The Annals of Statistics
1990, Vol. 18, No. 1, 141–171

ON MINIMAX ESTIMATION IN THE PRESENCE OF SIDE INFORMATION ABOUT REMOTE DATA

BY R. AHLSWEDE AND M. V. BURNASHEV

University of Bielefeld and Institute of Information Transmission, Moscow

We analyze the following model: One person, called "helper" observes an outcome $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ of the sequence $X^n = (X_1, \ldots, X_n)$ of i.i.d. RV's and the statistician gets a sample $y^n = (y_1, \ldots, y_n)$ of the sequence $Y^n(\theta, x^n)$ of RV's with a density $\prod_{t=1}^n f(y_t|\theta, x_t)$. The helper can give some (side) information about x^n to the statistician via an encoding function s_n: $\mathcal{X}^n \to \mathbb{N}$ with $\text{rate}(s_n) \stackrel{\text{def}}{=} (1/n)\log \#\ \text{range}(s_n) \le R$. Based on the knowledge of $s_n(x^n)$ and y^n the statistician tries to estimate θ by an estimator $\hat{\theta}_n$. For the maximal mean square error

$$e_n(R) \stackrel{\text{def}}{=} \inf_{\hat{\theta}_n} \inf_{s_n:\, \text{rate}(s_n) \le R} \sup_{\theta \in \Theta} E_\theta |\hat{\theta}_n - \theta|^2$$

we establish a Cramér–Rao type bound and, in case of a finite $\mathcal{X}$, prove asymptotic achievability of this bound under certain conditions. The proof involves a nonobvious combination of results (some of which are novel) for both coding and estimation.

1. Introduction. It is usually assumed in statistics that the statistician has free access to the data (samples). This assumption is not always justified. A scientist may be interested in the correlation between the values of physical measurements taken at places that are far apart. In this case data have to be communicated. Since there may be limitations on the capacities or permittable costs of the channels used, it becomes important to select suitable data or perform some more sophisticated data processing in order to meet some specified goals of the statistician. Whereas in computer science the communication aspect has already entered complexity considerations, for instance in parallel computing [Yao (1979)], it has been introduced only recently in statistics in the context of bivariate hypothesis testing [Ahlswede and Csiszár (1986)]. There data reduction is measured by the rate needed to transmit the reduced data *and* the performance of a best test based on those data. It was emphasized by these authors that this may be the beginning of a whole program, which also includes estimation problems.

As a further contribution to this problem we investigate here the effect of partial side information about remote data in estimating the distribution in a parametric family of distributions. It was our aim to establish the first results in this area and not to strive for the most general conditions on the family of distributions under which an asymptotic theory of estimation can be developed.

Received November 1986; revised April 1989.

AMS 1980 *subject classifications.* Primary 62A99, 62F12, 94A15; secondary 62N99.

Key words and phrases. Side information, Cramér–Rao-type inequality, efficiency, multiuser source coding, information measures.

We use the mean square error criterion and under certain familiar regularity conditions we establish a Cramér–Rao type bound and its asymptotic optimality. The characterization of this bound is in terms of a generalized Fisher information.

In the terminology of information theory it is not a single-letter characterization. This means that it involves not only product distributions and is, in general, not suited for a numerical evaluation. To find a single-letter characterization is a task of formidable mathematical difficulty. The situation is similar to the testing problem mentioned above, where the role of Fisher information is taken by the Kullback–Leibler divergence. However, in spite of the close connection between these information measures there are also differences to the effect that Fisher information allows a certain local single-letterization. This fact makes it possible to derive single-letter characterizations of our Cramér–Rao type bound for some classes of parametric families of distributions and for other more irregular classes to at least establish lower and upper bounds on its value.

In this paper we have focused on the need for data reduction in order to cut down the *communication* costs. Often it is also the case that data are available to the statistician only in an implicit form and they can be revealed only at high costs of *computation*. Those costs are to be considered in conjunction with errors suffered from statistical decisions. They can again be lowered by some kind of data reduction.

In summary it can be said that inclusion of the communication as well as the complexity aspect will challenge the body of classical statistics. Some demands can be met by modification of existing models and procedures; others require new concepts.

The organization of paper is as follows: In Section 2 we formulate our model and the estimation problem, which we investigate. In Section 3 we recall first the notion of Fisher information and some of its familiar properties. Then we introduce our related J function, which takes the role of Fisher information in our problem. It involves concepts from multiuser information theory [cf. Csiszár and Körner (1982)]. Some basic properties are established. In Section 4 we state and prove our Cramér–Rao type inequality, first for the unbiased and then for the biased case.

Before we state and prove our results on asymptotical achievability of this bound for finite $\mathcal{X}$ in Section 7, we present, in Section 5, our main auxiliary result on coding the side information and, in Section 6, we introduce and analyze the regularity conditions used. Finally, in Section 8 we discuss a case in which the J function "single-letterizes" and can be evaluated.

2. A model for parameter estimation in the presence of side information. We consider a one-parametric family of density functions depending on a nuisance parameter. Formally, we are given:

(2.1) A one-dimensional parameter space $\Theta = (a, b)$, $-\infty \leq a < b \leq \infty$.
(2.2) Two σ-finite measure spaces $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$.
(2.3) A probability density function p with respect to μ.

(2.4) A set $\{f(\cdot|\theta, x)\colon \theta \in \Theta,\ x \in \mathscr{X}\}$ of conditional probability densities with respect to ν.

Consider now a fixed sample size $n \in \mathbb{N} = \{1, 2, \dots\}$. One person, called the helper, shall observe a sample $x^n = (x_1, \dots, x_n) \in \mathscr{X}^n$ of the sequence $X^n = (X_1, \dots, X_n)$ of i.i.d. RV's with joint density $p^n = \prod_{t=1}^n p$ and the other person, called the statistician, shall observe a sample $y^n = (y_1, \dots, y_n) \in \mathscr{Y}^n$ of the sequence $Y^n(\theta) = (Y_1(\theta), \dots, Y_n(\theta))$ of random variables with a joint density in the family $\{\prod_{t=1}^n f(y_t|\theta, x_t)\colon \theta \in \Theta\}$.

The sequence $x^n = (x_1, x_2, \dots, x_n)$ of nuisance parameters is not known to the statistician. However, the helper may give him some (side) information about x^n. If he can transmit at a rate $R \geq 0$, then he can inform the statistician via encoding functions $s_n\colon \mathscr{X}^n \to \mathbb{N}$ with

$$\text{rate}(s_n) \triangleq \frac{1}{n}\log\|s_n\| \leq R. \tag{2.5}$$

Here $\|s_n\|$ denotes the cardinality of the range of s_n. It is convenient to introduce

$$\mathscr{S}_n(R) \triangleq \{s_n | s_n\colon \mathscr{X}^n \to \mathbb{N}, \text{rate}(s_n) \leq R\}, \qquad R \geq 0. \tag{2.6}$$

Based on the knowledge of $s_n(x^n)$ and y^n the statistician tries to estimate the parameter θ by means of an estimator

$$\hat{\theta} \triangleq \hat{\theta}_n(y^n, s_n(x^n)). \tag{2.7}$$

We use the maximal mean square error criterion

$$e_n(\hat{\theta}, s_n) \triangleq \sup_{\theta \in \Theta} E_\theta|\hat{\theta} - \theta|^2. \tag{2.8}$$

Since the helper and the statistician are cooperative we are led to study the minimax square error

$$e_n(R) \triangleq \inf_{s_n \in \mathscr{S}_n(R),\, \hat{\theta}} \sup_{\theta \in \Theta} E_\theta|\hat{\theta} - \theta|^2, \qquad R \geq 0. \tag{2.9}$$

In particular we are interested here in the asymptotical behavior of the function $e_n(R)$ when $n \to \infty$. We establish a Cramér–Rao-type bound and, in case of a finite $\mathscr{X}$, we prove asymptotic achievability of this bound under certain regularity conditions. In some cases this bound can be evaluated.

EXAMPLE 1 (Symmetric Bernoulli case). Let $\mathscr{X} = \{0, 1\}$, $\Theta = (0, 1)$, $\mathscr{Y} = \{0, 1\}$, and let $P_X(0) = P_X(1) = \frac{1}{2}$, $P(0|\theta, 0) = P(1|\theta, 1) = \theta$, $P(0|\theta, 1) = P(1|\theta, 0) = 1 - \theta$. Notice that without any side information about X, $P(0|\theta) = P(1|\theta) = \frac{1}{2}$ for all θ and it is impossible to estimate θ.

By our theory and the calculations in Section 8, upon division by n, the bound here is, according to (8.8), $[\theta(1 - \theta) + c(1 - c)](1 - 2c)^{-2}$, where c satisfies $1 - h(c) = R$ and h is the binary entropy function.

Notice that the term $\theta(1 - \theta)/n$ is the mean squared error for the usual estimator of θ when the X's are available to the statistician ($R = 1$, $c = 0$). The formula shows how the bound increases when R decreases. At the extreme $R = 0$ we get $c = \frac{1}{2}$ and therefore the value infinity.

EXAMPLE 2 (A Gaussian case). Sometimes no deep theory is needed. Suppose that $Y_t(\theta) = \theta + X_t + Z_t$, $-\infty < \theta < \infty$, where the Z_t are in $\mathcal{N}(0,1)$ and the X_t are i.i.d. and take finitely many values. If the statistician knows $X^n = (X_1, \ldots, X_n)$, then an optimal estimator has the form

$$\hat{\theta}_n = \frac{1}{n}\left(\sum_{t=1}^{n} Y_t - \sum_{t=1}^{n} X_t\right) \quad \text{and} \quad E_\theta(\hat{\theta}_n - \theta)^2 = \frac{1}{n}.$$

Now in case $R > 0$ the statistician can be informed about the value $\Sigma_{t=1}^n X_t$ with an accuracy $\sim e^{-cn}$, $c > 0$. Therefore, for *any* $R > 0$ we get $E_\theta(\hat{\theta}_n - \theta)^2 \sim 1/n$. Here the X_t's could be errors in measurements known to the helper.

3. On Fisher information, mutual information and the J function. We assume throughout this section that all functions defined do exist. Sufficient conditions for this are given later when needed. We present first properties of Fisher's information, which are either folklore or else easy to derive.

For a parametrized family $\{\mathcal{Y}(\theta)\colon \theta \in \Theta\}$ of random variables with ν densities $\{(f(\cdot|\theta)\colon \theta \in \Theta\}$ such that $f_\theta(y|\theta) = [\partial f(y|\theta)]/\partial\theta$ exists for ν-almost all $y \in \mathcal{Y}$, the Fisher information at θ is defined by

$$I(\theta; Y) = I(\theta) = \int_{\mathcal{Y}} \frac{|f_\theta(y|\theta)|^2}{f(y|\theta)}\nu(dy). \tag{3.1}$$

Similarly, if $\{(Y(\theta), Z(\theta))\colon \theta \in \Theta\}$ has the conditional densities $f(y|\theta, z)$, then for $z \in \mathcal{Z}$,

$$I(\theta; Y|Z = z) = \int_{\mathcal{Y}} \frac{|f_\theta(y|\theta, z)|^2}{f(y|\theta, z)}\nu(dy|z) \tag{3.2}$$

and

$$I(\theta; Y|Z) = \int_{\mathcal{Z}} I(\theta; Y|Z = z) f(z|\theta)\nu(dz) \tag{3.3}$$

is the Fisher information of Y about θ conditionally on Z. We also use the shorter notation $I(\theta|z)$ [resp. $I(\theta|Z)$] if the meaning is clear from the context.

LEMMA 1. *Let* $\{(Y(\theta), X, V)\colon \theta \in \Theta\}$ *be random variables, where* (X, V) *does not depend on* $Y(\theta)$. *Then*

(a) $I(\theta; Y|X) \geq I(\theta; Y)$,
(b) $I(\theta; Y|XV) \geq I(\theta; Y|V)$,

if these quantities exist.

Of course (a) is a special case of (b) and (b) can be derived from (a). Since under the stipulated conditions $I(\theta; X) = 0$ and since Fisher information is nonnegative, this is a consequence of the next lemma.

Fisher information has an additivity property, which is a direct consequence of the multiplicative property of conditional densities.

LEMMA 2. *Let* $\{(Z_1(\theta), \dots, Z_n(\theta), V(\theta): \theta \in \Theta\}$ *be random variables, where* V *may depend on* θ. *Then with* $Z^{t-1} = (Z_1, \dots, Z_{t-1})$,

$$\text{(a)} \qquad I(\theta; Z_1, \dots, Z_n|V) = \sum_{t=1}^{n} I\big(\theta; Z_t|V, Z^{t-1}\big),$$

if the quantities are defined. In particular, if $Z_1(\theta), \dots, Z_n(\theta)$ *are independent for all* $\theta \in \Theta$, *then*

$$\text{(b)} \qquad I(\theta; Z_1, \dots, Z_n) = \sum_{t=1}^{n} I(\theta; Z_t).$$

We draw attention to the fact that Lemma 1(a) is generally not true, if X does depend on θ. The situation is similar as for mutual information, where conditioning does not necessarily increase its value.

EXAMPLE 3. Let X, Y be binary random variables such that

$$P_{XY}(00|\theta) = \theta, \qquad P_{XY}(11|\theta) = 1 - \theta,$$
$$P_{XY}(01|\theta) = P_{XY}(10|\theta) = 0.$$

Then $I(\theta; X) = I(\theta; Y) = 1/\theta + 1/(1 - \theta)$, but also $I(\theta; XY) = 1/\theta + 1/(1 - \theta)$. By Lemma 2, therefore,

$$I(\theta; X|Y) = I(\theta; XY) - I(\theta; Y) = 0$$

and by symmetry also

$$I(\theta; Y|X) = 0.$$

Another extremal case occurs in

EXAMPLE 4.

$$P_{XY}(00|\theta) = P_{XY}(11|\theta) = \frac{1 - \theta}{2},$$

$$P_{XY}(01|\theta) = P_{XY}(10|\theta) = \frac{\theta}{2}.$$

Here

$$P_X(0|\theta) = P_X(1|\theta) = P_Y(0|\theta) = P_Y(1|\theta) = \tfrac{1}{2}$$

and thus $I(\theta; X) = I(\theta; Y) = 0$, whereas

$$I(\theta; X|Y) = I(\theta; Y|X) = I(\theta; XY) = \frac{1}{1 - \theta} + \frac{1}{\theta}.$$

Next we recall the definition of Shannon's mutual information. For a pair of RV's (X, Y), *where* X *takes only finitely many values*, the mutual information is

$$(3.4) \qquad I(X \wedge Y) = \sum_x P_X(x) \int \log \frac{dP_{Y|X}(y|x)}{d\big(\sum_x P_X(x) P_{Y|X}(y|x)\big)}\, dP_{Y|X}(y|x).$$

For $Y \equiv X$ we get the entropy $H(X) = I(X \wedge X)$ as a special case. Furthermore, for finite-valued Y we can also write $I(X \wedge Y) = H(X) - H(X|Y)$, where $H(X|Y) = H(X, Y) - H(Y)$ is the conditional entropy. There is also a conditional mutual information $I(X \wedge Y|Z) = H(X|Z) - H(X|YZ)$. These quantities have additivity properties similar to those of Fisher information [cf. Csiszár and Körner (1982)].

Before we can introduce the J function, which plays the same role for our estimation problem with side information as $\inf_{\theta \in \Theta} I(\theta; Y)$ does in classical minmax estimation theory, we have to recall a few definitions familiar in multiuser information theory, in particular also in coding problems involving side information [cf. Ahlswede and Körner (1975); Ahlswede (1979)].

Let U be a RV with values in a *finite* set $\mathcal{U}$, which for every $\theta \in \Theta$ has a joint distribution $P_{UX^nY^n}(\theta)$ with X^n, $Y^n(\theta)$. We use the abbreviation $U \multimap X^n \multimap Y^n$, if for every $\theta \in \Theta$ the triple $(U, X^n, Y^n(\theta))$ forms a Markov chain in this order.

It is convenient to have the definitions

$$\mathcal{M}_n = \{U: U \multimap X^n \multimap Y^n, U \text{ finite valued}\} \tag{3.5}$$

and for any $R \geq 0$,

$$\mathcal{M}_n(R) = \{U: U \in \mathcal{M}_n, I(X^n \wedge U) \leq Rn\}, \tag{3.6}$$

where $I(X^n \wedge U)$ is the mutual information between X^n and U. Define now for $R \geq 0$,

$$J_n(R) = \sup_{U \in \mathcal{M}_n(R)} \inf_{\theta \in \Theta} \frac{1}{n} I(\theta; Y^n|U) \tag{3.7}$$

and

$$J(R) = \lim_{n \to \infty} J_n(R). \tag{3.8}$$

Here the existence of the limit readily follows from the subadditivity of $nJ_n(R)$ in n, which can be shown by considering auxiliary variables U, which are pairs of independent variables U', U''.

Even though presently we do not have an example, it seems that in the language of information theory, $J_n(R)$ does in general not single-letterize, that is, $J_n(R) > J_1(R)$ may happen. This makes it usually impossible to even approximately calculate $J(R)$. However, we have a method by which $J(R) = J_1(R)$ can be shown in some cases.

For the analysis of $J_n(R)$ we use the functions

$$J^n(\theta, R) = \sup_{U \in \mathcal{M}_n(R)} \frac{1}{n} I(\theta; Y^n|U), \tag{3.9}$$

$$J^n(R) = \inf_{\theta \in \Theta} J^n(\theta, R). \tag{3.10}$$

They have two nice properties, which we now prove and later use for *finite* $\mathcal{X}$. However, they extend also to general RV's X.

LEMMA 3. *For every* $\theta \in \Theta$, $J^n(\theta, R)$ *is concave* $(\cap)$ *in* R.

PROOF. We follow an argument that is by now standard. There is no loss in generality if we assume $n = 1$. First notice that the constraint $I(U \wedge X) \leq R$ can equivalently be written as $H(X|U) \geq H(X) - R$. Let now

$$\big(U_t, X_t, (Y_t(\theta))_{\theta\in\Theta}\big),\ t = 1,2,$$

be two families of random variables with

$$P_{X_tY_t(\theta)} = P_X P_{Y(\theta)|X} \quad \text{for } t = 1,2.$$

Further let T be a random variable with values in $\{1,2\}$, which is independent of all other variables.

If $U_t \ominus X_t \ominus Y_t(\theta)$, $\theta \in \Theta$, $t = 1,2$, then also

$$(T, U_T) \ominus X_T \ominus Y_T(\theta), \qquad \theta \in \Theta,$$

and

$$\big(H(X_T|U_T, T),\ I(\theta; Y_T|U_T, T)\big) = \sum_{t=1,2} P_T(t)\big(H(X_t|U_t),\ I(\theta; Y_t|U_t)\big).$$

Varying P_T over all distributions on $\{0,1\}$ we get all points on the line segment

$$\big[\big(H(X_1|U_1), I(\theta; Y_1|U_1)\big), \big(H(X_2|U_2), I(\theta; Y_2|U_2)\big)\big]$$

and thus the concavity of J^n in R. □

Next we establish a less obvious property

LEMMA 4 (Local single-letterization).

(a) $J^n(\theta, R) = J^1(\theta, R)$ for all θ, R, n.
(b) $J^n(R) = J^1(R)$ for all R, n.

PROOF. It is clear from the definitions (3.9) and (3.10) that (a) implies (b). The relation $J^n(\theta, R) \geq J^1(\theta, R)$ readily follows by choosing $U = (U_1, \ldots, U_n)$ as a sequence of independent random variables. The issue is the reverse inequality. Lemma 2 gives the decomposition

$$I(\theta; Y^n|U) = \sum_{t=1}^{n} I\big(\theta; Y_t|Y^{t-1}U\big). \tag{3.11}$$

For a fixed θ define now $U_t = (U, Y^{t-1}(\theta))$ and notice that

$$U_t \ominus X_t \ominus Y_t(\theta).$$

We verify first that

$$H(X^n|U) \leq \sum_{t=1}^{n} H(X_t|U_t). \tag{3.12}$$

Indeed,

$$H(X^n|U) = \sum_{t=1}^{n} H(X_t|U, X^{t-1})$$

and since

$$H(X_t|U, X^{t-1}, Y^{t-1}) \le H(X_t|U, Y^{t-1})$$

it remains to be seen that

$$H(X_t|U, X^{t-1}) = H(X_t|U, X^{t-1}, Y^{t-1})$$

or equivalently that $I(X_t \wedge Y^{t-1}|U, X^{t-1}) = 0$. Now

$$\begin{aligned} I(X_t \wedge Y^{t-1}|U, X^{t-1}) &\le I(X^n \wedge Y^{t-1}|U, X^{t-1}) \le I(UX^n \wedge Y^{t-1}|X^{t-1}) \\ &= I(X^n \wedge Y^{t-1}|X^{t-1}) + I(U \wedge Y^{t-1}|X^n) = 0 \end{aligned}$$

because the first and second summand vanish by the independence structure and Markov property, respectively.

Now by the definition of J^1,

$$I(\theta; Y_t|U_t) \le J^1(\theta; I(X_t \wedge U_t)).$$

Thus by (3.11) and the concavity of J^1,

$$\frac{1}{n} I(\theta; Y^n|U) \le \frac{1}{n} \sum_{t=1}^{n} J^1(\theta, I(X_t \wedge U_t)) \le J^1\left(\theta, \frac{1}{n} \sum_{t=1}^{n} I(X_t \wedge U_t)\right).$$

Finally, by (3.12),

$$\frac{1}{n} \sum_{t=1}^{n} I(X_t \wedge U_t) \le \frac{1}{n} I(X^n \wedge U) \le R$$

and

$$\frac{1}{n} I(\theta; Y^n|U) \le J^1(\theta, R)$$

follow, because J^1 is monotone increasing in R. □

REMARK. The standard proof for single-letterization of entropies [cf. Ahlswede and Körner (1975); Csiszár and Körner (1982)] is based on a "dual trick." Instead of U_t one uses $V_t = (U, X^{t-1})$. Thus $H(X^n|U) = \sum_{t=1}^{n} H(X_t|V_t)$ has immediately the right decomposition and $H(Y^n|U)$ remains to be analyzed. This approach would in the present situation have the advantage that V_t does not depend on θ and could lead to a single-letterization of $J_n(R)$. Unfortunately we do not know whether

$$I(\theta; Y^n|U) \le \sum_{t=1}^{n} I(\theta; Y_t|V_t).$$

This would imply $J_n(R) = J_1(R)$.

However for the Kullback–Leibler divergence the analogous result is not true [Ahlswede and Csiszár (1986)] and very likely, in general, $J_n(R) \ne J_1(R)$. In any

case, this is the main problem left open in our investigations. Next we present a sufficient condition for $J_n(R) = J_1(R)$ to hold.

LEMMA 5. *If* $J_1(R) = J^1(R)$, *then* $J_n(R) = J_1(R)$.

PROOF. Since

$$\begin{aligned} J^n(R) &= \inf_{\theta} \sup_{U \in \mathscr{M}_n(R)} \frac{1}{n} I(\theta; Y^n|U) \\ &\geq \sup_{U \in \mathscr{M}_n(R)} \inf_{\theta} \frac{1}{n} I(\theta; Y^n|U) \\ &= J_n(R) \end{aligned}$$

and since by Lemma 4(b), $J^n(R) = J^1(R)$, the assumption ensures $J_1(R) \geq J_n(R)$. The reverse inequality is obviously true. □

In Section 8 we show that for the symmetric Bernoulli case (Example 1) the condition $J_1(R) = J^1(R)$ holds. In the light of the fact that the second derivative of the Kullback–Leibler divergence is one-half the Fisher information, it is very remarkable that for the hypothesis testing problem with side information for two members of that Bernoulli family the relevant divergences do not single-letterize. This is exactly the example of Ahlswede and Csiszár (1982).

EXAMPLE 5. Locations families

$$\{ f(y|x, \theta) = f(y - \theta|x) \colon -\infty < \theta < \infty \}$$

are an important class of problems for which the answer has the desired single-letter characterization, because $I(\theta; Y|X = x)$ is independent of θ and the hypothesis of Lemma 5 is trivially satisfied.

4. The informational inequality. We refer to our Cramér–Rao-type inequality (Theorems 1 and 2) also as "the informational inequality" and to its bound as "the informational bound."

To simplify matters, we consider first the unbiased case. An estimator $\hat{\theta}$ is unbiased for an encoding function s_n if

$$E_\theta \hat{\theta}(Y^n, s_n(X^n)) = \theta \quad \text{for all } \theta \in \Theta. \tag{4.1}$$

Needless to say, it is essentially impossible to decide whether such estimators exist. However, their study makes the role of the function J_n transparent. The informational inequality for our estimation problem with side information can readily be derived from the classical Cramér–Rao inequality with the help of well-known properties of the Fisher information (see Section 3). We also use here, for the biased case, a form of the Cramér–Rao inequality that is contained in Theorem 2.1 of Ibragimov and Khas'minskii (1975a).

PROPOSITION 1. *Suppose that the density function $f(y|\theta)$ is absolutely continuous in θ for almost all y and that*

(i) *$I(\theta)$ exists for $\theta \in \Theta$,*
(ii) *$I(\theta)$ is positive and locally bounded.*

If for the estimator $\hat{\theta}$,

$$E_\theta|\hat{\theta}(Y) - \theta|^2 \quad \text{is locally bounded,}$$

then

$$E_\theta|\hat{\theta}(Y) - \theta|^2 \geq \frac{(1 - b'(\theta))^2}{I(\theta)} + b^2(\theta), \qquad \theta \in \Theta,$$

where

$$b(\theta) = E\hat{\theta}(Y) - \theta.$$

We call $(s_n, \hat{\theta})$ regular, if $I(\theta; Y^n|s_n(X^n))$ is positive and $E_\theta|\hat{\theta}(Y^n, s_n(X^n)) - \theta|^2$ is locally bounded. We introduce

$$\mathscr{R}_n(R) = \text{set of all regular } \left(s_n, \hat{\theta}\right) \text{ with } s_n \in \mathscr{S}_n(R) \text{ [as defined in (2.6)]} \tag{4.2}$$

and, similarly,

$$\mathscr{R}_n^*(R) = \left\{\left(s_n, \hat{\theta}\right) \in \mathscr{R}_n(R)\colon \hat{\theta} \text{ unbiased for } s_n\right\}. \tag{4.3}$$

In order to make Proposition 1 applicable to our unbiased case we have to ensure (i) and (ii). This can be achieved by the following conditions:

(C1) $I(\theta; XY)$ exists for $\theta \in \Theta$, is positive and locally bounded.

(C2) For every $y \in \mathscr{Y}$ the function $f(y|\theta, x)$ is *uniformly in* x absolutely continuous on compact subsets of Θ.

If f satisfies (C2), then also for $U \in \mathscr{M}_n$ the conditional density $p(y^n|\theta, u)$ has a derivative p_θ for $\nu^n = \prod_1^n \nu$-almost all y^n and every $u \in \mathscr{U}$. We can therefore define

$$I(\theta; Y^n|U = u) = \int_{\mathscr{Y}^n} \frac{|p_\theta(y^n|\theta, u)|^2}{p(y^n|\theta, u)} \nu^n(dy^n) \tag{4.4}$$

and

$$I(\theta; Y^n|U) = \sum_u P_U(u) I(\theta; Y^n|U = u). \tag{4.5}$$

Our first result is

THEOREM 1. *If* (C1) *and* (C2) *hold, then for $R \geq 0$ and every $(s_n, \hat{\theta}) \in \mathscr{R}_n^*(R)$,*

$$\sup_{\theta \in \Theta} E_\theta|\hat{\theta}(Y^n, s_n(X^n)) - \theta|^2 \geq \frac{1}{nJ_n(R)}. \tag{4.6}$$

Classically one can derive from Proposition 1 the asymptotic form of the Cramér–Rao inequality for the biased case. The derivation given in Cencov (1972) is adaptable to our model with side information. Technically we make the derivation somewhat more elegant by extracting its essence in the form of an elementary analytical inequality, which we now state and prove.

PROPOSITION 2. *Let* g: $[a, b] \to \mathbb{R}$ *be absolutely continuous and let* G: $[a, b] \to \mathbb{R}^+$ *be bounded, that is,* $\lambda^2 = \sup_{r \in [a, b]} G(r) < \infty$. *Then we have*

$$\sup_{r \in [a, b]} \left[\frac{(1 + g'(r))^2}{G(r)} + g^2(r) \right] \geq \left[\frac{b - a}{(b - a)\lambda + 2} \right]^2.$$

PROOF. If γ^2 denotes the left side of this inequality, then obviously

$$\gamma \geq \max\{|g(a)|, |g(b)|\}$$

and

$$\lambda\gamma \geq 1 + g'(r) \quad \text{for } r \in [a, b].$$

Therefore

$$-2\gamma \leq g(b) - g(a) = \int_a^b g'(r)\, dr \leq (b - a)(\lambda\gamma - 1)$$

and thus

$$\gamma \geq \frac{b - a}{2 + (b - a)\lambda}. \qquad \square$$

In the biased case we use a condition, which is much stronger than (C1). In terms of the modulus of continuity,

$$\Omega(\delta) \triangleq \sup_{U \in \mathcal{M}_1} \sup_{\theta, \theta' : |\theta - \theta'| \leq \delta} |I(\theta; YU) - I(\theta'; YU)|, \tag{4.7}$$

it can be stated as

(C3) $\lim_{\delta \to 0} \Omega(\delta) = 0$.

THEOREM 2. *If* (C2) *and* (C3) *hold, then for* $R \geq 0$ *and every* $(s_n, \hat{\theta}) \in \mathcal{R}_n(R)$,

$$\liminf_{n \to \infty} \inf_{(s_n, \hat{\theta}) \in \mathcal{R}_n(R)} \sup_{\theta \in \Theta} E_\theta \left| \hat{\theta}(Y^n, s_n(X^n)) - \theta \right|^2 n J_n(R) \geq 1.$$

Here $J_n(R)$ *can be replaced by* $J(R)$.

REMARK. The uniformity in x required in (C2) is no issue for finite $\mathcal{X}$. In this case one also can show by the so-called Support Lemma [Ahlswede and Körner (1975); see also Csiszár and Körner (1982)] that in all our formulas the variables U can be assumed to take at most $|\mathcal{X}| + 1$ values.

PROOF OF THEOREM 1. Since $s_n(X^n) \in \mathscr{M}_n$, by (C2) and its consequence (4.5) $I(\theta; Y^n s_n(X^n))$ exists except if it takes an infinite value. Since by Lemma 2,

$$I(\theta; Y^n s_n(X^n)) \le I(\theta; Y^n X^n s_n(X^n)) = nI(\theta; YX),$$

(C1) implies that it is even locally bounded. By assumption, $(s_n, \hat{\theta}) \in \mathscr{R}_n^*(R)$ and, therefore, Proposition 1 applies and yields

$$E_\theta|\hat{\theta}(Y^n, s_n(X^n)) - \theta| \ge \frac{1}{I(\theta; Y^n, s_n(X^n))}, \qquad \theta \in \Theta. \tag{4.8}$$

Again by Lemma 2,

$$I(\theta; Y^n s_n(X^n)) = I(\theta; s_n(X^n)) + I(\theta; Y^n|s_n(X^n)).$$

Since X^n and a fortiori also $s_n(X^n)$ do not depend on θ, we also have $I(\theta; s_n(X^n)) = 0$. Therefore,

$$I(\theta: Y^n s_n(X^n)) = I(\theta; Y^n|s_n(X^n)), \qquad \theta \in \Theta, \tag{4.9}$$

and we can rewrite (4.8) in the form

$$E_\theta|\hat{\theta} - \theta|^2 \ge \frac{1}{I(\theta; Y^n|s_n(X^n))}, \qquad \theta \in \Theta, \tag{4.10}$$

which implies

$$\sup_\theta E_\theta|\hat{\theta} - \theta|^2 \ge \frac{1}{\inf_\theta I(\theta; Y^n|s_n(X^n))}. \tag{4.11}$$

Since any finite-valued function of X^n is in $\mathscr{M}_n$ and therefore also $\mathscr{S}_n(R) \subset \mathscr{M}_n(R)$, we conclude with (4.11),

$$\begin{aligned}\inf_{(s_n, \hat{\theta}) \in \mathscr{R}_n^*(R)} \sup_{\theta \in \Theta} E_\theta|\hat{\theta} - \theta|^2 &\ge \inf_{s_n \in \mathscr{S}_n(R)} \frac{1}{\inf_{\theta \in \Theta} I(\theta; Y^n|s_n(X^n))} \\ &\ge \frac{1}{\sup_{U \in \mathscr{M}_n(R)} \inf_\theta I(\theta; Y^n|U)} = \frac{1}{nJ_n(R)}\end{aligned}$$

[by (3.7)]. □

PROOF OF THEOREM 2. By the arguments that led to (4.10) we derive with Proposition 1 in the biased case for $(s_n, \hat{\theta}) \in \mathscr{R}_n(R)$,

$$\begin{aligned}&E_\theta|\hat{\theta}(Y^n, s_n(X^n)) - \theta|^2 \\ &\ge \frac{(1 + b_n'(\theta, s_n))^2}{I(\theta; Y^n|s_n(X^n))} + b_n^2(\theta, s_n), \qquad \theta \in \Theta,\end{aligned} \tag{4.12}$$

where

$$b_n(\theta, s_n) = E_\theta\hat{\theta}(Y^n, s_n(X^n)) - \theta.$$

Apply now Proposition 2 with the choices

$$r = \theta, \qquad g = b_n, \qquad G = I, \qquad a = \theta_0, \qquad b = \theta_0 + n^{-1/2}. \tag{4.13}$$

Thus for (4.12) for $\theta_0 \in \Theta$,

$$\sup_{\theta \in [\theta_0, \theta_0 + n^{-1/2}]} E_\theta \left| \hat{\theta}(Y^n, s^n) - \theta \right|^2 \geq \frac{1}{\sup_{\theta \in [\theta_0, \theta_0 + d]} I(\theta; Y^n | s_n) + 2 \cdot n^{1/2}}. \tag{4.14}$$

Since by Lemma 2,

$$I(\theta; Y^n | s_n(X^n)) = \sum_{t=1}^{n} I(\theta; Y_t | Y^{t-1} s_n(X^n))$$

and since $Y^{t-1} s_n(X^n) \leftarrow\!\!\!\circ\!\!\!- X_t \leftarrow\!\!\!\circ\!\!\!- Y_t$, from definition (4.7),

$$\left| I(\theta; Y^n | s_n) - I(\theta'; Y^n | s_n) \right| \leq n\Omega(n^{-1/2}), \quad \text{for } \theta, \theta' \in \left[\theta_0, \theta_0 + n^{-1/2}\right]. \tag{4.15}$$

This and (4.14) imply

$$\sup_\theta E_\theta \left| \hat{\theta}(Y^n, s_n) - \theta \right|^2 \geq \frac{1}{\inf_\theta I(\theta; Y^n | s_n) + n\Omega(n^{-1/2}) + 2n^{1/2}}. \tag{4.16}$$

Now we continue as in the proof of Theorem 1:

$$\begin{aligned} &\inf_{(s_n, \hat{\theta}) \in \mathscr{R}_n(R)} \sup_\theta E_\theta \left| \hat{\theta}(Y^n, s^n) - \theta \right|^2 \\ &\geq \inf_{s_n \in \mathscr{S}_n(R)} \frac{1}{\inf_\theta I(\theta; Y^n | s_n) + n\Omega(n^{-1/2}) + 2n^{1/2}} \\ &\geq \frac{1}{\inf_{U \in \mathscr{M}_n(R)} I(\theta; Y^n | U) + n\Omega(n^{-1/2}) + 2n^{1/2}} \\ &= \frac{1}{nJ_n(R) + n\Omega(n^{-1/2}) + 2n^{1/2}}. \end{aligned}$$

This and (C3) imply the result. □

5. Encoding the side information. We assume here that $\mathscr{X}$ is finite. Furthermore we require that

$$I(\theta; Y | x) < \infty \quad \text{for } x \in \mathscr{X}, \theta \in \Theta. \tag{5.1}$$

Thus also for any $U \in \mathscr{M}_n$,

$$I(\theta; Y^n | U) < \infty, \tag{5.2}$$

because $I(\theta; Y^n | U) = I(\theta; Y^n U) \leq I(\theta; Y^n X^n) = nI(\theta; Y | X) < \infty$.

We emphasize that we make no further assumptions in this section, which is devoted to the proof of the following basic result.

PROPOSITION 3. *Suppose that $\mathcal{X}$ is finite and* (5.1) *holds in our model. If U is a finite-valued random variable, which satisfies*

$$U \leftarrow\!\!\!\cdot\, X \leftarrow\!\!\!\cdot\, Y(\theta), \qquad \theta \in \Theta,$$

then for any $\rho, \delta > 0$ there is an $n_0(\rho, \delta)$ such that for every $n \geq n_0(\rho, \delta)$ there exists an encoding function s_n: $\mathcal{X}^n \to \mathbb{N}$ with

(i) $I(X^n \wedge s_n(X^n)) \leq n(I(X \wedge U) + \rho)$,
(ii) $I(\theta; Y^n | s_n(X^n)) \geq n(1 - \delta) I(\theta; Y|U)$ *for all* $\theta \in \Theta$.

The result immediately implies that $J_1(R)$ can be achieved arbitrarily closely by suitable encoding functions. By taking X^r in the role of X and letting r tend to infinity we see that also $J(R)$ can thus be achieved.

Constructions of encoding functions meeting Proposition 3(i) can be given by the approaches familiar from source coding with side information [Ahlswede and Körner (1975); Ahlswede (1979); Csiszár and Körner (1982)]. The issue is to establish Proposition 3(ii). This requires subtle continuity considerations due to the fact that we are now dealing with Fisher information for families of non-finite-valued random variables, whereas in source coding one usually deals with conditional entropies and in Ahlswede and Csiszár (1986) with divergences of finite-valued random variables. Our approach continues the program of Ahlswede (1979) to exploit the fact that our model is invariant under permutations of $1, 2, \ldots, n$. We thus obtain novel results (Lemmas 7 and 8). In particular their analogue for mutual information may be useful in information theory.

We recall now some standard definitions and results [cf. Ahlswede and Csiszár (1986); Ahlswede (1979); Csiszár and Körner (1982)]. Then we present our auxiliary results, and finally we prove Proposition 3.

A. Preliminaries on typical sequences. The *type* P_{x^n} of a sequence $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ 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 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

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

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

$$\mathcal{T}_P^n \triangleq \{x^n | P_{x^n} = P\} \tag{5.3}$$

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

$$\mathcal{T}_V^n(x^n) \triangleq \{y^n | P_{x^n, y^n}(x, y) = P_{x^n}(x) V(x|y) \text{ for all } x \in \mathcal{X},\ y \in \mathcal{Y}\} \tag{5.4}$$

denotes the set of sequences in $\mathcal{Y}^n$ *V-generated* by x^n. Given a random variable X and a positive number η, we call $P \in \mathcal{P}_n$ an (x, η)-*essential type* if

$$\max_x |P(x) - P_X(x)| \le \eta, \qquad P(x) = 0 \text{ whenever } P_X(x) = 0. \tag{5.5}$$

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

$$\max_{x,y} |P_{x^n}(x)V(y|x) - P_{x^n}(x)P_{Y|X}(y|x)| \le \eta, \tag{5.6}$$

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

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

$$\mathcal{T}^n_{X,\eta} \triangleq \bigcup_{(X,\eta)\text{-ess } P} \mathcal{T}^n_P; \quad \mathcal{T}^n_{Y|X,\eta}(x^n) = \bigcup_{(x^n, Y|X, \eta)\text{-ess } V} \mathcal{T}^n_V(x^n). \tag{5.7}$$

The following basic inequalities are noted:

$$|\mathcal{P}_n| \le (n+1)^{|\mathcal{X}|}, \qquad |\mathcal{V}_n(P)| \le (n+1)^{|\mathcal{X}||\mathcal{Y}|}, \tag{5.8}$$

$$\Pr\{X^n \in \mathcal{T}^n_{X,\eta}\} \ge 1 - \frac{|\mathcal{X}|}{4n\eta^2}, \tag{5.9}$$

$$\Pr\{Y^n \in \mathcal{T}^n_{Y|X,\eta}(x^n) | X^n = x^n\} \ge 1 - \frac{|\mathcal{X}|\,|\mathcal{Y}|}{4n\eta^2} \quad \text{if } P_X^n(x^n) > 0. \tag{5.10}$$

B. Novel auxiliary results.

LEMMA 6 (Equivalence). *For $U \in \mathcal{M}_1$ and for any $u^m \in \mathcal{U}^m$ and any conditional type $V \in \mathcal{V}_m(P_{u^m})$,*

$$I(\theta; Y^m | X^m \in \mathcal{T}^m_V(u^m), U^m = u^m) = I(\theta; Y^m | X^m \in \mathcal{T}^m_V(u^m)).$$

PROOF. Let us use the abbreviation $A = \mathcal{T}^m_V(u^m)$. Then for the conditional density

$$f(Y^m = y^m | \theta, X^m \in A, U^m = u^m)$$

$$= \frac{1}{P^n_{X|U}(A|u^m)} \sum_{x^m \in A} f(y^m|\theta, x^m, u^m) P^n_{X|U}(x^m|u^m) \tag{5.11}$$

and therefore by the Markov property

$$I(\theta; Y^m|X^m \in A, U^m = u^m)$$

$$= \int_{y^m} \frac{\left[P^m_{X|U}(A|u^m)^{-1}\sum_{x^m \in A} f_\theta(y^m|\theta, x^m)P^m_{X|U}(x^m|u^m)\right]^2}{P^m_{X|U}(A|u^m)^{-1}\sum_{x^m \in A} f(y^m|\theta, x^m)P^m_{X|U}(x^m|u^m)} d\nu(y^m). \tag{5.12}$$

Since for $x^m \in A$,

$$P^m_{X|U}(A|u^m)^{-1}P^m_{X|U}(x^m|u^m) = P^m_X(A)^{-1}P^m_X(x^m), \tag{5.13}$$

and since

$$I(\theta; Y^m|X^m \in A)$$

$$= \int_{\mathcal{Y}^m} \frac{\left[P^m_X(A)^{-1}\sum_{x^m \in A} f_\theta(y^m|\theta, x^m)P^m_X(x^m)\right]^2}{P^m_X(A)^{-1}\sum_{x^m \in A} f_\theta(y^m|\theta, x^m)P^m_X(x^m)} d\nu(y^m), \tag{5.14}$$

Lemma 6 follows by comparison of the two quantities. □

LEMMA 7 (Monotonicity). *For every nonempty* $B \subset \mathcal{T}^m_V(u^m)$, *where* $u^m \in \mathcal{U}^m$ *and* $V \in \mathcal{V}_m(P_{u^m})$, *we have*

$$I(\theta; Y^m|X^m \in B) \geq I(\theta; Y^m|X^m \in \mathcal{T}^m_V(u^m)).$$

PROOF. Define the sets of components

$$\mathcal{K}(u^m, u) = \{t|1 \leq t \leq m, u_t = u\}, \qquad u \in \mathcal{U}, \tag{5.15}$$

and let $\mathcal{G}(u^m, u)$ be the group of permutations of the elements in $\mathcal{K}(u^m, u)$. The direct product

$$\mathcal{G}(u^m) = \prod_{u \in \mathcal{U}} \mathcal{G}(u^m, u) \tag{5.16}$$

is a group of permutations of $1, 2, \ldots, m$. For $\pi \in \mathcal{G}(u^m)$ and $x^m \in \mathcal{X}^m$, respectively $C \subset \mathcal{X}^m$, set

$$\pi(x^m) = (x_{\pi(1)}, \ldots, x_{\pi(m)}) \tag{5.17}$$

and

$$\pi(C) = \{\pi(x^m)|x^m \in C\}. \tag{5.18}$$

Denote the equidistribution on $\mathcal{T}^m_V(u^m)$ by Q and the equidistribution on C by Q_C. Observe that

$$Q = \frac{1}{|\mathcal{G}(u^m)|} \sum_{\pi \in \mathcal{G}(u^m)} Q_{\pi(B)}. \tag{5.19}$$

By concavity of I (Lemma 1),

$$I(\theta; Y^m|X^m \in \mathcal{T}^m_V(u^m)) \leq \frac{1}{|\mathcal{G}(u^m)|} \sum_{\pi \in \mathcal{G}(u^m)} I(\theta; Y^m|X^m \in \pi(B))$$

and since by the invariance of the model,

$$I(\theta; Y^m|X^m \in \pi(B)) = I(\theta; Y^m|X^m \in B),$$

the result follows. □

For the description of our next auxiliary result it is convenient to associate with every type $P \in \mathscr{P}_n(\mathscr{X})$ its *absolute* type $(n(1), \dots, n(a))$, where for $\mathscr{X} = \{1, \dots, a\}$,

$$n(x) = P(x)n, \qquad x \in \mathscr{X}. \tag{5.20}$$

Instead of the notation $\mathscr{T}_P^n$, *we use now also the notation* $\mathscr{T}^n(n(1), \dots, n(a))$. *We say that* $(n(1), \dots, n(a))$ *extends* $(m(1), \dots, m(a))$, *if*

$$n(i) \geq m(i) \quad \text{for } i \in \mathscr{X}. \tag{5.21}$$

LEMMA 8. *If* $(s(1), \dots, s(a))$ *extends* $(r(1), \dots, r(a))$, *then*

$$I(\theta; Y^r|X^r \in \mathscr{T}^r(r(1), \dots, r(a))) \geq I(\theta; Y^s|X^s \in \mathscr{T}^s(s(1), \dots, s(a))) - \sum_{i=1}^{a} (s(i) - r(i))I(\theta; Y|X = i).$$

PROOF. By Lemma 7 we have for any i with $s(i) \geq 1$,

$$\begin{aligned} &I(\theta; Y^s|X^s \in \mathscr{T}^s(s(1), \dots, s(a))) \\ &\quad \leq I\big(\theta; Y^s|X^{s-1} \in \mathscr{T}^{s-1}(s(1), \dots, s(i-1), s(i) - 1, \\ &\qquad\qquad s(i+1), \dots, s(a)), X_s = i\big) \\ &\quad = I\big(\theta; Y^{s-1}|X^{s-1} \in \mathscr{T}^{s-1}(s(1), \dots, s(i-1), s(i) - 1, \\ &\qquad\qquad s(i+1), \dots, s(a))\big) \\ &\qquad + I(\theta; Y_s|X_s = i) \end{aligned}$$

by the memoryless character of our model. The desired inequality follows by applying this step $s - r$ times. □

PROOF OF PROPOSITION 3. Clearly we can assume that, for some $\gamma > 0$,

$$H(U) - I(X \wedge U) > \gamma, \tag{5.22}$$

because otherwise $I(X \wedge U) = H(U)$ and then, U being a deterministic function of X, the choice $s_n(X^n) = U^n$ would do.

We now decompose the set of components $\{1, 2, \dots, n\}$ into the sets $\{1, \dots, m\}$ and $\{m + 1, \dots, m + l\}$, $n = m + l$. l and m will be specified below. We compose our encoding function s_n from two functions s_m' and s_l'', where s_m' will be defined on $\prod_{t=1}^m \mathscr{X}_t$, s_l'' is the identity map on $\prod_{t=m+1}^n \mathscr{X}_t$ and

$$s_n(x^n) = (s_m'(x_1, \dots, x_m), s_l''(x_{m+1}, \dots, x_n)). \tag{5.23}$$

The reason for this approach will become apparent below. We describe now s_m' by the construction of Ahlswede and Körner (1975), page 633.

It closely resembles Feinstein's maximal code construction in the formulation of chapter 3 in Wolfowitz (1978). The main difference is that for the purpose of approximation it uses codes with *large* error probabilities ($\lambda \to 1$), whose decoding sets essentially exhaust the output space. The properties stated below involve standard entropy bounds for cardinalities of sets of typical sequences [see Csiszár and Körner (1982) and Wolfowitz (1978)].

Suppose that $\varepsilon < \gamma$. We specify $\eta_1 = c_1(1/\sqrt{n})$ and $\eta_2 = c_2(1/\sqrt{n})$. Then for any λ, $0 < \lambda < 1$, and suitable constants c_1, c_2, there is a system of pairs $\{(v_j, D_j)\}_{j=1}^{N}$ with the properties:

(P1) $v_j \in \mathscr{T}^m_{U,\eta_1}$ for $j = 1, 2, \ldots, N$.

(P2) $D_j \subset \mathscr{T}^m_{X|U,\eta_2}(v_j)$ for $j = 1, 2, \ldots, N$ and $D_j \cap D_{j'} = \phi$ for $j \neq j'$.

(P3) $P^m_{X|U}(D_j|v_j) \geq 1 - \lambda$ for $j = 1, 2, \ldots, N$.

(P4) $P^m_X(D_0) \leq (1 - \lambda)P^m_U(\mathscr{U}^m - \{v_1, \ldots, v_N\}) + \lambda P^m_U(\{v_1, \ldots, v_N\})$, where $D_0 = \mathscr{X}^m - \bigcup_{j=1}^{N} D_j$.

(P5) $(1/m)\log(N + 1) \leq I(X \wedge U) + \varepsilon$ for large m.

It follows from (P1) and (P5), the choice $\varepsilon < \gamma$ and from (5.22) that $P^m_U(\{v_1, \ldots, v_N\}) \to 0$ when $m \to \infty$. (P4) and the fact that we can choose λ arbitrarily close to 1 imply that $P^m_X(D_0)$ can be made arbitrarily small for m sufficiently large. Define now

$$s'_m: \mathscr{X}^m \to \left\{(j, V): 0 \leq j \leq N, V \in \mathscr{V}_n\left(P_{v_j}\right)\right\}$$

by

$$s'_m(x^m) = (j, V), \quad \text{if } x^m \in D_j \cap \mathscr{T}^m_V(v_j). \tag{5.24}$$

It follows now from (5.8) and (P5) with the choice $\varepsilon = \rho/4$ that

$$I(X^m \wedge s'_m(X^m)) \leq m(I(X \wedge U) + \rho/2). \tag{5.25}$$

Therefore, also

$$I(X^n \wedge s_n(X^n)) \leq m(I(X \wedge U) + \rho/2) + l \log|\mathscr{X}|$$

and with the choice

$$l = n \min\left(\frac{\rho}{2\log|\mathscr{X}|}, \delta\right) \tag{5.26}$$

(i) holds.

We verify now (ii). By our definitions

$$\begin{aligned} I(\theta; Y^n|s_n(X^n)) &= I(\theta; Y^m|s'_m(X^m)) + I(\theta; Y_{m+1}, \ldots, Y_n|s''_l(X_{m+1}, \ldots, X_n)) \\ &= \sum_{j,V} P^m_X\left(D_j \cap \mathscr{T}^m_V(v_j)\right) I\left(\theta; Y^m|X^m \in D_j \cap \mathscr{T}^m_V(v_j)\right) \\ &\quad + P^m_X(D_0) I(\theta; Y^m|X^m \in D_0) + lI(\theta; Y|X). \end{aligned}$$

Furthermore, by Lemmas 7 and 6,

$$I\big(\theta; Y^m|X^m \in D_j \cap \mathscr{T}_V^m(v_j)\big) \geq I\big(\theta; Y^m|X^m \in \mathscr{T}_V^m(v_j), U^m = v_j\big)$$

and we can conclude that

$$\begin{aligned} I(\theta; Y^n|s_n(X^n)) &\geq lI(\theta; Y|X) \\ &\quad + \sum_{j,V}{}^* P_X^m\big(D_j \cap \mathscr{T}_V^m(v_j)\big) I\big(\theta; Y^m|X^m \in T_V^m(v_j), U^m = v_j\big), \end{aligned} \tag{5.27}$$

when the asterisk indicates summation over those V which are $(v_j, X|U, \eta_2)$-essential [recall (5.6)].

Notice that by (P2),

$$\bigcup_V{}^* D_j \cap T_V^m(v_j) = D_j \quad \text{for } j = 1, \ldots, N,$$

and since $P_X^m(D_0) \to 0 (\lambda \to 1,\ m \to \infty)$, we also have

$$\sum_{j,V}{}^* P_X^m\big(D_j \cap \mathscr{T}_V^m(v_j)\big) \to 1 \quad (\lambda \to 1,\ m \to \infty). \tag{5.28}$$

On the other hand, again by concavity of Fisher's information,

$$\sum_V P_{X|U}^m\big(\mathscr{T}_V(v_j)|v_j\big) I\big(\theta; Y^m|X^m \in \mathscr{T}_V(v_j), U^m = v_j\big) \geq I\big(\theta; Y^m|U^m = v_j\big).$$

Since

$$\begin{aligned} I\big(\theta; Y^m|X^m \in \mathscr{T}_V(v_j), U^m = v_j\big) &\leq m \max_x I(\theta; Y|X = x) \\ &\leq m\tau I(\theta; Y|X), \end{aligned}$$

where

$$\tau = \max_{x:\ P_X(x)>0} P_X(x)^{-1},$$

and since by (5.10),

$$\sum_V{}^* P_{X|V}^m\big(\mathscr{T}_V(v_j)|v_j\big) \geq 1 - \frac{|\mathscr{U}|\,|\mathscr{X}|}{4m\eta_2^2},$$

we can conclude that

$$\begin{aligned} &\sum_V{}^* P_{X|U}^m\big(\mathscr{T}_V(v_j)|v_j\big) I\big(\theta; Y^m|X^m \in \mathscr{T}_V(v_j), U^m = v_j\big) \\ &\quad \geq I\big(\theta; Y^m|U^m = v_j\big) - \frac{|\mathscr{U}|\,|\mathscr{X}|\tau}{4\eta_2^2} I(\theta; Y|X). \end{aligned} \tag{5.29}$$

Notice that (5.27) and (5.29) would readily establish (ii) if the coefficients in the two sums were equal. This not being the case we circumvent the difficulty by exploiting the idea that the information quantities do not change very much as long as V varies over $(v_j, X|U, \eta_2)$ essential types. Technically we do this with

the help of Lemma 8 by making the comparison via longer sequences of length $k = m + l/2$. Consider any u^k for which $P_u k$ is (U, η_1)-essential and let W be $(u^k, X|U, \eta_2)$-essential. For fixed u define $s(x) = P_{u^k}(u)W(x|u)k$ and $r(x) = P_{v_j}(u)V(x|u)m$. Thus $s = P_{u^k}(u)k$, $r = P_{v_j}(u)m$ and, for η_1, η_2 sufficiently small, Lemma 8 implies

$$I\big(\theta; Y^r|X^r \in \mathscr{T}^r_{V(\cdot|u)}, U^r = (u, \dots, u)\big)$$
$$\geq I\big(\theta; Y^s|X^s \in \mathscr{T}^s_{W(\cdot|u)}, U^s = (u, \dots, u)\big)$$
$$- \sum_{i=1}^{a} \big(P_{u^k}(u)W(i|u)k - P_{v_j}(u)V(i|u)m\big)I(\theta; Y|X = i)$$

and hence for $\eta = \eta_1 + \eta_2$,

$$I\big(\theta; Y^m|X^m \in \mathscr{T}^m_V(v_j), U^m = v_j\big) \geq I\big(\theta; Y^k|X^k \in \mathscr{T}^k_W(u^k), U^k = u^k\big)$$
$$- \sum_{i=1}^{a} P_X(i)(k - m)I(\theta; Y|X = i)$$
$$- \sum_{i=1}^{a} \eta(k - m)I(\theta; Y|X = i).$$

Since this holds for *all* essential V and W, and since (5.29) holds with m, v_j, V replaced by k, u^k, W, we derive from (5.27)

$$I\big(\theta; Y^n|s_n(X^n)\big) \geq lI(\theta; Y|X) + I\big(\theta; Y^k|U^k = u^k\big) - (k - m)I(\theta; Y|X)$$
$$- \eta\tau|\mathscr{X}|(k - m)I(\theta; Y|X) - \frac{|\mathscr{U}|\,|\mathscr{X}|}{4\eta_2^2}\tau I(\theta; Y|X).$$

Obviously we have also

$$I\big(\theta; Y^k|U^k = u^k\big) \geq kI(\theta; Y|U) - k\eta\tau I(\theta; Y|X)$$

and thus for η small and $n \geq n_0(\eta)$,

$$I\big(\theta; Y^n|s_n(X^n)\big) \geq kI(\theta; Y|U) \geq n(1 - \delta)I(\theta; Y|U). \qquad \square$$

6. Regularity conditions for achievability of the informational bound. Among the basic work on the asymptotic theory of estimators we mention the important contributions Le Cam (1953, 1956, 1970), where one also finds a very good historical account, and Ibragimov and Khas'minskii (1972, 1973, 1975b). Various sets of regularity conditions have been considered in the extensive literature. The presence of side information requires additional uniformity conditions. Our aim here is not to have great generality but to have reasonable conditions under which our novel bounds can be established with not too much mathematical effort.

Suitable in our case are those conditions on the set of densities which can be lifted to the case with generic variables

$$(Y_1, \dots, Y_r), s_r(X_1, \dots, X_r).$$

For our Cramér–Rao-type inequality (Theorems 1 and 2 in Section 4) we need, for instance, that the Fisher information $I(\theta; Y^n, s_n(X^n))$ is locally bounded in Θ. This property follows from the condition (C1). As in the classical situation asymptotic achievability of the informational bound is guaranteed only under stronger regularity conditions. In the choice of our conditions here we follow closely those of Ibragimov and Khas'minskii (1972, 1973, 1975b).

Relevant for us are their conditions in groups I, II and III of (1972) and groups III, IV and V of (1975b). We now present our substitutes. Asterisks identify those related to the groups in (1975b).

We begin with the conditions relating to those in (1972).

Whereas their conditions I_1, I_2, and I_3 are already incorporated in our model, we have to substitute for I_4:

(I_4) There exists an $\mathscr{X}' \subset \mathscr{X}$ of positive measure with

$$\int_{\mathscr{Y}} \left| \int_{\mathscr{X}'} f(y|\theta, x)p(x)\mu(dx) - \int_{\mathscr{X}'} f(y|\theta', x)p(x)\mu(dx) \right| \nu(dy) > 0 \quad \text{for } \theta \neq \theta'.$$

The conditions in our group II are:

(II_1) For every $y \in \mathscr{Y}$ the function $f(y|\theta, x)$ is *uniformly in* x absolutely continuous on compact subsets of Θ and, for ν-almost all y and every x, all points $\theta \in \Theta$ are Lebesgue points of the function $\partial f(y|\theta, x)/\partial\theta$.

(II_2) For all $\theta \in \Theta$, $x \in \mathscr{X}$ the Fisher information at θ conditional on x exists, i.e.,

$$I(\theta|x) \triangleq \int_{\mathscr{Y}} \frac{|f_\theta(y|\theta, x)|^2}{f(y|\theta, x)} \nu(dy) < \infty.$$

The integrand is assumed to vanish wherever $f(y|\theta, x) = 0$.

(II_3) $I(\theta|U)$ is a continuous function of θ in Θ for all $U \in \mathscr{M}_1$.

(II_4) There exists a nonnegative number p such that

$$\sup_\theta (1 + |\theta|)^{-p} I(\theta|x) < \infty \quad \text{for all } x \in \mathscr{X}.$$

(II_5) For some $\delta > 0$,

$$\int_{\mathscr{Y}} \frac{|f_\theta(y|\theta, x)|^{2+\delta}}{f(y|\theta, x)^{1+\delta}} \nu(dy)$$

is locally bounded in θ for all $x \in \mathscr{X}$.

Our condition in group III is:

(III_1) There exists a $\delta > 0$ such that for $\theta_0 \in \Theta$,

$$\sup_{\theta \in \Theta} |\theta - \theta_0|^\delta \sup_{x, x' \in \mathscr{X}} \int \sqrt{f(y|\theta, x) f(y|\theta_0, x')}\, \nu(dy) < \infty.$$

Our first auxiliary result implies that properties (II_1)–(II_5) extend to encoding functions. This will be used for the proof of Theorems 3 and 4 in Section 7.

LEMMA 9. *Let $\mathcal{X}$ be finite and assume that (II_1)–(II_5) hold for $\{f(y|\theta, x)\}$. For $U \in \mathcal{M}_n$ define*

$$f(y^n|\theta, u) = \sum_{x^n} \prod_{t=1}^{n} f(y_i|\theta, x_t) P_{X^n|U}(x^n|u). \tag{6.1}$$

Then (II_1)–(II_5) hold for $\{f(y^n|\theta, u)\}$.

PROOF. (i) (II_1) holds again, because $f(y^n|\theta, u)$ is a polynomial in functions satisfying (II_1).

(ii)
$$\begin{aligned} I(\theta; Y^n|U) &= I(\theta; Y^nU) \le I(\theta; Y^nX^nU) \\ &= I(\theta; Y^n|X^nU) \quad \text{(by Lemmas 1, 2)} \\ &= nI(\theta; Y|X) \quad \text{(by the Markov property and Lemma 2)} \end{aligned}$$

and

$$n\sum_x P_X(x) I(\theta; Y|x) < \infty, \quad \text{by assumption.}$$

(iii)
$$I(\theta; Y^n|U) = \sum_{t=1}^{n} I\big(\theta; Y_t|Y^{t-1}U\big)$$

Since $Y^{t-1}U \ominus Y_t \ominus X_t$, the summands are continuous in θ and thus also $I(\theta; Y^n|U)$.

(iv) Since $I(\theta; Y^n|U) \le nI(\theta; Y|X)$, (II_4) holds again.

(v) It suffices to consider the case $n = 2$, because we can interate the argument. It is clear from (6.1) that it suffices to establish (II_5) first for $\prod_{t=1}^{2} f(y_t, |\theta, x_t)$ and then for convex combinations of such functions. The result therefore follows from the following two inequalities (a) and (b).

Let f_1, f_2 be two density functions with

$$\int \frac{|f_i'|^{2+\delta}}{f_i^{1+\delta}}\, d\nu_i < \infty \quad \text{for } i = 1, 2.$$

Then

(a)
$$\frac{|(f_1 f_2)'|^{2+\delta}}{(f_1 f_2)^{1+\delta}}\, d\nu_1 \times d\nu_2 \le 2^{1+\delta} \sum_{i=1}^{2} \int \frac{|f_i'|^{2+\delta}}{f_i^{1+\delta}}\, d\nu_i.$$

This follows with the well-known inequality $|a+b|^\rho < 2^{\rho-1}(|a|^\rho + |b^\rho|)$, $\rho \geq 2$, as follows:

$$\int \frac{|f_1' f_2 + f_1 f_2'|^{2+\delta}}{(f_1 f_2)^{1+\delta}}\, d\nu_1 \times d\nu_2$$

$$\leq 2^{1+\delta} \int \frac{|f_1' f_2|^{2+\delta} + |f_1 f_2'|^{2+\delta}}{(f_1 f_2)^{1+\delta}}\, d\nu_1 \times d\nu_2$$

$$= 2^{1+\delta}\left[\int \frac{|f_1'|^{2+\delta}}{f_1^{1+\delta}} f_2\, d\nu_1 \times d\nu_2 + \int \frac{|f_2'|^{2+\delta}}{f_2^{1+\delta}} f_1\, d\nu_1 \times d\nu_2\right]$$

$$= 2^{1+\delta}\left[\int \frac{|f_1'|^{2+\delta}}{f_1^{1+\delta}}\, d\nu_1 + \int \frac{|f_2'|^{2+\delta}}{f_2^{1+\delta}}\, d\nu_2\right].$$

Further, for $0 < \lambda < 1$ with the same inequality,

$$\int \frac{|\lambda f' + (1-\lambda)g'|^{2+\delta}}{(\lambda f + (1-\lambda)g)^{1+\delta}}\, d\nu$$

$$\leq 2^{1+\delta} \int \frac{|\lambda f'|^{2+\delta} + |(1-\lambda)g'|^{2+\delta}}{(\lambda f + (1-\lambda)g)^{1+\delta}}\, d\nu \tag{b}$$

$$\leq 2^{1+\delta}\left[\int \frac{|\lambda f'|^{2+\delta}}{(\lambda f)^{1+\delta}}\, d\nu + \int \frac{|(1-\lambda)g'|^{2+\delta}}{(\lambda g)^{1+\delta}}\, d\nu\right],$$

because $f, g \geq 0$. □

Next we show that condition (III_1) extends to encoding functions.

LEMMA 10. *Let $\mathcal{X}$ be finite. If (III_1) holds for $\{f(y|\theta, x)\}$, then it holds for $\{f(y^n|\theta, u)\}$ as defined in* (6.1).

PROOF. Since for nonnegative reals $\sqrt{\Sigma_i a_i} \leq \Sigma_i \sqrt{a_i}$, clearly

$$\int \sqrt{\sum_{x^n} f(y^n|\theta, x^n) p(x^n|u) \sum_{x'^n} f(y^n|\theta_0, x'^n) p(x'^n|u')}\, \nu(dy^n)$$

$$\leq \sum_{x^n, x'^n} \int \sqrt{f(y^n|\theta, x^n) f(y^n|\theta_0, x'^n)}\, \nu(dy^n)$$

$$\leq |\mathcal{X}|^{2n}\left(\max_{x, x'} \int \sqrt{f(y|\theta, x) f(y|\theta_0, x')}\, \nu(dy)\right)^n.$$

Therefore

$$\sup_{\theta\in\Theta}|\theta-\theta_0|^{\delta n}\int\sqrt{f(y^n|\theta,u)f(y^n|\theta_0,u')}\,\nu(dy^n)$$

$$\le\left(|\mathscr{X}|^2\sup_{\theta\in\Theta}|\theta-\theta|^{\delta}\max_{x,x'}\int\sqrt{f(y|\theta,x)f(y|\theta_0,x')}\,\nu(dy)\right)^n<\infty. \qquad \square$$

In order to obtain results on *uniform* asymptotic efficiency on a closed interval $A\subset\Theta$, we shall apply Theorem 3.1 of Ibragimov and Khas'minskii (1975b) in the next section. For this we have to extend conditions I–V of that paper to the case of side information. For the conditions I and II this is done already with our conditions in group I. The conditions III–V are to be replaced by the following conditions III*–V*.

For $d>0$ define

$$B(d,\theta,x)=\left\{y:\left|\frac{f_\theta(y|\theta,x)}{f(y|\theta,x)}\right|>d\right\} \tag{6.2}$$

and for a closed interval $A\subset\Theta$ set

$$\mathbf{I}_A(x)=\inf_{\theta\in A}I(\theta;Y|X=x). \tag{6.3}$$

The conditions in group III* are

$$0<\mathbf{I}_A(x)\qquad\text{for } x\in\mathscr{X}. \tag{III$_1^*$}$$

$$\sup_{\theta\in A}I(\theta|x)^{-1}E_\theta\left\{\left(\frac{f_\theta(\cdot|\theta,x)}{f(\cdot|\theta,x)}\right)^2 1_{B(d,\theta,x)}\right\}\to 0 \text{ as } d\to\infty \tag{III$_2^*$}$$

for all $x\in\mathscr{X}$.

We come now to the more restrictive conditions IV* and V*, which are not of single-letter type:

$$\sup_{\theta,\,\Delta\in A;\,|\theta-\Delta|\le\varepsilon}\int_{\mathscr{Y}^r}\left(\frac{\partial\left(\sum_{x^r}f(y^r|\theta,x^r)P(x^r)\right)^{1/2}}{\partial\theta}-\frac{\partial\left(\sum_{x^r}f(y^r|\Delta,x^r)P(x^r)\right)^{1/2}}{\partial\theta'}\right)d\nu(y^r)\to 0 \text{ as } \varepsilon\to 0 \tag{IV*}$$

for every r and every probability distribution P on $\mathscr{X}^r$.

$$\inf_{\theta,\,\Delta\in A;\,|\theta-\Delta|>\varepsilon}\int_{\mathscr{Y}^r}\left(\left(\sum_{x^r}f(y^r|\theta,x^r)P(x^r)\right)^{1/2}-\left(\sum_{x^r}f(y^r|\Delta,x^r)P(x^r)\right)^{1/2}\right)^2 d\nu(y^r)>0 \tag{V*}$$

for every $\varepsilon>0$ and all probability distributions P on $\mathscr{X}^r$.

We have seen already in Lemmas 9 and 10 that the condition (III_1) extends to coding functions. There is no essential loss if we consider later only coding functions which reproduce X_t for one component t. Therefore for these functions we have, by (III_1^*),

$$I(\theta|s_r) > 0 \quad \text{for all } \theta \in A. \tag{6.4}$$

By our next result these coding functions also satisfy (III_2^*).

LEMMA 11. (II_3), (II_5), *and* (III_1^*) *imply* (III_2^*) *for finite* $\mathscr{X}$.

PROOF. By Hölder's inequality,

$$\int_{B(d,\theta,x)} \frac{f_\theta(y|\theta,x)^2}{f(y|\theta,x)}\nu(dx)$$

$$\leq \left(\int_{\mathscr{Y}} \frac{|f_\theta(y|\theta,x)|^{2+\delta}}{f(y|\theta,x)^{1+\delta}}\nu(dy)\right)^{2/(2+\delta)} \left(\int_{B(d,\theta,x)} f(y|\theta,x)\nu(dy)\right)^{\delta/(2+\delta)}.$$

Since the first term to the right is by assumption (II_5) bounded in A and since by (III_1^*), $\sup_{\theta\in A} I(\theta|x)^{-1} = \mathbf{I}_A^{-1}(x) < \infty$, it suffices to show that

$$\lim_{d\to\infty}\left(\int_{B(d,\theta,x)} f(y|\theta,x)\nu(dy)\right)^{\delta/(2+\delta)} = 0.$$

Now notice that by definition of $B(d,\theta,x)$,

$$\int_{B(d,\theta,x)} f(y|\theta,x)\nu(dy) \leq \int_B \frac{|f_\theta(y|\theta,x)|}{d}\nu(dy)$$

$$\leq \int_B \frac{(f_\theta(y|\theta,x))^2}{f(y|\theta,x)d^2}\nu(dy) \leq \frac{I(\theta|x)}{d^2}$$

and the result follows with (II_3), since $\mathscr{X}$ is finite. □

We are thus left only with (IV*) and (V*) as non-single-letter conditions. Nevertheless we use them, because we can think only of single-letter conditions, which are more restrictive. Also, sometimes they can be verified without much effort. We discuss an important case.

EXAMPLE 6 ($\mathscr{X}$ and $\mathscr{Y}$ finite). In this case for $\{p(y|\theta)\colon y \in \mathscr{Y}, \theta \in A\}$,

$$\sup_{\theta,\,\Delta\in A;\,|\theta-\Delta|\leq\varepsilon} \sum_{y\in\mathscr{Y}}\left(\frac{p_\theta(y|\theta)}{\sqrt{p(y|\theta)}} - \frac{p_\theta(y|\Delta)}{\sqrt{p(y|\Delta)}}\right)^2 \to 0 \text{ as } \varepsilon \to 0,$$

is *equivalent* to the continuity of $(p_\theta(y|\theta))/\sqrt{p(y|\theta)}$ in the compact set A for

all $y \in \mathcal{Y}$. Here we use the notation

$$p_\theta(y|\Delta) = \frac{\partial p(y|\theta)}{\partial \theta}\bigg|_{\theta=\Delta}.$$

One readily verifies that with the continuity of

$$\frac{p_\theta(y|\theta)}{\sqrt{p(y|\theta)}}, \qquad y \in \mathcal{Y},$$

and of

$$\frac{q_\theta(z|\theta)}{\sqrt{q(z|\theta)}}, \qquad z \in Z \text{ finite},$$

also $(p(y|\theta)q(z|\theta))_\theta / \sqrt{p(z|\theta)q(z|\theta)}$ is continuous on A for all $(y, z) \in \mathcal{Y} \times \mathcal{Z}$. Obviously also for

$$\mathcal{Z} \triangleq \mathcal{Y}, \qquad \frac{\lambda p_\theta(y|\theta) + (1-\lambda)q_\theta(y|\theta)}{\sqrt{p(y|\theta) + (1-\lambda)q(y|\theta)}}$$

is continuous on A for all $y \in \mathcal{Y}$. Therefore, (IV_1^*) holds for $\{p(y|\theta, x)\colon x \in \mathcal{X}, y \in \mathcal{Y}, \theta \in A\}$, if $(p_\theta(y|\theta, x))/\sqrt{p(y|\theta, x)}$ is continuous on A for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$.

For the present example the single-letter condition

$$\inf_{\theta, \Delta \in A;\, |\theta - \Delta| > \varepsilon} \int_{\mathcal{Y}} \left(\left(\sum_x p(y|\theta, x)p(x) \right)^{1/2} - \left(\sum_x p(y|\Delta, x)p(x) \right)^{1/2} \right)^2 > 0 \tag{6.5}$$

for every $\varepsilon > 0$ and all probability distributions implies (V*), if $p(y|\theta, x)$ is continuous in θ for all x, y. This readily follows from the fact that for two stochastic matrices W, V with

$$\sum_{x_1, x_2} W(y_1|x_1)W(y_2|x_2)p(x_1, x_2) = \sum_{x_1, x_2} V(y_1|x_1)V(y_2|x_2)p(x_1, x_2)$$

for all y_1, y_2, necessarily (by summation)

$$\sum_{x_1} W(y_1|x_1)p(x_1) = \sum_{x_1} V(y_1|x_1)p(x_1) \quad \text{for all } y_1.$$

7. Asymptotic achievability of the informational bound in case of a finite $\mathcal{X}$. In order to avoid unpleasant technicalities in the handling of the side information we assume here that X takes only finitely many values.

We have to find for all large n suitable encoding functions and suitable estimators. For the encoding functions we only provide an existence proof (Proposition 3 in Section 5). We shall also always use suitable encoding functions repeatedly.

A. Asymptotic efficiency for an encoding function. To fix ideas let us first use any encoding function s_r repeatedly. Thus we are in the familiar case of i.i.d. drawings $(\tilde{X}_i, \tilde{Y}_i)_{i=1}^{\infty}$, where

$$\tilde{X}_i = s_r(X_{(i-1)r+1}, \dots, X_{ir}), \tag{7.1}$$

$$\tilde{Y}_i = (Y_{(i-1)r+1}, \dots, Y_{ir}). \tag{7.2}$$

Since $I(\theta; \tilde{X}, \tilde{Y}) = I(\theta; \tilde{Y}|\tilde{X})$ and since by Lemmas 9 and 10 in Section 6 the properties in groups I, II, and III extend to s_r, we can apply Theorem 5.1 of Ibragimov and Khas'minskii (1973) and get:

THEOREM 3. *Assume that the conditions in groups* I, II, *and* III *hold. For* $((\tilde{X}_i, \tilde{Y}_i))_{i=1}^{\infty}$ *as defined in* (7.1) *and* (7.2) *and* $n = l \cdot r$ *we have for the MLE* $\hat{\theta}$:

(a) $\sqrt{n}\,(\hat{\theta}(\tilde{X}^l, \tilde{Y}^l) - \theta)$ *is asymptotically normal with parameters* $(0, ((1/r)I(\theta|s_r(X^r))^{-1})$.

(b) *For all* $\alpha > 0$,

$$\lim_{n\to\infty} n^{\alpha/2} E\left|\hat{\theta}(\tilde{X}^l, \tilde{Y}^l) - \theta\right|^{\alpha} = \left(\frac{2}{(1/r)I(\theta|s_r)}\right)^{\alpha/2} \frac{\Gamma\left(\frac{1}{2}(\alpha+1)\right)}{\sqrt{\pi}}.$$

In particular for $\alpha = 2$,

(c) $$\lim_{n\to\infty} nE\left|\hat{\theta}(\tilde{X}^l, \tilde{Y}^l) - \theta\right|^2 = \frac{r}{I(\theta|s_r)}, \qquad \theta \in \Theta,$$

that is, $\hat{\theta}$ *is asymptotically efficient for* s_r.

B. Uniform asymptotic efficiency for an encoding function. We have explained in Section 6 that it suffices to study encoding functions s_r, which satisfy (6.4). Application of Theorem 3.1 of Ibragimov and Khas'minskii (1975b) gives the following result.

THEOREM 4. *Assume that the conditions in groups* I *and* II *as well as the conditions* (III_1), (IV), *and* (V) *hold. For* $n = l \cdot r$ *we have for the MLE* $\hat{\theta}$:

(a) $\hat{\theta}(\tilde{X}^l, \tilde{Y}^l)$ *is consistent in the closed interval* $A \subset \Theta$.

(b) *For all* $l \geq l_0(k)$ *this estimator has a moment of positive integral order* k *and for any function* $h(z)$ *growing no faster than a power function as* $|z| \to \infty$, *the following relation is satisfied uniformly in* $\theta \in A$:

$$\lim_{l\to\infty} E_\theta h\left(\left(\hat{\theta}(\tilde{X}^l, \tilde{Y}^l) - \theta\right)\sqrt{lI(\theta|s_r)}\right) = \frac{1}{\sqrt{2\pi}} \int h(z) e^{-z^2/2}\, dz.$$

(c) *In particular, for* $h(z) = z^2$,

$$\lim_{l\to\infty} E_\theta\left(\hat{\theta}(\tilde{X}^l, \tilde{Y}^l) - \theta\right)^2 \cdot l = \frac{1}{I(\theta|s_r)}$$

uniformly in $\theta \in A$.

C. Achievability of the informational bound for a closed interval. Let $\varepsilon, \delta, \rho$ be small numbers, which we further specify soon. By definition of $J(R)$ there exists an m and a $U \in \mathscr{M}_m(R - \varepsilon)$ such that

$$J(R - \varepsilon) \leq \frac{1}{m} I(\theta|U) + \frac{\eta}{3} \quad \text{for all } \theta \in A.$$

By Proposition 3 there exists an $r = m \cdot k$ and an s_r with

$$(1 - \delta)\frac{1}{m} I(\theta|U) \leq \frac{1}{r} I(\theta|s_r) \quad \text{for } \theta \in A$$

and

$$I(X^r \wedge s_r(X^r)) \leq k(I(X \wedge U) + \rho) \leq k(m(R - \varepsilon) + \rho).$$

Thus

$$\text{rate}(s_r) \leq R - \varepsilon + \frac{\rho}{k}$$

and

$$\frac{1}{r} I(\theta|s_r) \geq (1 - \delta)\left[J(R - \varepsilon) - \frac{\eta}{3}\right] \quad \text{for } \theta \in A.$$

Choosing ε such that $J(R - \varepsilon) \geq J(R) - \eta/3$, $\rho < \varepsilon$ and finally δ such that $(1 - \delta)(J(R) - 2\eta/3) \geq J(R) - \eta$, we arrive at the inequalities

$$\text{rate}(s_r) \leq R, \tag{7.3}$$

$$\frac{1}{r} I(\theta|s_r) \geq J(R) - \eta. \tag{7.4}$$

Write n in the form $n = lr + j$, $0 \leq j < r$. By ignoring the last j observations we define an estimator $\hat{\theta}_n$ by

$$\hat{\theta}_n = \hat{\theta}(\tilde{X}^l, \tilde{Y}^l). \tag{7.5}$$

These relations and Theorem 4 imply

THEOREM 5. *Under the assumptions of Theorem* 4, *for any rate* $R > 0$ *and any* η, $0 < \eta < J(R)$, *there is an estimator* $\hat{\theta}_n$ *based on an encoding function of rate* R *such that*

$$\limsup_{n \to \infty} \sup_{\theta \in A} E_\theta(\hat{\theta}_n - \theta)^2 n \leq \frac{1}{J(R) - \eta}.$$

REMARKS.

1. We have proved asymptotic achievability of $J(R)$ within an arbitrarily small accuracy η. By a proper modification of the scheme to the case of a sequence $\{s_{r_i}\}_{i=1}^{\infty}$ of independent but nonidentically distributed coding functions one can establish, with the help of Theorem 3.1 in Ibragimov and Khas'minskii

(1975b), exact asymptotic achievability. A formal proof requires lengthy calculations without any new ideas.

2. The results of this paper were announced at the NRW Colloquium on Statistics held in Bielefeld, May 31–June 1, 1984. They have been presented at the IVth International Vilnius Conference on Probability Theory and Statistics held on June 24–29, 1985.
3. In independent work, "Estimation via encoded information," Z. Zhang and T. Berger have also considered the problem of parameter estimation for bivariate distributions under communication constraints. They have included also the case where *both* marginal distributions can be made available to the statistician only at limited rates. In this greater generality they show, under certain regularity conditions, the existence of a sequence of *unbiased* estimators with variances converging to 0 at a speed $O(1/n)$. In the gaussian case they prove uniformity of this convergence in θ. Since the paper contains no result on efficiency or even a Cramér–Rao-type inequality, the overlap with our paper is negligible.

8. J single-letterizes in the symmetric Bernoulli case. Recall Example 1 in Section 2. We shall show that in case of side information,

$$\sup_{U\in\mathscr{M}_1(R)} \inf_{\theta} I(\theta; Y|U) = \inf_{\theta} \sup_{U\in\mathscr{M}_1(R)} I(\theta; Y|U) \tag{8.1}$$

and that therefore, by Lemma 5,

$$J(R) = \lim_{n\to\infty} J_n(R) = J_1(R),$$

which is the desired single-letterization. The identity (8.1) is an immediate consequence of

$$I\left(\tfrac{1}{2}; Y|U\right) \le I(\theta; Y|U) \quad \text{for all } \theta \text{ and all } U \ominus X \ominus Y(\theta),\ \theta \in \Theta, \tag{8.2}$$

which we now prove.

For any U with values in $\mathscr{U} = \{u_1, \dots, u_a\}$ we define

$$p_i = P_U(u_i), \qquad q_i = P_{X|U}(0|u_i).$$

Since $P_X(0) = P_X(1) = \frac{1}{2}$, necessarily

$$\sum_i p_i q_i = \tfrac{1}{2}. \tag{8.3}$$

The constraint $I(X \wedge U) \le R$ takes the form

$$\sum_i p_i h(q_i) \ge H(X) - R, \tag{8.4}$$

where h is the binary entropy function.

$I(\theta; Y|U)$ can readily be calculated as follows,

$$I(\theta; Y|U = u_i) = \sum_y \frac{P_\theta(y|\theta, u_i)^2}{P(y|\theta, u_i)},$$

$$P(0|\theta, u_i) = \theta q_i + (1-\theta)(1-q_i),$$

$$P(1|\theta, u_i) = (1-\theta)q_i + \theta(1-q_i),$$

and thus

$$I(\theta; Y|U = u_i) = \frac{(2q_i-1)^2}{\theta q_i + (1-\theta)(1-q_i)} + \frac{(1-2q_i)^2}{(1-\theta)q_i + \theta(1-q_i)}$$

$$= \frac{(2q_i-1)^2}{[\theta q_i + (1-\theta)(1-q_i)][(1-\theta)q_i + \theta(1-q_i)]}$$

$$= \frac{(2q_i-1)^2}{f(\theta, i)},$$

if we use the abbreviation

$$f(\theta, i) = [\theta(2q_i-1) + (1-q_i)][\theta(1-2q_i) + q_i]. \tag{8.5}$$

Hence

$$I(\theta; Y|U) = \sum_i p_i \frac{(2q_i-1)^2}{f(\theta, i)}. \tag{8.6}$$

Now for (8.2) to hold it suffices to show that for all i, $f(\theta, i)$ takes its maximum at $\theta = \frac{1}{2}$. Clearly,

$$\frac{df(\theta, i)}{d\theta} = (2q_i-1)[\theta(1-2q_i) + q_i] + [\theta(2q_i-1) + (1-q_i)](1-2q_i)$$

and

$$\left.\frac{df(\theta, i)}{d\theta}\right|_{\theta=1/2} = (2q_i-1)\left[\frac{1-2q_i}{2} + q_i - \frac{2q_i-1}{2} - 1 + q_i\right] = 0.$$

Furthermore,

$$\frac{d^2f(\theta, i)}{d\theta^2} = -(2q_i-1)^2 \cdot 2 < 0, \quad \text{for } q_i \neq \tfrac{1}{2} \text{ and for } q_i = \tfrac{1}{2} f(\theta, i)$$

is independent of θ. In any case, $\theta = \frac{1}{2}$ is a maximal value of $f(\theta, i)$ and thus (8.2) holds.

Finally, we mention that by a somewhat lengthy calculation it can be shown that for rate R an optimal choice of U is specified by

$$\mathcal{U} = \{1,2\}, \qquad P_U = \left(\tfrac{1}{2}, \tfrac{1}{2}\right), \qquad P_{X|U} = \begin{pmatrix} c & 1-c \\ 1-c & c \end{pmatrix}, \tag{8.7}$$

where c is a solution of $1 - h(c) = R$.

Furthermore, in this case

$$I(\theta; Y|U)^{-1} = [\theta(1-\theta) + c(1-c)](1-2c)^{-2}. \tag{8.8}$$

REFERENCES

AHLSWEDE, R. (1979). Coloring hypergraphs: A new approach to multi-user source coding. I. *J. Combin. Inform. System Sci.*, **1** 76–115; (1980). II. **5** 220–268.

AHLSWEDE, R. and CSISZÁR, I. (1986). Hypothesis testing with communication constraints. *IEEE Trans. Inform. Theory* **IT-32** 533–542.

AHLSWEDE, R. and KÖRNER, J. (1975). Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. Inform. Theory* **IT-21** 629–637.

CENCOV, N. N. (1972). *Statistical Decision Rules and Optimal Inference*. Nanka, Moscow (in Russian; transl. by AMS).

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press, Princeton, N.J.

CSISZÁR, I. and KÖRNER, J. (1982). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic, New York.

EDGEWORTH, F. Y. (1908–09). On the probable errors of frequency constants. *J. Roy. Statist. Soc.* **71** 381–397, 499–512; **72** 81–90.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

IBRAGIMOV, I. A. and KHAS'MINSKII, R. Z. (1972). Asymptotic behaviour of statistical estimators in the smooth case I. Study of the likelihood ratio. *Theory Probab. Appl.* **17** 445–462.

IBRAGIMOV, I. A. and KHAS'MINSKII, R. Z. (1973). Asymptotic behaviour of some statistical estimators II. Limit theorems for the a posteriori and Bayes' estimators. *Theory Probab. Appl.* **18** 76–91.

IBRAGIMOV, I. A. and KHAS'MINSKII, R. Z. (1975a). Information-theoretic inequalities and super-efficient estimates. *Problems Inform. Transmission* **9** 216–227.

IBRAGIMOV, I. A. and KHAS'MINSKII, R. Z. (1975b). Properties of maximum likelihood and Bayes' estimators for non-identically distributed observations. *Theory Probab. Appl.* **20** 689–697.

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22** 79–86.

LE CAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. Univ. Calif. Publ. Statist. I, 277–330.

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypothesis. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 129–156.

LE CAM, L. (1970). On the assumption used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41** 803–826.

RAO, J. C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** 81–91.

WOLFOWITZ, J. (1978). Coding Theorems of Information Theory. Springer, 3rd edition.

YAO, A. C. (1979). Some complexity questions related to distributive computing. *11th ACM Symp. Theory of Computing* 209–213.

FAKULTÄT FÜR MATHEMATIK
UNIVERSITÄT BIELEFELD
UNIVERSITÄTSSTRASSE 1
4800 BIELEFELD 1
WEST GERMANY

INSTITUTE OF INFORMATION TRANSMISSION
ACADEMY OF SCIENCES USSR
ERMOLOVOY 19
101447 MOSCOW
SOVIET UNION