Hypothesis Testing with Communication Constraints

RUDOLF AHLSWEDE AND I. CSISZÁR

Abstract—A new class of statistical problems is introduced, involving the presence of communication constraints on remotely collected data. Bivariate hypothesis testing, H_0: P_{XY} against H_1: $P_{\overline{XY}}$, is considered when the statistician has direct access to Y data but can be informed about X data only at a prescribed finite rate R. For any fixed R the smallest achievable probability of an error of type 2 with the probability of an error of type 1 being at most ϵ is shown to go to zero with an exponential rate not depending on ϵ as the sample size goes to infinity. A single-letter formula for the exponent is given when $P_{\overline{XY}} = P_X \times P_Y$ (test against independence), and partial results are obtained for general $P_{\overline{XY}}$. An application to a search problem of Chernoff is also given.

I. Introduction

IN THE simplest hypothesis testing problem

$$H_0\colon P = (P(x))_{x\in\mathcal{X}},$$
$$H_1\colon Q = (Q(x))_{x\in\mathcal{X}}, \qquad \mathcal{X} \text{ finite},$$

the statistician has to decide on the basis of a sample of size n between H_0 and H_1, of which only one is true. Often his task is to find a test with a minimal probability of an error of type 2 for a prescribed probability of an error of type 1, i.e., to find $B \subset \mathcal{X}^n$ with $P^n(B) \geq 1 - \epsilon$ and $Q^n(B) = \beta(n, \epsilon)$ where $\epsilon \in (0, 1)$ is given and

$$\beta(n, \epsilon) \triangleq \min_A \{Q^n(A) | A \subset \mathcal{X}^n,\ P^n(A) \geq 1 - \epsilon\}.$$

The exponential rate of convergence to zero of $\beta(n, \epsilon)$ as n goes to infinity has been determined by Stein [5].

Stein's Lemma: For any $\epsilon \in (0, 1)$

$$\lim_{n\to\infty} \frac{1}{n} \log \beta(n, \epsilon) = -D(P\|Q).$$

Here

$$D(P\|Q) \triangleq \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

is the familiar Kullback–Leibler informational divergence [11], [12] called simply *divergence* in the sequel.

Manuscript received September 6, 1983; revised October 29, 1985. This work was supported in part by the Deutsche Forschungsgemeinschaft. This paper was presented at the Colloquim on Information Theory, Oberwolfach, Germany, 1982, and at the International Symposium on Information Theory, Tashkent, USSR, 1984.

R. Ahlswede is with the Faklutät für Mathematik der Universität Bielefeld, D-4800 Bielefeld, Universitätstrasse 1, W. Germany.

I. Csiszár was with the University of Bielefeld, Bielefeld, W. Germany. He is now with the Mathematics Institute of the Hungarian Academy of Sciences, H-1364 Budapest, P.O.B. 127, Hungary.

IEEE Log Number 8608093.

It is commonly understood in statistics that the data (samples) are known to the statistician. We add here a new dimension to the problem by assuming that the statistician does not have direct access to the data; rather, he can be informed about them only at a prescribed finite rate. In the problem formulated earlier, this assumption is not a significant constraint if the data are collected at a single location. In fact, the transmission of one bit then enables the statistician to make an optimal decision in the sense of minimizing the probability of an error of type 2 for a prescribed probability of an error of type 1; the information to be transmitted is simply whether or not the observed sample belongs to B as described earlier. New mathematical problems, similar to those in multiuser Shannon theory (see [7]), arise for testing multivariate hypotheses if the different variables are measured at different locations.

In this paper we consider the simplest problem of this kind, namely, bivariate hypothesis testing when one of the variables is measured remotely, and information about it is transmitted over a noiseless channel of finite capacity. Mathematically, we are led to seemingly important connections between statistics and multiuser source coding theory. In another direction, Maljutov and his coworkers [4, appendix] have found connections between the design of screening experiments and multiway channels. A more intensive exchange of ideas between information theory and statistics should extend the frontiers in both areas and give further support to Fisher's thesis that "statistics is data reduction."

Of course, numerous papers are devoted to this general theme. The familiar concept of statistical sufficiency relates to data reduction. Models based on an information-theoretic point of view can be found in [12] and, for instance, also in the work of Perez (cf. [13] and the references therein) where the notion of ϵ-sufficiency plays the role of a measure for data reduction.

The novelty of our approach is to measure data reduction (or compression) by the *rate* needed to transmit the reduced data *and* the performance of the best test based on those data. Let us emphasize that here data compression is meant in a wider sense than in standard source coding or rate distortion theory. In particular, the original data are not required to be recoverable in any sense. Rather, the only requirement on the code, in addition to the rate constraint, is that a good test between the given hypotheses could be constructed based on the encoded

0018-9448/86/0700-0533$01.00

data. An application of our results to a search problem of Chernoff [6] will be discussed in Section V.

II. Statement and Discussion of Results

Throughout this paper we restrict attention to distributions on finite sets. The distribution and joint distribution of the random variables X, Y taking values in finite sets $\mathcal{X}, \mathcal{Y}$ will be denoted by P_X, P_Y, and P_{XY}, respectively. $X^n = (X_1, \cdots, X_n)$ and $Y^n = (Y_1, \cdots, Y_n)$ will denote samples with joint distribution $P_{X^nY^n} \triangleq P_{XY}^n$ where

$$P_{XY}^n(x^n, y^n) \triangleq \prod_{i=1}^n P_{XY}(x_i, y_i),$$
$$x^n = (x_1, \cdots, x_n),\ y^n = (y_1, \cdots, y_n). \quad (2.1)$$

The cardinality of a finite set A and of the range of a function f will be denoted by $|A|$ and $\|f\|$, respectively.

Test Against Independence with One-sided Data Compression

First we will consider a special case for which a complete solution is available, namely, that of testing the hypothesis of a given bivariate distribution P_{XY} against the alternative of independence given by

$$\mathrm{H}_0\colon P_{XY} = \left(P_{XY}(x, y)\right)_{x\in\mathcal{X}, y\in\mathcal{Y}},$$
$$\mathrm{H}_1\colon P_X \times P_Y = \left(P_X(x)P_Y(y)\right)_{x\in\mathcal{X}, y\in\mathcal{Y}}.$$

Notice that while the alternate choice $\mathrm{H}_0 = P_X \times P_Y$, $\mathrm{H}_1 = P_{XY}$ is more frequent in statistics (test *of* independence), our setup (test *against* independence) is also reasonable. Further, it will lead to an interesting application in search theory (cf. Section V). In the present case the divergence appearing in Stein's lemma is equal to the *mutual information* $I(X \wedge Y)$:

$$D(P_{XY}\|P_X \times P_Y) = \sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)}$$
$$= I(X \wedge Y). \quad (2.2)$$

Suppose that the statistician observes Y samples directly and can be informed about X samples indirectly, via encoding functions of rate R, that is, instead of the sample X^n, he is given only $f(X^n)$ where

$$\frac{1}{n} \log \|f\| \le R.$$

Then, for the probability of an error of type 1 not exceeding a fixed $\epsilon \in (0, 1)$, we are interested in the asymptotic behavior of the smallest possible probability of an error of type 2, defined as

$$\beta_R(n, \epsilon) \triangleq \min_f \{\beta(n, \epsilon, f) | \log \|f\| \le nR\} \quad (2.3)$$

where

$$\beta(n, \epsilon, f) \triangleq \min_A \left\{P_{f(X^n)} \times P_{Y^n})(A) | A \subset f(\mathcal{X}^n) \times \mathcal{Y}^n,\right.$$
$$\left. P_{f(X^n)Y^n}(A) \ge 1 - \epsilon\right\}. \quad (2.4)$$

Obviously, β_R is monotonically decreasing in both n and ϵ. Define for $k = 1, 2, \cdots$

$$\theta_k(R) \triangleq \sup_f \left\{\frac{1}{k} D(P_{f(X^k)Y^k} \| P_{f(X^k)} \times P_{Y^k}) | \log \|f\| \le kR\right\} \quad (2.5)$$

and

$$\theta(R) \triangleq \sup_k \theta_k(R). \quad (2.6)$$

Theorem 1: For every $R \ge 0$ we have

a) $\limsup_{n\to\infty} \frac{1}{n} \log \beta_R(n, \epsilon) \le -\theta(R)$ for *all* $\epsilon \in (0, 1)$

b) $\lim_{\epsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log \beta_R(n, \epsilon) \ge -\theta(R)$.

Proof: a) Application of Stein's lemma to

$$\mathrm{H}_0\colon P_{f(X^k)Y^k}, \quad \mathrm{H}_1\colon P_{f(X^k)} \times P_{Y^k}, \quad k \text{ fixed}$$

yields

$$\limsup_{l\to\infty} \frac{1}{lk} \log \beta_R(lk, \epsilon) \le -\theta_k(R)$$

for every $\epsilon \in (0, 1)$. Since for $lk \le n < (l + 1)k$ we have

$$\beta_R((l + 1)k, \epsilon) \le \beta_R(n, \epsilon) \le \beta_R(lk, \epsilon),$$

it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_R(n, \epsilon) \le -\theta_k(R) \quad (2.7)$$

for every $\epsilon \in (0, 1)$. Since k was arbitrary, this proves assertion a).

b) For every function f defined on $\mathcal{X}^n$ and every $A \subset f(\mathcal{X}^n) \times \mathcal{Y}^n$, we have

$$D(P_{f(X^n)Y^n} \| P_{f(X^n)} \times P_{Y^n}) \ge \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta} \quad (2.8)$$

where

$$\alpha \triangleq P_{f(X^n)Y^n}(A), \quad \beta \triangleq \left(P_{f(X^n)} \times P_{Y^n}\right)(A). \quad (2.9)$$

By (2.3) and (2.4) we can choose f and A such that

$$\log \|f\| \le nR \quad \alpha \ge 1 - \epsilon \quad \beta = \beta_R(n, \epsilon).$$

Then (2.5), (2.6), and (2.8) give

$$\theta(R) \ge \theta_n(R) \ge \frac{1}{n} D(P_{f(X^n)Y^n} \| P_{f(X^n)} \times P_{Y^n})$$
$$\ge -\frac{1 - \epsilon}{n} \log \beta_R(n, \epsilon) - \frac{h(\alpha)}{n}$$

where

$$h(\alpha) \triangleq -\alpha \log \alpha - (1 - \alpha) \log (1 - \alpha).$$

This completes the proof.

Remark 1: An implicit assumption underlying the definition (2.3) of $\beta_R(n, \epsilon)$ is that *any* encoding function f of a rate not exceeding R can be used to transmit information

about the X sample. It might be more realistic to restrict attention to the block codes of block length k much less than the sample size n, i.e., to functions f obtained by concatenation from a function f_k defined on $\mathcal{X}^k$ as

$$f(x_1, \cdots, x_n) \triangleq \big(f_k(x_1, \cdots, x_k), \cdots, f_k(x_{(l-1)k+1}, \cdots, x_{lk})\big), \qquad lk \le n < (l+1)k$$

where $\log \|f_k\| \le kR$. Such a restriction has, however, no significant effect on the result. In fact, if f is so restricted, part a) of Theorem 1 holds with $\theta_k(R)$ instead of $\theta(R)$ by the same proof. This is an arbitrarily small difference if a sufficiently large k is admitted. Of course, the converse part b) is not affected by an additional restriction on f. Moreover, under this restriction the converse can be obtained directly from Stein's lemma, even in the strong form (for every $\epsilon \in (0,1)$ rather than for $\epsilon \to 0$).

Next we consider two questions: 1) how can one give a single-letter characterization of the quantity $\theta(R)$ and 2) can one improve Theorem 1 to the statement

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_R(n, \epsilon) = -\theta(R)$$

for *all* $\epsilon \in (0,1)$? The answer to the first question can be obtained as an immediate consequence of the Ahlswede–Körner solution [1] to the problem of source coding with side information. In fact, by (2.2) we have

$$\frac{1}{k} D\big(P_{f(X^k)Y^k} \| P_{f(X^k)} \times P_{Y^k}\big) = \frac{1}{k} I\big(f(X^k) \wedge Y^k\big) = H(Y) - \frac{1}{k} H\big(Y^k | f(X^k)\big);$$

thus $\theta(R)$ defined by (2.5) and (2.6) can be written as

$$\theta(R) = H(Y) - \inf_{k,f} \left\{ \frac{1}{k} H\big(Y^k | f(X^k)\big) \,\middle|\, \log \|f\| \le kR \right\}. \tag{2.10}$$

The problem of giving a single-letter characterization of the infimum in (2.10) is a special case of "entropy characterization problems" playing a fundamental role in multiterminal source-coding theory (cf. [7]). The solution to this problem was a key step in [1]; the infimum appearing in (2.10) was shown to equal the infimum of $H(Y|U)$ for all random variables U such that $U \ominus X \ominus Y$ (ie., U, X, Y form a Markov chain) and $I(U \wedge X) \le R$. Moreover, here the range $\mathcal{U}$ of U may be supposed to satisfy the constraint $|\mathcal{U}| \le |\mathcal{X}| + 1$. Thus we obtain Theorem 2 from (2.10).

Theorem 2: For every $R \ge 0$

$$\theta(R) = \max_U \{ I(U \wedge Y) | U \ominus X \ominus Y, \; I(U \wedge X) \le R, |\mathcal{U}| \le |\mathcal{X}| + 1 \}.$$

Later we also prove that the answer to the second question formulated earlier is "yes," even in a more general context (cf. Theorem 6). Thus the following sharpening of Theorem 1 is true.

Theorem 3: For every $R \ge 0$ and $\epsilon \in (0,1)$

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_R(n, \epsilon) = -\theta(R).$$

Remark 2: The reader might wonder what happens if not only the X data but also the Y data are compressed. Application of Stein's lemma leads to the following problem: find a single-letter characterization for

$$\theta(R_X, R_Y) \triangleq \sup_{k,f,g} \left\{ \frac{1}{k} I\big(f(X^k) \wedge g(Y^k)\big) \,\middle|\, \log \|f\| \le kR_X, \log \|g\| \le kR_Y \right\}.$$

This problem appears to be of formidable mathematical complexity.

General Bivariate Hypotheses with One-Sided Data Compression

Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ and $\{(\bar{X}_i, \bar{Y}_i)\}_{i=1}^{\infty}$ be two sequences of independent pairs of random variables having bivariate distribution P_{XY} and $P_{\bar{X}\bar{Y}}$, respectively. The hypotheses to be tested are that the first resp. second sequence is being observed:

$$\mathrm{H}_0: P_{XY}, \qquad \mathrm{H}_1: P_{\bar{X}\bar{Y}}.$$

As before, we suppose that the statistician observes Y samples directly and can be informed about X samples indirectly, via encoding functions of rate R. Generalizing (2.3) and (2.4), define

$$\beta(n, \epsilon, f) \triangleq \min_A \{ P_{f(\bar{X}^n)\bar{Y}^n}(A) | A \subset f(\mathcal{X}^n) \times \mathcal{Y}^n, \; P_{f(X^n)Y^n}(A) \ge 1 - \epsilon \} \tag{2.11}$$

$$\beta_R(n, \epsilon) \triangleq \min_f \{ \beta(n, \epsilon, f) | \log \|f\| \le nR \}. \tag{2.12}$$

Again, we are interested in the limiting behavior of $\beta_R(n, \epsilon)$, the smallest probability of an error of type 2 achievable when X data are compressed to rate R and the permissible probability of an error of type 1 is ϵ. Similarly to (2.5) and (2.6), we now define

$$\theta_k(R) \triangleq \sup_f \left\{ \frac{1}{k} D\big(P_{f(X^k)Y^k} \| P_{f(\bar{X}^k)\bar{Y}^k}\big) \,\middle|\, \log \|f\| \le kR \right\} \tag{2.13}$$

and

$$\theta(R) \triangleq \sup_k \theta_k(R). \tag{2.14}$$

Theorem 4: Both assertions of Theorem 1 remain valid for the present β_R and $\theta(R)$.

Proof: The proof of Theorem 1 literally applies to this more general case.

Notice that Remark 1 also applies in the present situation. Some simple properties of $\theta(R)$ are stated in Lemma 1.

Lemma 1: a) $\theta(R) = \lim_{k\to\infty} \theta_k(R)$, $R \ge 0$; b) $\theta(R)$ is monotonically increasing and concave for $R \ge 0$; and c) $\theta(R)$ is continuous for *positive* R.

Proof: a) By time-sharing we get the subadditivity property

$$(k+l)\theta_{k+l}(R) \geq k\theta_k(R) + l\theta_l(R).$$

This implies assertion a).

b) The monotonicity of θ_k extends to the limit θ. Further, again by time-sharing, for $R_1, R_2 \geq 0$ and every k

$$\theta_{2k}\left(\frac{R_1+R_2}{2}\right) \geq \frac{1}{2}(\theta_k(R_1)+\theta_k(R_2)).$$

Hence, also in the limit

$$\theta\left(\frac{R_1+R_2}{2}\right) \geq \frac{1}{2}(\theta(R_1)+\theta(R_2)).$$

c) A concave function can have discontinuities only at the boundary.

Remark 3: Clearly, $\theta(0) = D(P_Y\|P_{\overline{Y}})$. Since for $R > 0$, $D(P_X\|P_{\overline{X}})$ contributes to $\theta(R)$, this function does have a discontinuity at $R = 0$, at least if

$$D(P_X\|P_{\overline{X}}) > D(P_Y\|P_{\overline{Y}}).$$

Now we turn to the two questions formulated after Remark 1. As to the first one, in the present more general case we have only a partial result, a single-letter lower bound to $\theta(R)$.

Theorem 5: For $R > 0$ let U be any random variable satisfying $I(U \wedge X) \leq R$ and the Markov condition $U \leftrightarrow X \leftrightarrow Y$. Then for $\theta(R)$ defined by (2.13) and (2.14), we have

$$\theta(R) \geq D(P_X\|P_{\overline{X}}) + D(P_{UY}\|P_{U\tilde{Y}})$$

where $\tilde{Y}$ denotes a random variable with $U \leftrightarrow X \leftrightarrow \tilde{Y}$ whose conditional distribution given X is the same as that of $\overline{Y}$ given $\overline{X}$.

Corollary: For every $R > 0$,

$$\theta(R) \geq D(P_X\|P_{\overline{X}}) + D(P_Y\|P_{\overline{Y}}).$$

We originally believed that the lower bound in Theorem 5 was tight when optimized for U. Unfortunately, this is not generally true (cf. a counterexample in Section III). Still, the bound is tight for $R \geq H(X)$. In fact, $U = X$ may then be taken, and by the easily checked identity

$$D(P_X\|P_{\overline{X}}) + D(P_{XY}\|P_{X\overline{Y}}) = D(P_{XY}\|P_{\overline{XY}})$$

we get $\theta(R) \geq D(P_{XY}\|P_{\overline{XY}})$. Clearly, the strict inequality is impossible here. Of course, it is intuitively obvious that the rate constraint does not matter when $R \geq H(X)$, and it is easy to prove directly that

$$\lim_{n\to\infty} \frac{1}{n}\log \beta_R(n,\epsilon) = -D(P_{XY}\|P_{\overline{XY}}), \qquad \text{if } R \geq H(X).$$

The answer to the second question is positive.

Theorem 6: For β_R and $\theta(R)$ defined by (2.11)–(2.14) we have

$$\lim_{n\to\infty} \frac{1}{n}\log \beta_R(n,\epsilon) = -\theta(R)$$

for all $R \geq 0$ and $\epsilon \in (0,1)$, provided that $P_{\overline{XY}}(x,y) > 0$ for every $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

Actually, we expect this result to hold in general but do not have yet a complete proof without the positivity assumption on $P_{\overline{XY}}$. Notice that Theorem 6 implies Theorem 3. In fact, for $P_{\overline{XY}} = P_X \times P_Y$ our positivity assumption reduces to

$$P_X(x) > 0, \qquad P_Y(y) > 0 \text{ for all } x \in \mathcal{X},\ y \in \mathcal{Y},$$

and this can be assumed without restricting generality.

Theorems 5 and 6 will be proved in Sections III and IV. The proofs are rather similar and rely on techniques familiar in multiuser Shannon theory. In particular, we will use the covering lemma from [3] and the blowing up lemma from [2]. The missing step to a complete solution of our problem appears to be comparable in difficulty to problems arising in multiterminal source coding. As discussed in [7], one encounters entropy characterization problems whose prototype was the one solved in [1]. The problem we are facing now, namely, that of getting a single-letter characterization of $\theta(R)$ defined by (2.13) and (2.14), is similar in nature and may be termed a *divergence characterization problem*.

The independence of the limit in Theorem 6 of the error threshold ϵ is a "strong converse" in the terminology of the Shannon theory. It is remarkable that it could be proved without having a single-letter formula for the limit because in the literature of the Shannon theory, strong converses are not available for problems to which a single-letter solution is not known.

III. Lower Bound to $\theta(R)$

First we recall some basic facts about types and typical sequences. The *type* P_{x^n} of a sequence $x^n = (x_1,\cdots,x_n) \in \mathcal{X}^n$ is a distribution on $\mathcal{X}$ where $P_{x^n}(x)$ is the relative frequency of x in x^n. The *joint type* P_{x^n,y^n} of two sequences $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$ is a distribution on $\mathcal{X}\times\mathcal{Y}$, defined similarly. We denote by $\mathcal{P}_n$ the set of all possible types of sequences $x^n \in \mathcal{X}^n$, and for a given $P \in \mathcal{P}_n$, we denote by $\mathcal{V}_n(P)$ the set of all stochastic matrices $V = (V(y|x))_{x\in\mathcal{X},\,y\in\mathcal{Y}}$ such that

$$V(y|x) \in \left\{0, \frac{1}{nP(x)}, \frac{2}{nP(x)}, \cdots, \right\}, \qquad \text{for all } x \in \mathcal{X},\ y \in \mathcal{Y}.$$

For $P \in \mathcal{P}_n$

$$\mathcal{T}_P^n \triangleq \{x^n | P_{x^n} = P\} \tag{3.1}$$

denotes the set of sequences of type P in $\mathcal{X}^n$, and for $x^n \in \mathcal{X}^n$, $V \in \mathcal{V}_n(P_{x^n})$,

$$\mathcal{T}_V^n(x^n) \triangleq \{y^n | P_{x^n,y^n}(x,y) = P_{x^n}(x)V(x|y) \qquad \text{for all } x \in \mathcal{X},\ y \in \mathcal{Y}\} \tag{3.2}$$

denotes the set of sequences in $\mathcal{Y}^n$ that are *V-generated* by x^n.

Given a random variable X and a positive number η, we call $P \in \mathscr{P}_n$ an (X, η)-*essential type* if

$$\max_x |P(x) - P_X(x)| \le \eta,$$

$$P(x) = 0 \text{ whenever } P_X(x) = 0. \quad (3.3)$$

The conditional distribution of a random variable Y given X is the stochastic matrix $P_{Y|X}$, defined by

$$P_{Y|X}(y|x) \triangleq \Pr\{Y = y | X = x\}$$

(and arbitrary if $P_X(x) = 0$).

For $x^n \in \mathscr{X}^n$ with $P_X^n(x^n) > 0$, we call $V \in \mathscr{V}_n(P_{x^n})$ $(x^n, Y|X, \eta)$-essential if

$$\max_{x,y} |P_{x^n}(x)V(y|x) - P_{x^n}(x)P_{Y|X}(y|x)| \le \eta,$$

$$V(y|x) = 0 \text{ whenever } P_{Y|X}(y|x) = 0. \quad (3.4)$$

The set of (X, η)-*typical* sequences in $\mathscr{X}^n$ and the set of sequences in $\mathscr{Y}^n$ $(Y|X, \eta)$-*generated* by x^n are defined by

$$\mathscr{T}_{X,\eta}^n \triangleq \bigcup_{(X,\eta)\text{-ess. } P} \mathscr{T}_P^n$$

$$\mathscr{T}_{Y|X,\eta}^n(x^n) \triangleq \bigcup_{(x^n, Y|X, \eta)\text{-ess. } V} \mathscr{T}_V^n(x^n). \quad (3.5)$$

We will use the following well-known facts (see, for example, [7, sec. 1.2]). In (3.7) $\vartheta_n(P)$ denotes an "exponentially negligible" factor or, more exactly,

$$(n+1)^{-|\mathscr{X}|} \le \vartheta_n(P) \le 1.$$

Similarly, in (3.8)

$$(n+1)^{-|\mathscr{X}||\mathscr{Y}|} \le \vartheta_n(P, V) \le 1.$$

It follows that

$$|\mathscr{P}_n| \le (n+1)^{|\mathscr{X}|}, \qquad |\mathscr{V}_n(P)| \le (n+1)^{|\mathscr{X}||\mathscr{Y}|}, \quad (3.6)$$

$$|\mathscr{T}_P^n| = \vartheta_n(P)\exp[nH(P)], \qquad P \in \mathscr{P}_n, \quad (3.7)$$

and

$$|\mathscr{T}_V^n(x^n)| = \vartheta_n(P, V)\exp[nH(V|P)]; \qquad x^n \in \mathscr{T}_P^n,\ V \in \mathscr{V}_n(P), \quad (3.8)$$

where

$$H(V|P) \triangleq \sum_x P(x)H(V(\cdot|x)) = -\sum_{x,y} P(x)V(y|x)\log V(y|x). \quad (3.9)$$

Further,

$$\Pr\{X^n \in \mathscr{T}_{X,\eta}^n\} \ge 1 - \frac{|\mathscr{X}|}{4n\eta^2} \quad (3.10)$$

and

$$\Pr\{Y^n \in \mathscr{T}_{Y|X,\eta}^n(x^n) | X^n = x^n\} \ge 1 - \frac{|\mathscr{X}||\mathscr{Y}|}{4n\eta^2}, \quad \text{if } P_X^n(x^n) > 0. \quad (3.11)$$

As immediate consequences of (3.7) and (3.8), we also have for any sequence of independent and identically distributed (i.i.d.) pairs $\{(\bar{X}_i, \bar{Y}_i)\}_{i=1}^{\infty}$ with generic distribution $P_{\bar{X}\bar{Y}}$,

$$\Pr\{\bar{X}^n \in \mathscr{T}_P^n\} = \vartheta_n(P)\exp[-nD(P\|P_{\bar{X}})], \qquad P \in \mathscr{P}_n \quad (3.12)$$

and

$$\Pr\{\bar{Y}^n \in \mathscr{T}_V^n(x^n) | \bar{X}^n = x^n\} = \vartheta_n(P, V)\exp[-nD(V\|P_{\bar{Y}|\bar{X}}|P)] \quad \text{if } P_X^n(x^n) > 0,\ x^n \in \mathscr{T}_P^n,\ V \in \mathscr{V}_n(P) \quad (3.13)$$

where

$$D(V\|P_{\bar{Y}|\bar{X}}|P) = \sum_x P(x)D(V(\cdot|x)\|P_{\bar{Y}|\bar{X}}(\cdot|x)) = \sum_{x,y} P(x)V(y|x)\log\frac{V(y|x)}{P_{\bar{Y}|\bar{X}}(y|x)}. \quad (3.14)$$

We notice that if $P_{XY}(x, y) = 0$ whenever $P_{\bar{X}\bar{Y}}(x, y) = 0$, then (3.12) and (3.13) imply, by continuity and (3.6), that to any $\delta > 0$, $\eta_0 > 0$ an n_0 exists such that for $0 < \eta < \eta_0$ and $n \ge n_0$

$$\exp[-n(D(P_X\|P_{\bar{X}}) + \delta)] \le \Pr\{\bar{X}^n \in \mathscr{T}_{X,\eta}^n\} \le \exp[-n(D(P_X\|P_{\bar{X}}) - \delta)] \quad (3.15)$$

unless $\mathscr{T}_{X,\eta}^n = \varnothing$ (which may happen if $\eta < 1/n$) and

$$\exp[-n(D(P_{Y|X}\|P_{\bar{Y}|\bar{X}}|P_X) + \delta)] \le \Pr\{\bar{Y}^n \in \mathscr{T}_{Y|X,\eta}^n(x^n) | \bar{X}^n = x^n\} \le \exp[-nD(P_{Y|X}\|P_{\bar{Y}|\bar{X}}|P_X) - \delta)] \quad (3.16)$$

for every (X, η)-typical $x^n \in \mathscr{X}^n$, unless $\mathscr{T}_{Y|X,\eta}^n = \varnothing$.

As a final preparation to the proof of Theorem 5 (as well as of Theorem 6), we state a covering lemma from [3, Part 2]. For any permutation π of the integers $1, \cdots, n$ and $x^n = (x_1, \cdots, x_n) \in \mathscr{X}^n$, we write

$$\pi(x^n) = (x_{\pi(1)}, \cdots, x_{\pi(n)})$$

$$\pi(C) = \{\pi(x^n) | x^n \in C\}, \qquad C \subset \mathscr{X}^n.$$

Covering Lemma: For any type $P \in \mathscr{P}_n$, set $C \subset \mathscr{T}_P^n$, and integer $N > |C|^{-1}|\mathscr{T}_P^n|\log|\mathscr{T}_P^n|$, N permutations $\pi_1, \cdots, \pi_N$ exist of the integers $1, \cdots, n$ such that

$$\bigcup_{i=1}^N \pi_i(C) = \mathscr{T}_P^n.$$

Now we can prove the following.

Proposition 1: Suppose sets $C \subset \mathscr{X}^n$ and $D \subset \mathscr{Y}^n$ exist such that for each $x^n \in C$

$$\Pr\{Y^n \in D | X^n = x^n\} \ge 1 - \frac{\epsilon}{2}$$

$$\Pr\{\bar{Y}^n \in D | \bar{X}^n = x^n\} \le \gamma \quad (3.17)$$

and

$$|C \cap \mathscr{T}_P^n| > \exp[n(H(X) - R + \delta)] \quad \text{for each } (X, \eta)\text{-essential } P \in \mathscr{P}_n \quad (3.18)$$

where $\eta = (|\mathscr{X}|/2n\epsilon)^{1/2}$. Then $\beta_R(n, \epsilon)$ defined by (2.11)

and (2.12) satisfies

$$\beta_R(n,\epsilon) \le \gamma \exp\left[-n\left(D(P_X\|P_{\bar{X}}) - \delta\right)\right] \quad (3.19)$$

provided that $n \ge n_0(\delta,\epsilon)$.

Proof: Apply the covering lemma to $C \cap \mathcal{T}_P^n$ in the role of C, for each (X,η)-essential $P \in \mathcal{P}_n$. Since for such P (3.7) and (3.8) imply (if n is sufficiently large)

$$|C \cap \mathcal{T}_P|^{-1}|\mathcal{T}_P^n| \log |\mathcal{T}_P^n| < \exp\left[n(R - \delta/2)\right] - 1$$

for each (X,η)-essential P, one can select permutations $\pi_{1,P}, \cdots, \pi_{N,P}$ such that

$$\bigcup_{i=1}^{N} \pi_{i,P}(C) \supset \mathcal{T}_P^n \qquad N \le \exp\left[n(R - \delta/2)\right]. \quad (3.20)$$

Let $\pi_1, \cdots, \pi_M$ be all the permutations so selected as P runs over the (X,η)-essential types. Then (3.20) implies by (3.5) and (3.6) that

$$\bigcup_{i=1}^{M} \pi_i(C) \supset \mathcal{T}_{X,\eta}^n \qquad M \le (n+1)^{|\mathcal{X}|}\exp\left[n(R - \delta/2)\right]. \quad (3.21)$$

Now consider the function $f: \mathcal{X}^n \to \{0,1,\cdots,M\}$ defined by

$$f(x^n) \triangleq \begin{cases} 0, & \text{if } x^n \notin \mathcal{T}_{X,\eta}^n \\ \text{smallest } i \text{ with } x^n \in \pi_i(C), & \text{if } x^n \in \mathcal{T}_{X,\eta}^n \end{cases}. \quad (3.22)$$

Then $\|f\| \le \exp(nR)$ (if n is sufficiently large); thus (3.19) will be proved (cf. (2.11) and (2.12)) if we find $A \subset \{0,1,\cdots,M\} \times \mathcal{Y}^n$ such that

$$P_{f(X^n)Y^n}(A) \ge 1 - \epsilon$$

$$P_{f(\bar{X}^n)\bar{Y}^n}(A) \le \gamma \exp\left[-n\left(D(P_X\|P_{\bar{X}}) - \delta\right)\right]. \quad (3.23)$$

We claim that

$$A \triangleq \bigcup_{i=1}^{M} \{i\} \times \pi_i(D)$$

satisfies (3.23). To see this, notice that by construction

$$P_{f(X^n)Y^n}(A) = \Pr\left\{(f(X^n), Y^n) \in A\right\} = \sum_{i=1}^{M} \sum_{x^n \in f^{-1}(i)} P_X^n(x^n) \cdot \Pr\left\{Y^n \in \pi_i(D) | X^n = x^n\right\}. \quad (3.24)$$

A significant observation is that

$$\Pr\left\{Y^n \in \pi_i(D) | X^n = x^n\right\} = \Pr\left\{Y^n \in D | X^n = \pi_i^{-1}(x^n)\right\}$$

because of the i.i.d. property of $\{(X_i, Y_i)\}_{i=1}^{\infty}$. By (3.22), for $x^n \in f^{-1}(i)$, $i = 1,\cdots,M$ we have $\pi_i^{-1}(x^n) \in C$. Thus by assumption (3.17) the last conditional probability is at least $1 - \epsilon/2$. Hence (3.24) gives

$$P_{f(X^n)Y^n}(A) \ge P_X^n\left(\bigcup_{i=1}^{M} f^{-1}(i)\right)(1 - \epsilon/2) = P_X^n(\mathcal{T}_{X,\eta}^n)(1 - \epsilon/2). \quad (3.25)$$

This establishes the first part of (3.23) because of (3.10) and the definition $\eta \triangleq (|\mathcal{X}|/2n\epsilon)^{1/2}$.

The same reasoning that led to (3.25) also gives

$$P_{f(\bar{X}^n)\bar{Y}^n}(A) \le P_{\bar{X}}^n(\mathcal{T}_{X,\eta}^n)\gamma. \quad (3.26)$$

Hence by (3.15) and (3.6) we get the second part of (3.23), completing the proof of Proposition 1.

Proof of Theorem 5: We will show that Proposition 1 implies

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_{R'}(n,\epsilon) \le -D(P_X\|P_{\bar{X}}) - D(P_{UY}\|P_{U\bar{Y}}) \quad (3.27)$$

for every $\epsilon \in (0,1)$ and $R' > R$. Because of Theorem 4 and Lemma 1 c), this will prove Theorem 5.

Let $\mathcal{U}$ designate the set of possible values of U; we may suppose that $P_U(u) > 0$ for each $u \in \mathcal{U}$. Fix $\eta > 0$ sufficiently small as specified later. Pick for every n a $u^n \in \mathcal{T}_{U,\eta}^n$, and set

$$C_n \triangleq \mathcal{T}_{X|U,\eta}^n(u^n), \qquad D_n \triangleq \mathcal{T}_{Y|U,\eta^*}^n(u^n), \qquad \eta^* \triangleq (|\mathcal{X}| + 1)\eta. \quad (3.28)$$

Then

$$\mathcal{T}_{Y|UX,\eta}^n(u^n, x^n) \subset D_n, \qquad \text{if } x^n \in C_n.$$

Hence by the Markov property $U \ominus X \ominus Y$ and by (3.11)

$$\begin{aligned} &\Pr\left\{Y^n \in D_n | X^n = x^n\right\} \\ &\quad = \Pr\left\{Y^n \in D_n | U^n = u^n, X^n = x^n\right\} \\ &\quad \ge \Pr\left\{Y^n \in \mathcal{T}_{Y|UX,\eta}^n(u^n, x^n) | U^n = u^n, X^n = x^n\right\} \\ &\quad \ge 1 - \frac{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}{4n\eta^2}, \qquad \text{if } x^n \in C_n. \end{aligned} \quad (3.29)$$

Notice further that $\Pr\{\bar{Y}^n \in D_n | \bar{X}^n = x^n\}$ is constant for $x^n \in \mathcal{T}_V^n(u^n)$ if V is fixed; denote its value by γ_V. Thus by the Markov property $U \ominus X \ominus \tilde{Y}$ and the identity $P_{\bar{Y}|\bar{X}} = P_{\tilde{Y}|X}$, we have

$$\begin{aligned} &\Pr\left\{\tilde{Y}^n \in D_n | U^n = u^n\right\} \\ &\quad = \sum_{x^n} \Pr\left\{\bar{Y}^n \in D_n | \bar{X}^n = x^n\right\} \Pr\left\{X^n = x^n | U^n = u^n\right\} \\ &\quad \ge \gamma_V \Pr\left\{X^n \in \mathcal{T}_V^n(u_n) | U^n = u^n\right\}. \end{aligned} \quad (3.30)$$

Fixing an arbitrary $\delta > 0$, (3.13) gives for each $(u^n, X|U,\eta)$-essential V and n sufficiently large that

$$\Pr\left\{X^n \in \mathcal{T}_V^n(u^n) | U^n = u^n\right\} \ge \exp(-n\delta)$$

provided that η has been chosen sufficiently small. Similarly,

$$\Pr\left\{\tilde{Y}^n \in D_n | U^n = u^n\right\} \le \exp\left[-n\left(D(P_{Y|U}\|P_{\tilde{Y}|U}|P_U) - \delta\right)\right]$$

by (3.16). Thus (3.30) gives for each $(x^n, X|U,\eta)$-essential V

$$\gamma_V \le \exp\left[-n\left(D(P_{Y|U}\|P_{\tilde{Y}|U}|P_U) - 2\delta\right)\right],$$

that is

$$\Pr\{\bar{Y}^n \in D_n | \bar{X}^n = x^n\} \le \exp\left[-n\left(D(P_{Y|U}\|P_{\bar{Y}|U}|P_U) - 2\delta\right)\right], \quad \text{if } x^n \in C_n. \tag{3.31}$$

To apply Proposition 1, we still need (3.18) for C_n and R' in the role of C and R, recalling that the η of (3.18) is $\eta_n \triangleq (|\mathcal{X}|/2n\epsilon)^{1/2}$ rather then our present fixed η. Clearly, for n sufficiently large, $u^n \in \mathcal{T}^n_{U,\eta}$ can be selected in such a way that the types of the sequences $x^n \in \mathcal{T}^n_{X|U,\eta}(u^n)$ include all (X, η_n)-essential types $P \in \mathcal{P}_n$, that is, to each (X, η_n)-essential P a $(u^n, X|U, \eta)$-essential V exists such that $\mathcal{T}^n_V(u^n) \subset \mathcal{T}^n_P$. Then by (3.8)

$$|C_n \cap \mathcal{T}^n_P| \ge |\mathcal{T}^n_V(u^n)| \ge (n+1)^{-|\mathcal{Y}||\mathcal{X}|}\exp\left[nH(V|P_u n)\right]. \tag{3.32}$$

Since u^n is (U, η)-typical and V is $(u^n, X|U, \eta)$-essential, here $H(V|P_{u^n}) \ge H(X|U) - \delta$ if η has been suitably chosen. Thus using the assumption $I(U \wedge X) \le R$, (3.32) gives for a sufficiently large n

$$\begin{aligned}|C_n \cap \mathcal{T}^n_P| &\ge \exp\left[n(H(X|U) - 2\delta)\right] \\ &\ge \exp\left[n(H(X) - R - 2\delta)\right] \\ &\ge \exp\left[n(H(X) - R' + \delta)\right], \quad \text{if } R' > R + 3\delta.\end{aligned} \tag{3.33}$$

The relations (3.29), (3.31), and (3.33) show that Proposition 1 is applicable, yielding

$$\beta_{R'}(n, \epsilon) \le \exp\left[-n\left(D(P_{Y|U}\|P_{\bar{Y}|U}|P_U) - 2\delta\right)\right] \cdot \exp\left[-n\left(D(P_X\|P_{\bar{X}}) - \delta\right)\right] \tag{3.34}$$

if n is sufficiently large and $R' > R + 3\delta$. Since here

$$\begin{aligned}&D(P_{Y|U}\|P_{\bar{Y}|U}|P_U) \\ &\quad = \sum_{u,y} P_U(u)P_{Y|U}(y|u)\log\frac{P_{Y|U}(y|u)}{P_{\bar{Y}|U}(y|u)} \\ &\quad = \sum_{u,y} P_{UY}(u,y)\log\frac{P_{UY}(u,y)}{P_{U\bar{Y}}(u,y)} \\ &\quad = D(P_{UY}\|P_{U\bar{Y}}),\end{aligned}$$

and $\delta > 0$ is arbitrarily small, (3.34) proves (3.27) and thereby Theorem 5. Finally, we show that the lower bound in Theorem 5 is, in general, not tight.

Example: Let $\mathcal{X} = \mathcal{Y} = \{0,1\}$, $P_X(x) = P_{\bar{X}}(x) = 1/2$ for $x \in \mathcal{X}$,

$$P_{Y|X}(y|x) \triangleq W(y|x) = \begin{cases}1, & x = y \\ 0, & x \ne y\end{cases},$$

and

$$P_{\bar{Y}|\bar{X}}(y|x) \triangleq W_\epsilon(y|x) = \begin{cases}1-\epsilon, & x = y \\ \epsilon, & x \ne y\end{cases}.$$

Choose $R = 1/2$. Then for $k = 2$, taking the indicator function of $A = \{(0,0),(1,1)\}$ as f, (2.13) gives

$$\begin{aligned}\theta_2\left(\frac{1}{2}\right) &\ge \frac{1}{8}\sum_{y^2}\sum_{x^2\in A} W(y^2|x^2)\log\frac{\sum_{x^2\in A}W(y^2|x^2)}{\sum_{x^2\in A}W_\epsilon(y^2|x^2)} \\ &\quad + \frac{1}{8}\sum_{y^2}\sum_{x^2\in A^c} W(y^2|x^2)\log\frac{\sum_{x^2\in A^c}W(y^2|x^2)}{\sum_{x^2\in A^c}W_\epsilon(y^2|x^2)} \\ &= -\frac{1}{8}\Big[\log\sum_{x^2\in A} W_\epsilon(00|x^2) + \log\sum_{x^2\in A} W_\epsilon(11|x^2) \\ &\quad + \log\sum_{x^2\in A^c} W_\epsilon(10|x^2) + \log\sum_{x^2\in A^c} W_\epsilon(01|x^2)\Big] \\ &= -\frac{1}{2}\log\left((1-\epsilon)^2 + \epsilon^2\right).\end{aligned}$$

Thus for the choice $\epsilon = 1/4$ we have

$$\begin{aligned}\theta\left(\tfrac{1}{2}\right) \ge \theta_2\left(\tfrac{1}{2}\right) &\ge -\tfrac{1}{2}\log\left((1-\epsilon)^2 + \epsilon^2\right) \\ &= \tfrac{1}{2} - \tfrac{1}{2}\log 5 \sim 0.339,\end{aligned}$$

whereas our computer value for $\max D(P_{UY}\|P_{U\bar{Y}})$ subject to the constraints in Theorem 5 is ~ 0.187. This value is assumed already for $|\mathcal{U}| = 2$ for the parameters

$$\mathcal{P}_{X|U} = \begin{pmatrix}\alpha & 1-\alpha \\ 0 & 1\end{pmatrix}, \quad \alpha \sim 0.773 \quad P_U(0) = (2\alpha)^{-1}.$$

IV. Independence of ϵ of the Exponent

In this section we will prove Theorem 6, using Proposition 1 and the blowing up lemma [2]. We state the latter in its uniform version, [7, lemma 1.5.4], although for the present purpose the original version in [2] would suffice.

Blowing Up Lemma

To any finite sets $\mathcal{X}, \mathcal{Y}$ and sequence $\epsilon_n \to 0$, a sequence of positive integers l_n with $l_n/n \to 0$ and a sequence $\gamma_n \to 1$ exist such that for any i.i.d. sequence of pairs of random variables (X_i, Y_i) with values in $\mathcal{X}$ and $\mathcal{Y}$ and for every n, $x^n \in \mathcal{X}^n$, $F \subset \mathcal{Y}^n$

$$\Pr\{Y^n \in F | X^n = x^n\} \ge \exp\{-n\epsilon_n\}$$

implies

$$\Pr\{Y^n \in \Gamma^{l_n}F | X^n = x^n\} \ge \gamma_n.$$

Here $\Gamma^l F$ denotes the Hamming l-neighborhood of F, i.e.,

$$\Gamma^l F \triangleq \{\bar{y}^n \in \mathcal{Y}^n | d_H(y^n, \bar{y}^n) \le l \text{ for some } y^n \in F\},$$

$$d_H(y^n, \bar{y}^n) \triangleq |\{i: y_i \ne \bar{y}_i\}|,$$

$$y^n = (y_1, \cdots, y_n), \quad \bar{y}^n = (\bar{y}_1, \cdots, \bar{y}_n).$$

To prove Theorem 6, only the case $R > 0$ has to be considered. Because of Theorem 4 and Lemma 1 c), it suffices to prove the following.

Proposition 2: Under the assumption of Theorem 6, to any $0 < \lambda < \epsilon < 1$ and $\alpha > 0$ an n_0 exists such that

$$\frac{1}{n}\log\beta_R(n,\epsilon) \geq \frac{1}{n}\log\beta_{R'}(n,\lambda) - \alpha$$

whenever $n \geq n_0$ and $R' \geq R + \alpha$.

Proof: Consider a function f defined on $\mathcal{X}^n$ with $\log\|f\| \leq nR$ and a set $A \subset f(\mathcal{X}^n) \times \mathcal{Y}^n$ such that

$$P_{f(X^n)Y^n}(A) \geq 1 - \epsilon \qquad P_{f(\bar{X}^n)\bar{Y}^n}(A) = \beta_R(n,\epsilon) \quad (4.1)$$

(cf. (2.11) and (2.12)). We may assume that the range of f is $f(\mathcal{X}^n) = \{1,\cdots,M\}$

$$M \leq \exp(nR). \quad (4.2)$$

Thus

$$A = \bigcup_{i=1}^{M} \{i\} \times G_i, \qquad G_i \subset \mathcal{Y}^n,\ i = 1,\cdots,M.$$

Then (4.1) means that

$$\Pr\{Y^n \in G_{f(X^n)}\} \geq 1 - \epsilon$$
$$\Pr\{\bar{Y}^n \in G_{f(\bar{X}^n)}\} = \beta_R(n,\epsilon). \quad (4.3)$$

Fix $\delta \in (0,(1-\epsilon)/2)$ and take $\eta \triangleq n^{-1/3}$. We first show that a set $E \subset \mathcal{X}^n$ exists with

$$\Pr\{X^n \in E\} \geq \frac{1-\epsilon}{2M}, \qquad E \subset \mathcal{T}^n_{X,\eta} \quad (4.4)$$

such that for every $x^n \in E$ $f(x^n) = i_0$ (for example) and with $F \triangleq G_{i_0}$

$$\Pr\{Y^n \in F | X^n = x^n\} \geq \delta \quad (4.5)$$

and

$$\Pr\{\bar{Y}^n \in F | \bar{X}^n = x^n\} \leq \beta_R(n,\epsilon)\exp\left[n\left(D(P_X\|P_{\bar{X}}) + 2\delta\right)\right]. \quad (4.6)$$

Next we will "blow up" E and F to obtain sets C and D satisfying the hypotheses of Proposition 1, with λ and R' in the roles of ϵ and R and with

$$\gamma = \beta_R(n,\epsilon)\exp\left[n\left(D(P_X\|P_{\bar{X}}) + 4\delta\right]\right.$$

Then the proof will be completed by application of Proposition 1.

For $x^n \in \mathcal{T}^n_{X,\eta}$ write

$$s(x^n) \triangleq \Pr\{Y^n \notin G_{f(x^n)} | X^n = x^n\}$$
$$t(x^n) \triangleq \frac{\Pr\{\bar{X}^n = x^n\}}{\Pr\{X^n = x^n\}}\Pr\{\bar{Y}^n \in G_{f(x^n)} | \bar{X}^n = \bar{x}^n\}$$

and set

$$B \triangleq \{x^n \in \mathcal{T}^n_{X,\eta} | s(x^n) \leq 1 - \delta, \quad t(x^n) \leq \beta_R(n,\epsilon)\exp(n\delta)\}.$$

Since by (4.3)

$$\sum_{x^n \in \mathcal{T}^n_{X,\eta}} \Pr\{X^n = x^n\}s(x^n) = \Pr\{X^n \in \mathcal{T}^n_{X,\eta}, Y^n \notin G_{f(X^n)}\} \leq \epsilon,$$

$$\sum_{x^n \in \mathcal{T}^n_{X,\eta}} \Pr\{X^n = x^n\}t(x^n) = \Pr\{\bar{X}^n \in \mathcal{T}^n_{X,\eta}, \bar{Y}^n \in G_{f(\bar{X}^n)}\} \leq \beta_R(n,\epsilon),$$

we get—using (3.10)—that

$$\Pr\{X^n \in B\} \geq \Pr\{X^n \in \mathcal{T}^n_{X,\eta}\} - \frac{\epsilon}{1-\delta} - \exp(-n\delta) \geq \frac{1-\epsilon}{2} \quad (4.7)$$

if n is sufficiently large.

Let $i_0 \in \{1,\cdots,M\}$ be a value of f maximizing $\Pr\{X^n \in B,\ f(X^n) = i\}$, and set $E \triangleq B \cap f^{-1}(i_0)$. Then (4.7) implies (4.4), and by the definition of B we have for every $x^n \in E$ with $F \triangleq G_{i_0}$

$$\Pr\{Y^n \in F | X^n = x^n\} = 1 - s(x^n) \geq \delta \quad (4.8)$$

$$\Pr\{\bar{Y}^n \in F | \bar{X}^n = x^n\} = \frac{\Pr\{X^n = x^n\}}{\Pr\{\bar{X}^n = x^n\}}t(x^n) \leq \frac{\Pr\{X^n = x^n\}}{\Pr\{\bar{X}^n = x^n\}}\beta_R(n,\epsilon)\cdot\exp(n\delta). \quad (4.9)$$

Here (4.8) is just the desired (4.5), while (4.9) implies (4.6) because for $x^n \in \mathcal{T}^n_{X,\eta}$ with $\eta \triangleq n^{-1/3}$

$$\frac{\Pr\{X^n = x^n\}}{\Pr\{\bar{X}^n = x^n\}} = \prod_{x\in\mathcal{X}}\left(\frac{P_X(x)}{P_{\bar{X}}(x)}\right)^{nP_{x^n}(x)} = \exp\left[n\sum_{x\in\mathcal{X}}P_{x^n}(x)\log\frac{P_X(x)}{P_{\bar{X}}(x)}\right] \leq \exp\left[n\left(D(P_X\|P_{\bar{X}}) + \delta\right)\right]$$

if n is sufficiently large.

Now we blow up E and F and take

$$C \triangleq \Gamma^k E \cap \mathcal{T}^n_{X,\eta} \qquad D \triangleq \Gamma^{k+l}F \quad (4.10)$$

with k and l to be specified later. We then check the hypotheses of Proposition 1 for C and D in (4.10).

Notice first that (4.4) implies

$$|E \cap \mathcal{T}^n_{\bar{P}}| \geq \frac{1-\epsilon}{2M}|\mathcal{T}^n_{\bar{P}}| \quad \text{for some } (X,\eta)\text{-essential } \bar{P} \in \mathcal{P}_n. \quad (4.11)$$

Let $P \in \mathcal{P}_n$ be any other (X,η)-essential type. Then

$$\max_x |P(x) - \bar{P}(x)| \leq 2\eta = 2n^{-1/3},$$

hence for $k \triangleq [2n^{2/3}|\mathcal{X}\|]$, say, the Hamming k-neighborhood of every $x^n \in \mathcal{T}^n_{\bar{P}}$ intersects $\mathcal{T}^n_P$. Since for this k the cardinality of the Hamming k-neighborhood of an $x^n \in \mathcal{X}^n$ is less than $\exp(n\delta)$ if n is sufficiently large, then

$$|E \cap \mathcal{T}^n_{\bar{P}}| \leq |\Gamma^k(\Gamma^k E \cap \mathcal{T}^n_P)| \leq |\Gamma^k E \cap \mathcal{T}^n_P|\exp(n\delta).$$

Hence by (4.10), (4.11), and (3.7) we obtain

$$|C \cap \mathcal{T}_P^n| = |\Gamma^k E \cap \mathcal{T}_P^n| \geq \frac{1}{M}\exp\left[n(H(X) - 2\delta)\right]$$

for every (X, η)-essential $P \in \mathcal{P}_n$ (4.12)

if n is sufficiently large.

Consider now any $\bar{x}^n = (\bar{x}_1, \cdots, \bar{x}_n) \in C$ and pick an $x^n = (x_1, \cdots, x_n) \in E$ with $d_H(x^n, \bar{x}^n) \leq k$. Then to each $y^n = (y_1, \cdots, y_n) \in F$ take a $\bar{y}^n = (\bar{y}_1, \cdots, \bar{y}_n) \in \Gamma^k F$ such that $\bar{y}_i = y_i$ if $x_i = \bar{x}_i$ and $\bar{y}_i$ maximizes $P_{Y|X}(y|\bar{x}_i)$ otherwise. Then, clearly,

$$\Pr\{Y^n = \bar{y}^n | X^n = \bar{x}^n\} \geq |\mathcal{Y}|^{-k} \Pr\{Y^n = y^n | X^n = x^n\}.$$

Since for fixed $\bar{x}^n$ and x^n at most $|\mathcal{Y}|^k$ different $y^n \in F$ can lead to the same $\bar{y}^n \in \Gamma^k F$, it follows that

$$\Pr\{Y^n \in \Gamma^k F | X^n = \bar{x}^n\} \geq |\mathcal{Y}|^{-2k} \Pr\{Y^n \in F | X^n = x^n\}.$$

On account of (4.5), this gives

$$\Pr\{Y^n \in \Gamma^k F | X^n = \bar{x}^n\} \geq |\mathcal{Y}|^{-2k}\delta,$$

for every $\bar{x}^n \in C$. (4.13)

Since $k = [2n^{2/3}|\mathcal{X}|]$, the right side of (4.13) can be written as $\exp(-n\epsilon_n)$ with $\epsilon_n \to 0$. Take l_n and γ_n to these ϵ_n in the blowing up lemma; in particular, for any fixed $\xi > 0$ and n sufficiently large, $l_n < \xi n$ and $\gamma_n > 1 - (\lambda/2)$. It follows that for sufficiently large n an $l < \xi n$ exists such that (4.13) implies

$$\Pr\{Y^n \in \Gamma^{k+l} F | X^n = \bar{x}^n\} \geq 1 - \frac{\lambda}{2},$$

for every $\bar{x}^n \in C$. (4.14)

Finally, for any $\bar{x}^n \in C$ and $x^n \in E$ with $d_H(x^n, \bar{x}^n) \leq k$, assign to each $\bar{y}^n \in \Gamma^{k+l} F$ a $y^n \in F$ with $d_H(y^n, \bar{y}^n) \leq k + l$. Then

$$\Pr\{\bar{Y}^n = y^n | \bar{X}^n = \bar{x}^n\} \leq \Pr\{\bar{Y}^n = y^n | \bar{X}^n = x^n\} p^{-(2k+l)} \quad (4.15)$$

where

$$p \triangleq \min_{x,y} P_{\bar{Y}|\bar{X}}(y|x) > 0;$$

(this is where we need the positivity hypothesis of Theorem 6).

By our choice of k and l, here $p^{-(2k+l)} \leq \exp(n\delta)$, and also the number of different $\bar{y}^n \in \Gamma^{k+l} F$ to which the same $y^n \in F$ is assigned is less than $\exp(n\delta)$ if n is large, provided that $\xi > 0$ has been chosen sufficiently small. Thus (4.15) and (4.6) give

$$\Pr\{\bar{Y}^n \in \Gamma^{k+l} F | X^n = \bar{x}^n\} \leq \Pr\{Y^n \in F | \bar{X}^n = x^n\} \exp(2n\delta) \leq \beta_R(n, \epsilon) \exp\left[nD(P_X \| P_{\bar{X}}) + 4\delta)\right],$$

for every $\bar{x}^n \in C$. (4.16)

Equations (4.12), (4.14), and (4.16) mean that the hypotheses of Proposition 1 are fulfilled for C and D in (4.10), with λ in the role of ϵ, with $\gamma = \beta_R(n, \epsilon)\exp[n(D(P_X \| P_{\bar{X}}) + 4\delta)]$, and any $R' > R + 3\delta$ in the role of R (recall (4.2), and that our present $\eta \triangleq n^{-1/3}$ is larger than the η in (3.18)).

Thus Proposition 1 gives

$$\beta_{R'}(n, \lambda) \leq \beta_R(n, \epsilon)\exp(5n\delta)$$

if n is sufficiently large and $R' > R + 3\delta$. This completes the proof of Proposition 2 and thereby of Theorem 6.

V. Identification in a Large Population

Chernoff [6] suggested the following model for the identification of an element of a large population in the presence of noise. Suppose that N items $X_1, \cdots, X_N$ are stored in a library and that these items may be regarded as independent observations from a distribution P_X.

Let Y be a new observation which with prior probability $\pi > 0$ is a "noisy version" of one of the items X_i stored in the library, while with prior probability $1 - \pi$ it does not correspond to any one of the items. Here "Y is a noisy version of X_i" means that the joint distribution of these random variables is P_{XY}, while otherwise this joint distribution is $P_X \times P_Y$. When $Y = y$ is observed, Chernoff's model calls for searching in a subset $\delta(y)$ of the range of the X_i's for the item to which Y corresponds. A cost $c > 0$ is incurred for each $X_j \in \delta(y)$, and a cost $k > c$ is incurred if the "true" X_i is not in $\delta(y)$.

Let L denote the number of those X_j's, $X_j \in \delta(y)$, that do not correspond to Y. Then the expected total cost is

$$\begin{aligned} C &= cEL + \pi(c \Pr\{X \in \delta(Y)\} + k \Pr\{X \notin \delta(Y)\}) \\ &= c(N - \pi)EP_X(\delta(Y)) \\ &\quad + \pi(k - c)\Pr\{X \notin \delta(Y)\} + \pi c. \end{aligned} \quad (5.1)$$

The "search regions" $\delta(Y)$ should be chosen so as to minimize C. It readily follows from (5.1) that the minimal expected cost $C^* = \min C$ satisfies

$$\pi c \leq C^* \leq \pi k. \quad (5.2)$$

The expected cost (5.1) can also be expressed in terms of the error probabilities of first and second type of a (nonrandomized) test for the hypothesis P_{XY} against the alternative $P_X \times P_Y$. In fact, define a one-to-one correspondence between such tests and specifications of search regions $\delta(y)$ by letting the test accept the null hypothesis P_{XY} if and only if the sample point (x, y) is such that $x \in \delta(y)$. Thus, denoting by ϵ and β the error probabilities of first and second type of such a test, (5.1) may be written as

$$C = c(N - \pi)\beta + \pi(k - c)\epsilon + \pi c. \quad (5.3)$$

This model is well-suited for deriving asymptotic results in the case where the role of the X_i and Y is played by n-tuples of random variables $X_i^n = X_{i1}, \cdots, X_{in}$ and $Y^n = Y_1, \cdots, Y_n$ such that the pairs $(X_{i1}, Y_1), \cdots, (X_{in}, Y_n)$ are n independent drawings from the joint distribution P_{XY} or $P_X \times P_Y$, depending on whether Y^n represents a noisy version of X_i^n or not. We shall refer to this case as Chernoff's model for n-tuples and denote the minimum expected cost for this model by C_n^*. Since the joint asymptotic behavior of the error probabilities of first and

second type for tests with sample size $n \to \infty$ between simple hypotheses is well understood (Hoeffding [10], Csiszár and Longo [8]), (5.3) enables us to get tight bounds on C_n^* when n is large. In particular, it follows from (5.3), simply by Stein's lemma, that for arbitrary $0 < \pi \leq 1$, $k > c > 0$, and $\eta > 0$, $\delta > 0$ we have

$$C_n^* < \pi c + \eta \qquad \text{for } \frac{1}{n} \log N < I(X \wedge Y) - \delta \quad (5.4)$$

$$C_n^* > \pi k - \eta \qquad \text{for } \frac{1}{n} \log N > I(X \wedge Y) + \delta \quad (5.5)$$

if n is sufficiently large. Let us make a few comments at this point.

a) The formula for expected cost in [6] contains a slight error. In fact, (2.2) there is incorrect because the conditional probability of $X_j \in \delta(y)$ given that $Y = y$ equals the unconditional probability only for the X_j's not corresponding to Y. This error does not substantially affect the results of [6], except that the expected cost C is not exactly equal to a linear combination of the error probabilities of a hypothesis test (as stated in [6, (2.6)]), rather, an additive constant πc also enters.

b) Chernoff [6] implicitly assumed that for each $X_j \in \delta(y)$, one could unambiguously determine, presumably by using additional information, whether it was the true item of which $Y = y$ was a noisy observation; he interpreted c as the cost of such a determination. Whether this assumption is justified or not, the collection of items $X_j \in \delta(y)$ may be considered as a "list decision" about the true X_i. It is natural to measure the goodness of a list decision rule by a linear combination of the expected number EL of incorrect items on the list and of the probability $\Pr\{X \notin \delta(Y)\}$ that the correct item is not on the list. Thus we recover formula (5.1), up to the constant term πc.

c) The mathematical problem that Chernoff's model leads to is formally equivalent to a channel-coding problem involving random codes with list decoding. In fact, consider a random code of block-length one for a channel with a transmission-probability matrix $P_{Y|X}$ encoding the messages $1, \cdots, N$ by independent random variables $X_1, \cdots, X_N$ with common distribution P_X. Use list decoding specified by a family of sets $\delta(y)$ so that the decoder, when observing y, prints the list of those messages j for which $X_j \in \delta(y)$. Then supposing for simplicity that $\pi = 0$, the terms EL and $\Pr\{X \notin \delta(y)\}$ in (5.1) are just the expected erroneous list size and the probability of list decoding error. Channel codes with list decoding have been studied with respect to these performance criteria by Forney [9] for block length $n \to \infty$ (rather than $n = 1$). Of course, random codes of block length n correspond in the foregoing sense to Chernoff's model for n-tuples.

Chernoff also raised in [6] the problem of data compression, suggesting that it might be possible to store a compressed version of the items in the library without much adverse effect on identification. As an application of our results, we now describe an asymptotic solution to this problem within the context of Chernoff's model for n-tuples.

Let $\mathcal{X}$ and $\mathcal{Y}$ be the (finite) sets of possible values of the random variables X_{ij} and Y_j, respectively. A compression of the library items is a mapping $f: \mathcal{X}^n \to \mathcal{Z}$, where $\mathcal{Z}$ is some finite set. For the compressed items $Z_i \triangleq f(X_i^n)$ $(1 \leq i \leq N)$ any specification of search regions $\delta(y^n)$, $y^n \in \mathcal{Y}^n$, gives rise to an expected cost defined as in (5.1), with Z_i and Y^n in the role of X_i and Y. Let us denote by $C_n^*(f)$ the minimum expected cost for a given f, and by $C_n^*(R)$ the minimum of $C_n^*(f)$ for all $f: \mathcal{X}^n \to \mathcal{Z}$ with $\|f\| \leq \exp(nR)$.

Theorem 7: For any positive η and δ an n_0 exists (also depending on π, k, c) such that for $n \geq n_0$ we have

$$C_n^*(R) < \pi c + \eta, \quad (5.6)$$

if $(1/n) \log N < I(U \wedge Y) - \delta$ for some random variable U with

$$U \ominus X \ominus Y \qquad I(U \wedge X) \leq R \qquad |\mathcal{U}| \leq |\mathcal{X}| + 1, \quad (5.7)$$

and, on the other hand,

$$C_n^*(R) > \pi k - \eta \quad (5.8)$$

if $(1/n) \log N > I(U \wedge Y) + \delta$ for every U with the property (5.7).

Proof: The result follows immediately from the representation (5.3) of expected cost and Theorems 2 and 3.

References

[1] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 629–637, Nov. 1975.

[2] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Z. Wahrscheinlichkeitsth. verw. Gebiete*, vol. 34, pp. 157–177, 1976.

[3] R. Ahlswede, "Coloring hypergraphs: A new approach to multi-user source coding, Part 1," *J. Combinatorics, Inform. Syst. Sci.* vol. 4, pp. 76–115, 1979.
——, "Coloring hypergraphs: A new approach to multi-user source coding: Part 2," *ibid.*, vol. 5, pp. 220–268, 1980.

[4] R. Ahlswede and R. Wegener, *Suchprobleme*. Stuttgart, W. Germany: Teubner Verlag, 1979. Russian edition with appendix by Maljutov, Moscow, USSR: MIR, 1982.

[5] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493–507, 1952.

[6] ——, "The identification of an element of a large population in the presence of noise," *Ann. Statist.*, vol. 8, pp. 1179–1197, 1980.

[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1982 and Budapest, Hungary: Akademiai Kiado, 1981.

[8] I. Csiszár and G. Longo, "On the error exponent for source coding and for testing simple statistical hypotheses," *Studia Sci. Math. Hungar.*, vol. 6, pp. 181–191, 1971.

[9] G. D. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 206–220, Mar. 1968.

[10] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–400, 1965.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[12] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[13] A. Perez, "Discrimination rate loss in simple statistical hypotheses by unfitted decision procedures," in *Studies in Probability and Related Topics: Papers in Honour of Octav Onicescu*. Nagard, 1983, pp. 381–390.