On some testing of hypothesis problems with information constraints *

Burnashev M. V.[†], Amari S.[‡], Han T. S.[§]

Abstract

A problem of hypothesis testing is considered where some part of data can not be directly observed. Our helper observes those data and can send us some limited amount of information about them. What kind of that information allows us to make the best statistical inferences ? In particular, what is the minimal information sufficient in order to get the same results as if we could observe directly all data ? Some upper bounds for that minimal amount of information and some related results are obtained.

Key words and phrases: hypothesis testing, error probability, rate of transmission, critical rate.

I. Introduction. General description of the problem

In this paper we would like to draw specialist's attention to one rather new and essentially open class of problems and to present some particular results on that matter. Although those problems have been started to investigate in information theory [1 - 5], they have statistical essence and knowledge of information theory is not needed for their understanding.

*The research described in this publication was made possible in part by Grant N 98-01-04108 from the Russian Fund for Fundamental Research.

†Institute for Problems of Information Transmission, RAS, Bolshoi Karetnii 19, 101447, Moscow, Russia.

‡RIKEN Brain Science Institute, Wako-shi, Hirosawa 2-1, Saitama 351-0198, Japan.

§Graduate School of Information Systems, University of Electro–Communications, Chfugaoka 1 - 5 - 1, Chofu, Tokyo 182, Japan.

For that purpose we consider some particularly interesting cases of the following general problem.

Let a "statistician" should make a certain statistical inferences concerning the system state (for example, to estimate some unknown parameter, to test some hypotheses, etc.). There are two sets of data (observations): set $\mathcal{A}$ ("available") and set $\mathcal{R}$ ("remote").

Statistician directly observes all data from the set $\mathcal{A}$.

He can not observe directly data from the remote set $\mathcal{R}$, but his "helper" can observe them. Moreover the helper is allowed to send statistician some limited amount of information about those data.

The problem is: what kind of information (limited) about those remote data should send the helper that will allow statistician to make the best possible statistical inferences (for example, to get the minimum mean–square error for parameter estimate, etc.) ?

There are many practical situations when we meet this kind of problem. For example, in some applications the set $\mathcal{R}$ can be regarded as some "nuisance noise" that has "contaminated" already the data from the set $\mathcal{A}$, and therefore we would like to "remove" (as much as possible) that "contamination" in order to improve our statistical inferences.

We will deal below with discrete–time models and moreover, by "limited amount of information" we will mean that the helper can send information with communication rate not exceeding some prescribed value $R > 0$ (see Example 1 below). Of course, if R is such large that the helper can simply resend to statistician all data from $\mathcal{R}$ then we come back to traditional statistical problem (that is not of our interest here). For that reason it is natural to assume that R is small enough in order to avoid such primitive resending.

Nevertheless, even with such assumption there are some cases (sometimes, probably, natural) when the optimal solution can be obtained quite easily. Certainly, it will always be the case when data from sets $\mathcal{A}$ and $\mathcal{R}$ both represent independent observations of the same phenomenon. The following particular example illustrates such case.

Example 1. Let all data $\{x_1, \ldots, x_n\}$ from the set $\mathcal{A}$ and all data $\{x_{n+1}, \ldots, x_{n+m}\}$ from the set $\mathcal{R}$ have the form

$$x_i = \theta + \xi_i \, ; \qquad i = 1, \ldots, n+m \, ; \quad -\infty < a < \theta < b < \infty \, ,$$

where $\theta \in (a,b) \in R^1$ is unknown parameter (to be estimated) and ξ_i ; $i = 1,\ldots,n+m$, are i.i.d. random variables with some known density $f(z)$ (with respect to Lebeque measure on R^1). We assume that function $f(z)$ is sufficiently smooth such that Fisher information $I(f)$ is well–defined, i.e.

$$\int_{-\infty}^{\infty} \frac{f'^2(z)}{f(z)}\, dz = I(f) < \infty$$

Let $\hat{\theta}_{\mathcal{A}}(n)$ and $\hat{\theta}_{\mathcal{R}}(m)$ be maximum–likelihood estimates for parameter θ based on data from $\mathcal{A}$ and $\mathcal{R}$, respectively. Denote

$$\eta_{\mathcal{A}}(n) = (\hat{\theta}_{\mathcal{A}}(n) - \theta)\sqrt{nI}\,, \qquad \eta_{\mathcal{R}}(m) = (\hat{\theta}_{\mathcal{R}}(m) - \theta)\sqrt{mI}\,.$$

If now $n, m \to \infty$ then (under some mild additional assumptions about $f(z)$) both $\eta_{\mathcal{A}}(n)$ and $\eta_{\mathcal{R}}(m)$ are asymptotically independent $(0,1)$–Gaussian random variables and, moreover, all their moments converge to corresponding moments of that Gaussian r.v. [6]. Consider now the estimate

$$\hat{\theta}(n+m) = \frac{n\hat{\theta}_{\mathcal{A}}(n) + m\hat{\theta}_{\mathcal{R}}(m)}{n+m}\,.$$

and denote $\eta(n+m) = (\hat{\theta}(n+m) - \theta)\sqrt{(n+m)I}$. Random variable $\eta(n+m)$ is asymptotically also $(0,1)$–Gaussian r.v. and therefore we get

$$E_\theta \left(\hat{\theta}(n+m) - \theta\right)^2 (n+m)I \to 1\,; \qquad n, m \to \infty\,.$$

It is known (Cramer–Rao bound) that we can not get asymptotically smaller mean–square error. On the other hand, in order to construct the estimate $\hat{\theta}(n+m)$ statistician should get from his helper the value of $\hat{\theta}_{\mathcal{R}}(m)$ with precision $O\left((n+m)^{-1}\right)$. It means that he should get the total amount of information of the order of $\log((b-a)(n+m))$. Therefore for finite $b-a$ the rate of transmission R has the order of $(\ln(n+m))/n$. Now if $(\log m)/n \to 0$, $n \to \infty$, then for large n transmission rate R can be made arbitrary small. It means that for arbitrary small $R > 0$ we can easily achieve asymptotically the best performance. $\Box$.

Situation becomes much more difficult (and more interesting) when there is a sufficiently strong dependence between data from sets $\mathcal{A}$ and $\mathcal{R}$. We consider mainly the case when neither statistician, nor helper can make any good statistical inferences based only on their own data (in other words, there is a very strong dependence between data from $\mathcal{A}$ and $\mathcal{R}$).

Next example (its model will serve as basic for us) illustrates such case.

Example 2. We consider the device called "binary symmetrical" channel (BSC) and representing the most popular and interesting model in information theory [7, 8]. Its input and output alphabets are binary, i.e. input and output channel signals take values 0 and 1. It is characterized by certain transition probability $0 < p \le 1/2$. It means that if x_i and y_i are input and output signals at any moment i then

$$P(y_i = 1|x_i = 0) = P(y_i = 0|x_i = 1) = p\,;$$

$$P(y_i = 0|x_i = 0) = P(y_i = 1|x_i = 1) = 1 - p = q\,.$$

Moreover, that channel performs independently at different time moments, i.e. for any collection of time moments $i_1, \ldots, i_n$ and any collection $(x_{i_1}, \ldots, x_{i_n}, y_{i_1}, \ldots, y_{i_n})$ the following condition is fullfilled

$$P(y_{i_1}, \ldots, y_{i_n}|x_{i_1}, \ldots, x_{i_n}) = P(y_{i_1}|x_{i_1}) \ldots P(y_{i_n}|x_{i_n})\,.$$

(Such channels are called "memoryless"). In particular, if at the channel input there is a sequence of zeros then at the channel output the sequence of independent Bernoulli random variables with success probability p is observed.

Unknown parameter for us will be transition probability $0 < p \le 1/2$ which we will need to estimate or to test some hypotheses about it. Moreover, statistician observes only the channel output $\mathcal{A} = (y_1, \ldots, y_n)$ and the helper observes only the channel input $\mathcal{R} = (x_1, \ldots, x_n)$. We assume also that there is no any prior information about input block $(x_1, \ldots, x_n)$.

It is clear that if statistician knows nothing about input block $(x_1, \ldots, x_n)$ then he can not make any reasonable conclusions on unknown value p.

The fact that helper may send to statistician information with rate not exceeding the prescribed value $R > 0$ means that they are allowed to partition the input space $E^n = \{0, 1\}^n$ on $N \le 2^{Rn}$ arbitrary parts $\{X_1, \ldots, X_N\}$ and helper informs staistician only to which part X_i belongs the input block x^n. It is clear that only the case $N < 2^n$, i.e. $R < 1$, is interesting (otherwise helper can simply resend the value x^n).

For example, helper can exactly inform statistician on the first Rn values $x_1, \ldots, x_{Rn}$ (but then he will send no information on other values x_i). Such simple partition method of input space E^n (on cylinder sets $\{X_i\}$) is not generally optimal. From statistician point of view input data $(x_1, \ldots, x_n)$ represent a very severe nuissance parameter.

We can say also that transmission of optimal limited information about block x^n means optimal "compression" of full information about block x^n. Of course, that optimal "compression" depends on prior information on transition probability p and quality criteria used.

Remark. It is clear that the problem will not be changed if statistician observes channel input and helper observes channel output. We will use later both variants of that problem statement.

In the paper for the BSC model from example 2 we consider traditional testing of two simple hypotheses problem on parameter p is considered. In the second part of the paper we will investigate parameter p estimation problem.

In both problems we will demonstrate some partitions $\{X_1, \ldots, X_N\}$ and decision making (estimation) methods that are asymptotically (when $n \to \infty$), probably, close to optimal ones. Unfortunately, we were not able yet to show that it is not possible to perform better and it remains an open problem.

We limit ourselves here to BSC (i.e. independent Bernoulli random variables with unknown parameter p) for the following reasons:

1. For a person sufficiently familiar with information theory it is rather clear that in interesting cases some function similar to the *reliability function* of the channel [7,8] should be presented in the solution. From reliability function point of view BSC is a very illustrative example (i.e. it contains all essential problems; all other channel are treated using essentially methods developed for BSC; there are still only some lower and upper bounds for reliability function of BSC; etc.).

2. All statistical quantities (e.g. Kullback–Leibler information, Fisher information, etc.) have very simple analitical form and geometrical meaning for BSC. For that reason in BSC case all main difficulties of the problem considered will be clearly seen and they will not be additionally complifyed by more technical type questions.

We can repeat also a well–known claim: "show us how to deal with BSC (or Bernoulli distributions) and we will show you how to do the same for much broader class of channels (distributions)".

Below we mean $\log x = \log_2 x$, $\exp_2 x = 2^x$. For any finite set A by $|A|$ its cardinality is denoted. For any function $f(x), x \in A$ by $|f|$ cardinality of the set $f(A)$ is denoted. In order to distinquish input and output alphabets $E = \{0, 1\}$ we denote them E_{in} and E_{out}, respectively.

II. Testing of two simple hypotheses

1. Statement of the problem and dual problem

We consider BSC with some crossover probability p to be tested. We assume that p satisfies one of two hypotheses: $H_0 : p = p_0$ or $H_1 : p = p_1$, where $0 < p_0 < p_1 \le 1/2$.

We denote by P and Q conditional output distributions for H_0 and H_1, respectively. Therefore, probabilities to get output block $y^n = (y_1, \ldots, y_n)$ provided that input block was $x^n = (x_1, \ldots, x_n)$ are given, respectively, by

$$P(y^n|x^n) = (1-p_0)^{n-d(x^n,y^n)} p_0^{d(x^n,y^n)}$$

and

$$Q(y^n|x^n) = (1-p_1)^{n-d(x^n,y^n)} p_1^{d(x^n,y^n)} ,$$

where $d(x^n, y^n)$ is the Hamming distance between blocks x^n and y^n (i.e. the minimal number of noncoinciding components on the whole length n).

We are interested in testing of those hypotheses when we observe only the channel output and we get from helper only some limited information about the input block. We consider minimax statement of the problem.

To be specific, assume that we are allowed to partition the input space E_{in}^n on N parts $\{X_1, \ldots, X_N\}$. After that we observe channel output $y^n \in E_{\text{out}}^n$ and helper informs us only to which part X_i belongs the input block x^n. On the basis of observed y^n and index of X_i we make a decision in favor of one of hypotheses H_0 or H_1. In order to avoid overcomplification we consider only nonrandomized decision methods (then problem essence and results remain the same). Then general decision method can be described as follows. For any partition element X_i we choose some set $\mathcal{A}(X_i) \subset E_{\text{out}}^n$ and then depending on observation y^n make a decision ($\mathcal{A}^c = E_{\text{out}}^n \setminus \mathcal{A}$):

$$y^n \in \mathcal{A}(X_i) \Longrightarrow H_0 \, ; \quad y^n \in \mathcal{A}^c(X_i) \Longrightarrow H_1 \, .$$

Define error probabilities of the 1–kind α_n and the 2–kind β_n as

$$\alpha_n = \Pr\left(H_1|H_0\right) = \max_{i=1,\ldots,N} \max_{x^n \in X_i} P\left(\mathcal{A}^c(X_i)|x^n\right) ,$$

$$\beta_n = \Pr\left(H_0|H_1\right) = \max_{i=1,\ldots,N} \max_{x^n \in X_i} Q\left(\mathcal{A}(X_i)|x^n\right) .$$

Let $\gamma > 0$ be some given constant. We demand that the 1–kind of error probability satisfyes condition

$$\alpha_n = \Pr\left(H_1|H_0\right) \leq 2^{-\gamma n} . \tag{1}$$

We are interested in the minimal possible (over all partitions of input space and all decisions) 2–kind of error probability and we want to minimize (overall partitions of the input set and all decisions) the second kind of error probability $\inf \beta_n$.

We consider asymptotic situation when $n \to \infty$ and $N = 2^{Rn}$, where $0 < R < 1$ – some prescribed constant [1]. Then for the best criteria we denote

$$e(\gamma, R) = \lim_{n \to \infty} \frac{1}{n} \log_2 \frac{1}{\inf \beta_n} > 0 , \tag{2}$$

where inf is taken over all methods satisfying condition (1).

Our aim is to find (or to get good bounds) for the function $e(\gamma, R)$.

It will be convinient for us to consider also the following dual problem (without helper). Let some constant $0 < r < 1$ is given and we are allowed to choose in advance any set $\mathcal{X} \subset E_{\mathrm{in}}^n$ consisting of $X = 2^{rn}$ input blocks. Let we know also that input block may be only from the set $\mathcal{X}$.

Now, knowing the set $\mathcal{X}$, we observe the channel output y^n and consider the testing of hypotheses H_0 and H_1 problem. For a chosen set $\mathcal{A}$ depending on observation y^n we make the decision:

$$y^n \in \mathcal{A} \Longrightarrow H_0 ; \quad y^n \in \mathcal{A}^c \Longrightarrow H_1 ,$$

and define 1–kind and 2–kind error probabilities as

$$\alpha_n = \max_{x^n \in \mathcal{X}} P\left(\mathcal{A}^c|x^n\right) , \quad \beta_n = \max_{x^n \in \mathcal{X}} Q\left(\mathcal{A}|x^n\right) .$$

Let now for the 1–kind error probability condition (1) is fullfilled and we want to choose set $\mathcal{X}$ of cardinality $X = 2^{rn}$ and decision method in order to achieve the minimal possible 2–kind error probability $\inf \beta_n$. For such dual problem we can similar to (2) define function $e_2(\gamma, R)$. The following result establishes a simple relation between functions $e(\gamma, R)$ and $e_2(\gamma, R)$.

[1] In order to simplify formulas we don't use integer part sign of value 2^{Rn}

Proposition 1. *The following relation holds true*

$$e(\gamma, 1-R) = e_2(\gamma, R)\,; \qquad 0 \le R \le 1\,, \quad \gamma > 0\,. \tag{3}$$

In order to prove Proposition 1 we will need simple "covering" lemma (certainly, known).

Lemma 1. *Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_X\} \subset E^n$ be any set of cardinality X. Then there exist $K = n2^n/X$ "shifts" $\{\mathbf{y}_1, \ldots, \mathbf{y}_K\} \subset E^n$ such that sets $\mathcal{X} + \mathbf{y}_i\,; i = 1, \ldots, K$, cover all space E^n.*

Proof. We choose all K shifts randomly and independently (with returns). Then for any $K > n2^n \ln 2/X$ we have

$$\Pr\left\{\text{there exists some noncovered point } \mathbf{x} \in E^n\right\} \le$$

$$\le 2^n \Pr\left\{\text{point } \mathbf{0} \text{ is not covered}\right\} = 2^n \left(1 - X2^{-n}\right)^K \le$$

$$\le \exp\left\{-XK2^{-n} + n\ln 2\right\} < 1\,.$$

Therefore among such randomly chosen $K > n2^n \ln 2/X$ shifts there exist a collection, satisfying Lemma 1. $\square$

Proof of Proposition 1. Let the set $\mathcal{X}$ of cardinality $\sim 2^{Rn}$ be the best for dual problem, i.e. it gives 2–kind error probability $\sim 2^{-ne_2(\gamma,R)}$. Due to Lemma 1 the whole input space E_{in}^n can be covered by $N \sim 2^{(1-R)n}$ shifted versions of set $\mathcal{X}$ (each of them has the same "testing performance"). Reducing some elements of that covering, we can construct partition of the space E_{in}^n on $\sim 2^{(1-R)n}$ parts. Since we consider minimax statement of problem, "testing performance" of each part will be not be worse than for the original set $\mathcal{X}$, from where follows the inequality

$$e(\gamma, 1-R) \ge e_2(\gamma, R)\,; \qquad 0 \le R \le 1\,, \quad \gamma > 0\,.$$

Let now in the original problem we are given some partition $\{X_1, \ldots, X_N\}$ $N \sim 2^{(1-R)n}$, giving 2–kind error probability $\sim 2^{-ne(\gamma,1-R)}$. Then there exists some partition element X_i of cardinality $\sim 2^{Rn}$, for which in the dual problem 2–kind error probability also does not exceed $2^{-ne(\gamma,1-R)}$ from where the opposite inequality follows

$$e(\gamma, 1-R) \le e_2(\gamma, R)\,; \qquad 0 \le R \le 1\,, \quad \gamma > 0\,.$$

that completes proof of proposition 1. $\square$

Therefore due to Proposition 1 it is sufficient to investigate the function $e_2(\gamma, R)$. But first we remind some results for known input block.

2. Known input block

Assume first that we know input block x^n and we observe output block y^n. Without loss of generality we may assume that x^n is all–0 block.

It is clear that for the optimal test the decision set in favor of p_0 is a ball $S(\tau n, \mathbf{0})$ of some radius $\tau(\gamma) \geq p_0 n$ centered at zero. Performing only with exponential on n values we have for coefficient $\tau(\gamma)$ condition

$$2^{-\gamma n} \simeq \alpha_n \simeq \binom{n}{\tau n}(1-p_0)^{n-\tau n}p_0^{\tau n} \simeq 2^{nh(\tau)}(1-p_0)^{n-\tau n}p_0^{\tau n}\,,$$

$$h(x) = x\log(1/x) + (1-x)\log(1/(1-x))\,,$$

or

$$\gamma = \tau\log\frac{\tau}{p_0} + (1-\tau)\log\frac{1-\tau}{1-p_0} \equiv D(\tau||p_0)\;. \tag{4}$$

Since we want to have the 2–kind error probability β_n also small we need to have $p_0 \leq \tau \leq p_1$. The function $D(\tau||p_0)$ is $\cup$–convex on τ and monotonically increasing for $\tau \geq p_0$. Therefore γ should satisfy condition

$$0 \leq \gamma \leq D(p_1||p_0)\;.$$

For such γ the value $\tau(\gamma)$ is given as the unique root (for $p_0 \leq \tau$) of the equation (4).

For the 2–kind error probability β_n we have

$$\beta_n \simeq \binom{n}{\tau n}(1-p_1)^{n-\tau n}p_1^{\tau n}$$

or

$$\frac{1}{n}\log\frac{1}{\beta_n} \simeq e(\gamma) = D(\tau||p_1) = \gamma_2\,, \tag{5}$$

It is convenient to consider $p_0 \leq \tau \leq p_1$ as a parameter through which both error probabilities can be expressed (see (4) and (5)).

Remark. The function $D(x||y)$ is the divergence for two binomial r.v.'s with parameters x and y, respectively. In other words, it gives the best possible exponential rate for the 2–kind error probability with fixed 1–kind error probabilty (so its exponent rate is equal to 0) when testing two simple hypotheses: $H_0 : p = x$ against $H_1 : p = y$.

Examples.

1) Let $\gamma = 0$, then $\tau = p_0$ and $\gamma_2 = D(p_0||p_1)$.

2) Let $\gamma_2 = 0$, then $\tau = p_1$ and $\gamma = D(p_1||p_0)$.

3) Let $\gamma = \gamma_2$, then τ is the unique root of the equation $D(\tau||p_0) = D(\tau||p_1)$, from where it follows

$$\tau = \left(\log \frac{1-p_0}{1-p_1}\right) \Big/ \left(\log \frac{p_1(1-p_0)}{p_0(1-p_1)}\right) \tag{6}$$

and

$$\gamma = \gamma_2 = D(\tau||p_0) = D(\tau||p_1)\ .$$

3. Unknown input block and critical rate

As was already shown, if we know input block and $\alpha_n \leq 2^{-\gamma n}$ then the best exponent for the 2–kind error probability $e(\gamma)$ is given by formulas (4)–(5).

If we know only that input block belongs to some set $\mathcal{X}$ of cardinality $X \sim 2^{rn}$ then for the best chosen such set $\mathcal{X}$ the exponent of the 2–kind error probability is defined by function $e_2(\gamma, r)$. It is clear that

$$e_2(\gamma, r) \leq e(\gamma) ; \quad \gamma \geq 0 , \quad 0 \leq r \leq 1 . \tag{7}$$

Function $e_2(\gamma, r)$ is monotonically nonincreasing on r and $e_2(\gamma, 0) = e(\gamma)$. Therefore regarding function $e_2(\gamma, r)$ the following question immediately arises: does there exist $r > 0$ such that equality in (7) is fullfilled and, if so, what is the maximal such rate $r_{\text{crit}}(\gamma)$? Formally, define $r_{\text{crit}}(\gamma)$ as

$$r_{\text{crit}}(\gamma) = \sup \{ r : e_2(\gamma, r) = e(\gamma) \} ; \quad \gamma \geq 0 . \tag{8}$$

In other words, what is the maximal cardinality 2^{rn} of the best set $\mathcal{X}$ for which we can achieve the same asymptotical efficiency as for known input block (although we don't know input block) ?

Similarly we introduce critical rate $R_{\text{crit}}(\gamma)$ for the original problem

$$R_{\text{crit}}(\gamma) = \inf \{ R : e(\gamma, R) = e(\gamma) \} ; \quad \gamma \geq 0 . \tag{9}$$

Due to proposition 1 we have

$$R_{\text{crit}}(\gamma) = 1 - r_{\text{crit}}(\gamma) ; \quad \gamma \geq 0 . \tag{10}$$

Remark. The value $r_{\text{crit}}(\gamma)$ is similar to the channel capacity C, and function $e_2(\gamma, r)$ is similar to realiability function $E(r)$ in information theory [7, 8]. Exact form of the realiability function $E(r)$ is not known till now. Therefore complete investigation of function $e_2(\gamma, r)$ (for $r > r_{\text{crit}}(\gamma)$) looks rather difficult problem.

III. Estimates for $r_{\text{crit}}(\gamma)$ and $e(\gamma, R)$

1. Lower bound for $r_{\text{crit}}(\gamma)$ (randomly chosen set $\mathcal{X}$)

As before, let measure P corresponds to p_0, measure Q corresponds to p_1 and $0 < p_0 < p_1 \leq 1/2$. We consider all sets $\mathcal{X}$ of cardinality $X \sim 2^{Rn}$ on E_{in}^n. Let also some decision rule is chosen such that 1–kind error probability for each set $\mathcal{X}$ does not exceed given value α_n. Then each $\mathcal{X}$ has its own 2–kind error probability $\beta_n(\mathcal{X})$. It is clear that there exists some set $\mathcal{X}$ for which the value $\beta_n(\mathcal{X})$ does not exceed the averaged (over all sets $\mathcal{X}$) value $E\beta_n(\mathcal{X})$. Therefore if we are able to calculate (or upperbound) the value $E\beta_n(\mathcal{X})$ then it will give a certain lower bound for $e_2(\gamma, r)$ and $r_{\text{crit}}(\gamma)$.

Such random choice method (with possible modifications) represents in information theory the most universal tool for getting various existance theorems [7, 8].

In order to realize that approach we choose as set $\mathcal{X}$ of cardinality $X \sim 2^{Rn}$ on E_{in}^n randomly and equiprobably X different points $\{x_1, \ldots, x_X\}$ and let y be our observation.

As the acceptance region $\mathcal{A}(\tau)$ in favor of p_0 we use

$$\mathcal{A}(\tau) = \left\{ y : \min_{i=1,\ldots,X} d(y, x_i) \leq \tau n \right\} ,$$

where value $p_0 \leq \tau \leq p_1$ will be chosen later.

In order to investigate such test performance without loss of generality we may assume that true value of block x is $x_1 = \mathbf{0}$.

If hypothesis p_0 is valid then for 1–kind error probability we have

$$\alpha_n \leq P\left\{ w(y) > \tau n | p_0, x_1 \right\} \simeq$$

$$\simeq \binom{n}{\tau n} (1 - p_0)^{(1-\tau)n} p_0^{\tau n} \simeq \exp_2\{-nD(\tau || p_0)\} .$$

Let now hypothesis p_1 is valid. If $w(y) \leq \tau n$ then we accept that decision error takes place. If $w(y) > \tau n$ then we can make decision error only if in a sphere of radius τn centered at y there is some point x_i. Now for averaged 2–kind error probability $E\beta_n$ we have $(M = |E^n| = 2^n,\ V$–cardinality of ball of radius τn $E_{\text{in}}^n)$

$$E\beta_n \leq P\left\{ w(y) \leq \tau n | p_1, x_1 \right\} + 1 - \binom{M - V}{X - 1} \Big/ \binom{M - 1}{X - 1} =$$

$$= \exp_2\{-nD(\tau||p_1)\} + 1 - \prod_{i=1}^{V-1}\left[1 - \frac{(X-1)}{(M-V-i)}\right] \le$$

$$\le \exp_2\{-nD(\tau||p_1)\} + 1 - \left[1 - \frac{(X-1)}{(M-2V+1)}\right]^{V-1} \le$$

$$\le \exp_2\{-nD(\tau||p_1)\} + \frac{XV}{(M-2V)} \simeq$$

$$\simeq \exp_2\{-nD(\tau||p_1)\} + \exp_2\{-[1-h(\tau)-r]n\}\ .$$

Therefore there exists a set $\mathcal{X}$ of cardinality $X \sim 2^{rn}$ for which under decision rule described the following inequalities are fullfilled

$$\alpha_n \le \exp_2\{-nD(\tau||p_0)\}\,;$$

$$\beta_n \le \exp_2\left\{-n\min\left\{D(\tau||p_1), 1-h(\tau)-r\right\}\right\}\ .$$

Therefore for function $e_2(\gamma, r)$ the following lower bound is valid

$$e_2(\gamma, r) \ge \min\ \{D(\tau||p_1), 1-h(\tau)-r\}\,; \quad 0 < r < 1\,, \tag{11}$$

where the value $p_0 \le \tau \le p_1$ for $0 \le \gamma \le D(p_1||p_0)$ is defined as the unique root of equation $\gamma = D(\tau||p_0)$.

In particular, if we want to be fullfilled relation $\beta_n \simeq D(\tau||p_1)$ (as for $X = 1$) then it is sufficient to have

$$r \le 1 - h(\tau) - D(\tau||p_1)\ .$$

Last result can be formulated in the following form.

Proposition 2. *For crirical rate $r_{\rm crit}(\gamma)$ in dual problem the lower bound holds true*

$$r_{\rm crit}(\gamma) \ge 1 - h(\tau) - D(\tau||p_1)\ ; \qquad 0 \le \gamma \le D(p_1||p_0)\,. \tag{12}$$

where the value $p_0 \le \tau \le p_1$ is defined as the unique root of equation $\gamma = D(\tau||p_0)$, $0 \le \gamma \le D(p_1||p_0)$.

Remark. 1) Estimates (11)–(12) remain valid even when we test composite hypothesis $H_0 : p \le p_0$ against simple alternative $H_1 : p = p_1$.

2) Let $\tau = p_1$, i.e. $\gamma = D(p_1||p_0)$. Then $D(\tau||p_1) = 0$ and bound (12) takes the form:

$$r_{\text{crit}}(D(p_1||p_0)) \geq 1 - h(p_1)\,.$$

That bound is defined by a "sphere packing" of space E_{in}^n by balls of radius $p_1 n$! The reason that we knowing only the set $\mathcal{X}$ of cardinality $X \sim 2^{rn}$ are able to achieve the same performance as if we would know input block x is the following. For good set $\mathcal{X}$ (almost all randomly chosen sets $\mathcal{X}$ are such) knowing output block y and set $\mathcal{X}$ it is possible with small error probability to identify which of input blocks $x \in \mathcal{X}$ was really used (under any hypotheses).

2. Case $p_1 = 1/2$

In the special case $p_1 = 1/2$ it turns out to be possible to find the function $e_2(\gamma, r)$.

If $p_1 = 1/2$ then for any x^n, y^n we have $Q(y^n|x^n) = 2^{-n}$. Therefore if in the dual problem $\mathcal{A}$ is the set of making decision in favor of p_0 then for the 2–kind error probability we have $\beta_n = |\mathcal{A}|2^{-n}$. Due to simplicity of that expression it is more convinient now to fix exponent rate of the 2–kind error probability $0 \leq \gamma_2 \leq D(p_0||1/2) = 1 - h(p)$ (i.e. $\beta_n \simeq 2^{-\gamma_2 n}$) and to investigate the best exponent rate of the 1–kind error probability $e_1(\gamma_2, r)$. For given value γ_2 for the cardinality of set $\mathcal{A}$ we have

$$|\mathcal{A}| \simeq 2^{(1-\gamma_2)n} .$$

On the other hand, for each input block $x_i \in \mathcal{X}$ optimal region of making decision in favor of p_0 is a ball of some radius τn centerd at x_i. Therefore for the optimal set $\mathcal{X}$ acception region $\mathcal{A}$ should contain "almost completely" each of those balls. It is clear that in such set $\mathcal{X}$ all points $\{x_i\}$ should be maximally close to each other (i.e. $\mathcal{X}$ is a ball) and $\mathcal{A}$ is also a ball concentric with it. Therefore we have

$$r_{\text{crit}}(\gamma) = 0, \;\gamma \geq 0 . \tag{13}$$

Let νn and μn be radiuses of balls $\mathcal{X}$ and $\mathcal{A}$, respectively. Since cardinality $|\mathcal{X}| \simeq 2^{rn}$ then ν and μ are defined from relations

$$r = h(\nu)\,; \; 1 - \gamma_2 = h(\mu)\,; \quad 0 \leq r < 1/2\,, \quad p_0 \leq \mu < 1/2 . \tag{14}$$

We may assume that balls $\mathcal{X}$ and $\mathcal{A}$ are centered at zero. If hypothesis p_0 is valid anf input block x has weight νn then due to large numbers law output block y has (with probability close to 1) weight $(p_0 + \nu(1 - 2p_0))n$. Therefore in order to have the 1–kind error probability α_n small it is necessary that μ satisfyes condition

$$\mu \geq p_0 + \nu(1 - 2p_0) . \tag{15}$$

From that condition and (14) follows

$$e_1(\gamma_2, r) = 0\,; \quad r \geq r_0(\gamma_2, p_0) = h\left(\frac{h^{-1}(1-\gamma_2) - p_0}{1 - 2p_0}\right) . \tag{16}$$

Now for $r < r_0(\gamma_2, p_0)$ we evaluate the 1–kind error probability (it will define function $e_1(\gamma, r)$). The 1–kind error takes place if outpur block has weight greater than μn. We may assume that input block x has ones on the first νn positions and zeros on remaining $(1-\nu)n$ positions. Let in denotes the number of errors on the first νn positions and jn denotes the number of errors on remaining $(1-\nu)n$ positions. Then erroneous decision takes place if $\nu - i + j \geq \mu$. Therefore denoting $z = (1-p_0)/p_0 > 1$ we have

$$\alpha_n \simeq (1-p_0)^n \max \binom{\nu n}{in}\binom{(1-\nu)n}{jn} z^{-(i+j)n} \simeq \tag{17}$$

$$\simeq (1-p_0)^n \exp_2\left\{n \max \left[\nu h\left(\frac{i}{\nu}\right) + (1-\nu)h\left(\frac{j}{1-\nu}\right) - (i+j)\log z\right]\right\} .$$

where maximum is taken over $0 \leq i \leq \nu$; $0 \leq j \leq 1-\nu$; $\nu - i + j \geq \mu$.

It is not difficult to check that at the point of maximum of right side of (17) equality $\nu - i + j = \mu$ holds true (otherwise condition (15) is violated). Therefore from (17) we get

$$e_1(\gamma_2, r) = \log\frac{1}{1-p_0} - \max_{0 \leq i \leq \nu} f(i)\,; \tag{18}$$

$$f(i) = \nu h\left(\frac{i}{\nu}\right) + (1-\nu)h\left(\frac{\mu-\nu+i}{1-\nu}\right) - (2i+\mu-\nu)\log z\,, \quad z = \frac{1-p_0}{p_0}\,.$$

It is easy to check that function $f(i)$ is $\cap$–convex on i and attains its maximum inside the interval $(0,\nu)$. Therefore optimal value i_0 is the unique root of equation

$$\log\frac{\nu-i}{i} + \log\frac{1-\mu-i}{\mu-\nu+i} = 2\log z\,,$$

from where denoting $u = (1-2p_0)/p_0^2 = z^2 - 1$ we get

$$i_0 = \frac{\sqrt{[u(\mu-\nu)+1]^2 + 4u\nu(1-\mu)} - u(\mu-\nu) - 1}{2u}\,, \quad u = \frac{1-2p_0}{p_0^2}\,. \tag{19}$$

These results can be formulated in the following form.

Proposition 3. *If $p_1 = 1/2$ then $r_{\rm crit}(\gamma) = 0$, $\gamma \geq 0$, and the best exponent $e_1(\gamma_2, r)$ of the 1–kind error probability is given by formula*

$$e_1(\gamma_2, r) = \begin{cases} 0, & r \geq r_0(\gamma_2, p_0); \\ -\log(1-p_0) - f(i_0)\,, & 0 \leq r < r_0(\gamma_2, p_0); \end{cases} \tag{20}$$

where $\mu, \nu, i_0, r_0(\gamma_2, p_0)$ are defined in (14), (16), (15) (19).

3. Useful counterexample

Unfortunately, we were not able yet to get good lower bounds for critical rate $R_{\text{crit}}(\gamma)$ (or upper bound for $r_{\text{crit}}(\gamma)$).

The following counterexample demonstrates some problems arising when trying to get such results.

We consider the following variant of dual problem. Let $0 < p_0 < p_1 < 1$ be fixed and it is known that input block x^n belongs to some set $\mathcal{X} \subset E_{\text{in}}^n$ of cardinality $X \simeq 2^{rn}$. What is the maximal growth rate $r_{\max}$ of cardinality of the best set $\mathcal{X}$ such that we can test those hypotheses if we demand only that both error probabilities vanish ? Answer is very simple: $r_{\max} = 1$.

Indeed, we choose as set $\mathcal{X} = \{x_1, \ldots, x_X\}$ all points on a sphere of some radius $rn < n/2$. Then $X \simeq 2^{nh(r)}$. Since input block has weight rn then due to large numbers law output block with probability close to 1 will have weight $(r + p(1 - 2r))n$. Therefore for large n we will be able to test hypotheses $p = p_0$ and $p = p_1$ with small error probabilities for any $r < 1/2$. It means that $r_{\max} = 1$.

REFERENCES

1. *Ahlswede R., Csiszár I.* Hypothesis testing with communication constraints. – IEEE Trans. on Inform. Theory, 1986, v. IT–32, No. 4, p. 533–542.

2. *Zhang Z., Berger T.* Estimation via compressed information. – IEEE Trans. on Inform. Theory, 1988, v. IT–34, No. 2, p. 198–211.

3. *Han T. S., Kobayashi K.* Exponential–type error probabilities for multiterminal hypothesis testing. – IEEE Trans. on Inform. Theory, 1989, v. IT–35, No. 1, p. 2–14.

4. *Ahlswede R., Burnashev M. V.* On Minimax estimation in the presence of side information about remote data. – The Annals of Statistics, 1990, v. 18, No. 1, p. 141–171.

5. *Han T. S., Amari S.* Parameter estimation with multiterminal data compression. – IEEE Trans. on Inform. Theory, 1965, v. IT–41, No. 6, p. 1802–1833.

6. *Ibragimov I. A., Has'minskii R. Z.* Statistical Estimation. Asymptotic Theory. Springer–Verlag, New York–Berlin–Heidelberg, 1981.

7. *Fano R. M.* Transmission of Information. A statistical theory of communication. MIT&Wiley, New York–London, 1961.

8. *Gallager R. G.* Information Theory and Reliable Communication. Wiley, New York–London–Sydney–Toronto, 1968.

9. *Ahlswede R., Althöfer I.* The asymptotic behavior of diameters in the average. – Journal of Combinatorial Theory, Series B, 1994, v. 61, No. 2, p. 167–177.