

Conditional Entropy and Error Probability

Siu-Wai Ho and Sergio Verdú
 Dept. of Electrical Engineering
 Princeton University
 Princeton, NJ 08544, U.S.A.
 Email: {siuho, verdu}@princeton.edu

Abstract—Fano’s inequality relates the error probability and conditional entropy of a finitely-valued random variable X given another random variable Y . It is not necessarily tight when the marginal distribution of X is fixed. In this paper, we consider both finite and countably infinite alphabets. A tight upper bound on the conditional entropy of X given Y is given in terms of the error probability and the marginal distribution of X . A new lower bound on the conditional entropy for countably infinite alphabet is also found. The equivalence of the reliability criteria of vanishing error probability and vanishing conditional entropy is established in wide generality.

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 serves to show that reliability in the sense of vanishing error probability implies reliability in the sense of vanishing equivocation. The fact that reliability in the sense of vanishing equivocation implies reliability in the sense of vanishing error probability was shown by Feder and Merhav in [2]. However, both Fano’s inequality and [2] assume finite alphabets. The results in this paper enable to extend the equivalence of both senses of reliability in the general case of possibly countably infinite alphabets.

The equivalence of 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 $\mathcal{X} = \{1, 2, \dots\}$, let

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

If \mathcal{X} is a finite set, Fano’s inequality [3] relates the conditional entropy of X given Y and the error probability ε by

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

where

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

for $0 < x < 1$ and $h(0) = h(1) = 0$. Now, suppose Y takes values on \mathcal{Y} which is not necessarily equal to \mathcal{X} . Consider

$$\hat{\varepsilon} = \min_{f: \mathcal{Y} \rightarrow \mathcal{X}} \mathbb{P}[X \neq f(Y)] \quad (4)$$

$$= \sum_y P_Y(y) (1 - \max_x P_{X|Y}(x|y)) \quad (5)$$

$$\leq 1 - \max_x P_X(x) \quad (6)$$

where the minimum in (4) is achieved by the maximum a posteriori (MAP) estimator and (6) holds by the suboptimal choice $f(y) = \operatorname{argmax}_x P_X(x)$. Note that (2) still holds if ε is replaced by $\hat{\varepsilon}$ [2].

If \mathcal{X} is countably infinite, (2) no longer gives an upper bound on $H(X|Y)$. In fact, it is possible that $H(X|Y)$ does not tend to 0 as ε tends to zero. This can be explained by the discontinuity of entropy [4] (see also examples in [5] and [6, Example 2.49]). Therefore, it is interesting to explore a generalized Fano’s inequality that can be used to determine under what condition $\varepsilon \rightarrow 0$ implies $H(X|Y) \rightarrow 0$ for countably infinite alphabets. 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

$$\max_{P_{Y|X}: \mathbb{P}[X \neq Y] = \varepsilon} H(X|Y) < \varepsilon \log(|\mathcal{X}| - 1) + h(\varepsilon). \quad (7)$$

This motivates the search for strengthened versions of Fano’s inequality for which the left side of (7) is achieved with equality.

Section II introduces several truncated distributions derived from a given distribution, which are useful throughout the development. Our main results on the upper and lower bounds of conditional entropy are given in Section III, while Section IV shows bounds for the entropy of the conditional distribution. All the logarithms are in the same base unless specified otherwise.

II. TRUNCATED DISTRIBUTIONS

The bounds in this paper are given in terms of various truncated distributions obtained from P_X .¹ For any given P_X and $0 < \eta \leq 1 - P_X(1)$, find a real number θ such that

$$q_i = \begin{cases} \theta & \text{if } P_X(i+1) > \theta \\ P_X(i+1) & \text{if } P_X(i+1) \leq \theta \end{cases} \quad (8)$$

¹For simplicity, we assume $P_X(i) \geq P_X(j)$ if $i < j$ in this paper.

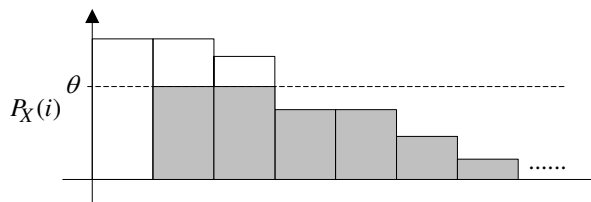


Fig. 1. An example demonstrating the θ and q_i in (8). The shaded region is $\{q_i\}$ and the total area is equal to η .

for $i \geq 1$ and $\eta = \sum_i q_i$. An example is shown in Figure 1. Define the following probability distributions

$$\mathcal{Q}(P_X, \eta) = \{\eta^{-1}q_1, \eta^{-1}q_2, \dots\}, \quad (9)$$

and

$$\mathcal{R}(P_X, \eta) = \{1 - \eta, q_1, q_2, \dots\}. \quad (10)$$

For any given P_X and $0 < \eta \leq 1$, find a real number ν such that

$$s_i = \begin{cases} \nu & \text{if } P_X(i) \geq \nu \\ P_X(i) & \text{if } P_X(i) < \nu \end{cases} \quad (11)$$

for $i \geq 1$ and $\eta = \sum_i s_i$. Define a probability distribution

$$\mathcal{S}(P_X, \eta) = \{\eta^{-1}s_1, \eta^{-1}s_2, \dots\}. \quad (12)$$

The definitions of \mathcal{Q} and \mathcal{S} are very similar except that $P_X(1)$ is not used when we construct \mathcal{Q} .

Suppose P_X and $0 < \eta \leq 1$ are given. Define a probability distribution

$$\mathcal{T}(P_X, \eta) = \{t_1, t_2, \dots, t_K\}. \quad (13)$$

If $\eta \leq P_X(1)$, let $t_1 = 1$ and $t_2 = \dots = t_K = 0$. Otherwise, there exists an integer $K \geq 2$ dependent on both P_X and η such that

$$\sum_{i=1}^{K-1} P_X(i) < \eta \leq \sum_{i=1}^K P_X(i). \quad (14)$$

Let

$$t_i = \begin{cases} \eta^{-1}P_X(i) & \text{if } 1 \leq i \leq K-1 \\ 1 - \eta^{-1} \sum_{i=1}^{K-1} P_X(i) & \text{if } i = K \end{cases} \quad (15)$$

so that $\sum_i t_i = 1$.

In the case of a finite alphabet with cardinality $|\mathcal{X}| = M$ and $\eta > 1 - P_X(M)$, we define

$$\mathcal{U}(P_X, \eta) = \{u_1, u_2, \dots, u_M\} \quad (16)$$

as follows:

$$u_i = \begin{cases} P_X(i) & \text{if } i \leq M-2 \\ P_X(M-1) + P_X(M) - 1 + \eta & \text{if } i = M-1 \\ 1 - \eta & \text{if } i = M. \end{cases} \quad (17)$$

Define

$$\mathcal{W}(P_X) = \{w_1, \dots, w_{l+1}\}, \quad (18)$$

where $l = \lfloor \frac{1}{P_X(1)} \rfloor$ and

$$w_i = \begin{cases} P_X(1) & \text{if } i = 1, \dots, \ell \\ 1 - \ell P_X(1) & \text{if } i = \ell + 1. \end{cases} \quad (19)$$

III. BOUNDS ON CONDITIONAL ENTROPY

Theorem 1: Let X and Y be random variables taking values in the same, possibly countably infinite, alphabet. Suppose $\varepsilon = \mathbb{P}[X \neq Y] \leq 1 - P_X(1)$, then

$$H(X|Y) \leq \varepsilon H(\mathcal{Q}(P_X, \varepsilon)) + h(\varepsilon) \quad (20)$$

Proof:

$$\begin{aligned} H(X|Y) &= H(X) - I(X; Y) \\ &\leq H(X) - \min_{P_{Y|X}: \mathbb{P}[X \neq Y] \leq \varepsilon} I(X; Y) \\ &= \varepsilon H(\mathcal{Q}(P_X, \varepsilon)) + h(\varepsilon), \end{aligned} \quad (21)$$

where (21) holds for $\varepsilon \leq 1 - P_X(1)$ as shown in [7]. ■

The next result generalizes Theorem 1 upper bounding $H(X|Y)$ in terms of $\hat{\varepsilon} = \min_f \mathbb{P}[X \neq f(Y)]$.

Theorem 2: Let X and Y be random variables taking values in the same, possibly countably infinite, alphabet. Then

$$H(X|Y) \leq \hat{\varepsilon} H(\mathcal{Q}(P_X, \hat{\varepsilon})) + h(\hat{\varepsilon}). \quad (22)$$

Proof: Let $f^* = \operatorname{argmin}_{f^*} \mathbb{P}[X \neq f^*(Y)]$ and $Z = f^*(Y)$. Then $\mathbb{P}[X \neq Z] = \hat{\varepsilon} \leq 1 - P_X(1)$ and

$$H(X|Y) = H(X|Y, Z) \quad (23)$$

$$\leq H(X|Z) \quad (24)$$

$$\leq \hat{\varepsilon} H(\mathcal{Q}(P_X, \hat{\varepsilon})) + h(\hat{\varepsilon}), \quad (25)$$

where (25) follows from Theorem 1. ■

Comparing Theorem 1 with Fano's inequality, $\log(|\mathcal{X}| - 1)$ is replaced by $H(\mathcal{Q}(P_X, \varepsilon))$. For $\varepsilon < 1 - P_X(1)$, the upper bound in (20) is tight when $P_{XY}(k, l)$ is given by

$$\begin{cases} (P_X(k) - \theta)\delta_{k,l} + \theta \frac{P_X(l) - \theta}{1 - \theta - \varepsilon} & \text{if } P_X(k) > \theta, P_X(l) > \theta, \\ P_X(k) \frac{P_X(l) - \theta}{1 - \theta - \varepsilon} & \text{if } P_X(k) \leq \theta, P_X(l) > \theta, \\ 0 & \text{if } P_X(l) \leq \theta, \end{cases} \quad (26)$$

where θ depends on ε and P_X through (8). For example, suppose $P_X = \{0.5, 0.3, 0.1, 0.1\}$ and $\varepsilon = 0.4$. Then $\theta = 0.2$. For the joint distribution

$$P_{XY} = \begin{bmatrix} 0.45 & 0.05 & 0 & 0 \\ 0.15 & 0.15 & 0 & 0 \\ 0.075 & 0.025 & 0 & 0 \\ 0.075 & 0.025 & 0 & 0 \end{bmatrix} \quad (27)$$

obtained in (26), we have $\mathbb{P}[X \neq Y] = \varepsilon$ and the equality holds in (20). For $\varepsilon = 1 - P_X(1)$, the upper bound in (20) is tight when $Y = 1$ is a constant.

It is readily checked that for the joint distribution in (26),

$$\min_f \mathbb{P}[X \neq f(Y)] = \mathbb{P}[X \neq Y] = \varepsilon. \quad (28)$$

If $Y = 1$ is a constant, then (28) is also true with $\varepsilon = 1 - P_X(1)$. Hence the bound in (22) is tight. Since the upper bounds on $H(X|Y)$ given in Theorems 1 and 2 are tight, for given P_X and ε , they are in general tighter than Fano's inequality which depends only on ε and on the cardinality of X . If P_X has M atoms and $\varepsilon \leq (M - 1) \min_i P_X(i)$, then the new bounds reduce to Fano's inequality (2).

Theorem 3: Let X and Y be random variables taking values in the same finite alphabet with cardinality M and assume that $\varepsilon = \mathbb{P}[X \neq Y] \geq 1 - P_X(M)$. Then

$$H(X|Y) \leq H(\mathcal{U}(P_X, \varepsilon)), \quad (29)$$

where \mathcal{U} is defined in (16) and the bound is tight in the sense that the left side of (7) is equal to the right side of (29).

Proof: Let $r_{ik} = \mathbb{P}[Y = k|X = i]$ so that

$$H(X|Y) = H(X, Y) - H(Y) \quad (30)$$

$$= - \sum_{ik} (r_{ik} P_X(i)) \log(r_{ik} P_X(i)) + \quad (31)$$

$$\sum_k \left(\sum_i r_{ik} P_X(i) \right) \log \left(\sum_i r_{ik} P_X(i) \right),$$

which is a concave function of r_{ik} . Using Lagrange multipliers, we can solve $\max_{r_{ik}} H(X|Y)$ subject to

$$\sum_i r_{ii} P_X(i) = 1 - \varepsilon, \quad (32)$$

$$\sum_k r_{ik} = 1 \quad \text{for all } i \text{ and,} \quad (33)$$

$$r_{ik} \geq 0 \quad \text{for all } i \text{ and } k. \quad (34)$$

Let $r_{ik}^* = \operatorname{argmax}_{r_{ik}} H(X|Y)$. Then the joint distribution $P_{XY}^*(i, k) = r_{ik}^* P_X(i)$ has the marginal distribution P_Y^* :

$$P_Y^*(1) = P_Y^*(2) = \dots = P_Y^*(M - 2) = 0, \quad (35)$$

$$P_Y^*(M - 1) = \frac{P_X(M) - 1 + \varepsilon}{P_X(M - 1) + P_X(M) - 2 + 2\varepsilon}, \quad (36)$$

and

$$P_Y^*(M) = \frac{P_X(M - 1) - 1 + \varepsilon}{P_X(M - 1) + P_X(M) - 2 + 2\varepsilon}. \quad (37)$$

At the same time, the conditional probability distribution $P_{X|Y}^*(i|k)$ is given by

$$\begin{cases} P_X(i) & \text{if } i \leq M - 2 \\ 1 - \varepsilon & \text{if } i = k \\ P_X(M - 1) + P_X(M) - 1 + \varepsilon & \text{otherwise} \end{cases} \quad (38)$$

for $k = M - 1$ or M . For $k < M - 1$, the value of $P_{X|Y}^*(i|k)$ is unspecified as $P_Y^*(k) = 0$ in (35). Note that

$$H(X|Y = M - 1) = H(X|Y = M) = H(\mathcal{U}(P_X, \varepsilon)), \quad (39)$$

where $\mathcal{U}(P_X, \varepsilon)$ is specified in (16). Together with (35) – (37), we have shown (29) and the bound is tight. ■

Whenever the conditions in either Theorem 1 or Theorem 3 are not satisfied, namely,

$$1 - P_X(1) < \varepsilon < 1 - P_X(M), \quad (40)$$

where if $M = \infty$, the right side of (40) is trivially equal to 1, then we can always find Y such that $\mathbb{P}[X \neq Y] = \varepsilon$ and $H(X|Y) = H(X)$: just take Y independent of X and

$$P_Y(k) = 1 - P_Y(k + 1) = \frac{1 - \varepsilon - P_X(k + 1)}{P_X(k) - P_X(k + 1)}, \quad (41)$$

where k is such that (41) is between 0 and 1.

Next, we give an auxiliary result that will be used to show the sufficient condition for $H(X|Y) \rightarrow 0$.

Lemma 1: For any P_X with finite entropy, we have

$$\lim_{\eta \rightarrow 0} \eta H(\mathcal{Q}(P_X, \eta)) = 0. \quad (42)$$

Proof: For any $\mathcal{Q}(P_X, \eta)$ specified in (9), let

$$P_U = \left\{ \frac{P_X(1)}{1 - \eta}, \frac{P_X(2) - q_1}{1 - \eta}, \frac{P_X(3) - q_2}{1 - \eta}, \dots \right\} \quad (43)$$

and

$$P_W = \{0, \eta^{-1} q_1, \eta^{-1} q_2, \dots\} \quad (44)$$

with $H(P_W) = H(\mathcal{Q}(P_X, \eta))$. Then

$$P_X = \eta P_W + (1 - \eta) P_U \quad (45)$$

and

$$\begin{aligned} H(X) &= H((1 - \eta) P_U + \eta P_W) \\ &\geq (1 - \eta) H(P_U) + \eta H(P_W), \end{aligned} \quad (46)$$

where the last inequality follows from the concavity of entropy. Note that $\mathcal{Q}(P_X, \eta)$ is a function of η , and hence q_i is also a function of η . Since

$$\lim_{\eta \rightarrow 0} P_U(i) = P_X(i) \quad (47)$$

for all i , we can apply that entropy is lower semi-continuous [8] to those P_X with finite entropy and conclude

$$\lim_{\eta \rightarrow 0} H(P_U) \geq H(X). \quad (48)$$

By taking $\eta \rightarrow 0$ on the both sides of (46), we have

$$H(X) \geq \lim_{\eta \rightarrow 0} (1 - \eta) H(P_U) + \lim_{\eta \rightarrow 0} \eta H(P_W), \quad (49)$$

$$\geq H(X) + \lim_{\eta \rightarrow 0} \eta H(\mathcal{Q}(P_X, \eta)) \quad (50)$$

and therefore,

$$\lim_{\eta \rightarrow 0} \eta H(\mathcal{Q}(P_X, \eta)) = 0. \quad (51)$$

Suppose X is defined on a countable alphabet with finite $H(X)$. By taking $\mathbb{P}[X \neq Y] \rightarrow 0$ in Theorem 1, the following theorem follows from Lemma 1 (or from [5, Theorem 6]).

Theorem 4: Suppose X and Y_n take values in the same possibly countably infinite alphabet \mathcal{X} with finite $H(X)$. If

$$\lim_{n \rightarrow \infty} \mathbb{P}[X \neq Y_n] = 0, \quad (52)$$

then

$$\lim_{n \rightarrow \infty} H(X|Y_n) = 0. \quad (53)$$

The converse does not hold. For example, let $Y_n = \bar{X} \in \{0, 1\}$, then $H(X|Y_n) = 0$ and $\mathbb{P}[X \neq Y_n] = 1$. On the other hand, we will show that $\min_f \mathbb{P}[X \neq f(Y_n)] \rightarrow 0$ is a necessary and sufficient condition for $H(X|Y_n) \rightarrow 0$ after the following auxiliary results. In particular, the following lemma leads to a simple proof of the generalization of [2, Lemma 2] to countably infinite alphabets.

Lemma 2:

$$H(X) \geq H(\mathcal{W}(P_X)). \quad (54)$$

Here, $\mathcal{W}(P_X)$ is specified in (18).

Proof: We have

$$\sum_{i=1}^k w_i \geq \sum_{i=1}^k P_X(i) \quad (55)$$

for all k and with equality when $k = \infty$. Therefore, P_X is majorized by $\mathcal{W}(P_X)$. Since entropy is Schur-concave [9], (54) follows. ■

For finite alphabets, a tight lower bound on $H(X|Y)$ has been studied in [2], [10], [11]. In fact, Lemma 2 can extend the proof in [2, Theorem 1] and verify that the lower bound in these papers can also be applied to countably infinite alphabets. In the next theorem, we give a simple lower bound on $H(X|Y)$ which is enough for our purposes.

Theorem 5: For any X taking values in a possibly countably infinite alphabet, we have

$$2 \min_f \mathbb{P}[X \neq f(Y)] \log 2 \leq H(X|Y). \quad (56)$$

Proof: For convenience we assume within the proof that the logarithms are binary. Let

$$\alpha = P_X(1)\ell(\ell + 1) - \ell. \quad (57)$$

Then (18-19) is equal to

$$\begin{aligned} \mathcal{W}(P_X) &= \alpha \cdot \left\{ \frac{1}{\ell}, \frac{1}{\ell}, \dots, \frac{1}{\ell} \right\} + \\ &(1 - \alpha) \cdot \left\{ \frac{1}{\ell + 1}, \frac{1}{\ell + 1}, \dots, \frac{1}{\ell + 1} \right\}. \end{aligned} \quad (58)$$

From the concavity of entropy, we obtain

$$H(\mathcal{W}(P_X)) \geq \alpha \log \ell + (1 - \alpha) \log(\ell + 1) \quad (59)$$

$$\geq 2(1 - P_X(1)), \quad (60)$$

where (60) follows from (57) and the fact that

$$\log \ell \geq 2(1 - \ell^{-1}) \quad (61)$$

for any positive integer ℓ . Together with Lemma 2, we have

$$H(X) \geq 2(1 - P_X(1)). \quad (62)$$

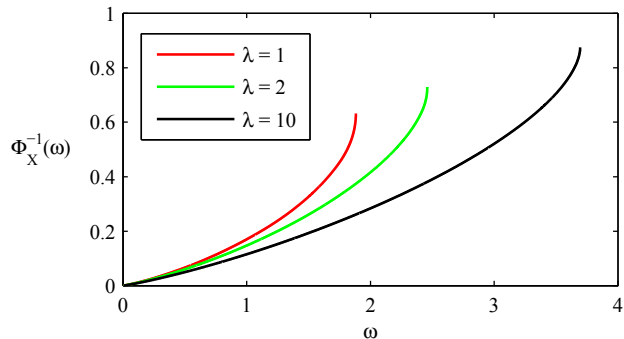


Fig. 2. A plot of $\Phi_X^{-1}(\omega)$ where P_X is a Poisson distribution.

Then

$$\begin{aligned} H(X|Y) &= \sum_y P_Y(y) H(X|Y=y) \\ &\geq \sum_y P_Y(y) \cdot 2 \left(1 - \max_{x'} P_{X|Y}(x'|y) \right) \end{aligned} \quad (63)$$

$$= 2 \min_f \mathbb{P}[X \neq f(Y)], \quad (64)$$

where (63) follows from (62). ■

Theorem 2, Lemma 1 and Theorem 5 enable us to conclude that reliability in the sense of vanishing error probability is equivalent to reliability in the sense of vanishing conditional entropy.

Theorem 6: For X defined on an arbitrary, possibly infinite, alphabet with finite entropy

$$\min_f \mathbb{P}[X \neq f(Y_n)] \rightarrow 0 \iff H(X|Y_n) \rightarrow 0. \quad (65)$$

We now proceed to obtain the tightest lower bound on error probability for a fixed P_X . Let

$$\Phi_X(\hat{\varepsilon}) = H(\mathcal{R}(P_X, \hat{\varepsilon})) \quad (66)$$

where \mathcal{R} is defined in (10). Since entropy is Schur-concave [9] and $\mathcal{R}(P_X, \hat{\varepsilon})$ is majorized by $\mathcal{R}(P_X, \delta)$ if $\hat{\varepsilon} > \delta$, $\Phi_X(\hat{\varepsilon})$ is a strictly increasing function. It can be verified that $\Phi_X(\hat{\varepsilon})$ is continuous and hence, $\Phi_X^{-1}(\cdot)$ exists. From Theorem 2 and $\Phi_X(\hat{\varepsilon}) = \hat{\varepsilon} H(\mathcal{Q}(P_X, \hat{\varepsilon})) + h(\hat{\varepsilon})$, we have

$$\Phi_X^{-1}(H(X|Y)) \leq \min_f \mathbb{P}[X \neq f(Y)]. \quad (67)$$

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

$$\min_{P_{Y|X}: H(X|Y)=\tau} \min_f \mathbb{P}[X \neq f(Y)] = \Phi_X^{-1}(\tau). \quad (68)$$

Although an analytical expression for Φ_X^{-1} is unknown, it can be readily found numerically (see Fig. 2). Note that both $\Phi_X^{-1}(\omega)$ and $\frac{d}{d\omega} \Phi_X^{-1}(\omega)$ are equal to 0 when $\omega = 0$.

As an application of (67), consider a random process $\{X_n\}_{n=-\infty}^{\infty}$, taking values on a finite or countably infinite alphabet. Consider the minimum prediction error

$$\hat{\varepsilon}_n = \min_f \mathbb{P}[X_n \neq f(X^{n-1})], \quad (69)$$

where $X^{n-1} = (X_1, \dots, X_{n-1})$. Then the predictability of the process [2] is defined as

$$\Pi = \lim_{n \rightarrow \infty} \hat{\varepsilon}_n. \quad (70)$$

It easily follows from the continuity of Φ that the entropy rate $\lim_{n \rightarrow \infty} H(X_n | X_1^{n-1})$ and the predictability satisfy

$$\lim_{n \rightarrow \infty} H(X_n | X_1^{n-1}) \leq \Phi_X(\Pi) \quad (71)$$

when the process is stationary, where X stands for the distribution of X_n . According to (68), for any P_X , there exists a process with first-order distribution P_X , whose predictability and entropy rate satisfy (71) with equality.

IV. BOUNDS ON $H(X|Y = y)$

We are going to show the tight bounds on $H(X|Y = y)$ which can be greater or less than $H(X)$. Consider any y such that $P_Y(y) > 0$. Let $\alpha = P_Y(y)$ and let V be a random variable such that $P_V(i) = P_{X|Y}(i|y)$. Let $v_i = P_V(i)$ so that

$$\sum_i v_i = 1, \quad (72)$$

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

$$v_i \leq \alpha^{-1} P_X(i). \quad (73)$$

Theorem 7: For any X taking values in a possibly countably infinite alphabet and $P_Y(y) > 0$,

$$\max_{P_{Y|X}: P_Y(y)=\alpha} H(X|Y = y) = H(\mathcal{S}(P_X, \alpha)), \quad (74)$$

where \mathcal{S} is defined in (12).

Proof: Using Lagrange multipliers, we can find $\max_V H(V)$ subject to the constraints in (72) and (73). The theorem follows from that

$$H(X|Y = y) \leq \max_V H(V) = H(\mathcal{S}(P_X, \alpha)). \quad (75)$$

■

Theorem 8: For X taking values in a possibly countably infinite alphabet and $P_Y(y) > 0$,

$$\min_{P_{Y|X}: P_Y(y)=\alpha} H(X|Y = y) = H(\mathcal{T}(P_X, \alpha)), \quad (76)$$

where \mathcal{T} is defined in (13).

Proof: Let $t_i = 0$ for $i > K$. It is readily checked that for all $k \geq 1$, $\sum_{i=1}^k t_i \geq \sum_{i=1}^k v_i$, where the equality holds when k is equal to the cardinality of X . Therefore, P_V is majorized by $\mathcal{T}(P_X, \alpha)$. Since entropy is Schur-concave [9], $H(\mathcal{T}(P_X, \alpha)) \leq H(P_V) = H(X|Y = y)$, which proves the theorem. ■

Consider the iid independent processes $\{Y_n \in \mathcal{A}\}$, $\{X_n(a), a \in \mathcal{A}\}$ and let $Z_n = X_n(Y_n)$. Suppose that an observer who has access to $\{Z_n\}_{n=1}^{\infty}$ and to $P_Y(a)$ wants to bound $H(X_n(a))$. Then, it can obtain a consistent estimate

of the distribution of Z_n , and apply the bounds (Theorem 7 and 8)

$$H(\mathcal{T}(P_Z, P_Y(a))) \leq H(X_n(a)) \leq H(\mathcal{S}(P_Z, P_Y(a))). \quad (77)$$

ACKNOWLEDGMENT

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

REFERENCES

- [1] C. E. Shannon, The Mathematical Theory of Communication, *Bell Tech. J.*, V. 27, pp.379-423, July 1948.
- [2] M. Feder and N. Merhav, "Relations Between Entropy and Error Probability," *IEEE Trans. Inform. Theory*, vol. 40, pp. 259-266, Jan 1994.
- [3] R. M. Fano, "Class notes for Transmission of Information, Course 6.574," MIT, Cambridge, MA, 1952.
- [4] S.-W. Ho and R. W. Yeung, "On the Discontinuity of the Shannon Information Measures," preprint.
- [5] T.S. Han and S. Verdú, "Generalizing the Fano Inequality," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1247-1250, Jul 1994.
- [6] R. W. Yeung, *A First Course in Information Theory*, Kluwer Academic/Plenum Publishers, 2002.
- [7] V. Erokhin, "ε-entropy of a discrete random variable," *Theory Probab. Its Applic.*, vol. 3, pp. 97-101, 1958.
- [8] F. Topsøe, "Basic Concepts, Identities and Inequalities – the Toolkit of Information Theory," *Entropy*, 3:162-190, Sept. 2001.
- [9] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
- [10] 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.
- [11] 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.