Correspondence

The Asymptotics of Posterior Entropy and Error Probability for Bayesian Estimation

Fumio Kanaya, *Member, IEEE*, and Te Sun Han, *Fellow, IEEE*

***Abstract*— We consider the Bayesian parameter estimation problem where the value of a finitary parameter X should be decided on the basis of i.i.d. sample Y^n of size n. In this context, the amount of missing information on X after observing Y^n may be evaluated by the posterior entropy, which is often called the equivocation or the conditional entropy, of X given Y^n, while it is well known that the minimum possible probability of error in estimating X is achieved by the maximum *a posteriori* probability (MAP) estimator. In this work, the focus is on the asymptotic relation between the posterior entropy and the MAP error probability as the sample size n becomes sufficiently large. It is shown that if the sample size n is large enough, the posterior entropy as well as the MAP error probability decay with n to zero at the identical exponential rate, and that the maximum achievable exponent for this decay is determined by the minimum Chernoff information over all the possible pairs of distinct parameter values. The results presented in this correspondence may be considered as a simpler derivation and also a generalization of the prior work of Rényi, Hellman, and Raviv.**

***Index Terms*— Bayesian parameter estimation, maximum *a posteriori* probability estimator, posterior entropy, equivocation, error probability, Fano's inequality, Chernoff information.**

I. Introduction

Let Y be a random variable ranging over a finite set $\mathcal{Y}$, and suppose its probability distribution depends on a parameter X which takes values also in a finite set $\mathcal{X}$. Let $p(\cdot \mid x)$ be the probability distribution of Y under the hypothesis that $X = x$, and $p(\cdot)$ the prior distribution of X.

Now consider the Bayesian parameter estimation problem where the value of the parameter X should be decided on the basis of sample Y^n of size n, each element of which we assume is drawn independently from the identical distribution of Y given X. Then the amount of missing information on X after observing Y^n may be evaluated by the posterior entropy, which often is called the equivocation or the conditional entropy, of X given Y^n, while it is well known that the minimum possible probability of error in estimating X is attained by the maximum *a posteriori* probability estimator (MAP estimator) that is specified as

$$\hat{x}(y^n) = \arg\max_{x \in \mathcal{X}} p(x \mid y^n) \, \forall y^n \in \mathcal{Y}^n.$$

In this context, the issue of primary concern has been the relation between the posterior entropy $H(X \mid Y^n)$ and the error probability $P_e(X \mid Y^n)$ of the MAP estimator.

It is well known that for every sample size n, the error probability $P_e^{(n)}$ of any possible estimator is lower bounded by Fano's inequality

$$H(X \mid Y^n) \le H_b(P_e^{(n)}) + P_e^{(n)} \log (|\mathcal{X}| - 1) \tag{1}$$

Manuscript received July 7, 1993; revised Mar. 10, 1995. The material in this correspondence was presented in part at ITW'93, Susono-Shi, Japan, June 4–8, 1993.

F. Kanaya is with Shonan Institute of Technology, Fujisawa 251, Japan.

T. S. Han is with the University of Electro-Communications, Chofu, 182 Japan.

IEEE Log Number 9414729.

where H_b denotes the binary entropy function and $|\cdot|$ the cardinality of a set. We denote by log the logarithm to the natural base e. On the other hand, for the *binary* parameter case ($|\mathcal{X}| = 2$), it is proved by Rényi [1] that for every sample size n

$$aH(X \mid Y^n)^2 \le \frac{1}{2} P_e(X \mid Y^n) \le H(X \mid Y^n)$$

where $a \ge 0$ is a constant. Also, for the *binary* case it is proved by Chernoff that for the sufficiently large sample size n

$$P_e(X \mid Y^n) \doteq e^{-nC}$$

where C is the Chernoff information (cf. [2]), and that C is the maximum achievable exponent. We use the notation that $A \doteq e^{-nB}$ to indicate

$$\lim_{n \to \infty} -\frac{1}{n} \log A = B.$$

However, for the case of an arbitrary *finitary* parameter ($|\mathcal{X}| \ge 2$), only an exponential upper bound, though not tight, was given by Arimoto [6, pp. 96–99] as follows: for every sample size n

$$H(X \mid Y^n) \le \frac{|\mathcal{X}|(|\mathcal{X}| - 1)}{2} e^{-nE}$$

$$P_e(X \mid Y^n) \le \frac{|\mathcal{X}|(|\mathcal{X}| - 1)}{2} e^{-nE}$$

where

$$E = \min_{x \ne \hat{x}} \{-\log d(x, \hat{x})\}$$

with $d(x, \hat{x})$ being the Bhattacharyya distance between the two distributions $p(\cdot \mid x)$ and $p(\cdot \mid \hat{x})$. Similar results had also been published by Rényi [3] in a Hungarian journal, and later, Hellman and Raviv [4] extended his results to the case with the Chernoff information. They also proved in [4] for the more than two parameter case ($|\mathcal{X}| \ge 2$) that

$$P_e(X \mid Y^n) \le \frac{1}{2} H(X \mid Y^n) \tag{2}$$

$$H(X \mid Y^n) \le K_4 e^{-nC} \tag{3}$$

where K_4 is a constant and

$$C = \min_{x \ne \hat{x}} C(x, \hat{x})$$

with $C(x, \hat{x})$ being the Chernoff information between the two distributions $p(\cdot \mid x)$ and $p(\cdot \mid \hat{x})$. Although in [4] it is out of concern to address the problem of determining the exact exponent of $P_e(X \mid Y^n)$ and $H(X \mid Y^n)$, one might intuitively conclude that

$$H(X \mid Y^n) \doteq P_e(X \mid Y^n) \doteq e^{-nC} \tag{4}$$

from the upper bounds (2), (3) as well as the fact that $P_e(X \mid Y^n)$ is not smaller than the error probability in the worst binary hypothesis test case over $x \ne \hat{x} \in \mathcal{X}$. However, the fact, which had been overlooked by Rényi, was also revealed in [4] that the Chernoff upper bound (3) on the posterior entropy need not hold unless every $p(\cdot \mid x)$ has the same support for all $x \in \mathcal{X}$. Thus in the general context, (4) still remains to be proved in a more rigorous manner. In order to establish (4) in general, as will be seen in the sequel, we do not have any recourse to upper bounds such as (2) and (3); instead, we

will invoke as a main tool a simple lower bound on $H(X \mid Y^n)$ which holds quite generally without any reservation, and is, in turn, used in combination with Fano's inequality (1) that again holds in full generality. In this respect, the intention of the present work is to rederive the prior results (4) of Rényi, Hellman, and Raviv in a direct and simple fashion for the general case.

It should also be pointed out that recently in related work [5], Feder and Merhav gave their attention to the relation between the entropy of a discrete random variable and the minimum attainable probability of error in guessing its value. They derived a tight upper bound on the MAP error probability in terms of the posterior entropy, which is in a sense the converse result to Fano's inequality.

II. Preliminaries

Before dealing with the general *finitary* parameter case, we briefly describe the Bayesian estimation for the *binary* parameter case. The setup is as follows: $Y_1, Y_2, \cdots, Y_n$ are i.i.d. according to either p_1 or p_2 depending on the hypothesis that $X = x_1$ or that $X = x_2$, respectively. Prior probabilities π_1 and π_2 are assigned to each of the hypotheses. Let $\mathcal{A}_n \subset \mathcal{Y}^n$ be a decision region for the hypothesis that $X = x_1$. Then the Bayesian probability of error is

$$P_e^{(n)} = \pi_1 \sum_{y^n \in \mathcal{A}_n^c} p_1(y^n) + \pi_2 \sum_{y^n \in \mathcal{A}_n} p_2(y^n)$$

where $\cdot^c$ denotes the complement of a set.

Let D^* denote the maximum achievable exponent in the Bayesian probability of error

$$D^* = \lim_{n \to \infty} \max_{\mathcal{A}_n \subset \mathcal{Y}^n} -\frac{1}{n} \log P_e^{(n)}.$$

Let $C(p_1, p_2)$ denote the Chernoff information.

$$C(p_1, p_2) = -\min_{0 \le \lambda \le 1} \log \left(\sum_{y \in \mathcal{Y}} p_1(y)^\lambda p_2(y)^{1-\lambda} \right).$$

Then we have the next theorem.

Theorem 1 (Chernoff Bound):

$$D^* = C(p_1, p_2)$$

which is achieved by adopting the MAP decision region

$$\mathcal{A}_n = \left\{ y^n : \frac{\pi_1 p_1(y^n)}{\pi_2 p_2(y^n)} > 1 \right\}.$$

For a modern proof of this theorem in the light of large deviation theory, refer to Cover and Thomas [2, pp. 312–315].

We now proceed to the case of an arbitrary *finitary* parameter, on which the focus of our work is placed. In this case, it is possible to consider the Chernoff information for each pair of distinct parameter values. For each pair of distinct $x, \hat{x} \in \mathcal{X}$, let

$$C(x, \hat{x}) = -\min_{0 \le \lambda \le 1} \log \left(\sum_{y \in \mathcal{Y}} p(y \mid x)^\lambda p(y \mid \hat{x})^{1-\lambda} \right) \tag{5}$$

and let D be its minimum

$$D = \min_{x \neq \hat{x}} C(x, \hat{x}).$$

Now we present the following two preliminary lemmas that are needed to prove our theorem. To begin with, the next lemma provides an exponential lower bound on the posterior entropy.

Lemma 1: Let δ be an arbitrary positive number. Then there exists an integer $n(\delta)$ such that for every integer $n \ge n(\delta)$

$$H(X \mid Y^n) \ge \rho e^{-n(D+\delta)}$$

where

$$\rho = (\log 2) \min_{x \neq \hat{x}} \{ p(x) + p(\hat{x}) \}.$$

Proof: Starting from the definition of conditional entropy, we have

$$H(X \mid Y^n) = \sum_{x \in \mathcal{X}} p(x) h(x) \tag{6}$$

where

$$h(x) = \sum_{y^n \in \mathcal{Y}^n} p(y^n \mid x) \log \left(1 + \sum_{x' \neq x} \frac{p(y^n \mid x') p(x')}{p(y^n \mid x) p(x)} \right).$$

Now we fix two distinct parameter values x and $\hat{x}$ arbitrarily taken from the set $\mathcal{X}$. It follows from the nonnegativity of $h(x)$

$$H(X \mid Y^n) \ge p(x) h(x) + p(\hat{x}) h(\hat{x}). \tag{7}$$

Define a subset of $\mathcal{Y}^n$ by

$$\mathcal{A}_n(x) = \left\{ y^n : \frac{p(y^n \mid x) p(x)}{p(y^n \mid \hat{x}) p(\hat{x})} > 1 \right\}.$$

Then, using the nonnegativity of $\log(1 + x)$ for $x \ge 0$ as well as its strictly increasing property, we obtain

$$\begin{aligned} h(x) &= \left\{ \sum_{y^n \in \mathcal{A}_n(x)} + \sum_{y^n \in \mathcal{A}_n^c(x)} \right\} p(y^n \mid x) \\ &\quad \cdot \log \left(1 + \sum_{x' \neq x} \frac{p(y^n \mid x') p(x')}{p(y^n \mid x) p(x)} \right) \\ &\ge \sum_{y^n \in \mathcal{A}_n(x)} p(y^n \mid x) \log \left(1 + \frac{p(y^n \mid \hat{x}) p(\hat{x})}{p(y^n \mid x) p(x)} \right). \end{aligned} \tag{8}$$

In the same way, we obtain for $\hat{x}$

$$h(\hat{x}) \ge \sum_{y^n \in \mathcal{A}_n^c(x)} p(y^n \mid \hat{x}) \log \left(1 + \frac{p(y^n \mid x) p(x)}{p(y^n \mid \hat{x}) p(\hat{x})} \right). \tag{9}$$

By using the inequality

$$\log(1 + x) \ge (\log 2) x \quad \forall x \in [0, 1]$$

we obtain from (8) and (9)

$$h(x) \ge (\log 2) \frac{p(\hat{x})}{p(x)} \sum_{y^n \in \mathcal{A}_n(x)} p(y^n \mid \hat{x})$$

$$h(\hat{x}) \ge (\log 2) \frac{p(x)}{p(\hat{x})} \sum_{y^n \in \mathcal{A}_n^c(x)} p(y^n \mid x).$$

Thus by combining the last two inequalities with (7)

$$\begin{aligned} H(X \mid Y^n) \ge (\log 2) \Bigg\{ & p(\hat{x}) \sum_{y^n \in \mathcal{A}_n(x)} p(y^n \mid \hat{x}) \\ & + p(x) \sum_{y^n \in \mathcal{A}_n^c(x)} p(y^n \mid x) \Bigg\}. \end{aligned} \tag{10}$$

Now we define the prior probabilities for the *binary* parameter taking values in $\{x, \hat{x}\}$ as

$$\pi(x) = \frac{p(x)}{p(x) + p(\hat{x})} \quad \text{and} \quad \pi(\hat{x}) = \frac{p(\hat{x})}{p(x) + p(\hat{x})}.$$

Then it follows from (10) that

$$H(X \mid Y^n) \geq (\log 2)\{p(x) + p(\hat{x})\}P_e^{(n)}(x, \hat{x}) \tag{11}$$

where we have put

$$P_e^{(n)}(x, \hat{x}) = \pi(\hat{x}) \sum_{y^n \in \mathcal{A}_n(x)} p(y^n \mid \hat{x}) + \pi(x) \sum_{y^n \in \mathcal{A}_n^c(x)} p(y^n \mid x).$$

Since it is evident from the definition of $\mathcal{A}_n(x)$ that $P_e^{(n)}(x, \hat{x})$ is the error probability for the MAP estimator for the *binary* parameter $\{x, \hat{x}\}$ with *a priori* probability distribution equal to $(\pi(x), \pi(\hat{x}))$, it follows from the above Chernoff bound Theorem 1 that for an arbitrary positive number δ, there exists an integer $n(\delta)$ such that

$$P_e^{(n)}(x, \hat{x}) \geq e^{-n(C(x,\hat{x})+\delta)} \; \forall n \geq n(\delta). \tag{12}$$

Thus by combining (11) and (12),

$$H(X \mid Y^n) \geq (\log 2)\{p(x) + p(\hat{x})\}e^{-n(C(x,\hat{x})+\delta)} \; \forall n \geq n(\delta). \tag{13}$$

Now recall that x and $\hat{x}$ are distinct but arbitrary. Then, by choosing them to attain

$$D = \min_{x \neq \hat{x}} C(x, \hat{x})$$

we obtain from (13)

$$H(X \mid Y^n) \geq (\log 2)\{p(x) + p(\hat{x})\}e^{-n(D+\delta)} \; \forall n \geq n(\delta).$$

Consequently, by defining

$$\rho = (\log 2) \min_{x \neq \hat{x}} \{p(x) + p(\hat{x})\}$$

we obtain the desired result

$$H(X \mid Y^n) \geq \rho e^{-n(D+\delta)} \; \forall n \geq n(\delta).$$

While the last lemma pertains to an asymptotic lower bound on the posterior entropy, the next lemma provides for the MAP error probability $P_e(X \mid Y^n)$ an exponential upper bound that is valid for each sample size n.

Lemma 2: For every positive integer n

$$P_e(X \mid Y^n) \leq \frac{|\mathcal{X}|(|\mathcal{X}| - 1)}{2} e^{-nD}.$$

Proof: First let a positive integer n be arbitrarily fixed, and let the MAP decision region for each $x \in \mathcal{X}$ be denoted by

$$\mathcal{A}_n(x) = \{y^n : p(x)p(y^n \mid x) > p(x')p(y^n \mid x') \; \forall x' \neq x\}.$$

Then we obtain for the MAP error probability

$$P_e(X \mid Y^n) = \sum_{x \in \mathcal{X}} \sum_{y^n \in \mathcal{A}_n^c(x)} p(x)p(y^n \mid x). \tag{14}$$

Now we choose x arbitrarily from the set $\mathcal{X}$ and keep it fixed. Obviously, for each $y^n \in \mathcal{A}_n^c(x)$ there exists a parameter value x' such that $x' \neq x$ and

$$\frac{p(x')p(y^n \mid x')}{p(x)p(y^n \mid x)} \geq 1.$$

Therefore, we have for any $|\mathcal{X}| - 1$ nonnegative numbers $\lambda_{xx'} (x' \neq x)$ that are arbitrarily fixed in the range [0, 1]

$$\sum_{x' \neq x} \left(\frac{p(x')p(y^n \mid x')}{p(x)p(y^n \mid x)} \right)^{1-\lambda_{xx'}} \geq 1 \; \forall y^n \in \mathcal{A}_n^c(x).$$

Hence

$$\begin{aligned}
&\sum_{y^n \in \mathcal{A}_n^c(x)} p(x)p(y^n \mid x) \\
&\leq \sum_{y^n \in \mathcal{A}_n^c(x)} p(x)p(y^n \mid x) \sum_{x' \neq x} \left(\frac{p(x')p(y^n \mid x')}{p(x)p(y^n \mid x)} \right)^{1-\lambda_{xx'}} \\
&= \sum_{x' \neq x} p(x)^{\lambda_{xx'}} p(x')^{1-\lambda_{xx'}} \\
&\quad \cdot \sum_{y^n \in \mathcal{A}_n^c(x)} p(y^n \mid x)^{\lambda_{xx'}} p(y^n \mid x')^{1-\lambda_{xx'}}.
\end{aligned} \tag{15}$$

On the other hand, by the assumption that samples are i.i.d., for each $x \in \mathcal{X}$

$$p(y^n \mid x) = \prod_{i=1}^{n} p(y_i \mid x).$$

Then we obtain

$$\begin{aligned}
&\sum_{y^n \in \mathcal{A}_n^c(x)} p(y^n \mid x)^{\lambda_{xx'}} p(y^n \mid x')^{1-\lambda_{xx'}} \\
&\leq \sum_{y^n \in \mathcal{Y}^n} p(y^n \mid x)^{\lambda_{xx'}} p(y^n \mid x')^{1-\lambda_{xx'}} \\
&= \sum_{y^n \in \mathcal{Y}^n} \prod_{i=1}^{n} p(y_i \mid x)^{\lambda_{xx'}} p(y_i \mid x')^{1-\lambda_{xx'}} \\
&= \prod_{i=1}^{n} \sum_{y_i \in \mathcal{Y}} p(y_i \mid x)^{\lambda_{xx'}} p(y_i \mid x')^{1-\lambda_{xx'}} \\
&= \left(\sum_{y \in \mathcal{Y}} p(y \mid x)^{\lambda_{xx'}} p(y \mid x')^{1-\lambda_{xx'}} \right)^n
\end{aligned} \tag{16}$$

where the first inequality follows from the nonnegativity of $p(y^n \mid x)^{\lambda_{xx'}} p(y^n \mid x')^{1-\lambda_{xx'}}$.

Since $0 \leq \lambda_{xx'} \leq 1$ are arbitrary, we can now choose the $0 \leq \lambda_{xx'}^* \leq 1$ that satisfy

$$\sum_{y \in \mathcal{Y}} p(y \mid x)^{\lambda_{xx'}^*} p(y \mid x')^{1-\lambda_{xx'}^*} = \min_{0 \leq \lambda \leq 1} \sum_{y \in \mathcal{Y}} p(y \mid x)^{\lambda} p(y \mid x')^{1-\lambda}.$$

Thus noting that the Chernoff information $C(x, x')$ between the two parameter values x and x' is defined by

$$C(x, x') = -\min_{0 \leq \lambda \leq 1} \log \left(\sum_{y \in \mathcal{Y}} p(y \mid x)^{\lambda} p(y \mid x')^{1-\lambda} \right)$$

we obtain from (15) and (16)

$$\sum_{y^n \in \mathcal{A}_n^c(x)} p(x)p(y^n \mid x) \leq \sum_{x' \neq x} p(x)^{\lambda_{xx'}^*} p(x')^{1-\lambda_{xx'}^*} e^{-nC(x,x')}. \tag{17}$$

Combining (17) with (14), we obtain for every n

$$P_e(X \mid Y^n) \leq \sum_{x} \sum_{x' \neq x} p(x)^{\lambda_{xx'}^*} p(x')^{1-\lambda_{xx'}^*} e^{-nC(x,x')}.$$

Finally, by letting

$$D = \min_{x \neq x'} C(x, x')$$

and by upper-bounding each $p(x)$ with one, we immediately obtain the desired result

$$P_e(X \mid Y^n) \leq \frac{|\mathcal{X}|(|\mathcal{X}| - 1)}{2} e^{-nD}.$$

III. Theorem and Proof

Now we are in a position to present our theorem. This theorem shows that the minimum Chernoff information over all the possible pairs of distinct parameter values determines the maximum achievable exponent not only for the MAP error probability but also for the posterior entropy. Since obviously the Chernoff information is not less than the Bhattacharyya distance, it is also shown by this theorem that Rényi's exponent is not the best achievable.

Theorem 2: For sufficiently large sample size n

$$H(X \mid Y^n) \doteq P_e(X \mid Y^n) \doteq e^{-nD}.$$

To prove this theorem, we need Fano's inequality as well as two lemmas presented in the previous section. Now we start with the posterior entropy part of this theorem.

Proof of $H(X \mid Y^n) \doteq e^{-nD}$: By Lemma 1, we have for every $n \geq n(\delta)$

$$-\frac{1}{n}\log H(X \mid Y^n) \leq D + \delta - \frac{1}{n}\log \rho.$$

Since $\delta - \frac{1}{n}\log\rho$ can be made arbitrarily small if we take n sufficiently large and let $\delta \to 0$, it follows immediately that

$$\limsup_{n\to\infty} -\frac{1}{n}\log H(X \mid Y^n) \leq D. \tag{18}$$

Here note that

$$H_b(x) \leq -2x\log x \; \forall x \in [0, 1/2]$$

and that $-2x\log x$ is strictly increasing in $[0, 1/e]$. Then, using Fano's inequality as well as Lemma 2, we have for sufficiently large n

$$H(X \mid Y^n) \leq \frac{|\mathcal{X}|(|\mathcal{X}|-1)}{2}\left\{2nD - \log\frac{|\mathcal{X}|^2(|\mathcal{X}|-1)}{4}\right\}e^{-nD}.$$

Hence for n large enough

$$-\frac{1}{n}\log H(X \mid Y^n) \geq D - \frac{1}{n} \cdot \log\left\{\frac{|\mathcal{X}|(|\mathcal{X}|-1)}{2}\left(2nD - \log\frac{|\mathcal{X}|^2(|\mathcal{X}|-1)}{4}\right)\right\}.$$

Since the last term on the right-hand side becomes arbitrarily close to zero as n goes to infinity, we obtain

$$\liminf_{n\to\infty} -\frac{1}{n}\log H(X \mid Y^n) \geq D. \tag{19}$$

Therefore, it follows immediately from (18) and (19) that

$$\lim_{n\to\infty} -\frac{1}{n}\log H(X \mid Y^n) = D.$$

Thus the proof is completed.

Now we proceed to the final task to prove the error probability part of our theorem.

Proof of $P_e(X \mid Y^n) \doteq e^{-nD}$: To prove by contradiction that

$$\limsup_{n\to\infty} -\frac{1}{n}\log P_e(X \mid Y^n) \leq D$$

we assume contrarily that

$$\limsup_{n\to\infty} -\frac{1}{n}\log P_e(X \mid Y^n) > D.$$

Then it follows from this assumption that there exists a strictly positive number α such that

$$\limsup_{n\to\infty} -\frac{1}{n}\log P_e(X \mid Y^n) = D + \alpha.$$

Thus we can choose an increasing subsequence $\{n_i, i = 1, 2, \cdots\}$ of positive integers diverging ultimately to infinity such that there exists an integer N for which we have

$$-\frac{1}{n_i}\log P_e(X \mid Y^{n_i}) \geq D + \frac{\alpha}{2} \; \forall n_i \geq N.$$

Then, noting that this implies $P_e(X \mid Y^{n_i}) \leq 1/e$ for every $n_i \geq N$, we can use Fano's inequality again to obtain for every $n_i > N$

$$-\frac{1}{n_i}\log H(X \mid Y^{n_i}) \geq D + \frac{\alpha}{2} - \frac{1}{n_i}\log\left\{2n_i\left(D + \frac{\alpha}{2}\right) + \log(|\mathcal{X}| - 1)\right\}.$$

Since the last term on the right-hand side of the above inequality gets arbitrarily close to zero as n_i diverge to infinity, we can choose an integer $n(\alpha)$ such that

$$\frac{1}{n_i}\log\left\{2n_i\left(D + \frac{\alpha}{2}\right) + \log(|\mathcal{X}| - 1)\right\} \leq \frac{\alpha}{4} \; \forall n_i \geq n(\alpha).$$

Hence, for every $n_i \geq \max(N, n(\alpha))$

$$-\frac{1}{n_i}\log H(X \mid Y^{n_i}) \geq D + \frac{\alpha}{4}.$$

On the other hand, we are assured by Lemma 1 with $\delta = \frac{\alpha}{5}$ that there exists an integer $n\left(\frac{\alpha}{5}\right)$ such that for every $n \geq n\left(\frac{\alpha}{5}\right)$

$$-\frac{1}{n}\log H(X \mid Y^n) \leq D + \frac{\alpha}{5}.$$

Thus by the last two inequalities we must have for every $n_i \geq \max\left(N, n(\alpha), n\left(\frac{\alpha}{5}\right)\right)$

$$D + \frac{\alpha}{5} \geq -\frac{1}{n_i}\log H(X \mid Y^{n_i}) \geq D + \frac{\alpha}{4}$$

which leads to

$$\frac{\alpha}{5} \geq \frac{\alpha}{4} > 0.$$

Since this is a contradiction, we have just proved

$$\limsup_{n\to\infty} -\frac{1}{n}\log P_e(X \mid Y^n) \leq D.$$

On the other hand, it is an immediate consequence of Lemma 2 that

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_e(X \mid Y^n) \geq D.$$

We thus have proved

$$\lim_{n\to\infty} -\frac{1}{n}\log P_e(X \mid Y^n) = D$$

which completes the proof.

IV. Conclusions

In the foregoing arguments, we have made two finiteness assumptions: the finiteness of the set $\mathcal{X}$ in which the parameter X takes values and the finiteness of the set $\mathcal{Y}$ in which the sample variable Y takes values. The finiteness of $\mathcal{X}$ plays the crucial role in establishing Theorem 2. However, the finiteness of $\mathcal{Y}$ can be dispensed with to establish Theorem 2, as will be seen by scrutinizing the details of the foregoing proofs. $\mathcal{Y}$ may be, for example, the set $\boldsymbol{R}$ of real numbers, in which case $p(\cdot \mid x)$ for each $x \in \mathcal{X}$ stands for the probability density function of Y under the hypothesis that $X = x$.

Acknowledgment

The authors wish to thank two anonymous reviewers whose comments helped improve the original manuscript, especially to one who brought reference [4] to their attention.

REFERENCES

[1] A. Rényi, "On some basic problems of statistics from the point of view of information theory," in *Proc. 5th Berkely Symp. on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 531–543.

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

[3] A. Rényi, "On the amount of information concerning an unknown parameter in a sequence of observations," *Magyar Tud. Akad. Mat. Kutató Int. Köyl*, vol. 9, pp. 617–625, 1965.

[4] M. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 368–372, July 1970.

[5] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Trans. Inform. Theory*, vol. 40, pp. 259–266, Jan. 1994.

[6] S. Arimoto, *Probability, Information and Entropy* (in Japanese). Tokyo: Morikita Pub. Co., 1980.

A Lower Bound on the Probability of Error in Multihypothesis Testing

H. Vincent Poor, *Fellow, IEEE*, and Sergio Verdú, *Fellow, IEEE*

***Abstract*—Consider two random variables X and Y, where X is finitely (or countably-infinitely) valued, and where Y is arbitrary. Let ϵ denote the minimum probability of error incurred in estimating X from Y. It is shown that**

$$\epsilon \geq \sup_{0 \leq \alpha \leq 1} (1-\alpha) P(\pi(X|Y) \leq \alpha)$$

where $\pi(X|Y)$ denotes the posterior probability of X given Y. This bound finds information-theoretic applications in the proof of converse channel coding theorems. It generalizes and strengthens previous lower bounds due to Shannon, and to Verdú and Han.

***Index Terms*—Hypothesis testing, probability of error, Shannon theory, Converse Channel Coding Theorem.**

I. INTRODUCTION

Consider two random variables X and Y, where Y is arbitrary and where X takes values in a finite or countably infinite set $\mathcal{X}$. The minimum-error-probability estimate of X conditioned on the observation of Y is given by

$$\hat{X} = \arg\left\{\max_{k \in \mathcal{X}} \pi(k|Y)\right\} \tag{1}$$

where

$$\pi(k|Y) \triangleq P(X = k|Y). \tag{2}$$

The minimum error probability incurred in testing among the values of X is thus given by

$$\epsilon = P\left(\hat{X} \neq X\right) \equiv 1 - E\left\{\max_{k \in \mathcal{X}} \pi(k|Y)\right\}. \tag{3}$$

Manuscript received September 8, 1994; revised May 16, 1995. This work was supported by the U. S. Army Research Office under Grant DAAH04-93-G-0219.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

IEEE Log Number 9414768.

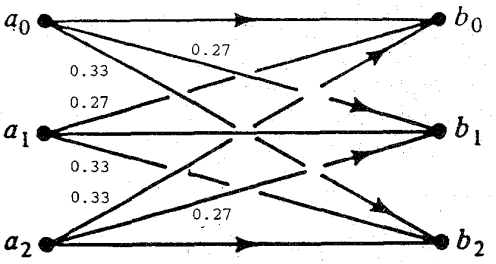

Fig. 1. A ternary hypothesis testing problem with $X \in \{a_0, a_1, a_2\}$ and with a ternary observation $Y \in \{b_0, b_1, b_2\}$.

The maximum occurring in the argument of the expectation of (3) often makes the mimimum error probability ϵ difficult to deal with directly. For this reason, bounds on ϵ are of interest in areas for which multihypothesis testing is of central importance, such as digital communications, information theory, and pattern recognition. Such bounds basically have two uses—computational and analytical. In particular, some error-probability bounds are useful because they are simpler to compute than is the actual error probability, and thereby provide a means for performance prediction in multihypothesis testing; whereas other such bounds are of use because they provide analytically tractable means of assessing the behavior of the error probability as various asymptotes are approached. Bounds of this latter type play an important role in the development and proof of coding theorems, and this correspondence presents a new lower bound of this type.

Two classical lower bounds on the multihypothesis error probability that have found use in proving coding theorems are the Fano and Shannon inequalities, which place lower bounds on ϵ in testing among $|\mathcal{X}| = M < \infty$ *equiprobable* hypotheses. In particular, the Fano inequality (e.g., [1]) is given by

$$\epsilon \geq 1 - \frac{I(X;Y) + \log 2}{\log M} \tag{4}$$

where $I(X;Y)$ denotes the mutual information between X and Y, defined as the expected value (over the joint distribution of X and Y) of the information density

$$i_{XY}(X;Y) = \log \frac{\pi(X|Y)}{P_X(X)} \tag{5}$$

where P_X denotes the probability mass function of X.

The Shannon bound [2] is expressed in terms of the cumulative probability distribution function (cdf) of the information density instead of its average; namely

$$\epsilon \geq \frac{1}{2} P\left(i_{XY}(X;Y) \leq \log \frac{M}{2}\right) \tag{6}$$

$$= \frac{1}{2} P\left(\pi(X|Y) \leq \frac{1}{2}\right) \tag{7}$$

where (7) readily follows from (5) and the assumption that the hypotheses are equiprobable, i.e.

$$P_X(k) = \frac{1}{M}, \quad k \in \mathcal{X}.$$

In the case of nonequiprobable and possibly countably infinitely valued X, the Fano inequality (4) has recently been generalized by Han and Verdú [3]; viz.

$$\epsilon \geq 1 + \frac{I(X;Y) + \log 2}{\log\left(\max_{k \in \mathcal{X}} P_X(k)\right)}. \tag{8}$$