fano’s inequality that can be used to determine sufficient conditions under which  $\varepsilon \rightarrow 0$  implies  $H(X|Y) \rightarrow 0$  in the case of countably infinite alphabets.

Section II introduces several truncated distributions derived from a given distribution, which are useful throughout the development. Our main results on the bounds of conditional entropy are given in Section III; upper bounds on conditional entropy with a given marginal distribution  $P_X$  are shown in Section III-A, further upper bounding with respect to  $P_X$  is analyzed in Section III-B, and lower bounds on conditional entropy are shown in Section III-C. Convenient lower bounds on error probability in terms of conditional entropy are shown in Section IV. We show that the Schur-concavity of entropy holds in the general case of countable alphabets, thus enabling elegant proofs of new results, as well as generalization of existing results that hold for finite alphabets. Relationships among reliability criteria of vanishing equivocation, vanishing

Manuscript received September 17, 2008; revised November 30, 2009. Date of current version November 19, 2010. The material in this paper was presented (in part) at the 2008 IEEE International Symposium on Information Theory, Toronto, ON, Canada, July 2008. This work was supported (in part) by the National Science Foundation under Grants CCF-0728445 and CCF-0635154, by the Croucher Foundation under the Croucher Foundation Fellowship, and by the Australian Research Council under the Australian Postdoctoral Fellowship.

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

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: verdu@princeton.edu).

Communicated by H. Yamamoto, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIT.2010.2080891

<sup>1</sup>The base for the logarithms, entropies and mutual information is arbitrary.

normalized equivocation, vanishing symbol error probability and block error probability are shown in Section V.

## II. TRUNCATED DISTRIBUTIONS

The bounds in this paper are given in terms of various truncated distributions on the positive integers obtained from a given probability mass function  $P$  defined on the set of positive integers; without loss of generality and for notational simplicity, we assume that for all  $n = 1, 2, \dots$

$$P(n) \geq P(n+1). \quad (5)$$

As we will see, the truncated distributions facilitate the formalization of the results and they appear naturally in those proofs that invoke the theory of majorization [9].

$\mathcal{Q}$ : If  $0 < \eta \leq 1 - P(1)$ , define (cf. Fig. 1)

$$\mathcal{Q}(P, \eta) = \{\eta^{-1}q_1, \eta^{-1}q_2, \dots\} \quad (6)$$

with

$$q_i = \min\{\theta, P(i+1)\} \quad (7)$$

where  $0 < \theta \leq P(1)$  is such that  $\mathcal{Q}(P, \eta)$  is a distribution, i.e.,

$$\eta = \sum_{i=2}^{\infty} \min\{\theta, P(i)\}. \quad (8)$$

Note, for future use, that

$$H(X|X \neq 1) = H(\mathcal{Q}(P_X, 1 - P_X(1))). \quad (9)$$

$\tilde{\mathcal{Q}}$ : For any subset of positive integers  $\mathcal{X}'$ , and

$$1 - \sum_{w \in \mathcal{X}'} P(w) \leq \eta \leq 1 - \max_{w \in \mathcal{X}'} P(w) \quad (10)$$

define (cf. Fig. 2)

$$\tilde{\mathcal{Q}}(P, \mathcal{X}', \eta) = \{\eta^{-1}\tilde{q}_1, \eta^{-1}\tilde{q}_2, \dots\} \quad (11)$$

with

$$\tilde{q}_i = \begin{cases} 0 & \text{if } i = \arg \max_{w \in \mathcal{X}'} P(w) \\ P(i) & \text{if } i \notin \mathcal{X}' \\ \min\{\tilde{\theta}, P(i)\} & \text{otherwise,} \end{cases} \quad (12)$$

where  $0 < \tilde{\theta} \leq \max_{w \in \mathcal{X}'} P(w)$  is such that  $\tilde{\mathcal{Q}}(P, \mathcal{X}', \eta)$  is a distribution

$$\eta = \sum_{i=1}^{\infty} \tilde{q}_i. \quad (13)$$

Note that if  $\mathcal{X}'$  is the set of positive integers, then  $\tilde{\mathcal{Q}}(P, \mathcal{X}', \eta) = \mathcal{Q}(P, \eta)$ .

$\mathcal{R}$ : If  $0 < \eta \leq 1 - P(1)$ , define

$$\mathcal{R}(P, \eta) = \{1 - \eta, q_1, q_2, \dots\}. \quad (14)$$

Note that

$$H(\mathcal{R}(P, \eta)) = \eta H(\mathcal{Q}(P, \eta)) + h(\eta). \quad (15)$$

$\tilde{\mathcal{R}}$ : For any given  $\mathcal{X}'$ , if

$$1 - \sum_{w \in \mathcal{X}'} P(w) \leq \eta \leq 1 - \max_{w \in \mathcal{X}'} P(w) \quad (16)$$

define

$$\tilde{\mathcal{R}}(P, \mathcal{X}', \eta) = \{\tilde{q}_1, \tilde{q}_2, \dots\} \quad (17)$$

where  $\tilde{q}_{w^*} = 0$  for  $w^* = \arg \max_{w \in \mathcal{X}'} P_X(w)$  is replaced by  $1 - \eta$ . Note that

$$H(\tilde{\mathcal{R}}(P, \mathcal{X}', \eta)) = \eta H(\tilde{\mathcal{Q}}(P, \mathcal{X}', \eta)) + h(\eta). \quad (18)$$

Furthermore, if  $\mathcal{X}'$  is the set of positive integers, then

$$\tilde{\mathcal{R}}(P, \mathcal{X}', \eta) = \mathcal{R}(P, \eta). \quad (19)$$

$\mathcal{S}$ : If  $0 < \eta \leq 1$ , define

$$\mathcal{S}(P, \eta) = \{s_1, s_2, \dots\} \quad (20)$$

where

$$s_i = \min\{\nu, \eta^{-1}P(i)\} \quad (21)$$

for  $\nu > 0$  chosen so that  $\mathcal{S}(P, \eta)$  is a distribution.

$\mathcal{T}$ : If  $\eta > P(1)$ , define

$$\mathcal{T}(P, \eta) = \{t_1, t_2, \dots, t_K\} \quad (22)$$

with

$$t_i = \begin{cases} \eta^{-1}P(i), & \text{if } 1 \leq i \leq K-1 \\ 1 - \eta^{-1} \sum_{i=1}^{K-1} P(i), & \text{if } i = K \end{cases} \quad (23)$$

where  $K \geq 2$  is chosen such that

$$\sum_{i=1}^{K-1} P(i) < \eta \leq \sum_{i=1}^K P(i). \quad (24)$$

If  $\eta \leq P(1)$ , then

$$\mathcal{T}(P, \eta) = \{1, 0, \dots, 0\}. \quad (25)$$

$\mathcal{W}$ : Define

$$\mathcal{W}(P) = \{w_1, \dots, w_{\ell+1}\} \quad (26)$$

where

$$\ell = \left\lfloor \frac{1}{P(1)} \right\rfloor \quad (27)$$

$$w_i = \begin{cases} P(1) & \text{if } i = 1, \dots, \ell \\ 1 - \ell P(1) & \text{if } i = \ell + 1. \end{cases} \quad (28)$$

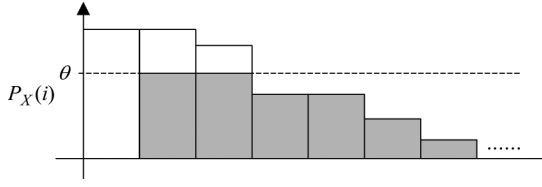


Fig. 1. Example illustrating  $\theta$  and  $q_i$  in (7). The shaded region is  $\{q_i\}$  and the total area is equal to  $\eta$ . Bins represent the positive integers.

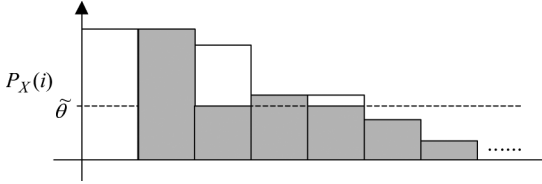


Fig. 2. Example illustrating  $\tilde{\theta}$  and  $\tilde{q}_i$  in (12) with  $\mathcal{X}'$  equal to the odd integers. The shaded region is  $\{\tilde{q}_i\}$  and the total area is equal to  $\eta$ .

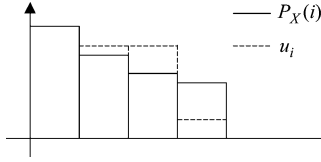


Fig. 3. Example illustrating (32), where  $M = 4$ ,  $u_1 = P_X(1)$ ,  $u_2 = u_3 = \vartheta$ ,  $u_4 = 1 - \eta$ .

It is easy to check that

$$\mathcal{W}(P) = \alpha U_\ell + (1 - \alpha) U_{\ell+1} \quad (29)$$

where  $U_n$  is the equiprobable distribution on  $(1, \dots, n)$  and

$$\alpha = P(1)\ell(\ell + 1) - \ell. \quad (30)$$

$\mathcal{U}$ : If  $P$  is defined on  $\{1, \dots, M\}$ , and  $\eta \geq 1 - P(M)$ , define (cf. Fig. 3)

$$\mathcal{U}(P, \eta) = \{u_1, u_2, \dots, u_M\} \quad (31)$$

with

$$u_i = \begin{cases} \max\{\vartheta, P(i)\}, & \text{if } i \leq M - 1 \\ 1 - \eta, & \text{if } i = M \end{cases} \quad (32)$$

where  $P(M - 1) \leq \vartheta$  is such that  $\mathcal{U}(P, \eta)$  is a distribution.

### III. BOUNDS ON CONDITIONAL ENTROPY

#### A. Upper Bounds on Conditional Entropy With Fixed $P_X$

The first bound tightens Fano's inequality replacing  $\log(|\mathcal{X}| - 1)$  by  $H(\mathcal{Q}(P_X, \varepsilon))$  (see (15)), and it is equivalent to showing the following expression for the rate distortion function [10], [11]

$$\min_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} I(X; Y) = H(X) - H(\mathcal{R}(P_X, \varepsilon)). \quad (33)$$

*Theorem 1:* Let  $X$  and  $Y$  be random variables taking values on the same, possibly countably infinite, alphabet. Suppose

$$\varepsilon \leq 1 - P_X(1). \quad (34)$$

Then

$$\begin{aligned} \max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) \\ = H(\mathcal{R}(P_X, \varepsilon)) \end{aligned} \quad (35)$$

$$\begin{aligned} = (1 - \varepsilon) \log \frac{1}{1 - \varepsilon} + (K - 1)\theta \log \frac{1}{\theta} \\ + \sum_{i=K+1}^{\infty} P_X(i) \log \frac{1}{P_X(i)} \end{aligned} \quad (36)$$

where the integer  $K$  and  $P_X(K + 1) \leq \theta < P_X(K)$  are chosen so that

$$\sum_{i=1}^K P_X(i) = (K - 1)\theta + 1 - \varepsilon. \quad (37)$$

If (34) holds with strict inequality, then the joint distribution that achieves the upper bound in (35) has  $Y$ -marginal:

$$\bar{P}_Y(\ell) = \begin{cases} \frac{P_X(\ell) - \theta}{1 - \varepsilon - \theta} & \ell = 1, \dots, K \\ 0, & \text{otherwise} \end{cases} \quad (38)$$

and conditional distribution (for  $\ell = 1, \dots, K$ )

$$\bar{P}_{X|Y}(k|\ell) = \begin{cases} P_X(k) & k = K + 1, \dots \\ \theta & k = 1, \dots, K, k \neq \ell \\ 1 - \varepsilon & k = \ell = 1, \dots, K. \end{cases} \quad (39)$$

If (34) holds with equality, then

$$\max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) = H(X) \quad (40)$$

achieved by  $Y = 1$ .

*Proof:* If (34) holds with equality, the result is elementary. Suppose otherwise. From (37), it is easy to verify that the solution in (38) and (39) is indeed valid

$$\sum_{\ell=1}^K \bar{P}_{X|Y}(k|\ell) \bar{P}_Y(\ell) = P_X(k) \quad k = 1, 2, \dots \quad (41)$$

and

$$\sum_{\ell=1}^K \bar{P}_{X|Y}(\ell|\ell) \bar{P}_Y(\ell) = 1 - \varepsilon. \quad (42)$$

We now choose an arbitrary  $P_{Y|X}$  such that

$$\varepsilon = \mathbb{P}[X \neq Y] \quad (43)$$

where the joint distribution is  $P_X P_{Y|X}$  and  $P_Y \ll \bar{P}_Y$ , which in turn implies  $P_{XY} \ll \bar{P}_{XY}$ . Consider the following identities where the expectation is with respect to  $P_X P_{Y|X}$

$$\begin{aligned}
& H(X|Y) + D(P_{X|Y} || \bar{P}_{X|Y} | P_Y) \\
&= \mathbb{E} \left[ \log \frac{1}{\bar{P}_{X|Y}(X|Y)} \right] \\
&= \left( \log \frac{1}{1-\varepsilon} \right) \sum_{\ell=1}^K P_Y(\ell) P_{X|Y}(\ell|\ell) \\
&\quad + \left( \log \frac{1}{\theta} \right) \left( \sum_{\ell=1}^K \sum_{k=1}^K P_Y(\ell) P_{X|Y}(k|\ell) \right. \\
&\quad \quad \left. - \sum_{\ell=1}^K P_Y(\ell) P_{X|Y}(\ell|\ell) \right) \\
&\quad + \sum_{\ell=1}^K \sum_{k=K+1}^{\infty} P_Y(\ell) P_{X|Y}(k|\ell) \log \frac{1}{P_X(k)} \\
&= (1-\varepsilon) \log \frac{1}{1-\varepsilon} \\
&\quad + \left( \log \frac{1}{\theta} \right) \left( \sum_{i=1}^K P_X(i) - (1-\varepsilon) \right) \\
&\quad + \sum_{i=K+1}^{\infty} P_X(i) \log \frac{1}{P_X(i)}
\end{aligned} \tag{44}$$

which is equal to (36) in view of (37). Because of the non-negativity of conditional relative entropy, we conclude that the maximum conditional entropy achievable by  $P_{Y|X}$  that satisfy (43) and  $P_Y \ll \bar{P}_Y$  is indeed (36). The fact that the condition  $P_Y \ll \bar{P}_Y$  does not incur loss of optimality can be seen from [11, Sec. 2.9] which is devoted to a full proof of this result. ■

Consider an example where  $P_X = \{0.5, 0.3, 0.1, 0.1\}$  and  $\varepsilon = 0.4$ . Then  $\theta = 0.2$  and the joint distribution defined by (38) and (39) is

$$\bar{P}_{XY} = \begin{bmatrix} 0.45 & 0.05 & 0 & 0 \\ 0.15 & 0.15 & 0 & 0 \\ 0.075 & 0.025 & 0 & 0 \\ 0.075 & 0.025 & 0 & 0 \end{bmatrix}. \tag{47}$$

If  $Y$  is restricted to take values on some proper subset of  $\mathcal{X}$ , the joint distribution achieving the upper bound in (35) may not be given by the joint distribution obtained from (38) and (39), namely

$$\bar{P}_{XY}(\ell) = \begin{cases} 0 & \ell = K+1, \dots \\ P_X(k) \frac{P_X(\ell)-\theta}{1-\varepsilon-\theta} & \ell = 1, \dots, K \\ \theta \frac{P_X(\ell)-\theta}{1-\varepsilon-\theta} & \ell = 1, \dots, K \\ (1-\varepsilon) \frac{P_X(\ell)-\theta}{1-\varepsilon-\theta} & k = 1, \dots, K \\ & k \neq \ell \\ & k = \ell = 1, \dots, K \end{cases} \tag{48}$$

The following result applies to this more general case.

*Theorem 2:* Let  $X$  and  $Y$  be random variables taking values on  $\mathcal{X}$  and  $\mathcal{X}'$ , respectively, where  $\mathcal{X}' \subseteq \mathcal{X}$ . Suppose

$$\varepsilon \leq 1 - \max_{w \in \mathcal{X}'} P_X(w). \tag{49}$$

Then

$$\max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) = H(\tilde{\mathcal{R}}(P_X, \mathcal{X}', \varepsilon)). \tag{50}$$

If (49) holds with strict inequality, then the joint distribution that achieves the upper bound in (50) is

$$\bar{P}_{XY}(k, \ell) = \begin{cases} \frac{[P_X(\ell)-\tilde{\theta}]^+}{1-\varepsilon-\tilde{\theta}} (1-\varepsilon), & k = \ell \\ \frac{[P_X(\ell)-\tilde{\theta}]^+}{1-\varepsilon-\tilde{\theta}} P_X(k) & k \in \mathcal{X} \setminus \mathcal{X}' \\ \frac{[P_X(\ell)-\tilde{\theta}]^+}{1-\varepsilon-\tilde{\theta}} \min\{\tilde{\theta}, P_X(k)\}, & \text{otherwise,} \end{cases} \tag{51}$$

where  $\tilde{\theta}$  is defined through (12) so that  $\tilde{\mathcal{R}}(P_X, \mathcal{X}', \varepsilon)$  is a distribution.

If (49) holds with equality, then

$$\max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) = H(X) \tag{52}$$

achieved by  $Y = \arg \max_{w \in \mathcal{X}'} P_X(w)$ .

*Proof:* Let  $E = \mathbf{1}\{X \in \mathcal{X} \setminus \mathcal{X}'\}$ . Then  $P_E(1) = \sum_{w \in \mathcal{X} \setminus \mathcal{X}'} P_X(w)$  and

$$\mathbb{P}[X \neq Y | E = 0] = \frac{\varepsilon - P_E(1)}{P_E(0)}. \tag{53}$$

Hence

$$H(X|Y) \tag{54}$$

$$= H(X|Y, E) + I(X; E|Y) \tag{55}$$

$$= P_E(0)H(X|Y, E=0) + P_E(1)H(X|Y, E=1) + I(X; E|Y) \tag{56}$$

$$\leq P_E(0)H\left(\mathcal{R}\left(P_{X|E=0}, \frac{\varepsilon - P_E(1)}{P_E(0)}\right)\right) + P_E(1)H(X|Y, E=1) + I(X; E|Y) \tag{57}$$

$$\leq P_E(0)H\left(\mathcal{R}\left(P_{X|E=0}, \frac{\varepsilon - P_E(1)}{P_E(0)}\right)\right) + P_E(1)H(X|E=1) + H(E) \tag{58}$$

where (57) follows from Theorem 1. Let  $P_V = \tilde{\mathcal{R}}(P_X, \mathcal{X}', \varepsilon)$ . Then

$$\begin{aligned}
H(V) &= \mathbb{P}[V \in \mathcal{X}'] H(V|V \in \mathcal{X}') \\
&\quad + \mathbb{P}[V \notin \mathcal{X}'] H(V|V \notin \mathcal{X}') \\
&\quad + H(\mathbf{1}\{V \in \mathcal{X}'\})
\end{aligned} \tag{59}$$

which is equivalent to the right side of (58).

It is readily verified that the maximum is achieved when  $P_{XY}$  has a joint distribution as shown in (51). ■

Let the probability distribution  $\tilde{R}_X = \tilde{\mathcal{R}}(P_X, \mathcal{X}', \varepsilon)$  and let  $K = \max\{i \in \mathcal{X}' : \tilde{R}_X(i) < P_X(i)\}$ . Then  $H(\tilde{\mathcal{R}}(P_X, \mathcal{X}', \varepsilon))$  in (50) becomes

$$\begin{aligned}
& (1-\varepsilon) \log \frac{1}{1-\varepsilon} + H(P_X) \\
& \quad + \sum_{i \in \mathcal{X}': i \leq K} \left( \tilde{\theta} \log \frac{1}{\tilde{\theta}} - P_X(i) \log \frac{1}{P_X(i)} \right)
\end{aligned} \tag{60}$$

where  $\tilde{\theta}$  is defined through (12).

Now, suppose  $Y$  takes values on  $\mathcal{Y}$  which is not necessarily a subset of  $\mathcal{X}$ . Define the minimum probability of error when guessing  $X$  given  $Y$  by

$$\hat{\varepsilon} = \min_{f: \mathcal{Y} \rightarrow \mathcal{X}} \mathbb{P}[X \neq f(Y)] \quad (61)$$

$$= \sum_y P_Y(y) (1 - \max_x P_{X|Y}(x|y)) \quad (62)$$

$$\leq 1 - \max_x P_X(x) \quad (63)$$

where the minimum in (61) is achieved by the maximum *a posteriori* (MAP) estimator and (63) holds by the suboptimal choice  $f(y) = \arg \max_x P_X(x)$ . Note that (2) still holds if  $\varepsilon$  is replaced by  $\hat{\varepsilon}$ .

Before we proceed to generalize Theorem 2, upper bounding  $H(X|Y)$  in terms of  $\hat{\varepsilon} = \min_f \mathbb{P}[X \neq f(Y)]$ , we pause to introduce some background on majorization theory [9]. Consider discrete probability distributions  $\mathcal{P} = \{p_i\}$  and  $\mathcal{Q} = \{q_i\}$  defined on the positive integers labeled in decreasing probabilities, i.e.,

$$p_i \geq p_{i+1} \quad (64)$$

$$q_i \geq q_{i+1}. \quad (65)$$

*Definition 1:*  $\mathcal{Q}$  is majorized by  $\mathcal{P}$  if for all  $k = 1, 2, \dots$

$$\sum_{i=1}^k q_i \leq \sum_{i=1}^k p_i. \quad (66)$$

If a real-valued functional  $f(\cdot)$  is such that  $f(\mathcal{P}) \leq f(\mathcal{Q})$  whenever  $\mathcal{Q}$  is majorized by  $\mathcal{P}$ , we say that  $f(\cdot)$  is Schur-concave.

It is well known that the entropy of finitely-valued random variables is a Schur-concave function. The following result strengthens and generalizes that result allowing for countably infinite random variables.

*Theorem 3:* If  $\mathcal{Q}$  is majorized by  $\mathcal{P}$ , then

$$H(\mathcal{Q}) - H(\mathcal{P}) \geq D(\mathcal{P} \parallel \mathcal{Q}). \quad (67)$$

Therefore, entropy is a Schur-concave function.

*Proof:* If  $\mathcal{Q}$  takes values on  $\{1, \dots, M\}$ , then

$$\begin{aligned} H(\mathcal{Q}) - H(\mathcal{P}) - D(\mathcal{P} \parallel \mathcal{Q}) &= \sum_{j=1}^M (q_j - p_j) \log \frac{1}{q_j} \\ &= \sum_{j=1}^{M-1} (q_j - p_j) \sum_{k=j}^{M-1} \left( \log \frac{q_{k+1}}{q_k} \right) \\ &\quad + \sum_{j=1}^M (q_j - p_j) \log \frac{1}{q_M} \end{aligned} \quad (68)$$

$$\begin{aligned} &= \sum_{k=1}^{M-1} \left( \log \frac{q_k}{q_{k+1}} \right) \left( \sum_{j=1}^k p_j - \sum_{j=1}^k q_j \right) \\ &\geq 0 \end{aligned} \quad (69)$$

$$\geq 0 \quad (70)$$

where the second term in the right side of (69) is equal to zero and the inequality follows from (65) and (66). If  $\mathcal{Q}$  is nonzero for all integers, we need to take  $\lim_{M \rightarrow \infty}$  in the foregoing expressions. We can also conclude that (71) holds because

$$\sum_{j=1}^M (q_j - p_j) \log \frac{1}{q_M} \geq \sum_{j=M+1}^{\infty} q_j \log q_M \quad (72)$$

$$\geq \sum_{j=M+1}^{\infty} q_j \log q_j \quad (73)$$

which vanishes if  $H(\mathcal{Q}) < \infty$ . ■

We proceed to generalize Theorem 2.

*Theorem 4:* Let  $X$  and  $Y$  be random variables taking values on possibly countably infinite alphabets  $\mathcal{X}$  and  $\mathcal{Y}$ , respectively. If  $\hat{\varepsilon} \leq 1 - P_X(1)$ , then

$$\max_{P_{Y|X}: \min_f \mathbb{P}[X \neq f(Y)] = \hat{\varepsilon}} H(X|Y) = H(\tilde{\mathcal{R}}(P_X, \mathcal{X}', \hat{\varepsilon})) \quad (74)$$

where  $\mathcal{X}' = \{1, 2, \dots, \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$  and the maximum is achieved by the same joint distribution as in (51). If  $|\mathcal{Y}| \geq |\mathcal{X}'|$ , (74) simplifies to

$$\max_{P_{Y|X}: \min_f \mathbb{P}[X \neq f(Y)] = \hat{\varepsilon}} H(X|Y) = H(\mathcal{R}(P_X, \hat{\varepsilon})) \quad (75)$$

and (51) is simplified as (48).

*Proof:* Without loss of generality and for notational simplicity, we assume that  $P_X(n) \geq P_X(n+1)$  for all  $n = 1, 2, \dots$ . Let  $f^* = \arg \min_f \mathbb{P}[X \neq f(Y)]$  and  $Z = f^*(Y)$ . Let  $\mathcal{Z} \subseteq \mathcal{X}$  be the range of  $f^*(Y)$ , so  $|\mathcal{Z}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}|\} = |\mathcal{X}'|$ . Then we have

$$\begin{aligned} H(X|Y) &= H(X|Y, Z) \\ &= H(X|Z) \leq H(\tilde{\mathcal{R}}(P_X, \mathcal{Z}, \hat{\varepsilon})) \end{aligned} \quad (76)$$

where the last inequality follows from Theorem 2. Furthermore, if  $\mathcal{Z}' \subseteq \mathcal{Z} \subseteq \mathcal{X}$ , Theorem 2 implies that  $H(\tilde{\mathcal{R}}(P_X, \mathcal{Z}', \hat{\varepsilon})) \leq H(\tilde{\mathcal{R}}(P_X, \mathcal{Z}, \hat{\varepsilon}))$ . Therefore, we can consider  $|\mathcal{Z}| = |\mathcal{X}'|$  in (76). Suppose  $\mathcal{Z} \neq \mathcal{X}'$ . Then there exist  $k$  with  $k \in \mathcal{X}'$  but  $k \notin \mathcal{Z}$  and  $z$  with  $z \in \mathcal{Z}$  but  $z > |\mathcal{X}'| \geq k$ . Let  $\hat{\mathcal{Z}} = \mathcal{Z} \cup \{k\} \setminus \{z\}$ . Since  $P_X(k) \geq P_X(z)$ , it can be verified that  $\tilde{\mathcal{R}}(P_X, \hat{\mathcal{Z}}, \hat{\varepsilon})$  is majorized by  $\tilde{\mathcal{R}}(P_X, \mathcal{Z}, \hat{\varepsilon})$  (i.e., the cumulative distribution function of the majorizing distribution is never below the other one) and hence,  $H(\tilde{\mathcal{R}}(P_X, \hat{\mathcal{Z}}, \hat{\varepsilon})) \geq H(\tilde{\mathcal{R}}(P_X, \mathcal{Z}, \hat{\varepsilon}))$ . Therefore

$$H(X|Y) \leq H(\tilde{\mathcal{R}}(P_X, \mathcal{X}', \hat{\varepsilon})). \quad (77)$$

Since  $\hat{\varepsilon} \leq 1 - P_X(1) \leq 1 - \max_{k \in \mathcal{X} \setminus \mathcal{X}'} P_X(k)$ ,  $f^*(y) = (y)$  for the joint distribution in (51). Therefore, the equality in (77) holds when the joint distribution is (51).

When  $|\mathcal{Y}| \geq |\mathcal{X}|$ ,  $\mathcal{X} = \mathcal{X}'$  so that (51) is equivalent to (48). Together with (19), (74) is equivalent to (75). ■

The right sides of (74) and (75) can be reexpressed as (60) and (36), respectively, with  $\varepsilon$  replaced by  $\hat{\varepsilon}$ .

Since the upper bounds on  $H(X|Y)$  given in Theorems 1 and 4 are tight for given  $P_X$  and  $\varepsilon$ , they are in general tighter than Fano's inequality which depends only on  $\varepsilon$  and on the cardinality of  $X$ . If  $X$  and  $Y$  take values on the same alphabet with  $M$  atoms and  $\varepsilon \leq (M-1)\min_i P_X(i)$ , then the bounds in Theorems 1 and 4 reduce to Fano's inequality (2). Otherwise, Fano's inequality is not tight.

*Example 1:* Consider  $P_X = \{0.4, 0.3, 0.2, 0.1\}$ . If  $\varepsilon = 0.1$  and  $\mathcal{Y} = \{1, 2, 3\}$ , then Fano's inequality gives  $H(X|Y) \leq 0.6275$  bits while using Theorem 2 yields  $\max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) = 0.469$  bits.

The following two theorems consider  $\varepsilon$  which does not satisfy the condition in Theorem 1.

*Theorem 5:* Let  $X$  and  $Y$  be random variables taking values on the same finite alphabet with cardinality  $M$  and assume that  $\varepsilon \geq 1 - P_X(M)$ . Then

$$\max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) = H(\mathcal{U}(P_X, \varepsilon)). \quad (78)$$

The joint distribution that achieves the upper bound in (78) is

$$\bar{P}_{XY}(k, \ell) = \begin{cases} \frac{[\theta - P_X(\ell)]^+}{\theta - 1 + \varepsilon} (1 - \varepsilon), & k = \ell \\ \frac{[\theta - P_X(\ell)]^+}{\theta - 1 + \varepsilon} \max\{\vartheta, P_X(k)\}, & k \neq \ell \end{cases} \quad (79)$$

where  $\vartheta$  is defined through (32) so that  $\mathcal{U}(P_X, \varepsilon)$  is a distribution.

*Proof:* Similar to [10], we can use Lagrange multipliers to get the necessary conditions on the joint distribution  $\bar{P}_{XY}$  with achieves the upper bound in (78). It can be verified that for  $\bar{P}_Y(m) > 0$

$$\bar{P}_{X|Y}(m|m) = 1 - \varepsilon \quad (80)$$

$$\bar{P}_{X|Y}(\ell|m) = \vartheta \quad \text{if } \bar{P}_Y(\ell) > 0 \quad (81)$$

$$\bar{P}_{X|Y}(\ell|m) = P_X(\ell) \quad \text{if } \bar{P}_Y(\ell) = 0 \quad (82)$$

where  $\vartheta$  is a constant to be determined. Together with  $\bar{P}_X = \bar{P}_{XY}$ , we have

$$\bar{P}_Y(\ell) = \frac{\vartheta - P_X(\ell)}{\vartheta - 1 + \varepsilon} \quad (83)$$

for those  $\ell$  with  $\bar{P}_Y(\ell) > 0$ . If  $\vartheta < 1 - \varepsilon$ ,  $\sum_x \bar{P}_{X|Y}(x|y) < 1$ . Therefore

$$\vartheta \geq P_X(\ell) \quad (84)$$

for those  $\ell$  with  $\bar{P}_Y(\ell) > 0$ . It can be verified that  $\mathcal{U}(P_X, \varepsilon)$  is majorized by all those distributions satisfying the conditions in (80)–(84), and hence  $H(X|Y) \leq H(\mathcal{U}(P_X, \varepsilon))$  with equality when the joint distribution is (79). ■

It is easy to check that the solution of (78) is equal to

$$\begin{aligned} H(\mathcal{U}(P_X, \varepsilon)) &= \sum_{i=1}^{K-1} P_X(i) \log \frac{1}{P_X(i)} \\ &\quad + (M-K)\vartheta \log \frac{1}{\vartheta} \\ &\quad + (1-\varepsilon) \log \frac{1}{1-\varepsilon} \end{aligned} \quad (85)$$

where the integer  $K$  and  $P_X(K+1) < \vartheta \leq P_X(K)$  are chosen so that

$$\varepsilon = (M-K)\vartheta + \sum_{i=1}^{K-1} P_X(i). \quad (86)$$

Whenever  $X$  and  $Y$  take values on the same, possibly countably infinite, alphabet but the conditions in either Theorem 1 or Theorem 5 are not satisfied, the maximal conditional entropy is given by the following result.

*Theorem 6:* Let

$$1 - P_X(1) \leq \varepsilon \leq 1 - P_X(M) \quad (87)$$

where if  $M = \infty$ , the right side of (87) is trivially equal to 1. Then

$$\max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) = H(X). \quad (88)$$

*Proof:* Since conditioning does not increase entropy, the result will follow if we can find  $Y$  independent of  $X$  such that  $\mathbb{P}[X \neq Y] = \varepsilon$ . Consider first the case of equiprobable  $X$  on  $\{1, \dots, M\}$ . Then  $Y$  equiprobable on  $\{1, \dots, M\}$  and independent of  $X$  achieves  $\mathbb{P}[X \neq Y] = 1 - \frac{1}{M} = \varepsilon$ , since both inequalities in (87) are equalities in this case.

If one of the inequalities in (87) is strict, then we can choose

$$P_Y(k) = 1 - P_Y(k+1) = \frac{1 - \varepsilon - P_X(k+1)}{P_X(k) - P_X(k+1)} \quad (89)$$

where  $k$  is such that (89) is between 0 and 1. Then

$$\begin{aligned} \mathbb{P}[X \neq Y] &= 1 - \sum_{\ell=1}^M P_X(\ell) P_Y(\ell) \\ &= 1 - P_X(k) P_Y(k) - P_X(k+1) P_Y(k+1) \end{aligned} \quad (90)$$

$$= 1 - P_X(k+1) - P_Y(k)[P_X(k) - P_X(k+1)] \quad (91)$$

$$= 1 - P_X(k+1) - P_Y(k)[P_X(k) - P_X(k+1)] \quad (92)$$

$$= \varepsilon. \quad (93)$$

When  $X$  and  $Y$  take values on the same alphabet, the maximum value of  $H(X|Y)$  has been characterized, for all possible values of  $\varepsilon$ , by Theorems 1, 5 and 6. We proceed to show a tight upper bound on  $H(X|Y = y)$  which can be greater than  $H(X)$ . The lower bound will be shown in Section III-C.

*Theorem 7:* Suppose that  $X$  takes values on a possibly countably infinite alphabet. Furthermore, suppose that  $Y$  is not a constant and  $y$  is such that  $P_Y(y) > 0$ . Then

$$\max_{P_{Y|X}: P_Y(y) = \alpha} H(X|Y = y) = H(\mathcal{S}(P_X, \alpha)) \quad (94)$$

where  $\mathcal{S}(P_X, \alpha) = \{s_1, s_2, \dots\}$  is defined in (20). The conditional distribution that achieves the maximum in (94) is

$$\bar{P}_{Y|X}(\ell|k) = \begin{cases} \frac{P_Y(y)s_k}{P_X(k)}, & \ell = y \\ 1 - \frac{P_Y(y)s_k}{P_X(k)}, & \ell = y + 1 \\ 0, & \text{otherwise.} \end{cases} \quad (95)$$

*Proof:* Let  $\alpha = P_Y(y) > 0$  and let  $V_y$  be a random variable such that  $P_{V_y}(i) = P_{X|Y}(i|y)$ . Then

$$\sum_i P_{V_y}(i) = 1 \quad (96)$$

and  $\alpha P_{V_y}(i) = P_{XY}(i, y) \leq P_X(i)$ . Therefore,

$$P_{V_y}(i) \leq \alpha^{-1} P_X(i) \quad (97)$$

for any  $y$  with  $P_Y(y) > 0$ . Using Lagrange multipliers, it can be verified that

$$H(X|Y = y) \leq \max_V H(V) = H(\mathcal{S}(P_X, \alpha)) \quad (98)$$

subject to the constraints in (96) and (97). It is easily verified that the maximum is achieved when  $P_{Y|X}$  has the conditional distribution shown in (95). ■

Let the probability distribution  $S_X = \mathcal{S}(P_X, \alpha)$  and let  $K = \max\{i : S_X(i) < P_X(i)\}$ . Then  $H(\mathcal{S}(P_X, \alpha))$  in (94) becomes

$$K\nu \log \frac{1}{\nu} + \sum_{i=K+1}^{\infty} (\alpha^{-1} P_X(i)) \log \frac{1}{\alpha^{-1} P_X(i)} \quad (99)$$

where  $\nu$  is defined through (21).

### B. Upper Bounds on Conditional Entropy Maximizing Over the Joint Distribution

The previous results assumed a fixed  $P_X$ . Now, we further upper bound  $H(X|Y)$  by supremizing not only over  $P_{Y|X}$  but over  $P_X \in \mathcal{P}$ , in the special case in which  $\mathcal{P}$  contains a distribution majorized by all others.

*Theorem 8:* Assume that there exists  $P_{X^*} \in \mathcal{P}$  which is majorized by every other distribution in  $\mathcal{P}$ . Then for  $\varepsilon \leq 1 - P_{X^*}(1)$

$$\max_{P_X \in \mathcal{P}, \mathbb{P}[X \neq Y_1] = \varepsilon} H(X|Y_1) = H(\mathcal{R}(P_{X^*}, \varepsilon)) \quad (100)$$

and

$$\max_{P_X \in \mathcal{P}, \min_f \mathbb{P}[X \neq f(Y_2)] = \hat{\varepsilon}} H(X|Y_2) = H(\mathcal{R}(P_{X^*}, \mathcal{X}', \hat{\varepsilon})) \quad (101)$$

where  $X$  and  $Y_1$  take values on  $\mathcal{X}$ ,  $Y_2$  takes values on  $\mathcal{Y}$  and  $\mathcal{X}' = \{1, 2, \dots, \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$ . The maximum values in (100) and (101) are achieved by the joint distributions (48) and (51), respectively, substituting  $P_X = P_{X^*}$  therein.

Theorem 8 is a direct consequence of Theorems 1 and 4 in view of the following result.

*Lemma 1:* If  $P_{X^*}$  is majorized by  $P_X$ , then  $\mathcal{R}(P_{X^*}, \eta)$  is majorized by  $\mathcal{R}(P_X, \eta)$  for any  $0 < \eta \leq 1 - P_X(1)$ .

*Proof:* See the Appendix. ■

The right side of (100) can be reexpressed as (36) with  $P_X$  replaced by  $P_{X^*}$ , and the right side of (101) can be reexpressed as (60) with  $P_X$  and  $\varepsilon$  replaced by  $P_{X^*}$  and  $\hat{\varepsilon}$ , respectively.

If the set of distributions  $\mathcal{P}$  is closed, then the set  $\hat{\mathcal{P}} = \mathcal{P} \cup P_{X^*}$ , where

$$P_{X^*}(1) = \min_{P \in \mathcal{P}} P(1) \quad (102)$$

$$P_{X^*}(a) = \min_{P \in \mathcal{P}} \sum_{i=1}^a P(i) - \sum_{i=1}^{a-1} P_{X^*}(i) \quad a > 1 \quad (103)$$

always satisfies the assumption in Theorem 8 due to the following lemma.

*Lemma 2:* For a closed but possibly uncountably infinite set of probability distributions  $\mathcal{P}$ ,  $P_{X^*}$  (defined in (102) and (103)) is majorized by  $P \in \mathcal{P}$  and  $P_{X^*}$  majorizes any  $P_{X'}$  which is majorized by all  $P \in \mathcal{P}$ .

*Proof:* See the Appendix. ■

The following theorem is a direct consequence of Theorem 8.

*Theorem 9:* For any closed set  $\mathcal{P}$  of probability distributions on  $\mathcal{X}$ , and  $\varepsilon \leq 1 - P_{X^*}(1)$ , where  $P_{X^*}$  is defined in (102) and (103)

$$H(X|Y_1) \leq H(\mathcal{R}(P_{X^*}, \mathbb{P}[X \neq Y_1])) \quad (104)$$

and

$$H(X|Y_2) \leq H(\mathcal{R}(P_{X^*}, \mathcal{X}', \min_f \mathbb{P}[X \neq f(Y_2)])) \quad (105)$$

where  $X$  and  $Y_1$  take values on  $\mathcal{X}$ ,  $Y_2$  takes values on  $\mathcal{Y}$  and  $\mathcal{X}' = \{1, 2, \dots, \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$ .

The right side of (104) can be reexpressed as (36) with  $P_X$  and  $\varepsilon$  replaced by  $P_{X^*}$  and  $\mathbb{P}[X \neq Y_1]$ , respectively. Similarly, the right side of (105) can be reexpressed as (60) with  $P_X$  and  $\varepsilon$  replaced by  $P_{X^*}$  and  $\min_f \mathbb{P}[X \neq f(Y_2)]$ , respectively.

### C. Lower Bounds on Conditional Entropy

We start by generalizing the lower bound on entropy of a distribution with a given maximal mass in [5, Lemma2], to countably infinite alphabets.

*Theorem 10:*

$$\min H(X) = h(\pi \lfloor \pi^{-1} \rfloor) + \pi \lfloor \pi^{-1} \rfloor \log \lfloor \pi^{-1} \rfloor \quad (106)$$

where the minimization is over all discrete random variables whose maximum probability mass is equal to  $\pi$ .

*Proof:* It is easy to check that the right side of (106) is equal to  $H(\mathcal{W}(P_X))$ . Therefore, the result is equivalent to

$$H(X) \geq H(\mathcal{W}(P_X)). \quad (107)$$

In view of Theorem 3, (107) follows from the fact that  $P_X$  is majorized by  $\mathcal{W}(P_X)$ : from (26) and (28), we have

$$\sum_{i=1}^k w_i \geq \sum_{i=1}^k P_X(i) \quad (108)$$

for all  $k$  (with equality when  $k = \infty$ ). ■

Denoting the convex hull of the function in the right side of (106) by  $\phi(\pi)$ , it is immediate to conclude that

$$\min H(X|Y) = \phi(\varepsilon) \quad (109)$$

where the minimum is over all joint distributions of  $X$  and  $Y$  such that the minimal error probability of guessing  $X$  on the basis of  $Y$  is equal to  $\varepsilon$ . In the special case of finite alphabets, (109) was found in [3]–[5].

In the next result, we give a simple lower bound on  $H(X|Y)$ .

*Theorem 11:* For any  $X$  taking values on a possibly countably infinite alphabet, we have

$$2 \min_f \mathbb{P}[X \neq f(Y)] \leq H(X|Y), \quad (110)$$

where the conditional entropy is measured in bits.

*Proof:* For convenience we assume within the proof that all entropies are in bits. With the notations in (26)–(30),

$$\begin{aligned} \mathcal{W}(P_X) &= \alpha \left( \frac{1}{\ell}, \frac{1}{\ell}, \dots, \frac{1}{\ell} \right) \\ &+ (1 - \alpha) \left( \frac{1}{\ell + 1}, \frac{1}{\ell + 1}, \dots, \frac{1}{\ell + 1} \right). \end{aligned} \quad (111)$$

From the concavity of entropy, we obtain

$$H(\mathcal{W}(P_X)) \geq \alpha \log_2 \ell + (1 - \alpha) \log_2(\ell + 1) \quad (112)$$

$$\geq 2(1 - P_X(1)) \quad (113)$$

where (113) follows from (30) and the fact that

$$\log_2 \ell \geq 2 - \frac{2}{\ell} \quad (114)$$

for any positive integer  $\ell$ . Together with Theorem 10, we have

$$H(X) \geq 2(1 - P_X(1)) \quad (115)$$

a bound which is weaker than [5, Lemma 2]. Then

$$\begin{aligned} H(X|Y) &= \sum_y P_Y(y) H(X|Y = y) \\ &\geq \sum_y P_Y(y) 2 \left( 1 - \max_{x'} P_{X|Y}(x'|y) \right) \end{aligned} \quad (116)$$

$$= 2 \min_f \mathbb{P}[X \neq f(Y)] \quad (117)$$

where (116) follows from (115). ■

Together with the fact that

$$\sum_w |P_X(w) - P_Y(w)| \leq 2\mathbb{P}[X \neq Y] \quad (118)$$

we can connect variational distance and conditional entropy in the following corollary.

*Corollary 12:* Let  $f^* = \arg \min_f \mathbb{P}[X \neq f(Y)]$  and  $Z = f^*(Y)$ . So  $Z$  is the function of  $Y$  that gives the maximum likelihood estimate of  $X$ . Then

$$\sum_w |P_X(w) - P_Z(w)| \leq H(X|Z) \quad (119)$$

where the conditional entropy is measured in bits.

*Proof:*

$$h(x|Z) \geq H(x|Y, F^*(y)) \quad (120)$$

$$= H(x|Y) \quad (121)$$

$$\geq 2\mathbb{P}[x \neq F^*(y)] \quad (122)$$

$$= 2\mathbb{P}[x \neq Z] \quad (123)$$

$$\geq \sum_w |P_x(w) - P_z(w)| \quad (124)$$

where (122) and (124) follows from Theorem 11 and (118), respectively. ■

We proceed to show a tight lower bound on  $H(X|Y = y)$ .

*Theorem 13:* For any  $X$  taking values on a possibly countably infinite alphabet and  $Y$  is not a constant with  $P_Y(y) > 0$

$$\min_{P_{Y|X}: P_{Y(y)=\alpha}} H(X|Y = y) = H(\mathcal{T}(P_X, \alpha)) \quad (125)$$

where  $\mathcal{T}$  is defined in (22). The conditional distribution that achieves the minimum in (125) is

$$\bar{P}_{Y|X}(\ell|k) = \begin{cases} \frac{P_Y(y)t_k}{P_X(k)}, & \ell = y \\ 1 - \frac{P_Y(y)t_k}{P_X(k)}, & \ell = y + 1 \\ 0, & \text{otherwise.} \end{cases} \quad (126)$$

*Proof:* Let  $\alpha = P_Y(y) > 0$  and let  $V_y$  be a random variable such that  $P_{V_y}(i) = P_{X|Y}(i|y)$ . Consider those  $t_i$  in  $\mathcal{T}(P_X, \alpha)$  and let  $t_i = 0$  for  $i > K$ . It is readily checked that for all  $k \geq 1$ ,  $\sum_{i=1}^k t_i \geq \sum_{i=1}^k P_{V_y}(i)$ , where the equality holds when  $k$  is equal to the cardinality of  $X$ . Therefore,  $P_{V_y}$  is majorized by  $\mathcal{T}(P_X, \alpha)$  and  $H(\mathcal{T}(P_X, \alpha)) \leq H(P_{V_y}) = H(X|Y = y)$ , according to Theorem 3. It is easily verified that the maximum is achieved when  $P_{Y|X}$  has a conditional distribution as shown in (126). ■

Note that  $H(\mathcal{T}(P_X, \alpha))$  in (125) can be reexpressed as

$$\begin{aligned} &\sum_{i=1}^{K-1} (\alpha^{-1} P_X(i)) \log \frac{1}{\alpha^{-1} P_X(i)} \\ &+ \left( 1 - \eta^{-1} \sum_{i=1}^{K-1} P(i) \right) \log \frac{1}{1 - \eta^{-1} \sum_{i=1}^{K-1} P(i)} \end{aligned} \quad (127)$$

where  $K$  is defined through (24). ■

#### IV. LOWER BOUNDS ON ERROR PROBABILITY

We now proceed to obtain the tightest lower bound on error probability for a fixed  $P_X$  after we show some properties of the function  $\Phi_X : [0, 1] \rightarrow [0, \infty)$  defined by

$$\Phi_X(a) = H(\mathcal{R}(P_X, a)) \quad (128)$$

$$= aH(\mathcal{Q}(P_X, a)) + h(a) \quad (129)$$

where  $\mathcal{R}$  is defined in (14). Note that from Theorem 1

$$\Phi_X(\varepsilon) = \max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) \quad (130)$$

$$= H(X) - \min_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} I(X; Y) \quad (131)$$

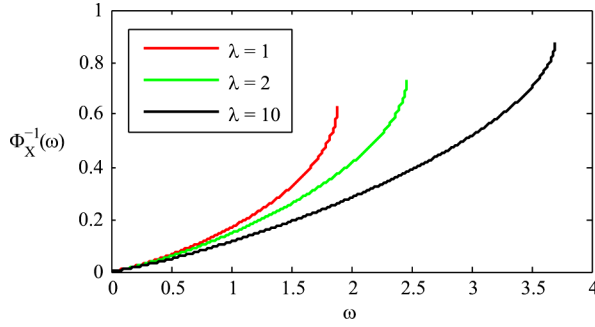


Fig. 4. Plot of  $\Phi_X^{-1}(\omega)$  where  $P_X$  is a Poisson distribution with mean  $\lambda$ .

for  $\varepsilon \leq 1 - P_X(1)$  and from Theorem 4

$$\begin{aligned} \Phi_X(\hat{\varepsilon}) &= \max_{P_{Y|X}: \min_f \mathbb{P}[X \neq f(Y)] = \hat{\varepsilon}} H(X|Y) \\ &= H(X) - \min_{P_{Y|X}: \min_f \mathbb{P}[X \neq f(Y)] = \hat{\varepsilon}} I(X; Y). \end{aligned} \quad (132)$$

So  $\Phi_X(a)$  can be seen as a constant minus a rate-distortion function with (1) and (61) being the distortion measures in (131) and (133), respectively. Since the rate-distortion function is convex [8], we have shown that  $\Phi_X(\cdot)$  is a concave function. Since entropy is Schur-concave from Theorem 3 and  $\mathcal{R}(P_X, \delta')$  is majorized by  $\mathcal{R}(P_X, \delta)$  if  $\delta' > \delta$ ,  $\Phi_X(\cdot)$  is a strictly increasing function. It can be verified that  $\Phi_X(\cdot)$  is continuous and hence,  $\Phi_X^{-1}(\cdot)$  exists. From Theorem 4 and (129), we have

$$\Phi_X^{-1}(H(X|Y)) \leq \min_f \mathbb{P}[X \neq f(Y)]. \quad (134)$$

Furthermore, for any given  $P_X$ , and  $0 \leq \tau \leq H(X)$

$$\min_{P_{Y|X}: H(X|Y) = \tau} \min_f \mathbb{P}[X \neq f(Y)] = \Phi_X^{-1}(\tau). \quad (135)$$

Although an analytical expression for  $\Phi_X^{-1}$  is unknown, it can be readily found numerically (see Fig. 4). Note that

$$\Phi_X^{-1}(0) = \frac{d}{d\omega} \Phi_X^{-1}(\omega)|_{\omega=0} = 0. \quad (136)$$

Furthermore,  $\Phi_X^{-1}$  is convex because  $\Phi_X$  is concave.

As an application of (134), consider a random process  $\{X_n\}_{n=-\infty}^{\infty}$ , taking values on a finite or countably infinite alphabet. Consider the minimum prediction error

$$\hat{\varepsilon}_n = \min_f \mathbb{P}[X_n \neq f(X^{n-1})] \quad (137)$$

where  $X^{n-1} = (X_1, \dots, X_{n-1})$ . Then the predictability of the process [5] is defined as

$$\hat{\varepsilon}_\infty = \lim_{n \rightarrow \infty} \hat{\varepsilon}_n. \quad (138)$$

It easily follows from the continuity of  $\Phi$  that the entropy rate  $\lim_{n \rightarrow \infty} H(X_n|X^{n-1})$  and the predictability satisfy

$$\lim_{n \rightarrow \infty} H(X_n|X^{n-1}) \leq \Phi_X(\hat{\varepsilon}_\infty) \quad (139)$$

when the process is stationary, where  $\Phi_X = \Phi_{X_k}$  since all  $X_k$  are identically distributed. According to (135), for any  $P_X$ ,

there exists a process with first-order distribution  $P_X$ , whose predictability and entropy rate satisfy (139) with equality. If the process is nonstationary but there exists  $X$  such that  $P_X$  is majorized by  $P_{X_i}$  for all  $i$  and  $H(X) < \infty$ , it follows from the continuity of  $\Phi$  and Lemma 1 that

$$\lim_{n \rightarrow \infty} H(X_n|X^{n-1}) \leq \Phi_X(\hat{\varepsilon}_\infty). \quad (140)$$

## V. VANISHING EQUIVOCATION VERSUS VANISHING ERROR PROBABILITY

### A. Vanishing Equivocation

Vanishing equivocation implies vanishing error probability in the general setup of possibly countably-valued random variables.

*Theorem 14:* For  $X_n$  taking values on a possibly countably infinite alphabet, we have

$$H(X_n|Y_n) \rightarrow 0 \implies \min_f \mathbb{P}[X_n \neq f(Y_n)] \rightarrow 0. \quad (141)$$

*Proof:* Corollary of Theorem 11. ■

Note that  $\min_f \mathbb{P}[X_n \neq f(Y_n)]$  cannot be replaced by  $\mathbb{P}[X_n \neq Y_n]$  in (141). For example, let  $Y_n = \bar{X}_n \in \{0, 1\}$ , then  $H(X_n|Y_n) = 0$  and  $\mathbb{P}[X_n \neq Y_n] = 1$ . In the remainder of this subsection, we study sufficient conditions under which vanishing error probability implies vanishing equivocation. The most well-known link between both criteria is the following result, which is the cornerstone of converse proofs in channel capacity.

*Theorem 15:* Let  $X_n$  be equiprobable on  $\mathcal{X} = \{1, 2, \dots, a^n\}$ . Then

$$\mathbb{P}[X_n \neq Y_n] \rightarrow 0 \implies \frac{1}{n} H(X_n|Y_n) \rightarrow 0. \quad (142)$$

*Proof:* Application of Fano's inequality (2) to the special case  $\mathcal{X} = \{1, 2, \dots, a^n\}$ . ■

When the alphabet grows with  $n$  (as in Theorem 15) the unnormalized equivocation need not vanish even if the error probability vanishes.

*Example 2:* Consider  $\mathcal{X} = \{1, 2, \dots, n+1\}$ . Let

$$P_{X_n} = \left\{ 1 - \frac{1}{\log n}, \frac{1}{n \log n}, \dots, \frac{1}{n \log n} \right\}$$

and  $Y = 1$ . Then,  $H(X_n|Y) = h\left(\frac{1}{\log n}\right) + 1$  bits, while  $\mathbb{P}[X_n \neq Y] = \frac{1}{\log n}$ . This is not due to the fact that  $Y$  is deterministic in Example 2.

*Example 3:* Let  $Y$  be an arbitrary random variable with  $P_Y(1) > 0$  and let

$$P_{W_n} = \left\{ 1 - \frac{\alpha}{\log n}, \frac{\alpha}{n \log n}, \dots, \frac{\alpha}{n \log n} \right\} \quad (143)$$

where  $\alpha$  is any positive finite number and  $n \geq \exp(\alpha)$ . Consider

$$X_n = \begin{cases} Y, & \text{if } Y \neq 1 \\ W_n, & \text{if } Y = 1. \end{cases} \quad (144)$$

Then

$$\varepsilon_n = \mathbb{P}[X_n \neq Y] \tag{145}$$

$$= P_Y(1)\mathbb{P}[X_n \neq Y|Y = 1] \tag{146}$$

$$= P_Y(1)\frac{\alpha}{\log n} \tag{147}$$

which tends to 0 as  $n \rightarrow \infty$ . However

$$H(X_n|Y) \geq P_Y(1)H(X_n|Y = 1) = P_Y(1)H(W_n) \tag{148}$$

which tends to an arbitrary value  $\alpha P_Y(1)$  as  $n \rightarrow \infty$ . Now, more insight can be gained from this example. Let  $Z = \mathbf{1}\{Y = 1\}$ , where  $\mathbf{1}$  is the indicator function and assume that  $H(Y) < \infty$ , then

$$H(X_n) = H(X_n|Z) + I(X_n; Z) \tag{149}$$

$$= \mathbb{P}[Y = 1]H(W_n) + H(Y) \tag{150}$$

$$+ I(X_n; Z) - h(P_Y(1)) \tag{151}$$

$$< \infty, \tag{152}$$

because  $H(W_n) \leq 1 + \alpha$  bits and  $I(X_n; Z) \leq 1$  bits. Furthermore, it is easy to verify that

$$\sum_x |P_{X_n}(x) - P_Y(x)| \rightarrow 0 \tag{153}$$

as  $n \rightarrow \infty$ . Therefore,  $H(X_n) < \infty$  and convergence of  $P_{X_n}$  are not sufficient to show  $H(X_n|Y_n) \rightarrow 0$ .

We now give two different sufficient conditions under which vanishing error probability implies vanishing unnormalized equivocation in the general setting of possibly countably infinite random variables.

*Theorem 16:* Suppose there exists a limiting distribution  $P_X$  such that

$$\lim_{n \rightarrow \infty} P_{X_n}(i) = P_X(i) \tag{154}$$

for all  $i$  and

$$\lim_{n \rightarrow \infty} H(X_n) = H(X) \tag{155}$$

is finite. Then

$$\min_f \mathbb{P}[X_n \neq f(Y_n)] \rightarrow 0 \implies H(X_n|Y_n) \rightarrow 0. \tag{156}$$

*Proof:* A simple consequence of Theorem 1 and the following lemma. ■

*Lemma 3:* Suppose there exists a limiting distribution  $P_X$  such that (154) and (155) are satisfied. If

$$\lim_{n \rightarrow \infty} \eta_n = 0 \tag{157}$$

then

$$\lim_{n \rightarrow \infty} \eta_n H(\mathcal{Q}(P_{X_n}, \eta_n)) = 0. \tag{158}$$

*Proof:* For any  $\mathcal{Q}(P_{X_n}, \eta_n) = \{\eta^{-1}q_1^{(n)}, \eta^{-1}q_2^{(n)}, \dots\}$  specified in (7), let

$$Q_n = \left\{ \frac{P_{X_n}(1)}{1 - \eta_n}, \frac{P_{X_n}(2) - q_1^{(n)}}{1 - \eta_n}, \frac{P_{X_n}(3) - q_2^{(n)}}{1 - \eta_n}, \dots \right\} \tag{159}$$

and

$$P_{W_n} = \{0, \eta_n^{-1}q_1^{(n)}, \eta_n^{-1}q_2^{(n)}, \dots\} \tag{160}$$

with

$$H(P_{W_n}) = H(\mathcal{Q}(P_{X_n}, \eta_n)). \tag{161}$$

Then

$$P_{X_n} = \eta_n P_{W_n} + (1 - \eta_n)Q_n \tag{162}$$

and the concavity of entropy implies that

$$H(X_n) \geq (1 - \eta_n)H(Q_n) + \eta_n H(P_{W_n}). \tag{163}$$

Due to (8) and (157)

$$0 \leq \lim_{n \rightarrow \infty} q_i^{(n)} \leq \lim_{n \rightarrow \infty} \eta_n = 0. \tag{164}$$

Together with (154), we have

$$\lim_{n \rightarrow \infty} Q_n(i) = \lim_{n \rightarrow \infty} P_{X_n}(i) = P_X(i) \tag{165}$$

for all  $i$ . Now we can recall that entropy is lower semi-continuous [12] at  $P_X$  with finite entropy and conclude

$$\lim_{n \rightarrow \infty} H(Q_n) \geq H(X). \tag{166}$$

By taking  $n \rightarrow \infty$  on both sides of (163), we apply (155) and (166) to conclude

$$H(X) \geq \lim_{n \rightarrow \infty} ((1 - \eta_n)H(Q_n) + \eta_n H(P_{W_n})) \tag{167}$$

$$\geq H(X) + \lim_{n \rightarrow \infty} \eta_n H(\mathcal{Q}(P_{X_n}, \eta_n)). \tag{168}$$

Therefore, the desired limit (158) is satisfied. ■

Now, we give another sufficient condition for  $H(X_n|Y_n) \rightarrow 0$ .

*Theorem 17:* Suppose that there exists a distribution  $P_X$  with finite entropy which is majorized by  $P_{X_n}$  for all sufficiently large  $n$ . Then

$$\min_f \mathbb{P}[X_n \neq f(Y_n)] \rightarrow 0 \implies H(X_n|Y_n) \rightarrow 0. \tag{169}$$

*Proof:*

$$\lim_{n \rightarrow \infty} \eta_n H(\mathcal{Q}(p_{x_n}, \eta_n)) \tag{170}$$

$$\leq \lim_{n \rightarrow \infty} \eta_n H(\mathcal{Q}(p_x, \eta_n)) \tag{171}$$

$$= 0 \tag{172}$$

where (171) and (172) follow from Lemma 1 and Lemma 3, respectively. Then, we can simply use (15) and Theorem 1 to claim the desired result. ■

Since  $\mathbb{P}[X_n \neq Y_n] \geq \min_f \mathbb{P}[X_n \neq f(Y_n)]$ , we can replace  $\min_f \mathbb{P}[X_n \neq f(Y_n)]$  by  $\mathbb{P}[X_n \neq Y_n]$  in Theorems 16 and 17.

*Theorem 18:* Suppose there exists  $P_X$  with  $H(X) < \infty$  such that:

- equations (154) and (155) are satisfied, or;
- $P_X$  is majorized by  $P_{X_n}$  for all sufficiently large  $n$ .

Then

$$\mathbb{P}[X_n \neq Y_n] \rightarrow 0 \implies H(X_n|Y_n) \rightarrow 0. \quad (173)$$

In Theorems 16–17, two sufficient conditions under which vanishing error probability implies vanishing equivocation have been shown. Together with Theorem 14, vanishing  $\min_f \mathbb{P}[X_n \neq f(Y_n)]$  is equivalent to vanishing  $H(X_n|Y_n)$  under these conditions. We now exhibit several examples where these conditions are satisfied.

*Example 4:* If  $X_n$  is stationary with distribution  $P_X$  and  $H(X) < \infty$ , then (154) and (155) are satisfied.

*Example 5:* If  $X_n$  takes values on a finite alphabet with size  $M$ , then  $P_X = \{\frac{1}{M}, \dots, \frac{1}{M}\}$  is majorized by all  $X_n$ .

*Example 6:* Suppose each  $W_n$  taking values on, possibly countably infinite, alphabets and  $P_{W_n}$  is selected from a finite set  $\mathcal{P} = \{P_{X_1}, P_{X_2}, \dots, P_{X_K}\}$  with  $H(X_i) < \infty$  for  $i \leq K$ . Define  $X^*$  according to (102) and (103) so that for any  $a$ , there exists  $i$  such that  $P_{X^*}(a) \leq P_{X_i}(a)$ . Together with  $\sum_{i=1}^K H(X_i) < \infty$ ,  $H(X^*) < \infty$ . Furthermore,  $P_{X^*}$  is majorized by  $P_{W_n}$  for all  $n$  from Lemma 2.

*Example 7:* If  $P_{X_n}$  is a Poisson (or Geometric) distribution with finite mean  $\lambda_n$  where  $\lambda_n \leq b$  for all  $n$ , choose  $X$  as Poisson (or Geometric) with mean  $b$  so that  $P_X$  is majorized by  $P_{X_n}$  for all  $n$ .

## B. Vanishing Normalized Equivocation

Consider a general discrete source vector  $S^k = (S_1, S_2, \dots, S_k)$ . After going through encoder, (possibly channel) and decoder the reproduced output is  $\hat{S}^k = (\hat{S}_1, \hat{S}_2, \dots, \hat{S}_k)$ . The average symbol error probability is defined as

$$\lambda_k = \frac{1}{k} \sum_{i=1}^k \mathbb{P}[S_i \neq \hat{S}_i] \quad (174)$$

while the block error probability is defined as

$$\mu_k = \mathbb{P}[S^k \neq \hat{S}^k]. \quad (175)$$

We are interested in determining sufficient conditions under which vanishing symbol error probability implies or requires vanishing normalized equivocation.

*Theorem 19:* Assume that  $S_1, \dots, S_k$  are independent. Then

$$\frac{1}{k} H(S^k | \hat{S}^k) \leq \Phi_{S^*}(\lambda_k) \quad (176)$$

where  $S^*$  is constructed from  $\mathcal{P} = \{P_{S_1}, \dots, P_{S_k}\}$  according to (102) and (103).

*Proof:*

$$\frac{1}{K} I(S^k; \hat{S}^k) \geq \frac{1}{K} \sum_{i=1}^k I(S_i; \hat{S}_i) \quad (177)$$

$$= \frac{1}{K} \sum_{i=1}^k H(S_i) - H(S_i | \hat{S}_i) \quad (178)$$

$$\geq \frac{1}{K} \sum_{i=1}^k H(S_i) - \phi_{S_i}(\mathbb{P}[S_i \neq \hat{S}_i]) \quad (179)$$

$$\geq \frac{1}{K} \sum_{i=1}^k H(S_i) - \frac{1}{K} \sum_{i=1}^k \phi_{S^*}(\mathbb{P}[S_i \neq \hat{S}_i]) \quad (180)$$

$$\geq \frac{1}{K} \sum_{i=1}^k H(S_i) - \phi_{S^*}(\lambda_k) \quad (181)$$

where (177) follows from the assumed independence of the  $S_i$ 's, (179) follows from (129) and Theorem 1, (180) follows from Lemma 1 and (181) follows from the concavity of  $\phi_{S^*}$ . ■

A sufficient condition for the reliability in the sense that vanishing average equivocation is equivalent to reliability in the sense of vanishing error probability

$$\lambda'_k = \frac{1}{k} \sum_{i=1}^k \min_f \mathbb{P}[S_i \neq f(\hat{S}_i)] \quad (182)$$

is shown in the next result.

*Theorem 20:* For any general discrete source, if there exists  $S$  such that  $H(S) < \infty$  and  $P_S$  is majorized by  $P_{S_i}$  for all  $i$ , then

$$\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k H(S_i | \hat{S}_i) = 0 \iff \lim_{k \rightarrow \infty} \lambda'_k = 0. \quad (183)$$

*Proof:*

$$\frac{1}{K} \sum_{i=1}^k H(S_i | \hat{S}_i) \quad (184)$$

$$\leq \frac{1}{K} \sum_{i=1}^k \phi_{S_i}(\min_f \mathbb{P}[S_i \neq F(\hat{S}_i)]) \quad (185)$$

$$\leq \frac{1}{K} \sum_{i=1}^k \phi_S(\min_f \mathbb{P}[S_i \neq F(\hat{S}_i)]) \quad (186)$$

$$\leq \phi_S(\lambda'_k), \quad (187)$$

where (185) follows from (131), (186) follows from Lemma 1 and (187) follows from the concavity of  $\phi_S$ . Therefore, the right side of (183) implies the left side. Together with Theorem 11, the theorem is shown. ■

From Theorem 11, we have

$$\lim_{k \rightarrow \infty} H(S^k | \hat{S}^k) = 0 \implies \lim_{k \rightarrow \infty} \mu_k = 0. \quad (188)$$

If  $S^k$  satisfies the conditions in Theorem 18, then

$$\lim_{k \rightarrow \infty} \mu_k = 0 \implies \lim_{k \rightarrow \infty} H(S^k | \hat{S}^k) = 0. \quad (189)$$

In general, the conditions in Theorem 18 may not be satisfied because  $H(S^k)$  may tend to infinity. We now show that the left side of (189) may not imply the right side for an arbitrary  $S^k$ .

*Example 8:* Consider  $S^k$  is uniformly distributed in  $\{0, 1, 2, \dots, A^k\}$  where integer  $A \geq 2$ . Let  $\varepsilon_k = k^{-\frac{1}{2}}$ . Then a joint distribution  $P_{S^k \hat{S}^k}$  can be constructed according to (48) with  $X$  corresponding to  $S^k$  and  $Y$  corresponding to  $\hat{S}^k$  for  $k \geq 2$  so that

$$H(S^k | \hat{S}^k) = \varepsilon_k H(Q(P_{S^k}, \varepsilon_k)) + h(\varepsilon_k) \quad (190)$$

$$= k^{\frac{1}{2}} \log A + h(\varepsilon_k) \quad (191)$$

which does not tend to 0 as  $k \rightarrow \infty$ , but  $\lim_{k \rightarrow \infty} \mu_k = \lim_{k \rightarrow \infty} \varepsilon_k = 0$ . On the other hand

$$\lim_{k \rightarrow \infty} \frac{1}{k} H(S^k | \hat{S}^k) = 0. \quad (192)$$

The relation between vanishing error probability and vanishing normalized equivocation is summarized in the next theorem.

*Theorem 21:* For any general discrete source, if there exists  $S$  such that  $H(S) < \infty$  and  $P_S$  is majorized by  $P_{S_i}$  for all  $i$ , then

$$\lim_{k \rightarrow \infty} \mu_k = 0 \implies \lim_{k \rightarrow \infty} \lambda_k = 0 \quad (193)$$

$$\implies \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k H(S_i | \hat{S}_i) = 0 \quad (194)$$

$$\implies \lim_{k \rightarrow \infty} \frac{1}{k} H(S^k | \hat{S}^k) = 0. \quad (195)$$

*Proof:* Implication (193) is very simple [13]. Similar to the argument from (184) to (187), (193) implies (194). Since

$$\begin{aligned} \frac{1}{k} H(S^k | \hat{S}^k) &= \frac{1}{k} \sum_{i=1}^k H(S_i | S^{i-1} \hat{S}^k) \\ &\leq \frac{1}{k} \sum_{i=1}^k H(S_k | \hat{S}_k) \end{aligned} \quad (196)$$

(194) implies (195). ■

The following example shows that the implications in Theorem 21 do not hold in reverse. ■

*Example 9:* Consider

$$\hat{S}^k = \begin{cases} S^k & \text{with probability } \frac{1}{2} \\ \bar{S}^k & \text{with probability } \frac{1}{2}. \end{cases} \quad (197)$$

Then  $\frac{1}{k} H(S^k | \hat{S}^k) = \frac{1}{k} \rightarrow 0$ , but  $H(S_i | \hat{S}_i) = 1$  for all  $i$  and  $\mu_k = \lambda_k = \frac{1}{2}$ .

## APPENDIX

*Proof of Lemma 1:* It is sufficient to show that  $\mathcal{Q}(P_{X^*}, \eta)$  is majorized by  $\mathcal{Q}(P_X, \eta)$ . Since  $P_{X^*}$  is majorized by  $P_X$ ,  $\eta \leq 1 - P_X(1) \leq 1 - P_{X^*}(1)$  so that  $\mathcal{Q}(P_{X^*}, \eta)$  is well defined. Let  $Q_{X^*}$  and  $Q_X$  be  $\mathcal{Q}(P_{X^*}, \eta)$  and  $\mathcal{Q}(P_X, \eta)$ , respectively. Let  $k$  be the smallest integer such that  $Q_{X^*}(k) = \eta^{-1} P_{X^*}(k+1)$ . Then  $Q_{X^*}(i) = \theta$  for  $i < k$  and  $Q_{X^*}(i) = \eta^{-1} P_{X^*}(i+1)$  for  $i \geq k$ , where  $\theta$  depends on  $\eta$  and  $P_{X^*}$  through (7). For any  $l \geq k$ , it follows from (7) that

$$\eta \sum_{i=l}^{\infty} Q_X(i) \leq \sum_{i=l+1}^{\infty} P_X(i). \quad (198)$$

Together with  $P_{X^*}$  is majorized by  $P_X$ , for  $l \geq k$

$$\sum_{i=l}^{\infty} Q_X(i) \leq \eta^{-1} \sum_{i=l+1}^{\infty} P_{X^*}(i) = \sum_{i=l}^{\infty} Q_{X^*}(i). \quad (199)$$

We now complete the proof by contradiction. Suppose there exists an integer  $l < k$  such that

$$\sum_{i=l}^{\infty} Q_X(i) > \sum_{i=l}^{\infty} Q_{X^*}(i). \quad (200)$$

By putting  $l = k$  into (199) together with (200)

$$\sum_{i=l}^{k-1} Q_X(i) > \sum_{i=l}^{k-1} Q_{X^*}(i) = (k-l)\theta. \quad (201)$$

Since  $Q_X(l) \geq Q_X(i)$  for  $i \in [l, k-1]$ ,  $Q_X(l) > \theta$ . Then

$$Q_X(1) \geq Q_X(2) \geq \dots \geq Q_X(l) > \theta. \quad (202)$$

Therefore

$$\sum_{i=1}^{\infty} Q_X(i) = \sum_{i=1}^{l-1} Q_X(i) + \sum_{i=l}^{\infty} Q_X(i) \quad (203)$$

$$> (l-1)\theta + \sum_{i=l}^{\infty} Q_X(i) \quad (204)$$

$$> (l-1)\theta + \sum_{i=l}^{\infty} Q_{X^*}(i) \quad (205)$$

$$= 1, \quad (206)$$

which contradicts that  $Q_X$  is a probability distribution. Thus

$$\sum_{i=l}^{\infty} Q_X(i) \leq \sum_{i=l}^{\infty} Q_{X^*}(i) \quad (207)$$

for  $l < k$ . Together with (199), the proof is completed. ■

*Proof of Lemma 2:* For any  $k$  and  $a \geq 1$

$$\sum_{i=1}^a P_{X^*}(i) = \min_{P \in \mathcal{P}} \sum_{i=1}^a P(i) \quad (208)$$

$$\leq \sum_{i=1}^a P_{X_k}(i). \quad (209)$$

Furthermore

$$\sum_{i=1}^{\infty} P_{X^*}(i) = \min_{P \in \mathcal{P}} \sum_{i=1}^{\infty} P(i) = 1. \quad (210)$$

In the following, we consider any  $a \geq 1$  and show  $P_{X^*}(a) \geq P_{X^*}(a+1)$ . Let

$$t_a = \arg \min_{P \in \mathcal{P}} \sum_{i=1}^a P(i) \quad (211)$$

where if  $a = 0$ , let  $t_k = 1$ . Let  $j = t_{a-1}$ ,  $k = t_a$  and  $m = t_{a+1}$ . Then

$$P_{X^*}(a+1) = \sum_{i=1}^{a+1} P_{X_m}(i) - \sum_{i=1}^a P_{X_k}(i) \quad (212)$$

$$\leq \sum_{i=1}^{a+1} P_{X_k}(i) - \sum_{i=1}^a P_{X_k}(i) \quad (213)$$

$$= P_{X_k}(a+1) \quad (214)$$

$$\leq P_{X_k}(a) \quad (215)$$

$$= \sum_{i=1}^a P_{X_k}(i) - \sum_{i=1}^{a-1} P_{X_k}(i) \quad (216)$$

$$\leq \sum_{i=1}^a P_{X_k}(i) - \sum_{i=1}^{a-1} P_{X_j}(i) \quad (217)$$

$$= P_{X^*}(a). \quad (218)$$

Therefore,  $P_{X^*}$  is majorized by  $P \in \mathcal{P}$ . Finally, suppose  $P_{X'}$  is majorized by all  $P \in \mathcal{P}$ . For any  $a \geq 1$

$$\sum_{i=1}^a P_{X'}(i) \leq \min_{P \in \mathcal{P}} \sum_{i=1}^a P(i) = \sum_{i=1}^a P_{X^*}(i). \quad (219)$$

Hence,  $P_{X^*}$  majorizes  $P_{X'}$ . ■

#### ACKNOWLEDGMENT

The authors would like to thank R. W. Yeung for valuable comments.

#### REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 623–656, Oct. 1948.
- [2] R. M. Fano, *Class Notes for Transmission of Information, Course 6.574*. Cambridge, MA: MIT, 1952.
- [3] V. A. Kovalevsky, "The problem of character recognition from the point of view of mathematical statistics," in *Character Readers and Pattern Recognition*. New York: Spartan, 1968.
- [4] D. L. Tebbe and S. J. Dwyer, III, "Uncertainty and probability of error," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 516–518, May 1968.
- [5] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Trans. Inform. Theory*, vol. 40, no. 1, pp. 259–266, Jan. 1994.
- [6] S.-W. Ho and R. W. Yeung, "On the discontinuity of the Shannon information measures," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5362–5374, Dec. 2009.
- [7] T. S. Han and S. Verdú, "Generalizing the Fano inequality," *IEEE Trans. Inform. Theory*, vol. 40, no. 7, pp. 1247–1250, Jul. 1994.
- [8] R. W. Yeung, *Information Theory and Network Coding*. Berlin/New York: Springer, 2008.
- [9] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*. New York: Academic, 1979.
- [10] V. Erokhin, "ε-entropy of a discrete random variable," *Theory Probab. Its Applic.*, vol. 3, pp. 97–101, 1958.
- [11] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
- [12] F. Topsøe, "Basic concepts, identities and inequalities—the toolkit of information theory," *Entropy*, vol. 3, pp. 162–190, Sep. 2001.
- [13] S. Verdú, *Class Notes for EE528 Information Theory*. Princeton, NJ: Princeton Univ., 2009.

**Siu-Wai Ho** (S'05–M'07) was born in Hong Kong. He received the B.Eng., M.Phil., and Ph.D. degrees in information engineering from The Chinese University of Hong Kong in 2000, 2003, and 2006, respectively.

During 2006–2008, he was a Postdoctoral Research Fellow in the Department of Electrical Engineering, Princeton University, Princeton, NJ. Since 2009, he has been a Research Fellow in the Institute for Telecommunications Research (ITR), University of South Australia (UniSA), Adelaide, Australia, where he holds the ITR Directors Fellowship. His current research interests include Shannon theory, data communications and recording systems, and biometric security systems.

Dr. Ho was a recipient of the Croucher Foundation Fellowship for 2006/2008, the 2008 Young Scientist Award from the Hong Kong Institution of Science, the UniSA Research SA Fellowship for 2010/2013, and the Australian Research Council Australian Postdoctoral Fellowship for 2010/2013.

**Sergio Verdú** received the telecommunications engineering degree from the Universitat Politècnica de Barcelona, Barcelona, Spain, in 1980, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign in 1984.

Since 1984, he has been a Member of the Faculty of Princeton University, Princeton, NJ, where he is the Eugene Higgins Professor of Electrical Engineering.

Dr. Verdú is the recipient of the 2007 Claude E. Shannon Award and the 2008 IEEE Richard W. Hamming Medal. He is a member of the National Academy of Engineering and was awarded a Doctorate Honoris Causa from the Universitat Politècnica de Catalunya in 2005. He is a recipient of several paper awards from the IEEE: the 1992 Donald Fink Paper Award, the 1998 Information Theory Outstanding Paper Award, an Information Theory Golden Jubilee Paper Award, the 2002 Leonard Abraham Prize Award, the 2006 Joint Communications/Information Theory Paper Award, and the 2009 Stephen O. Rice Prize from IEEE Communications Society. He has also received paper awards from the Japanese Telecommunications Advancement Foundation and from Eurasp. In 1998, Cambridge University Press published his book *Multiuser Detection*, for which he received the 2000 Frederick E. Terman Award from the American Society for Engineering Education. He served as President of the IEEE Information Theory Society in 1997. He is currently Editor-in-Chief of *Foundations and Trends in Communications and Information Theory*.