

MAXIMIZATION OF EMPIRICAL SHANNON INFORMATION IN TESTING SIGNIFICANT VARIABLES OF LINEAR MODEL

M. MALYUTOV AND H. SADAKA

ABSTRACT. Search for an unknown set $A, Card(A) = s$, of significant variables of a linear model with random IID discrete binary carriers and finitely supported IID noise is studied. Two statistics T_1, T_s , based on maximization of Shannon Information (SI) of the corresponding classes of joint empirical input-output distributions, are proposed inspired by the related study in Csiszar and Körner (1981). The first one compares sequences of values of each variable and of the output separately. The second one explores the relation between the subsets of the $(N \times t)$ design matrix corresponding to each subset of variables of given cardinality and the output sequence. Here N is the number of experiments and t is the total number of variables. Both statistics are shown to be asymptotically as efficient as the ML-test for the corresponding classes of joint empirical distributions in the artificial case when ML-test is applicable: if the unknown parameters $b_\lambda, \lambda \in A$, of the model and the distribution of errors are known. Our tests do not require this information. Therefore, they are asymptotically uniformly most efficient in the corresponding classes of tests. The second statistic is shown to provide asymptotically best rate of search for the set A of significant variables when $t \rightarrow \infty$, but requires about $t^s \log t$ cycles of computing. This may appear in accessible for actual computations in some applications. The first statistic requires only $t \log t$ cycles of computing operations and provides the best order of magnitude of the characteristics studied for the second class of tests.

INTRODUCTION. FORMULATION OF MAIN RESULTS.

The problem of search for a limited number of distinguishable elements of a large population is very popular in mathematics and related natural sciences. Here we restrict our attention only to *active problems* of this kind, namely, we assume a rational choice (or design) of search experiments. Under certain restriction on the design and analysis of such experiments and the pre-assigned error probability it is desirable to minimize the number of experiments.

We present here the construction of generally applicable search strategies and prove certain optimality of the performance of our search strategies under a linear model. This distinguishes our study from some applied ones (see e.g. recent volume Patel (1987), devoted to search construction in specific models, where even the notions are absent sometimes providing possibility of comparing procedures proposed with the optimal ones).

After the famous textbook of W. Feller on probability, it has been popular to begin exposition of successful search strategies from Dorfman's group (pooled) testing of blood. This approach is being intensively developed at present for applications in

The research was carried out within SFB 373 at Humboldt University, Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft.

medical screening. Less known is the dramatic history of Random Balance Method. Consider a linear model

$$z_i = \sum_{\substack{1 \leq \lambda \leq t \\ \lambda \in A}} b_\lambda x_i(\lambda) + e_i, \quad (1)$$

with random jointly IID carriers $x_i(\lambda) \in \mathcal{B} = \{-1, 1\}$, $x_i(\lambda) = 1$ with probability β , $0 < \beta < 1$, $\lambda = 1, \dots, t$; $i = 1, \dots, N$ and IID with finite support random errors e_i , $i = 1, \dots, N$. Assume that $b_\lambda = 0$ unless $\lambda \in A$ with cardinality $|A| = s \ll t$. For the detection of the set A the "Random Balance Method" (RBM) (F. E. Satterthwaite, T. A. Budne, *Technometrics*, **1**, No 2, (1959)), was proposed and applied successfully to numerous cases of finding disorders of industrial production.

In the discussion contained in the same issue of the journal, RBM was unanimously rejected by the leading applied statisticians of that time. They seemed to overlook the difference of the set-up studied from that of the estimation of parameters of the model requiring a new type of design and analysis. Actually, the visual inspection of scatter diagrams of data N -sequences $(X^N(\lambda), Z^N)$ for each $\lambda = 1, \dots, t$, was proposed by the authors of RBM, although some rank statistics (e.g. Budne test) were also applied.

The main point in their arguments was probably Fisher's idea of randomization: when the value of a variable is fixed, other variables' contribution to the output is a pure noise under the random design. Hence the influence of each variable on the output can be estimated in this noisy background. The relevant Bahadur efficiency of linear and other rank tests was studied for the search of A in Nikitina (1971), Shcherbakova (1992), but they failed to find the asymptotically best statistic from this class.

Use of rank statistics seemed natural because the parameters of noise caused by other variables were unknown suggesting some non-parametricity of tests. It was precisely with the justification of RBM that mathematical study of the planning of screening experiments began in Kolmogorov laboratory of Moscow State University, USSR. Its main ideas were revived by simulation and early theoretical papers attributed to simpler models (Meshalkin (1970), Malyutov, Pinsker (1972), Malyutov and Freidlina (1973), Freidlina (1975), summarized in survey Malyutov (1977)), where RBM was at last been related to earlier combinatorial approach of P. Erdős and A. Renyi to search problems without experimental errors.

The paper of Erdős and Renyi (1963) was one of their series devoted to the theory of search in various simple models. In this particular one they partly solved a problem of identifying special elements (which we call further false) of a set $[t]$ containing t elements supposed to be numerated by numbers from 1 to t by the following tests which we call *weighings*. Any sequence of subsets of $[t]$ can be chosen and numbers of false elements in these subsets become known after weighing. It can be interpreted as a search for false coins (FC) (say, copper) with known identical weight which is e.g. smaller than the weight of a real coin (say, golden). The construction of asymptotically effective static search for FC in a large sample if no restrictions are imposed on the number of FC is described in Lindström (1975). He formulated also new problems, particularly on the best rate of sequential search (in terms of its maximal length) for a limited number of FC in a large sample.

He invented a general principle of concatenated construction of successful search strategies for unique FC in a collection of subsets (which is the second phase of the general class of sequential strategies and was extended afterwards at least in two Ph.D. theses) and conjectured the lower bound for the minimal duration $N_{s,t}$ of static errorless search

$$\lim_{t \rightarrow \infty} \frac{N_{s,t}}{\log \binom{t}{s}} \geq h(s), \quad (2)$$

where $h(s)$ is the entropy of the binomial distribution of a number of successes in s trials with probability $\frac{1}{2}$ of a success.

Recall that if p_1, \dots, p_t are probabilities with sum 1 in a finite sample space, then the entropy of this distribution is $H = -\sum_{i=1}^t p_i \log p_i$. The bound (2) was proved in Dyachkov (1977) using one elegant result of Mateev (1978). Useful instruments for it were already prepared in Erdős and Renyi (1963). The random coding upper bound for the left-hand side of (2) is asymptotically two times worse (Dyachkov (1977)).

We modify some of Lindström's problems in that general linear model (1) (of which the search of false coins is a special case when all parameters b_λ , $\lambda \in A$ are identical) and we admit mean error probability (MEP) $\gamma > 0$ of search under the uniform prior distribution $U(s, t)$ of FC allocations.

Introduce the asymptotic rate $AR = \lim_{t \rightarrow \infty} \frac{\log t}{N(\gamma)}$ of a test T . Here $N(\gamma)$ is the minimal sample size for the probability of T to detect correctly subset A (under the uniform prior distribution of variables' subsets of cardinality s) to be not less than $1 - \gamma$, $0 < \gamma < 1$. It will be shown for the chosen tests that AR does not depend on γ , $0 < \gamma < 1$.

The ML-decision provides (see e.g. Gallager (1968)) the best AR under any fixed design when all the coefficients b_λ , $\lambda \in A$ and the distribution of errors are known. For the case when significant parameters b_λ , $\lambda \in A$ and the law of IID disturbances e_i , $i = 1, \dots, N$ are unknown ML-test is not applicable. Nevertheless, we find two tests T_i , $i = 1$ or s , with the following properties. Test T_s provides the maximal AR of search for the subset A , when $t \rightarrow \infty$ for any fixed $0 < MEP < 1$, it requires about $t^s \log t$ cycles of operations. The test T_1 has the best characteristics of the same kind inside the subset of the tests based on comparing functions of $(X^N(\lambda), Z^N)$, $\lambda = 1, \dots, t$. About $t \log t$ cycles of computing operations are required for these tests.

Remark: For $t \rightarrow \infty$ and MEP of the order e^{-dN} in some range $0 < d < D$ the first class of tests provide the maximal AR as well. This will be proved in our subsequent publications.

Both statistics for the case $t \rightarrow \infty$ are based on calculation of Kullback-Leibler divergence (KULD)

$$\mathcal{K}[\tau(X^N(A), Z^N), \tau(X^N(A)) \times \tau(Z^N)]$$

of the joint empirical distribution $\tau = \tau^N(A)$ of the output and a subset $X(A)$ of input variables of cardinality s or 1 from the corresponding product distributions with the same marginals. This particular case of KULD (called Empirical Shannon

Information (ESI) is

$$\mathcal{I}(\tau^N(A)) = \sum_{x(A) \in \mathcal{B}^{|A|}} \sum_{z \in \mathcal{Z}} \tau(x(A), z) \log(\tau(z|x(A))/\tau(z)),$$

where $\tau(\cdot)$ is the marginal empirical distribution of the output.

Introduce $\Lambda(s, t)$ —set of unordered s —tuples (without replacement) from $[t] = \{1, \dots, t\}$ and $U(s, t)$ —uniform prior distribution over $\Lambda(s, t)$. Test T_s chooses the subset \hat{A} maximizing $\mathcal{I}(\tau^N(A))$ among all subsets A of cardinality s , i.e.

$$T_s = \arg \max_{A \in \Lambda(s, t)} \mathcal{I}(\tau^N(A)).$$

Test T_1 chooses the subset of variables corresponding to s maximal values of $\mathcal{I}(\tau^{(N)}(\lambda))$, $\lambda = 1, \dots, t$ describing influence of each variable $x(\lambda)$ on output separately. We describe this definition by the following notation:

$$T_1 = \arg \max_{1 \leq \lambda \leq t} \mathcal{I}(\tau^N(\lambda)).$$

Remark: We consider only extreme cases $|A| = 1$ or s . The extension of these definitions and properties for the tests T_k , $1 < k < s$ is straightforward. We are going to deal with them in the next publication.

One of intuitive ideas behind the choice of the statistics T_* is that for significant variables (and subsets of these variables) our statistics are strictly positive while for non-significant variables (and sets) these statistics vanish for large samples. The less transparent idea which will become clear later is that the large deviation probabilities for these tests are unexpectedly easy to bound.

All cases of ambiguity of the above decisions are treated as errors of the tests. The similar study of the case of unknown but *a priori* bounded $|A|$ is postponed until the next publication as well as the discussion of yet unsolved problems connected to the choice of the optimal randomization parameter β . It is reduced in Malyutov and Mateev (1980) to that of finding the maximin strategy of the player in the game (6), which is solved only in the extreme cases when either *i*. all significant parameters b_λ are equal or

ii. the cardinality of the set of their distinct linear combinations with coefficients 1 or -1 is 2^s .

In both cases the optimal randomization parameter is, of course, $\frac{1}{2}$.

In general, maximin randomization parameter can be a mixture of not more than 2^s pure ones, if model (1.1) fails to be ordinary in terminology of Malyutov and Mateev (1980).

Meshalkin (1970), Malyutov and Pinsker (1972) (respectively Lindström (1975), Dyachkov (1977) and Mateev (1978)) proved for errorless observations in the optimally randomized case *ii* (respectively, *i*) that the uniqueness of the subset A of cardinality s that provides existence of the solution of the natural system of linear equations

$$\sum_{\lambda \in A} \theta_\lambda x_i(\lambda) = z_i, \quad i = 1, \dots, N$$

($\theta_\lambda \equiv \text{const}$ for the case *i*) takes place with the probability approaching 1 with exponential in N rate iff $AR < 1$ (respectively, $AR < s^{-1}h(s)$, where $h(s)$ is the entropy of the sum of s Bernoulli trials with probability $\frac{1}{2}$ of success, i.e., of Binomial $(s, \frac{1}{2})$). The subset A which provides the solution of the corresponding

system of linear equations may be regarded as an extreme case of ML-test for errorless observations.

It is straightforward to check that our test T_s requires the same AR as the optimal one in both cases. For the test T_1 it suffices $AR \leq 1$ in the case *i* and $AR < \gamma_s$,

$$\gamma_s = \sum_0^{s-1} \text{Bin}_{s-1}(z) \log \left(\frac{2(z+1)(s-z)}{s^2} \right),$$

in the case *ii* for its *MEP* to decay when $t \rightarrow \infty$.

For describing the maximal *AR* of tests in the presence of noise first introduce Shannon information between the output Z and a subset $X(A)$ of IID P_β -distributed input variables

$$I_\beta(X(A) \wedge Z) = E_{P_\beta} \log \frac{P_\beta(Z|X(A))}{P_\beta(Z)},$$

where $P_\beta(X_\alpha = 1) = \beta$, $P_\beta(X_\alpha = 0) = 1 - \beta$, $P_\beta(\cdot)$ is the marginal distribution of Z .

$$C(s) = \max_{\beta \in D} I_\beta(X(A) \wedge Z). \quad (3)$$

These quantities play the main role in describing the maximal rate of search for the separate testing of variables for significance ($|A| = 1$) and for ordinary (Malyutov, Mateev (1980), particularly for symmetric) models ($|A| = s$). The extreme models *i* and *ii* described above are ordinary ones. Moreover *i* remains ordinary even in the presence of arbitrary discrete IID noise. It is not yet established if the general model (1) is always ordinary.

It is worthwhile to emphasize that the distribution $P_\beta(Z|X(F))$ describes the noise created not only by observation errors but also by the randomly varying SV's not included in $x(F)$, $|F| < s$.

For general (not ordinary) models *AR* of ML- and of our test T_s , under $t \rightarrow \infty$ is described in terms of conditional Shannon information (CSI). Under the fixed sequence $[s] = \{1, \dots, s\}$ of SV let $\mathcal{V}(v, s)$ denotes the set of unordered subsets $V \subset [s]$ of the cardinality $|V| = v$, $\mathcal{V}(s) = \cup_{v=0}^{s-1} \mathcal{V}(v, s)$, $V^c = [s] \setminus V$, $x(V)$ be a function $x(i), i \in V$. CSI is

$$I_\beta(V) = I_{P_\beta}(Z \wedge X(V)|X(V^c)), \quad (4)$$

where

$$I_P(X \wedge Y|Z) = E_P \log \frac{P(X|YZ)}{P(X|Z)}, \quad (5)$$

$P(\cdot)$ being the joint density of RV X, Y, Z . Let $\mathcal{C}(s)$ be the value of the game, in which one of players chooses $\beta \in D = [0, 1]$ and the second chooses $V \in \mathcal{V}(s)$ with the pay-off function

$$J_\beta(V) = \frac{I_\beta(V)}{|V|},$$

i.e.,

$$\mathcal{C}(s) = \sup \min J_\mu(V), \quad (6)$$

sup is over the class D^* of probability distribution μ on D , min is over $V \subset \mathcal{V}(s)$. Malyutov (1979) found that sup in (6) can be replaced by the max over distributions on D , concentrated in not more than 2^s points. This follows from the characterization of maximin strategies with convex pay-off function (see e.g. Karlin (1959)). Moreover, it was stated in Malyutov (1979) that $\mathcal{C}(s)$ determines the maximal AR of static search in a very general discrete model including model (1). Now we are ready to formulate our main results. Defining

$$R(\mathbf{X}_t) = \frac{\log t}{N(\mathbf{X}_t)},$$

where $N(\mathbf{X}_t)$ is the number of rows in the design matrix \mathbf{X}_t we have **Theorem 1**

i. For any $\varepsilon > 0$ and discrete mixture μ of IID P_β -randomized designs \mathbf{X}_t with the limit rate

$$\min_{\lambda \in A} I_\mu(X(\lambda) \wedge Z) - \varepsilon,$$

ML-test based on s maximal values of the likelihood of $(X(\lambda), Z)$ and T_1 have MEP exponentially small in N when $t \rightarrow \infty$.

ii. Particularly, the assertion of the first part holds for the maximin randomization mixture $\mu^(1)$.*

Theorem 2. *i. For any $\varepsilon > 0$ and discrete mixture μ of IID P_β -randomized designs with the limit rate $\min_V J_\mu(v) - \varepsilon$, (see (6)) ML-test and the test of maximal ESI have MEP exponentially small in N when $t \rightarrow \infty$.*

ii. Particularly, the assertion of the first part holds if

$$\lim_{t \rightarrow \infty} R(\gamma_t) = \mathcal{C}(s).$$

and the maximin randomized design for the game (6) is chosen. $\mathcal{C}(s)$ is the maximal rate of any design and test admitting arbitrary small positive MEP.

Remark: Proof of the last statement of theorem 1 for ML-test uses some of the main ideas of the capacity region construction for MAC without feedback (see Csiszar and Körner (1981) and the end of this section), although it does not follow from the MAC theory. However, the lower bound for AR is proved almost along the lines of Csiszar and Körner (1981) (it does not follow from famous Fano inequality). Proof and its generalization for the case of unknown s and for unknown $T(\cdot)$ are published in Malyutov, Dyachkov (1980) (and in Malyutov (1983a)) in Russian, supplied by the Dyachkov's estimate from below for the same quantity in the case

$$|\log \gamma(t)| = o(\log t),$$

which is, unfortunately, not sharp. The proof of Malyutov is reproduced in German by Viemeister (1982) supplied by some minor generalizations. In Malyutov, Mateev (1980) the coincidence of (1.3) and (1.6) for ordinary models introduced there (including symmetric ones) is proved.

Solution of maximal empirical Shannon Information was studied in Csiszar and Körner (1981) for conventional Shannon scheme of Information Theory and essentially the same decision was studied in Dyachkov and Rashad (1989) for the search of SV's of symmetric discrete functions of s variables when $t \rightarrow \infty$. In Malyutov and Mateev (1980) the maximal AR of separate detection of SV's by ML-test is found.

In papers (18), (19) and (20) the theory of sequential search for significant variables was developed.

Outline of MAC theory. A discrete multiple access channel (MAC) with s senders, one receiver and (with or without) errorless feedback is described by the transition probability

$$P(z|x(1), x(2), \dots, x(s)) \quad (7)$$

of a symbol z from a finite alphabet \mathcal{Z} at the input of the receiver when senders transmit code symbols $x(1), \dots, x(s)$ respectively. If sender i chooses one of M_i messages, say, the j th one, he codes it into codeword

$$X_j = (x_1^{(j)}(i), \dots, x_N^{(j)}(i))$$

and all senders transmit their codewords synchronously, symbol by symbol. These symbols are converted into sequence z_1, \dots, z_N of conditionally independent RV's given all the codewords of the senders according to (1.3), where $z_n, x_n(j), j = 1, \dots, s$, are substituted instead of $z, x(j), j = 1, \dots, s$.

A decision rule must be chosen, mapping the set of sequences $\{z_1, \dots, z_N\}$ into the set of s -tuples of messages' indices, where N is fixed block size. A transmission, consisting of coding and a decision is called *erroneous* if at least one of the indices of messages is determined erroneously. For a general MAC, information-theoretic framework is equivalent to the following *statistical* one.

Let us numerate the sequences $\lambda_1, \dots, \lambda_s$ corresponding to different distributions and introduce the deterministic function $f(\lambda_1, \dots, \lambda_s)$ which takes values from the set \mathcal{Z} of indices of distinct distributions (i. e. we have glued together sequences $\lambda_1, \dots, \lambda_s$ having the same input distribution $P(\cdot|\lambda_1, \dots, \lambda_s)$ of the receiver).

Hence each MAC-scheme is equivalent to the superposition of a deterministic function $f(\cdot)$ and transformation of its values into distributions $P(\cdot|f(\lambda_1, \dots, \lambda_s))$ such that the function $f(\cdot)$ depends essentially on each of its arguments and all the distributions $P(\cdot|z)$ are distinct for $z \in \mathcal{Z}$.

The important advantage of MAC senders is that they know their messages and can use arbitrary codes whereas in the search the statistician does not know the true indices of the SV's and design. Thus the coding is indirect.

Let N be the codelength. The rate R_i of transmission for the i -th sender is defined as $R_i = \frac{\log M_i}{N}$. An s -tuple R_1, \dots, R_s is called ε -achievable, $0 < \varepsilon < 1$, if transmission with these rates and with mean error probability over \mathcal{P} not exceeding ε is possible. The capacity region is the set of all ε -achievable s -tuples for all $\varepsilon > 0$.

Let us outline for $s = 2$ and binary inputs the construction of the capacity region \mathcal{E} of a general MAC without feedback described e. g. in Csiszar and Körner (1981). Let X_i be independent random variables taking values 0,1 with probabilities $p_i, 1 - p_i, i = 1, 2$, and let the conditional distribution of the random variable Z be $P(\cdot|x_1, x_2)$. Consider the set $\mathcal{E}(p_1, p_2)$ of pairs R_1, R_2 satisfying the inequalities in terms of CSI (2)

$$\begin{aligned} 0 &\leq R_i \leq I(X_i \wedge Z|X_{1-i}), \\ 0 &\leq R_1 + R_2 \leq I(Z \wedge (X_1, X_2)). \end{aligned}$$

Then \mathcal{E} is the convex hull of all $E(p_1, p_2), 0 \leq p_i \leq 1, i = 1, 2$. It is straightforward to prove that AR of our strategy is the intersection of the capacity region described

above for $s = 2$ with the main diagonal \mathcal{D} where the rates of different senders coincide. Indeed, maximization on \mathcal{D} over the arbitrary distributions corresponds exactly to the intersection with \mathcal{D} of the convex hull of the sets corresponding to one-point randomizations and inner minimization over B is equivalent to the intersection of the corresponding sets determined by inequalities for R_i in terms of CSI.

Thus, we may use the well-known upper bound from the MAC-theory for the numbers of the points of support of the optimal distributions of randomization parameters.

NOTATIONS

We use the following abbreviations:

AR	for asymptotic rate,
C	for correct message,
CSI	for conditional Shannon information,
IID	for independent identically distributed RV,
MAC	for multiple access channel,
MEP	for mean error probability,
ML	Maximum Likelihood,
NSV	for nonsignificant variable,
RV	for random variable,
SI	for Shannon information,
SV	for significant variable,

and notations:

$:=$	means equality by definition,
X^N	is a sequence (X_1, \dots, X_N) ,
A^c	is the complement to the set A ,
$ A $	is the number of elements in a finite set A ,
A^*	is the set of Borel measures over a space A ,
ClA	is the closure of A ,
\mathcal{B}	is the set $\{0, 1\}$,
Z	is a RV with values z ,
R	is the rate of a design: $R \cdot N = \log t$,
\mathcal{I}	for Empirical Shannon information,
$[t]$	is the set $\{1, \dots, t\}$,
$\binom{t}{s}$	is a binomial coefficient,
$Bin_s(z)$	$\binom{s}{z} 2^{-s}$ (if we assume the probability of success is $\frac{1}{2}$),
$U(s, t)$	is the uniform distribution over $\binom{t}{s}$ allocations of SV's in $[t]$.

1. PRELIMINARIES

Entropy, Relative Entropy and Mutual Information. The concept of information is too broad to be captured completely by a single definition. However, for

any probability distribution, we define a quantity called *entropy*, which has many properties that agree with the intuitive notion of what a measure of information should be. This notion is extended to define *mutual information*, which is a measure of information one random variable contains about another. Entropy then becomes the self-information of a random variable.

Mutual information is a special case of a more general quantity called **relative entropy** or Kullback-Leiber divergence, which is a measure of the divergence of one probability distribution from another. In particular, relative entropy of a distribution with respect to the uniform distribution is entropy.

In the following section we give survey of the definitions and results on information measures. The proofs are contained in Cover and Thomas (1991).

Entropy.

Definition 1.1. The *entropy* $H(X)$ of a discrete random variable X with $p(X = j) = p_j$, $j = 1, 2, \dots, t$, is defined by

$$H(X) = - \sum_x p(x) \log p(x), \quad (8)$$

where we put $0 \log 0 = 0$.

Lemma 1.2. $0 \leq H(X) \leq \log t$.

Joint Entropy and Conditional Entropy.

Definition 1.3. The joint entropy $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = \sum_x \sum_y p(x, y) \log p(x, y), \quad (9)$$

Definition 1.4. If $(X, Y) \sim p(x, y)$, then the conditional entropy $H(Y|X)$ is defined as

$$H(Y|X) = \sum_x p(x) H(Y|X = x) \quad (10)$$

Theorem 1.5. (*Chain rule*):

$$H(X, Y) = H(X) + H(Y|X). \quad (11)$$

Definition 1.6. The *relative entropy or Kullback Leibler* divergence (KULD) of probability mass functions $p(x)$ from $q(x)$ is defined as

$$\mathcal{K}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}, \quad (12)$$

and ∞ , if $p(\cdot)$ is not absolutely continuous with respect to $q(\cdot)$.

Definition 1.7. Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass function $p(x)$ and $p(y)$. The *mutual information* $I(X \wedge Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$, i.e.,

$$I(X \wedge Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \quad (13)$$

we can write the definition of mutual information $I(X \wedge Y)$ as

$$I(X \wedge Y) = H(X) - H(X|Y). \quad (14)$$

Theorem 1.8. (*Chain rule for entropy*): Let X_1, X_2, \dots, X_N be drawn according to $p(x_1, x_2, \dots, x_N)$. Then

$$H(X_1, X_2, \dots, X_N) = \sum_{i=1}^N H(X_i | X_{i-1}, \dots, X_1). \quad (15)$$

Definition 1.9. The conditional mutual information of random variables X and Y given Z is defined by

$$I(X \wedge Y | Z) = H(X|Z) - H(X|Y, Z) = E_{p(x,y,z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)}.$$

Theorem 1.10. (*Chain rule for information*):

$$I(X_1, X_2, \dots, X_N \wedge Y) = \sum_{i=1}^N I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1). \quad (16)$$

Definition 1.11. The conditional relative entropy $\mathcal{K}(p(y|x)||q(y|x))$ is the mean of the relative entropies of the conditional probability mass function $p(y|x)$ with respect to $q(y|x)$ averaged over the probability mass function $p(x)$.

More precisely,

$$\mathcal{K}(p(y|x)||q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} = E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)} \quad (17)$$

Theorem 1.12. (*Chain rule for relative entropy*):

$$\mathcal{K}(p(x, y)||q(x, y)) = \mathcal{K}(p(x)||q(x)) + \mathcal{K}(p(y|x)||q(y|x)).$$

Corollary 1.13. If $y = f(x)$, where $f: X \rightarrow Y$ is injective, then

$$\mathcal{K}(p(y)||q(y)) \leq \mathcal{K}(p(x)||q(x)) \quad (18)$$

Theorem 1.14. (*Information inequality*): Let $p(x), q(x), x \in \mathcal{X}$, be two probability mass functions. Then

$$\mathcal{K}(p||q) \geq 0 \quad (19)$$

with equality if and only if

$$p(x) = q(x) \quad \text{for all } x. \quad (20)$$

Corollary 1.15. *For any two random variables, X, Y ,*

$$I(X \wedge Y) \geq 0, \quad (21)$$

with equality if and only if X and Y are independent.

Corollary 1.16.

$$\mathcal{K}(p(y|x)||q(y|x)) \geq 0, \quad (22)$$

with equality if and only if $p(y|x) = q(y|x)$ for all y and x with $p(x) > 0$.

Corollary 1.17.

$$I(X \wedge Y|Z) \geq 0, \quad (23)$$

with equality if and only if X and Y are conditionally independent given Z .

Theorem 1.18. *$H(X) \leq \log |\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of elements in the range of X , with equality if and only if X has the uniform distribution over \mathcal{X} .*

Definition 1.19. A function $f(x)$ is said to be convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \zeta \leq 1$,

$$f(\zeta x_1 + (1 - \zeta)x_2) \leq \zeta f(x_1) + (1 - \zeta)f(x_2). \quad (24)$$

A function f is said to be strictly convex if equality holds only if $\zeta = 0$ or 1 .

Theorem 1.20. *$\mathcal{K}(p||q)$ is convex in the pair (p, q) , i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then*

$$\begin{aligned} & \mathcal{K}(\zeta p_1 + (1 - \zeta)p_2 || \zeta q_1 + (1 - \zeta)q_2) \leq \\ & \leq \zeta \mathcal{K}(p_1 || q_1) + (1 - \zeta)\mathcal{K}(p_2 || q_2), \end{aligned} \quad (25)$$

for all $0 \leq \zeta \leq 1$.

THE METHOD OF TYPES

Let X_1, X_2, \dots, X_N be a sequence of N symbols from an alphabet $\mathcal{X} = \{\lambda_1, \lambda_2, \dots, \lambda_{|\mathcal{X}|}\}$.

Definition 1.21. The components of the vector type $\tau(\underline{X}^N)$ (or empirical probability distribution) of a sequence \underline{X}^N are the relative proportions of occurrences of each symbol in \underline{X}^N

$$\tau_{X^N}(\lambda) = \tau(\underline{X}^N) = \left(\frac{n_1(\lambda_1|\underline{X}^N)}{N}, \dots, \frac{n_{|\mathcal{X}|}(\lambda_{|\mathcal{X}|}|\underline{X}^N)}{N} \right)$$

where $n_*(\lambda|\underline{X}^N)$ is the number of times the symbol λ occurs in the sequence $\underline{X}^N \in \mathcal{X}^N$.

Definition 1.22. Let \mathcal{T}^N denote the set of types with denominator N .

Definition 1.23. If $\underline{t} \in \mathcal{T}^N$, then

$$T(\underline{t}) := \{\underline{X}^N \in \mathcal{X}^N : \underline{\tau}(\underline{X}^N) = \underline{t}\}. \quad (26)$$

The type class is sometimes called the composition class of \underline{t} .

Theorem 1.24.

$$|\mathcal{T}^N| \leq (N+1)^{|\mathcal{X}|}. \quad (27)$$

Theorem 1.25. If X_1, X_2, \dots, X_N are drawn i.i.d. with distribution $Q(\underline{X})$, then the probability of \underline{X}^N depends only on its type and is given by

$$Q^N(\underline{X}) = Q(\underline{X}^N) = \exp\{-N(H(\underline{\tau}) + \mathcal{K}(\underline{\tau}||Q))\}. \quad (28)$$

Corollary 1.26. If \underline{X}^N is in the type class of Q , then

$$Q^N(\underline{X}^N) = \exp\{-NH(Q)\}. \quad (29)$$

Theorem 1.27. (Size of a type class $T(\underline{t})$):

For any type $\underline{t} \in \mathcal{T}^N$,

$$\frac{1}{(N+1)^{|\mathcal{X}|}} \exp\{NH(\tau)\} \leq |T(\underline{t})| \leq \exp\{NH(\tau)\}. \quad (30)$$

Theorem 1.28. (Probability of type class): For any $\underline{t} \in \mathcal{T}^N$ and any probability distribution Q , the probability of the type class $T(\underline{t})$ under Q^N is approximately $\exp\{-N\mathcal{K}(\underline{t}||Q)\}$ in the sense that

$$\frac{1}{(N+1)^{|\mathcal{X}|}} \exp\{-N\mathcal{K}(\underline{t}||Q)\} \leq Q^N(T(\underline{t})) \leq \exp\{-N\mathcal{K}(\underline{t}||Q)\}. \quad (31)$$

Theorem 1.29. (Sanov's theorem): i. Let X_1, X_2, \dots, X_N be i.i.d. with distribution $Q(x)$. Let E be a set of distributions closed in the euclidean topology of \mathcal{X}^* . Then

$$Q^N(E) = Q^N(\tau_{X^N} \subset E) \leq (N+1)^{|\mathcal{X}|} \exp\{-N\mathcal{K}(t^*||Q)\}, \quad (32)$$

where

$$t^* = \arg \min_{t \in E} \mathcal{K}(t||Q) \quad (33)$$

is the distribution in E that is closest to Q in relative entropy.

ii. If in addition $E = (\text{Int}E)^c$ then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log Q^N(E) = -\mathcal{K}(t^*||Q). \quad (34)$$

Remark 1.30. We shall use only part i of Sanov's theorem in the proof of theorem 1 of our dissertation.

CONDITIONAL TYPES

In our main theorem 2 we use the following generalization of the above method of types (see Csiszar and Körner (1981)).

If \mathcal{X} and \mathcal{Y} are two finite sets, the joint type of a pair of sequences $\underline{X}^N \in \mathcal{X}^N$ and $\underline{Y}^N \in \mathcal{Y}^N$ is defined as the type of the sequence $\{(x_i, y_i)\}_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$. In other words, it is distribution $\tau_{\underline{X}^N, \underline{Y}^N}$ on $\mathcal{X} \times \mathcal{Y}$ denoted by

$$\tau_{\underline{X}^N, \underline{Y}^N}(\lambda, \alpha) = n(\lambda, \alpha | \underline{X}^N, \underline{Y}^N) \quad \text{for every } \lambda \in \mathcal{X}, \alpha \in \mathcal{Y}. \quad (35)$$

Joint types will often be given in terms of the type of \underline{X}^N and a stochastic matrix $M : \mathcal{X} \rightarrow \mathcal{Y}$ such that

$$\tau_{\underline{X}^N, \underline{Y}^N}(\lambda, \alpha) = \tau_{\underline{X}^N}(\lambda)M(\alpha|\lambda) \quad \text{for every } \lambda \in \mathcal{X}, \alpha \in \mathcal{Y}. \quad (36)$$

Notice that the joint type $\tau_{\underline{X}^N, \underline{Y}^N}$ uniquely determines $M(\alpha, \lambda)$ for those $\lambda \in \mathcal{X}$ which do occur in the sequence \underline{X}^N . For conditional probabilities of sequences $\underline{Y}^N \in \mathcal{Y}^N$ given a sequence $\underline{X}^N \in \mathcal{X}^N$, the matrix M in the equation above will play the same rule as the type of \underline{Y}^N does for unconditional probabilities.

Definition 1.31. We say that $\underline{Y}^N \in \mathcal{Y}^N$ has conditional type M given $\underline{X}^N \in \mathcal{X}^N$ if

$$n(\lambda, \alpha | \underline{X}^N, \underline{Y}^N) = n(\lambda | \underline{X}^N)M(\alpha|\lambda) \quad \text{for every } \lambda \in \mathcal{X}, \alpha \in \mathcal{Y}. \quad (37)$$

For any given $\underline{X}^N \in \mathcal{X}^N$ and stochastic matrix $M : \mathcal{X} \rightarrow \mathcal{Y}$, the set of sequences $\underline{Y}^N \in \mathcal{Y}^N$ having conditional type M given \underline{X}^N will be called the M -shell of \underline{X}^N , $\mathcal{T}_M(\underline{X}^N)$.

Let \mathcal{M}^N denote the family of such matrices M . Then

$$|\mathcal{M}^N| \leq (N+1)^{|\mathcal{X}||\mathcal{Y}|}. \quad (38)$$

Remark 1.32. The conditional type of \underline{Y}^N given \underline{X}^N is not uniquely determined if some $\lambda \in \mathcal{X}$ do not occur in \underline{X}^N . Nevertheless, the set $\mathcal{T}_M(\underline{X}^N)$ containing \underline{Y}^N is unique.

Notice that the conditional type is a generalization of types. In fact, if all the components of the sequence \underline{X}^N equal x , conditional type coincides with the set of sequences of type $M(\cdot|x)$ in \mathcal{Y}^N .

In order to formulate the basic size and probability estimates for M -shells, it will be convenient to introduce some notations.

The average of the entropies of the rows of a stochastic matrix $M : \mathcal{X} \rightarrow \mathcal{Y}$ with respect to a distribution τ on \mathcal{X} will be denoted by

$$H(M|\tau) := \sum_{x \in \mathcal{X}} \tau(x)H(M(\cdot|x)). \quad (39)$$

The analogous average of the Kullback-Leibler divergences of the corresponding rows of stochastic matrix $M_1 : \mathcal{X} \rightarrow \mathcal{Y}$ from $M_2 : \mathcal{X} \rightarrow \mathcal{Y}$ will be denoted by

$$\mathcal{K}(M_1||M_2|\tau) := \sum_{x \in \mathcal{X}} \tau(x)\mathcal{K}(M_1(\cdot|x)||M_2(\cdot|x)). \quad (40)$$

Notice that $H(M_1|\tau)$ is the conditional entropy $H(X|Y)$ of RV's X and Y such that X has distribution τ and Y has conditional distribution M on X . The quantity $\mathcal{K}(M_1||M_2|\tau)$ is called conditional Kullback-Leibler divergence (conditional relative entropy).

The counterpart of (theorem 1.27) for M -shell is

Theorem 1.33. *For every $\underline{X}^N \in \mathcal{X}^N$ and stochastic matrix $M : \mathcal{X} \rightarrow \mathcal{Y}$ such that $\mathcal{T}_M(\underline{X}^N)$ is non-void, we have*

$$\begin{aligned} \frac{1}{(N+1)^{|\mathcal{X}||\mathcal{Y}|}} \exp\{NH(M|\tau_{\underline{X}^N})\} &\leq \\ &\leq |\mathcal{T}_M(\underline{X}^N)| \leq \exp\{NH(M|\tau_{\underline{X}^N})\}. \end{aligned} \quad (41)$$

Proof

This is an easy consequence of theorem 1.24. In fact, $|\mathcal{T}_M(\underline{X}^N)|$ depends on \underline{X}^N only through $\tau_{\underline{X}^N}$. Hence we may assume that \underline{X}^N is juxtaposition of sequences \underline{X}_λ^N , $\lambda \in \mathcal{X}$ where \underline{X}_λ^N consists of $n(\lambda|\underline{X}^N)$ identical elements of λ . In this case $\mathcal{T}_M(\underline{X}^N)$ is the cartesian product of the sets of sequences of type $M(\cdot|\lambda)$ in $\mathcal{Y}^{n(\lambda|\underline{X}^N)}$, with λ running over those elements of \mathcal{X} which occur in \underline{X}^N . Thus, theorem 1.28 gives

$$\begin{aligned} \prod_{\lambda \in \mathcal{X}} \frac{1}{(n(\lambda|\underline{X}^N) + 1)^{|\mathcal{Y}|}} \exp\{n(\lambda|\underline{X}^N)H(M(\cdot|\lambda))\} &\leq \\ &\leq |\mathcal{T}_M(\underline{X}^N)| \leq \prod_{\lambda \in \mathcal{X}} \exp\{n(\lambda|\underline{X}^N)H(M(\cdot|\lambda))\}, \\ \frac{1}{(N\tau_{\underline{X}^N}(\lambda) + 1)^{|\mathcal{X}||\mathcal{Y}|}} \exp\{N \sum_{\lambda \in \mathcal{X}} \tau_{\underline{X}^N}(\lambda)H(M(\cdot|\lambda))\} &\leq \\ &\leq |\mathcal{T}_M(\underline{X}^N)| \leq \exp\{N \sum_{\lambda \in \mathcal{X}} \tau_{\underline{X}^N}(\lambda)H(M(\cdot|\lambda))\}, \end{aligned}$$

then

$$\begin{aligned} \frac{1}{(N+1)^{|\mathcal{X}||\mathcal{Y}|}} \exp\{NH(M|\tau_{\underline{X}^N})\} &\leq \\ &\leq |\mathcal{T}_M(\underline{X}^N)| \leq \exp\{NH(M|\tau_{\underline{X}^N})\}. \end{aligned}$$

Theorem 1.34. *For every $\underline{X}^N \in \mathcal{X}^N$ and stochastic matrix $M_1 : \mathcal{X} \rightarrow \mathcal{Y}$, $M_2 : \mathcal{X} \rightarrow \mathcal{Y}$ such that $|\mathcal{T}_{M_1}(\underline{X}^N)|$ is non-void,*

$$M_2(\underline{Y}^N|\underline{X}^N) = \exp\{-N(\mathcal{K}(M_1||M_2|\tau_{\underline{X}^N}) + H(M_1|\tau_{\underline{X}^N}))\} \quad (42)$$

if $\underline{Y}^N \in |\mathcal{T}_{M_1}(\underline{X}^N)|$

$$\begin{aligned} \frac{1}{(N+1)^{|\mathcal{X}||\mathcal{Y}|}} \exp\{-N\mathcal{K}(M_1||M_2|\tau_{\underline{X}^N})\} &\leq \\ &\leq M_2(\mathcal{T}_{M_1}(\underline{X}^N)|\underline{X}^N) \leq \exp\{-N\mathcal{K}(M_1||M_2|\tau_{\underline{X}^N})\}. \end{aligned} \quad (43)$$

Proof

$$\begin{aligned}
M_2(\underline{Y}^N | \underline{X}^N) &= \prod_{\lambda \in \mathcal{X}} \prod_{\alpha \in \mathcal{Y}} M_2(\alpha | \lambda)^{n(\lambda, \alpha | \underline{X}^N, \underline{Y}^N)} \\
&= \prod_{\lambda \in \mathcal{X}} \prod_{\alpha \in \mathcal{Y}} M_2(\alpha | \lambda)^{N \tau_{\underline{X}^N}(\lambda) M_1(\alpha | \lambda)} \\
&= \prod_{\lambda \in \mathcal{X}} \prod_{\alpha \in \mathcal{Y}} \exp\{N \tau_{\underline{X}^N}(\lambda) M_1(\alpha | \lambda) \ln M_2(\alpha | \lambda)\} \\
&= \exp\{N \sum_{\lambda \in \mathcal{X}} \tau_{\underline{X}^N}(\lambda) \{- \sum_{\alpha \in \mathcal{Y}} M_1(\alpha | \lambda) \ln \frac{M_1(\alpha | \lambda)}{M_2(\alpha | \lambda)} + \\
&\quad + \sum_{\alpha \in \mathcal{Y}} M_1(\alpha | \lambda) \ln M_1(\alpha | \lambda)\}\} \\
&= \exp\{-N \sum_{\lambda \in \mathcal{X}} \tau_{\underline{X}^N}(\lambda) \{\mathcal{K}(M_1(\alpha | \lambda) || M_2(\alpha | \lambda)) + \\
&\quad + H(M_1(\alpha | \lambda))\}\} \\
&= \exp\{-N(\mathcal{K}(M_1 || M_2 | \tau) + H(M_1 | \tau))\}.
\end{aligned}$$

For the second part, we have

$$\begin{aligned}
M_2(\mathcal{T}_{M_1}(\underline{X}^N) | \underline{X}^N) &= \sum_{\underline{Y}^N \in \mathcal{T}_{M_1}} M_2(\underline{Y}^N | \underline{X}^N) = \sum_{\underline{Y}^N \in \mathcal{T}_{M_1}} \prod_{i=1}^N \prod_{j=1}^N M_2(y_i | x_j) \\
&= \sum_{\underline{Y}^N \in \mathcal{T}_{M_1}} \prod_{\lambda \in \mathcal{X}} \prod_{\alpha \in \mathcal{Y}} M_2(\alpha | \lambda)^{n(\lambda, \alpha | \underline{X}^N, \underline{Y}^N)} \\
&= \sum_{\underline{Y}^N \in \mathcal{T}_{M_1}} \prod_{\lambda \in \mathcal{X}} \prod_{\alpha \in \mathcal{Y}} M_2(\alpha | \lambda)^{N \tau_{\underline{X}^N}(\lambda) M_1(\alpha | \lambda)} \\
&= \sum_{\underline{Y}^N \in \mathcal{T}_{M_1}} \exp\{-N(\mathcal{K}(M_1 || M_2 | \tau_{\underline{X}^N}) + \\
&\quad + H(M_1 | \tau_{\underline{X}^N}))\} \\
&= |\mathcal{T}_{M_1}(\underline{Y}^N)| \exp\{-N(\mathcal{K}(M_1 || M_2 | \tau_{\underline{X}^N}) + \\
&\quad + H(M_1 | \tau_{\underline{X}^N}))\},
\end{aligned}$$

using theorem 1.3.29, we get

$$\begin{aligned}
\frac{1}{(N+1)^{|\mathcal{X}||\mathcal{Y}|}} \exp\{-N\mathcal{K}(M_1 || M_2 | \tau_{\underline{X}^N})\} &\leq \\
&\leq M_2(\mathcal{T}_{M_1}(\underline{X}^N) | \underline{X}^N) \leq \exp\{-N\mathcal{K}(M_1 || M_2 | \tau_{\underline{X}^N})\}.
\end{aligned}$$

Theorem 1.35. (*Conditional Sanov's Theorem*): Let $\underline{X}^N \in \mathcal{X}^N$ be i.i.d. with distribution matrix $M_2 : \mathcal{X} \rightarrow \mathcal{Y}$. Let E be a set of distributions closed in the euclidean topology of \mathcal{Y}^* . Then

$$M_2(E | \underline{X}^N) \leq (N+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\{-N\mathcal{K}(M_1^* || M_2 | \tau_{\underline{X}^N})\} \quad (44)$$

where

$$M_1^* = \arg \min_{M_1 \in E} \mathcal{K}(M_1 || M_2 | \tau_{\underline{X}^N}). \quad (45)$$

is the distribution in E that is closest to the matrix distribution M_2 in terms of conditional relative entropy.

Proof

$$\begin{aligned}
M_2(E|\underline{X}^N) &= \sum_{M_1 \in E \in \mathcal{M}^N} M_2^N(\mathcal{T}_{M_1}(\underline{X}^N)|\underline{X}^N) \\
&\leq \sum_{M_1 \in E \cap \mathcal{M}^N} \exp\{-N\mathcal{K}(M_1||M_2|\tau_{\underline{X}^N})\} \\
&\leq \sum_{M_1 \in E \cap \mathcal{M}^N} \exp\{-N \min_{M_1 \in E \cap \mathcal{M}^N} \mathcal{K}(M_1||M_2|\tau_{\underline{X}^N})\} \\
&= \sum_{M_1 \in E \cap \mathcal{M}^N} \exp\{-N\mathcal{K}(M_1^*||M_2|\tau_{\underline{X}^N})\} \\
&\leq (N+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\{-N\mathcal{K}(M_1^*||M_2|\tau_{\underline{X}^N})\}.
\end{aligned}$$

Remark 1.36. We prove only the first part in the conditional Sanov's theorem since we shall use only it in our dissertation.

For simplicity we assume all the input variables are binary. The generalization to an arbitrary finite input alphabet is straightforward. For a sequence $\underline{x} = (x(1), \dots, x(t)) \in \mathcal{B}^t$, $\mathcal{B} = \{-1, 1\}$ of binary variables and a subset $A \subset [t] = \{1, \dots, s\}$ of cardinality $|A| = s \ll t$ denote by $\underline{x}(A)$ the subset of \underline{x} components with indices from A . Let the function $\eta(A) : \mathcal{B}^s \rightarrow Z$, $\eta(A) = \sum_{\lambda \in A} b_\lambda x(\lambda)$, $x(\lambda) \in \mathcal{B}$. the planning is static, i.e., the points of experiments $x_i \in X$, $i = 1, \dots, N$, are preassigned. The planning of the experiment will be denoted by the $(N \times t)$ -matrix $\|x_i(\lambda)\| = \underline{x}$, $i = 1, \dots, N$ and $\lambda = 1, \dots, t$. The rows of \underline{x} are x_i , and $x(\lambda)$ is the λ -th column.

In place of the value of the function η , the random variable z with a finite range Z is observed. The matrix of the conditional probabilities of distortions of η -the "noise" - is known:

$$\underline{P} = \|p(z|x)\|, \quad x \in X, \quad z \in Z.$$

Without loss of generality, we can assume that the distributions $P(.|x_1)$ and $P(.|x_2)$ are different for $x_1 \neq x_2$.

$$P(z|x) = \prod_1^N p(z_i|x_i).$$

The uniform distribution $U = U(s, t)$ is given on $\Lambda(s, t)$ -the set of all ordered s -sets λ . The decision about the true λ is obtained with the help of the decision function

$$d(\underline{x}, z) : X^N \times Z^N \rightarrow \Lambda(s, t).$$

We shall assume that an error occurs if the set of SV's is restored incorrectly. This error depends on the design, an unknown set of SV and a decision function. A static (non-sequential) strategy $\mathbf{S} = S_t$, consisting of a static design and a decision, is characterized by its sample size N and by its mean error probability MEP not less than $\gamma > 0$.

Denote by $\arg \max_{z \in Z} f(z)$ the set of s maximal values of function $f(z)$ over a finite set Z .

$$T_1 = \arg \max_{\lambda} \mathcal{I}(\tau^N(\lambda)).$$

The case of ambiguity of this definition is always regarded as the error of T_1 , hence any of definitions for the case of ambiguity can be applied.

$$T_s = \arg \max_{F \in \Lambda(s, t)} \mathcal{I}(\tau^N(F)).$$

(Here, and in sequel, for a closed (particularly finite) set Y , $\arg \max_Y \phi(y) = \{y_0 : \max_{y \in Y} \phi(y) = \phi(y_0)\}$); Moreover, if the maximum is attained for several λ , then one of them is chosen by trial in an equiprobable manner.

2. NOISELESS SEARCH

To expose the idea of the method, we consider first the case $e_i \equiv 0$. Assume first that the cardinality of the set $\{\sum_{\lambda \in A_s} b_{\lambda} x_i(\lambda) | x_i(\lambda) = \pm 1\}$ is 2^s . This is the maximal cardinality of the range of the function $\eta(b_1, \dots, b_s)$. (For the case $b_{\lambda} \equiv 1$, the cardinality of the set $\{\sum_{\lambda \in A_s} b_{\lambda} x_i(\lambda) | x_i(\lambda) = \pm 1\}$ is $s + 1$. This is the minimal cardinality of the range of the function $\eta(b_1, \dots, b_s)$). We consider the method of one by one detection of significant variables by test T_1 . We want to examine all pairs $(x^N(\lambda), z^N)$ where $x^N(\lambda)$ is the binary input column and z^N is the output column with components taking 2^s values.

Let us assume for definiteness that variables $x(1), \dots, x(s)$ are SV.

First case: Let $X(1)$ be significant variable (SV).

Second: $X(\lambda)$ is a nonsignificant (NSV.), $\lambda > s$.

The random balance method, (RBM) consists of inspections of scatter diagrams of data N -sequences $(x^N(\lambda), z^N)$.

In the left (right) side of the scatter diagrams corresponding to $x(\lambda) = \pm 1$ we have non-overlapping sets of outputs $y - b_{\lambda}$ ($y + b_{\lambda}$), respectively, with coefficients b_{λ} , $\lambda = 1, \dots, s$, y where $y \in A_{\lambda}$, A_{λ} is the set of linear combinations of significant variables different from $X(\lambda)$. The cardinality $|A_{\lambda}|$ is 2^{s-1} . Hence for each significant variable $X(\lambda)$ we have a separate partition of the outputs Z into two subsets $\{\pm b_{\lambda} + A_{\lambda}\}$ described visually by scatter diagrams. Let us first calculate the theoretical Shannon information, SI,

$$I(\pi(x(\lambda), z)) = \sum_{x(\lambda)} \sum_z \pi(x(\lambda), z) \log \frac{\pi(x(\lambda), z)}{\pi(x(\lambda), \cdot) \pi(\cdot, z)}$$

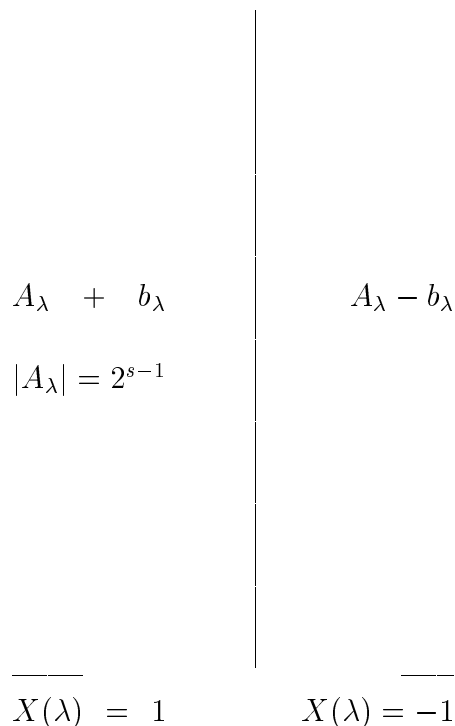


FIGURE 1. Outline of limiting scatter diagrams in the case of significant variables in the absence of noise

where $\pi(x(\lambda), \cdot) = \sum_z \pi(x(\lambda), z)$, $\pi(\cdot, z) = \sum_{x(\lambda)} \pi(x(\lambda), z)$, $\pi(x(\lambda), z)$ is the joint distribution over $x(\lambda)$, $\pi(x(\lambda), z) = 2^{-s}$ and $\pi(z)$ is the marginal distribution over Z , $\pi(z) = 2^{-s}$, $z \in Z$. We get

$$I(\pi(x(\lambda), z)) = 1/2 \sum_{y \in A_\lambda} 2^{-s+1} \log \frac{2^{-s+1}}{2^{-s}} = \log 2, \quad \lambda = 1, \dots, s.$$

Next we study the behaviour of the empirical Shannon information ESI . Case of NSV; if $x(\lambda)$ NSV, the distributions of $X^N(\lambda)$ and Z^N are independent. We have two subsets, each of cardinality 2^s , and the mutual Shannon information $I(\pi(x(\lambda), z)) = 0$ since the joint distribution $\pi(x(\lambda), z)$ over $x(\lambda)$ and z is uniform with weights 2^{-s} coinciding with the product of the marginal distributions $\pi(\cdot, z) \times \pi(x(\lambda), \cdot)$ i.e.,

$$\log \frac{\pi(x(\lambda), z)}{\pi(x(\lambda), \cdot)\pi(\cdot, z)} = \log 1 = 0.$$

Since we defined the acceptance region \mathcal{AR} of significant variables by the set of all $(x^N(\lambda), z^N)$ such that $I(\tau(x^N(\lambda), z^N)) \geq 1 - \varepsilon$, using Sanov's theorem for each NSV

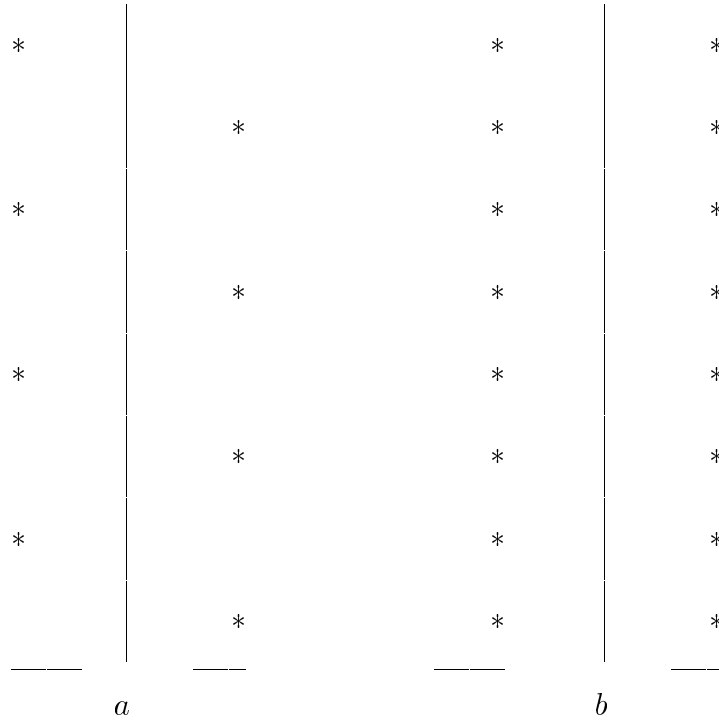


FIGURE 2. *Limiting scatter diagrams in case of three significant variables and $|\mathcal{Z}| = 2^3 = 8$. For significant (a) and non-significant (b) variables in the absence of noise*

(see theorem 1 below)

$$\begin{aligned}
 P(\mathcal{I}(\tau(x^N(\lambda), z^N)) \geq 1 - \varepsilon) &\leq (N + 1)^{2^{s+1}} \sum_{\tau(x^N(\lambda), z^N) \in \mathcal{AR}} 2^{\{-NI(\tau(x^N(\lambda), z^N))\}} \\
 &\leq (N + 1)^{2^{s+1}} \max 2^{\{-NI(\tau(x^N(\lambda), z^N))\}} \\
 &\approx 2^{\{-N \min I(\tau(x^N(\lambda), z^N))\}} \\
 &\approx 2^{\{-N(1-\varepsilon)(1+o(1))\}}.
 \end{aligned}$$

Our bound for AR follows easily from the above estimate by using Bonferroni bound.

Algorithm. We give a brief description of our algorithm which compute both empirical and theoretical Shannon informations.

1. Create a random $(N \times t)$ -matrix of 1's and -1 's.
2. Fix s , pick value's for $b_s(i)$ and e_i and compute z_i , $i = 1, \dots, N$.
3. Count the number of 1's and -1 's for each $X^N(k)$, $k = 1, \dots, t$, separately, and compute $\tau(X^N(k))$, count the number of similar outputs and compute its empirical distribution. Also count the number of various pairs $(X^N(k), Z^N)$ and compute $\tau(X^N(k), Z^N)$, $i = 1, \dots, N$.
4. Compute the ESI, $\mathcal{I}(\tau(X^N(k), Z^N))$, $k = 1, \dots, t$, then sort $\mathcal{I}(\tau(X^N(k), Z^N))$, $k \in [s]$ in ascending order and sort $\mathcal{I}(\tau(X^N(j), Z^N))$, $j > s$, in descending order. If $\mathcal{I}(\tau(X^N(1), Z^N)) > \mathcal{I}(\tau(X^N(s+1), Z^N))$, there is no error, else there is an error.

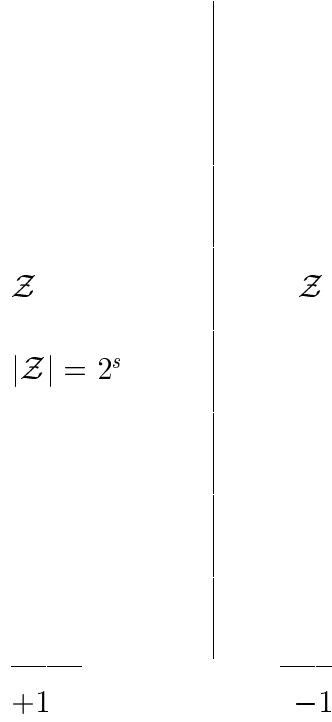


FIGURE 3. *Outline of limiting scatter diagrams in the case of non-significant variables in the absence of noise*

The following tables show the outputs of the above algorithm in the case when $e_i \equiv 0$. Each table contains number of runs, (e.g. 10 runs) which includes $\min_{k \in [s]} \mathcal{I}(\tau(X^N(k), Z^N))$, $\max_{j > s} \mathcal{I}(\tau(X^N(j), Z^N))$ and *results*.

Notice that for table 1, we choose $b_\lambda = (1, 2, 3)$, and for table 2 we choose $b_\lambda = (1, 1, 4)$. In each row in the above tables, we get

$$\min_{k \in [s]} \mathcal{I}(\tau(X^N(k), Z^N)), \geq \max_{j > s} \mathcal{I}(\tau(X^N(j), Z^N)),$$

and the conclusion is, we have no error.

Proof of Theorem 1

Let us introduce the following lemma.

Lemma 1:

The mean of MEP over the ensemble of designs equals the same mean of conditional error probabilities $\mathcal{P}_{[s]}$, when $[s]$ is the true set of SV's.

The proof is straightforward because of the symmetry of the ensemble of the design. As we have already mentioned, it was proved in Malyutov and Mateev (1980) that *AR* of separate search for SV's in model (1) using the most efficient ML-test when

$\min_{j \in [s]} \mathcal{I}(\tau(X^N, Z^N))$	$\max_{k > s} \mathcal{I}(\tau(X^N, Z^N))$	<i>results</i>
0.216595	0.042982	C
0.201501	0.046515	C
0.244194	0.059394	C
0.221705	0.036725	C
0.217343	0.048427	C
0.231038	0.037558	C
0.229712	0.039366	C
0.218441	0.041273	C
0.221563	0.048136	C
0.217657	0.046794	C

TABLE 1. *Algorithm outputs*, $N = 10$, $t = 200$, $s = 3$, $b_\lambda = (1, 2, 3)$, and $e_i \equiv 0$.

$\min_{j \in [s]} \mathcal{I}(\tau(X^N, Z^N))$	$\max_{k > s} \mathcal{I}(\tau(X^N, Z^N))$	<i>results</i>
0.130677	0.031557	C
0.148835	0.039004	C
0.142865	0.036245	C
0.151071	0.043865	C
0.155800	0.046079	C
0.133071	0.033853	C
0.138825	0.034960	C
0.154540	0.034736	C
0.154765	0.041052	C
0.161390	0.057084	C

TABLE 2. *Algorithm outputs*, $N = 10$, $t = 200$, $s = 3$, $b_\lambda = (1, 1, 4)$, and $e_i \equiv 0$.

$\min_{j \in [s]} \mathcal{I}(\tau(X^N, Z^N))$	$\max_{k > s} \mathcal{I}(\tau(X^N, Z^N))$	<i>results</i>
0.046953	0.061174	E
0.047368	0.039004	C
0.046411	0.060245	E
0.048365	0.061071	E
0.050800	0.046079	C
0.043061	0.043853	E
0.058825	0.054990	C
0.054550	0.070736	E
0.044765	0.040852	C
0.051390	0.047084	C

TABLE 3. *Algorithm outputs*, $N = 23$, $t = 180$, $s = 3$, $b_\lambda = (1, 2, 3)$, and e_i random between 0 and 3

it is applicable is

$$AR = \min_{1 \leq j \leq s} I_j = I_1,$$

where $I_j = I(X(j) \wedge Z)$, $j \in [s]$ and we assumed for definitness WLOG (without loss of generality) that $X(j)$ are SV's, $j = 1, \dots, s$, and $I_1 \leq I_2 \leq \dots \leq I_s$. The error of the test T_1 occurs when

$$\max_{k > s} \mathcal{I}(\tau(X^N(k), Z^N)) \geq \min_{j \in [s]} \mathcal{I}(\tau(X^N(j), Z^N)) \quad (46)$$

The event in (46) is included in the event

$$\bigcap_{T \geq 0} \left[\bigcup_{k > s} \{ \mathcal{I}(\tau(X^N(k), Z^N)) \geq T \} \cup_{j \in [s]} \{ \mathcal{I}(\tau(X^N(j), Z^N)) \leq T \} \right]. \quad (47)$$

We bound probability of (47) from above by

$$\sum_{k > s} P(\mathcal{I}(\tau(X^N(k), Z^N)) \geq I_1 - \varepsilon) + \sum_{j \in [s]} P(\mathcal{I}(\tau(X^N(j), Z^N)) \leq I_1 - \varepsilon) \quad (48)$$

where $\varepsilon > 0$ is arbitrary. Let us estimate from above the first sum in (48). It is easy to check that

$$\begin{aligned} \mathcal{K}(\tau(x(k), z) || \pi(x(k), \cdot) \pi(\cdot, z)) &= \mathcal{K}(\tau(x(k), z) || \tau(x(k)) \tau(z)) + \\ &+ \mathcal{K}(\tau(z) || \pi(\cdot, z)) + \mathcal{K}(\tau(x(k)) || \pi(x, \cdot)) \\ &= \mathcal{I}(\tau(x(k), z)) + \mathcal{K}(\tau(z) || \pi(\cdot, z)) + \\ &+ \mathcal{K}(\tau(x(k)) || \pi(x, \cdot)). \end{aligned} \quad (49)$$

By Sanov theorem the first sum in (48) does not exceed

$$\begin{aligned} (t - s) \exp\{-N \min_{\tau \in D_\varepsilon} \mathcal{K}(\tau(x(k), z) || \pi(x(k), \cdot) \pi(\cdot, z))\} \\ \leq (t - s) \exp\{-N(I_1 - \varepsilon)\}, \end{aligned} \quad (50)$$

where $D_\varepsilon = \{\tau : \mathcal{I}(\tau) \leq I_1 - \varepsilon\}$, in view of (48), (49) and nonnegativity of $\mathcal{K}[\cdot || \cdot]$. When $R = \frac{\ln t}{N} \leq I_1 - 2\varepsilon$ we have

$$(t - s) \exp\{-N(I_1 - \varepsilon)\} \leq \exp\{N((I_1 - 2\varepsilon) - I_1 - \varepsilon)\} = \exp\{-N\varepsilon\}.$$

By Sanov theorem, the second sum in (48) is bounded from above by

$$s \exp\{-N \min_{\tau \in ClD_\varepsilon^c} \mathcal{K}(\tau(x(\xi), z) || \pi(x, z))\}. \quad (51)$$

Now, we note that, by definition, $\pi(\cdot, \cdot) \notin ClD_\varepsilon^c$ for any $\varepsilon > 0$.

Since $\mathcal{K}(\tau || \pi)$ is a convex function of the pair (τ, π) (see Cover and Thomas(1991)), hence \mathcal{K} is jointly continuous in τ , and π , and

$\min_{\tau \in ClD_\varepsilon^c} \mathcal{K}(\tau || \pi) = \mathcal{K}(\tau^* || \pi)$ is attained at some points $\tau^* \in D_\varepsilon$ which does not coincide with π . Hence $\mathcal{K}(\tau^*(x(\xi), z) || \pi(x, z)) = \delta(\varepsilon) > 0$. We see that (51) does not exceed $s \exp\{-N\delta(\varepsilon)\}$. The total upper bound for (46) is

$$\exp\{-N\varepsilon\} + s \exp\{-N\delta(\varepsilon)\} \rightarrow 0 \text{ when } N \rightarrow \infty$$

for any rate $R \leq I_1 - 2\varepsilon$. Therefore, $AR > I_1 - 2\varepsilon$ for any $\varepsilon > 0$ and, consequently, $AR \geq I_1$. The proof is complete.

Here we prove the theorem on asymptotic optimality of test T_s . Recall that by Lemma 1 we need to bound from above the mean over the ensemble of designs of

the conditional error probability when the true set A is fixed. The proof is the generalization of the proof of theorem 1 involving some type of conditioning.

Proof of Theorem 2

We denote the set of unordered v -sets $\nu = (i_1, \dots, i_\nu)$, $1 \leq i_1 < i_2 < \dots < i_\nu < s$, $0 \leq \nu < s$ by V . For an arbitrary $\nu \in V$ we define $I(\nu)$ as follows:

$$I(\nu) = I(Z \wedge X([s] \setminus \nu) | X(\nu)).$$

Let us consider the subset $\Lambda(\nu)$, $\nu = (i_1, \dots, i_\nu) \in V$ of the set $\Lambda_0 = \Lambda(s, t) \setminus [s]$ that consists of those $\lambda \in \Lambda_0$ for which $\{\lambda_1, \dots, \lambda_s\} \cap [s] = \{i_1, \dots, i_\nu\}$.

It is obvious that $\Lambda(\nu) \cap \Lambda(\nu^*) = \emptyset$ if $\nu \neq \nu^*$ and $\Lambda_0 = \cup_V \Lambda(\nu)$.

We define

$$A(\nu) = \{(X^N, Z^N) : T_s(\mathcal{X}^N, Z^N) \in \Lambda(\nu)\}.$$

It is obvious that $A(\nu)$ is expressed in the form

$$A(\nu) = \{(\mathcal{X}^N, Z^N) : \cup_{\lambda \in \Lambda(\nu)} \mathcal{I}(\tau(X^N(\lambda), Z^N)) \geq \mathcal{I}(\tau(X^N([s]), Z^N))\}.$$

By the formula for the composite probability we get

$$P(A(\nu)) = \sum_{x^N(\nu), z} P(A(\nu) | x^N(\nu)) P(x^N(\nu)) \quad (52)$$

The following inclusion is valid for arbitrary T :

$$A(\nu) \subseteq \{\mathcal{I}(\tau(x^N([s]), z^N)) \leq T\} \cup \{\max_{\Lambda(\nu)} \mathcal{I}(\tau(x^N(\lambda), z^N)) \geq T\}.$$

Consequently, for fixed $x^N(\nu)$ the conditional probability of the event $A(\nu)$ can be estimated as

$$\begin{aligned} P(A(\nu) | x^N(\nu)) &\leq P(\mathcal{I}(\tau((x^N([s]), z^N))) \leq T | *) + \\ &+ P(\cup_{\Lambda(\nu)} \mathcal{I}(\tau((x^N(\lambda), z^N))) \geq T | *) \\ &\leq P(\mathcal{I}(\tau((x^N([s]), z^N))) \leq T | *) + \\ &+ \sum_{\Lambda(\nu)} P(\mathcal{I}(\tau((x^N(\lambda), z^N))) \geq T | *), \end{aligned}$$

where $T = T(*), * := x^N(\nu)$.

$$\begin{aligned} P(A(\nu)) &= \sum_{x^N(\nu), z} \left[P(\mathcal{I}(\tau((x^N([s]), z^N))) \leq T | *) + \right. \\ &+ \left. \sum_{\lambda \in \Lambda(\nu)} P(\mathcal{I}(\tau((x^N(\lambda), z^N))) \geq T | *) \right] P(x^N(\nu)). \quad (53) \end{aligned}$$

Let us estimate from above the second sum in square brackets in (53), $T = I(\nu) - \varepsilon$. we have the corresponding conditional identities

$$\begin{aligned}
& \mathcal{K}(\tau(x^N(\lambda), z^N) \|\pi(x^N(\lambda), \cdot)\pi(\cdot, z^N) | \tau_{x^N}(\nu)) = \\
&= \sum_{x \in \mathcal{X}} \tau_{x^N}(\nu) \{ \mathcal{K}(\tau(x^N(\lambda), z^N) | \tau_{x^N}(\nu)) \|\pi(x^N(\lambda), \cdot)\pi(\cdot, z^N) | \tau_{x^N}(\nu) \} = \\
&= \sum_{x \in \mathcal{X}} \tau_{x^N}(\nu) \{ \mathcal{K}(\tau(x^N(\lambda), z^N) \|\tau(x^N(\lambda))\tau(z^N) | \tau_{x^N}(\nu)) + \\
&+ \mathcal{K}(\tau(z^N) \|\pi(\cdot, z^N) | \tau_{x^N}(\nu)) + \mathcal{K}(\tau(x^N(\lambda)) | \tau_{x^N}(\nu)) \|\pi(x^N, \cdot) | \tau_{x^N}(\nu) \} = \\
&= \mathcal{K}(\tau(x^N(\lambda), z^N) \|\tau(x^N(\lambda))\tau(z^N) | \tau_{x^N}(\nu)) + \\
&+ \mathcal{K}(\tau(z^N) \|\pi(\cdot, z^N) | \tau_{x^N}(\nu)) + \mathcal{K}(\tau(x^N(\lambda)) \|\pi(x^N, \cdot) | \tau_{x^N}(\nu)) = \\
&= \mathcal{I}(\tau(x^N(\lambda), z^N) | \tau_{x^N}(\nu)) + \\
&+ \mathcal{K}(\tau(z^N) \|\pi(\cdot, z^N) | \tau_{x^N}(\nu)) + \mathcal{K}(\tau(x^N(\lambda)) \|\pi(x^N, \cdot) | \tau_{x^N}(\nu)). \tag{54}
\end{aligned}$$

By conditional Sanov theorem in view of (53), (54), the nonnegativity of $\mathcal{K}(\cdot \|\cdot)$, and the inequality $|\Lambda(\nu)| \leq t^{s-\nu}$, the second sum in square brackets in (53) does not exceed

$$\begin{aligned}
& t^{s-\nu} (N+1)^{|\mathcal{X}|^s |\mathcal{Z}|} \times \\
& \times \exp\left\{-N \min_{\tau \in D_\varepsilon} \mathcal{K}(\tau(x^N(\lambda), z^N) \|\pi(x^N(\lambda))\pi(z^N) | \tau_{x^N}(\nu))\right\}.
\end{aligned}$$

Here $D_\varepsilon = \{\tau : \mathcal{I}(\tau) \leq I(\nu) - \varepsilon\}$. Since the last bound does not depend on $\tau(x^N(\nu))$ we bound the sum in (53) over types of $x^N(\nu)$ from above by

$$t^{s-\nu} (N+1)^{|\mathcal{X}|^s |\mathcal{Z}|} \exp\{-N(I_\pi(\nu) - \varepsilon)\}. \tag{55}$$

When $R = \frac{\ln t}{N} \leq I(\nu) - 2\varepsilon$, (55) is bounded from above by

$$\exp\left\{-N\left[\varepsilon(s - \nu) - \frac{|\mathcal{X}|^s |\mathcal{Z}| \log(N+1)}{N}\right]\right\},$$

which is exponentially small for sufficiently large N .

Also by conditional Sanov theorem, the first term in (53) is bounded from above by

$$\sum_{\tau_{x^N}(\nu)} (N+1)^{|\mathcal{X}^s| |\mathcal{Z}|} \times \\ \times \exp\{-N \min_{CID_\varepsilon^c} \mathcal{K}(\tau(x^N([\cdot]), z^N) || \pi(x^N, z^N) | \tau_{x^N}(\nu))\} p(\tau_{x^N}(\nu)). \quad (56)$$

We note that, by definition, $\pi(\cdot, \cdot) \notin CID_\varepsilon^c$ for any $\varepsilon > 0$. Hence $\min_{CID_\varepsilon^c} \mathcal{K}(\tau(x^N(\lambda), z^N) || \pi(x^N, z^N) | \tau_{x^N}(\nu)) = \delta(\varepsilon) > 0$. because of the continuity of conditional KULD in all its variables. We see that (56) does not exceed

$$\exp\{-N(\delta(\varepsilon) - \frac{|\mathcal{X}^s| |\mathcal{Z}| \log(N+1)}{N})\}.$$

The total upper bound for (53) is

$$\exp\{-N(\delta(\varepsilon) - \frac{|\mathcal{X}^s| |\mathcal{Z}| \log(N+1)}{N})\} + \\ + \exp\{-N[\varepsilon(s - \nu) - \frac{|\mathcal{X}^s| |\mathcal{Z}| \log(N+1)}{N}]\} \rightarrow 0, \quad \text{when } N \rightarrow \infty$$

for any rate $R \leq I - 2\varepsilon$. Therefore, $AR > I - \varepsilon$ for any $\varepsilon > 0$, and consequently, $AR \geq I$. The proof is complete.

Remark Proof of the last statement of theorem 1 for ML-test uses some of the main ideas of the capacity region construction for MAC without feedback (see Csiszar and Körner (1981) and the end of this section), although it does not follow from the MAC theory. However, the lower bound for AR is proved almost along the lines of Csiszar and Körner (1981) (it does not follow from famous Fano inequality). Proof and its generalization for the case of unknown s and for unknown $T(\cdot)$ are published in Malyutov, Dyachkov (1980) (and in Malyutov (1983a)) in Russian, supplied by the Dyachkov's estimate from below for the same quantity in the case

$$|\log \gamma(t)| = 0(\log t),$$

which is, unfortunately, not sharp. The proof of Malyutov is reproduced in German by Viemeister (1982) supplied by some minor generalizations. In Malyutov, Mateev (1980) the coincidence of (3) and (6) for ordinary models introduced there (including symmetric ones) is proved. Solution of maximal empirical Shannon Information was studied in Csiszar and Körner (1981) for conventional Shannon scheme of Information Theory and essentially the same decision was studied in Dyachkov and Rashad (1989) for the search of SV's of symmetric discrete functions of s variables when $t \rightarrow \infty$. In Malyutov and Mateev (1980) the maximal AR of separate detection of SV's by ML-test is found.

REFERENCES

1. Budne, T.A. (1959). "Application of Random Balance Designs," *Technometrics*, 1, No 2.
2. Chernoff, H. (1956). "Large sample theory: parametric case," *Ann. Math. Statist.* Vol. 27, pp. 1 - 22.

3. Csiszar, I. and Körner, J. (1981). **Information Theory: Coding Theorems for Discrete Memoryless Systems**, Academic Press and Akadémiai Kiadó, Budapest.
4. Cover, T. M., Thomas, J. A. (1991). **Elements of Information Theory**. Wiley, New York.
5. Dyachkov, A. G., "On a Search Model of False Coins," in *Colloquia Mathematica Societatis Janos Bolyai: Topics in Information Theory*, Keszthely (1975), North-Holland, Amsterdam (1977).
6. Dyachkov, A. G. and Rashad, A. M.(1989). Universal Decoding for Random Design of Screening Experiments. *Microelectronics Reliability*, **29**, No 6, 965-971.
7. Erdős, P. and Renyi, A. (1963). "On two Problems of Information Theory," *Publ. Math. Inst. of Hung. Acad.of Sc.*, bf 8, 229-243.
8. Freidlina, V. L. (1975). "On One Problem of Screening Experimental Design," *Theory Probab. Appl.* **20**, 1.
9. Gallager, R. G. **Information theory and reliable communication**. Wiley, New York.
10. Karlin, S. (1959). *Mathematical Methods and Theory In Game, Theory Programming and Economics*. Addison-Wesley, Reading. MA.
11. Kullback, S. (1956). **Information Theory and Statistics**. Wiley, New York.
12. Lindström, B. (1975). Determining Subsets by Unramified Experiments. *A Survey of Statistical Design and Linear Models*, North-Holland Pub.Co, 407-418
13. Malyutov, M. B. "Mathematical Models and Results in the Theory of Screening Experiments." In: *Theoretical Problems of Experimental Design* (ed. M. B. Malyutov), Sov. Radio, 5-69 (In Russian).
14. Malyutov, M. B. (1983a). Information-Theoretic Methods of Design and Analysis of Experiments. *D. Sci. Thesis, Moscow Lomonosov University* (In Russian).
15. Malyutov, M. B. (1979). "On the Limit Rate of Screening Designs," *Theory Probab. and Appl.* **24**, 3 (In Russian).
16. Malyutov, M. B. and Dyachkov, "Coding of Requests in The Case of multiple Access," in *Proc. Fifth All-Union Seminar-Workshop on Computer Networks* (In Russian), *part 2*,: Abstracts of Reports, Moscow-Viadivostok (1980), pp. 62-64.
17. Malyutov, M. B. and Mateev, P. S. (1980). Planning of Screening Experiments for a Nonsymmetric Response Function. *Math. Zametki* **27**, 1, English translation by Plenum Publ. Co., 57-68.
18. Malyutov, M. B. and Pinsker, M. S. (1972). Note on the Simplest Model of the Random Balance Method. *Probabilistic Methods of Research*. Moscow University Press (ed. A. N. Kolmogorov). (In Russian).
19. Malyutov, M. B. and Tsitovich, I. I. (1996). On Sequential Search for Significant Variables of Unknown Function. *Proceeding of 6th-Lucacs Symposium*, 115-138, VSP, Netherlands

20. Malyutov, M. B. and Wynn, H. P. (1994). Screening of Significant Variables of an Additive Smooth Function. In: *Markov Processes and Related Fields*. Festschrift for E. B. Dynkin (ed. M. I. Freidlin), Birkhäuser, Boston.
21. Mateev, P. S. (1978). On Entropy of a Polynomial Distribution. *Theory Probab. and Appl.* **23**, 1, 196–198.
22. Meshalkin, L. D. (1970). To the Justification of Random Balance Method. *Industrial Laboratory*, **36**, No 3.
23. Patel, M. S., Editor (1987). Experiments in Factor Screening. *Communications in Statistics*. Theory and Methods. *Special Issue*, **vol. 16**, No. 10.
24. Renyi, A. (1965). On The Theory of Random Search. *Bull. Amer. Math. Soc.*, **71**, 6, 809-828.
25. Viemeister, J. (1982). Diplomarbeit, *Bielefeld University*.

DEPARTMENT OF MATHEMATICS, NORTHEASTERN UNIVERSITY, BOSTON, MA 02115

E-mail address: `mltv@neu.edu`, `sadaka@neu.edu`