KYBERNETIKA — VOLUME 10 (1974), NUMBER 3

Data Compression in Discriminating Stochastic Processes*

ALBERT PEREZ

In discriminating stochastic processes there arises a need of observation data reduction concerning the length of the realization to be considered as well as the variety (alphabet) of the instantaneous process states to be identified. In the paper a method for such data compression is given based on the theory of asymptotic discernibility of two stationary random processes as developed by the author for processes with memory.

1. INTRODUCTION

Let $\{\xi_n, n = 0, \pm 1, \pm 2, \ldots\}$ be a sequence of abstract valued random variables representing in every "instant" n the state of a stochastic system evolving according to a stationary discrete-time random process.

Let either P or Q be the probability measure induced by the above sequence on the corresponding infinite product space generated by the (measurable) space-alphabet of values of the ξ_n's. In other words, the stochatic system above may evolve either according to the stationary probability law P or according to the stationary probability law Q.

Let us denote by H_P and H_Q the respective statistical hypotheses occurring with the a priori probabilities p and q, $p + q = 1$, provided that these probabilities exist. Note that if p and q are both positive, their exact values are irrelevant for the asymptotic behaviour of the probability of error $e_n(P, Q)$ in discriminating H_P and H_Q on the base of a growing number n of observed successive random variables of the above sequence. For the sake of simplicity it is, thus, possible to take in the sequel $p = q = 1/2$ and restrict us to the study of the maximum likelihood error probabilities $e_{Pn}(P, Q)$ and $e_{Qn}(P, Q)$ corresponding to the statistical hypotheses H_P and H_Q,

* This paper is an extended version of our communication presented at *the Second Seminar on Experimental Modelling and Solution of Probability Problems* — Liblice near Prague, May 28 — June 1, 1973.

respectively. If, now, p and q exist and $p = q = 1/2$, then the minimal error probability $e_n(P, Q)$ is given by the mean value of the maximum likelihood error probabilities.

In his 1952 paper [1] Herman Chernoff has determined the asymptotic rate of convergence to zero of the above error probabilities for the case of a sequence of mutually independent and identically distributed random variables ξ_n, i.e. under the assumption that P and Q are stationary memoryless random processes. If P_0 and Q_0 are their one-dimensional restrictions, it namely holds

$$(1.1) \qquad \lim_n \frac{1}{n} \log e_n(P, Q) = \lim_n \frac{1}{n} \log e_{Pn}(P, Q) = \\ = \lim_n \frac{1}{n} \log e_{Qn}(P, Q) = \log H_{\alpha_0}(P_0, Q_0),$$

where

$$(1.2) \qquad H_{\alpha_0}(P_0, Q_0) = \min_{0 \leqq \alpha \leqq 1} H_\alpha(P_0, Q_0),$$

and

$$(1.3) \qquad H_\alpha(P_0, Q_0) = \text{alpha-entropy of } P_0 \text{ with respect to } Q_0 = \\ = \int \left(\frac{\mathrm{d}P_0}{\mathrm{d}W}\right)^\alpha \left(\frac{\mathrm{d}Q_0}{\mathrm{d}W}\right)^{1-\alpha} \mathrm{d}W,$$

with W a measure dominating P_0 and Q_0. In the case $\alpha = 0$ (resp. $\alpha = 1$) it is necessary to consider the definition (1.3) as the limit for $\alpha \downarrow 0$ (resp. for $\alpha \uparrow 1$).

In the case the sequence of the ξ_n's above represents a Markov chain stationary and ergodic with finite state-space $(1, 2, \ldots, s)$ and with transition probabilities $\{p_{ij}\}$, $i, j = 1, 2, \ldots, s$, under the statistical hypothesis H_P and $\{q_{ij}\}$, $i, j = 1, 2, \ldots, s$, under the statistical hypothesis H_Q, Koopmans [2] derived that the limits in (1.1) exist and are equal to $\log \hat{r}_0$ where

$$(1.4) \qquad \hat{r}_0 = \inf_{0 < \alpha < 1} r_\alpha$$

with r_α = maximal eigenvalue of the matrix $(p_{ij}^\alpha q_{ij}^{1-\alpha})_{i,j=1,\ldots,s}$.

This result was proved by Koopmans under the assumption that all the p_{ij}'s and q_{ij}'s are positive. However, the weaker assumption of ergodicity is sufficient and may be further weakened on the base of our result in paper [4].

In our paper [3] are obtained a lower bound $\hat{r}_1$ and an upper bound $\hat{r}_2$ of $\hat{r}_0$, namely,

$$(1.5) \qquad \hat{r}_2 = \max_{1 \leqq i \leqq s} \min_{0 \leqq \alpha \leqq 1} H_\alpha(\{p_{ij}\}_{j=1,\ldots,s}, \{q_{ij}\}_{j=1,\ldots,s}) = \\ = \max_{1 \leqq i \leqq s} \min_{0 \leqq \alpha \leqq 1} H_\alpha(i),$$

$$(1.6) \qquad \hat{r}_1 = \min_{0 \leqq \alpha \leqq 1} \sum_{i=1}^{s} w_i^{(\alpha)} H_\alpha(i),$$

where $\{w_i^{(\alpha)}\}_{i=1,\ldots,s}$ is the stationary distribution of a Markov chain with transition probabilities of the type $r_{ij} = p_{ij}^{\alpha} q_{ij}^{1-\alpha} / H_{\alpha}(i)$.

In paper [4] we give a generalization of the above results to the case of P and Q stationary processes not necessarily memoryless (as in the Chernoff's case) or of the Markov type (as in the Koopmans case).

It concerns conditions for the validity of the following statement (throughout $e_n(P, Q)$ may be replaced by the maximum likelihood error probabilities $e_{P_n}(P, Q)$ and $e_{Q_n}(P, Q)$ as well as by their sum): The limits below exist and are equal,

$$\lim_n \frac{1}{n} \log e_n(P, Q) = \lim_n \frac{1}{n} \log H_{\alpha_n}(P_{0,n}, Q_{0,n}), \tag{1.7}$$

where $P_{0,n}$ and $Q_{0,n}$ are the n-dimensional restrictions of the stationary probability measures P and Q, respectively, and

$$H_{\alpha_n}(P_{0,n}, Q_{0,n}) = H_{\alpha_n}^{(n)} = \min_{0 \leqq \alpha \leqq 1} H_{\alpha}(P_{0,n}, Q_{0,n}), \quad n = 1, 2, \ldots. \tag{1.8}$$

The right-hand limit figuring in (1.7) is what we call minimal alpha-entropy rate of the random process P with respect to the random process Q. In the sequel it will be denoted by $h_a(P, Q)$. In the Chernoff's case this rate is, of course, equal to $\log H_{\alpha_0}(P_0, Q_0)$ (cf. (1.1)) and in the Markov case this rate is equal to $\log \hat{r}_0$ (cf. (1.4)). Due to Koopmans we, thus, have in the Markov case a closed procedure for the calculation of the minimal alpha-entropy rate.

The equality (1.7) holds for general abstract alphabet stationary processes P and Q provided that they satisfy some general condition GC introduced in [4]. For the sake of brevity, we give here a simpler condition implying the condition GC and, thus, the equality (1.7).

Let $p_P(F/x_{-n,0})$ and $p_Q(F/x_{-n,0})$ be the conditional probabilities corresponding to P and Q, respectively, that $\xi_1 \in F$ given that $\xi_0 = x_0, \xi_{-1} = x_{-1}, \ldots, \xi_{-n} = x_{-n}$, where by $x_{-n,0}$ we denote the sequence $x_0, x_{-1}, \ldots, x_{-n}$. We assume that:

(1) $p_P(\cdot/x_{-n,0})$ and $p_Q(\cdot/x_{-n,0})$ are regular, i.e. they represent probability measures on the one-dimensional σ-algebra of subsets F of the state-space (alphabet) X_1 of ξ_1 for every $x_{-n,0} \in X_{-n,0} = X_0 \times X_{-1} \times \ldots \times X_{-n}$, where the X_i's are all equal to A, the common alphabet of the ξ_i's;

(2)
$$\lim_n p_P(F/x_{-n,0}) = p_P(F/x_{-\infty,0}) \quad \text{and}$$
$$\lim_n p_Q(F/x_{-n,0}) = p_Q(F/x_{-\infty,0})$$

exist uniformly in logarithmic sense for every F and for every $x_{-\infty,0} \in X_{-\infty,0}$.

Assumptions (1) and (2) represent the condition mentioned above under which the statement contained in (1.7) holds.

It is obvious that for Markov processes of arbitrary order this condition is fulfiled.

As to the method used for proving the above statement (1.7), we restrict us to two basic inequalities. The first inequality represents only a slight generalization of the Chernoff's one. However, the second inequality is completely new and is derived by applying our generalized Shannon-McMillan's limit theorem for entropy densities (cf. [5]). These two basic inequalities are:

$$e_{Pn}(P, Q) + e_{Qn}(P, Q) \leqq H_{\alpha_n}(P_{0,n}, Q_{0,n}) \,, \tag{1.9}$$

$$\underline{\lim}_{n} \frac{1}{n} \log e_n(P, Q) \geqq \min \{-h(R, P), -h(R, Q)\} \,, \tag{1.10}$$

where R is an ergodic probability measure such that, for every $n = 1, 2, \ldots$, the n-dimensional restriction $R_{0,n}$ of R is dominated by $P_{0,n}$ and $Q_{0,n}$ and $h(R, P)$ and $h(R, Q)$ are the Shannon entropy rates of R with respect to P and to Q.

Let us recall that $h(R, P)$ is defined as follows:

$$h(R, P) = \lim_{n} \frac{1}{n} H(R_{0,n}, P_{0,n}) \tag{1.11}$$

where $H(R_{0,n}, P_{0,n})$ is the Shannon (generalized) entropy of $R_{0,n}$ with respect to $P_{0,n}$ given by

$$H(R_{0,n}, P_{0,n}) = \int \log \frac{\mathrm{d}R_{0,n}}{\mathrm{d}P_{0,n}} \mathrm{d}R_{0,n} \,. \tag{1.12}$$

Similarly is defined $h(R, Q)$.

2. DATA COMPRESSION

In discriminating two stochastic processes P and Q on the base of a growing number of observations there arises a need of data compression concerning the length of the realization to be considered, i.e. the number of the successive ξ_i's of the underlying random sequence (cf. Introduction) to be observed, as well as the alphabet A of their possible values or states to be identified, i.e. the accuracy with which the realized values (in general, abstract) of the observed ξ_i's are measured. Both these types of compression, imposed by the boundedness of the capabilities (memory, time, capacities, etc.) at our disposal, imply in general a loss of discernibility of the statistical hypotheses H_P and H_Q. In particular, they lead to an increment of the maximum likelihood error probabilities $e_{Pn}(P, Q)$ and $e_{Qn}(P, Q)$ as well as of the minimal error probability $e_n(P, Q)$ in the case it has a sense.

If, thus, such data compression are inevitable, our aim is to perform the compression conformly to the capabilities at our disposal in such a way that the loss of discernibility connected with as compared to the unreduced case (i.e. with respect to an ideal observer) to be minimal or, at least, admissible, provided that the admissibility criterion may be satisfied in the frame of the existing capabilities. In this context, the question arises how to balance the two types of data reduction mentioned above.

In order to be more definite, let us assume, for instance, that the ξ_i's are vector valued with components $u_{i,1}, u_{i,2}, \ldots, u_{i,m}$, and with alphabet $A = A_1 \times A_2 \times \ldots \ldots \times A_m$ so that the component $u_{i,1}$ takes its values on A_1, the component $u_{i,2}$ on $A_2, \ldots$, the component $u_{i,m}$ takes its values on A_m.

As said in the Introduction, the validity of the two conditions (1) and (2) concerning the two alternative probability laws P and Q of the stationary random sequence under consideration implies the validity of the statement contained in (1.7). Let $h_a(P, Q)$ be the minimal alpha-entropy rate of the random process P with respect to the random process Q (i.e. the limit figuring on the right-hand side of (1.7)) and suppose that $h_a(P, Q)$ is different from zero, i.e. strictly negative. According to (1.7), it asymptotically holds

$$e_n(P, Q) \cong \exp\left[nh_a(P, Q) + no(1)\right] = H_{\alpha_n}(P_{0,n}, Q_{0,n}), \tag{2.1}$$

where "$\cong$" may be always replaced by "$\leqq$" and $e_n(P, Q)$ by every of the maximum likelihood error probabilities $e_{Pn}(P, Q)$ or $e_{Qn}(P, Q)$ or by their sum $e_{Pn}(P, Q) + e_{Qn}(P, Q)$ (cf. (1.9)).

Let us now suppose that instead of observing in every "instant" $i = 1, 2, \ldots$, the value taken by the corresponding $\xi_i = (u_{i,1}, u_{i,2}, \ldots, u_{i,m})$, we restrict us to observe the values only of certain of its components by rejecting the others. Let, thus, denote by ξ_i^j the random variable resulting from ξ_i by rejecting the component $u_{i,j}$ ($j = 1, 2, \ldots, m$). Similarly, let us denote by ξ_j^{ki} the random variable resulting from ξ_i by rejecting both the components $u_{i,j}$ and $u_{i,k}$ ($j \neq k; j, k = 1, 2, \ldots, m$), and so on. The corresponding reduced alphabets will be denoted by A^j, A^{jk}, and so on. The corresponding restrictions of P and Q will be denoted by P^j, Q^j, P^{jk}, Q^{jk}, $P^{j_1 j_2 \ldots j_r}$, $Q^{j_1 j_2 \ldots j_r}$, the latter in the case of r rejected components, $r < m$; $j_1, j_2, \ldots, j_r$ (all different) $= 1, 2, \ldots, m$.

It is possible to see that if the conditions (1) and (2) above are verified for P and Q, the same holds for their restrictions $P^{j_1 \ldots j_r}$, $Q^{j_1 \ldots j_r}$ and, thus, the corresponding statement contained in (1.7) remains valid (cf. (2.1)):

$$\begin{aligned} e_n(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r}) &\cong \exp e\left[nh_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r}) + no(1)\right] = \\ &= H_{\alpha_n(j_1 \ldots j_r)}(P_{0,n}^{j_1 \ldots j_r}, Q_{0,n}^{j_1 \ldots j_r}), \end{aligned} \tag{2.2}$$

where $h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r})$ is the minimal alpha-entropy rate of the reduced process $P^{j_1 \ldots j_r}$ with respect to the reduced process $Q^{j_1 \ldots j_r}$ and $\alpha_n(j_1, \ldots, j_r)$ is the α minimizing the alpha-entropy of their n-dimensional restrictions.

Due to the concavity of the function z^α for z nonnegative and α fixed between 0 and 1, the α-entropy after reduction is greater than or equal to the α-entropy before reduction. As a consequence, the same is the case for the minimal alpha-entropies and, thus, also for the minimal alpha-entropy rates, i.e.

$$h_a(P, Q) \leqq h_a(P^{j_1}, Q^{j_1}) \leqq \ldots \leqq h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r}). \tag{2.3}$$

The case of equality, unfortunately only exceptional, is the more favorable since if it holds for some alphabet reduction, it is possible to obtain the same asymptotic discernibility of the statistical hypotheses H_P and H_Q, i.e. the same rate of convergence to zero of the error probabilities as before reduction. In the sequel we shall assume that between the first and the last member of (2.3) a strict inequality holds so that, in order to obtain (asymptotically) the same level of the error probability in discriminating between H_P and H_Q after reduction as before reduction, it will be necessary to observe a sequence of random variables $\{\xi_i^{j_1 \ldots j_r}\}_{i=1}^{i=n'}$ of length n' sufficiently greater than the length n of the sequence $\{\xi_i\}_{i=1}^{i=n}$ to be observed before reduction, namely,

$$n h_a(P, Q) \cong n' h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r}), \tag{2.4}$$

where we suppose that not only $h_a(P, Q)$ but also $h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r})$ is strictly negative, (cf. (2.1) and (2.2)).

3. COMPARISON OF TWO VERSIONS OF ALPHABET REDUCTION

Let us consider two versions of the decision problem of discriminating the two statistical hypotheses H_P and H_Q:

In the first version the discrimination is based on the observation of the sequence of random variables $\{\xi_i^{j_1 \ldots j_r}\}_{i=1}^{i=n'}$.

In the second version the discrimination is based on the observation of the sequence of random variables $\{\xi_i^{k_1 \ldots k_s}\}_{i=1}^{i=n''}$.

We shall assume that both versions are admissible from the point of view of our capabilities (cf. section 2) and, moreover, that the error probability level in discriminating between H_P and H_Q is in both versions the same, say, to that given by the left-hand member of (2.4). This in particular means that between n' and n'' the following relation holds:

$$\frac{n'}{n''} \cong \frac{h_a(P^{k_1 \ldots k_s}, Q^{k_1 \ldots k_s})}{h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r})}. \tag{3.1}$$

In order to compare these two versions of our decision problem, let us introduce a cost function including namely "costs" of identifying, memorising, processing and waiting connected with the observed sequence of random variables.

For the sake of simplicity, we shall assume that the costs are proportional to the sequence length, i.e.

$$(3.2) \qquad c(\{\xi_i^{j_1 \ldots j_r}\}_{i=1}^{i=n'}) = n' \, C(j_1, \ldots, j_r) ,$$

$$(3.3) \qquad c(\{\xi_i^{k_1 \ldots k_s}\}_{i=1}^{i=n''}) = n'' \, C(k_1, \ldots, k_s) ,$$

where $C(j_1, \ldots, j_r)$ and $C(k_1, \ldots, k_s)$ are respectively the costs corresponding to one random variable of the type $\xi_i^{j_1 \ldots j_r}$ or $\xi_i^{k_1 \ldots k_s}$. These costs will be supposed positive.

If the cost (3.2) is smaller than the cost (3.3), that is (taking account of (3.1)) if the inequality

$$(3.4) \qquad \frac{-h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r})}{C(j_1, \ldots, j_r)} > \frac{-h_a(P^{k_1 \ldots k_s}, Q^{k_1 \ldots k_s})}{C(k_1, \ldots, k_s)}$$

holds we shall prefer the first version of the decision problem. Otherwise, we shall prefer the second version.

It is natural to suppose that

$$(3.5) \qquad C(0) \geqq C(j_1) \geqq C(j_1, j_2) \geqq \ldots \geqq C(j_1, j_2, \ldots, j_{m-1}) ,$$

where by $C(0)$ we denote the cost corresponding to one unreduced random variable, i.e. of the type ξ_i. The sign of equality in (3.5) will be only exceptional (cf. (2.3)).

We repeat that the preference relation (3.4) concerns the comparison of two admissible (i.e. compatible with the capabilities at our disposal) versions of the decision problem with the same level of error probability. Under these constraints, our aim is to choose such version of the alphabet compression, i.e. to reject such set of components indexed by $(j_1, j_2, \ldots, j_r)$, which maximalizes the ratio figuring at the left-hand side of (3.4) denoted by $R(j_1, \ldots, j_r)$,

$$(3.6) \qquad R(j_1, \ldots, j_r) = \frac{-h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r})}{C(j_1, \ldots, j_r)} .$$

If the number m of components is relatively large (cf. section 4), the maximalization of $R(j_1, \ldots, j_r)$ by considering all the admissible versions of $(j_1, \ldots, j_r)$ may be practically impossible because of the extremely large number of these versions with respect to the computing capacities. This situation leads us to apply the following approximate method:

We consider, in the first step, all the admissible versions of the type (j_1), i.e. rejecting one component, and we definitely reject the component indexed by j_1^0 for which $R(j_1^0)$ is maximum. In the second step, we consider all the admissible versions of the type (j_1^0, j_2), i.e. rejecting the component indexed by j_1^0 and a further component, and we definitely reject a second component indexed by j_2^0 for which $R(j_1^0, j_2^0)$ is maximum ... In the r-th step, we consider all the admissible versions of the type $(j_1^0, j_2^0, \ldots, j_{r-1}^0, j_r)$, i.e. rejecting the components indexed by $j_1^0, j_2^0, \ldots, j_{r-1}^0$ and a further component, and we definitely reject a r-th component indexed by j_r^0 for which $R(j_1^0, j_2^0, \ldots, j_{r-1}^0, j_r^0)$ is maximum. We continue in this way up to the $r \leqq \leqq m - 1$ for which there exists at least one admissible version. Finally, we choose a version $(j_1^0, \ldots, j_{r_0}^0)$ for which $R(j_1^0, \ldots, j_{r_0}^0)$ is maximum in the set of all the R's obtained above including, eventually, $R(0)$ if the case with no compression is admissible too.

Analogue considerations may be applied if the alphabet A is finite, having say m points, instead of being of the Cartesian product type corresponding to m components as before; the rejection of components in the process of compression is here replaced by the fusion of points. However, we shall not consider in this paper this case. Also we shall not consider the case of compression, i.e. suitable finite partition, of more general alphabets.

4. SPECIAL CASES

For the sake of simplicity, we shall assume in the sequel that the cost function $C(j_1, \ldots, j_r)$ depends only on $r = 0, 1, \ldots, m - 1$, i.e.

$$C(j_1, \ldots, j_r) = K(r)\,, \tag{4.1}$$

where, according to (3.5), $K(r)$ is a positive decreasing function of the number of rejected components r.

The maximal admissible level of the logarithm of the error probability in discriminating the statistical hypotheses H_P and H_Q will be given in terms of $h_a(P, Q)$ (i.e. of the minimal alpha-entropy rate of the process P with respect to the process Q) by $nh_a(P, Q)$ (cf. (2.1) and (2.4)).

The maximal admissible observation delay, i.e. the maximal admissible length of the observed sequence of random variables will be denoted by N. Obviously, if N is smaller than n it is impossible to obtain an admissible level of the error probability (in an asymptotic sense, of course) provided that P and Q satisfy the assumptions (1) and (2) of the Introduction so that the statement contained in (1.7) holds, what is assumed throughout the paper.

Case 1. The components $u_{i,1}, u_{i,2}, \ldots, u_{i,m}$ of the vector valued random variables ξ_i are supposed to be mutually independent and equally distributed. The cost function $K(r)$ is assumed to be linear,

$$K(r) = k\,.\,(m - r)\,, \tag{4.2}$$

k being the cost corresponding to one component.

Let us denote by $h_a(j)$ the minimal alpha-entropy rate of P with respect to Q as restricted to have the alphabet A_j of the j-th component, i.e.

$$h_a(j) = h_a(P^{1\ldots j-1,j+1\ldots m}, Q^{1\ldots j-1,j+1\ldots m})\,. \tag{4.3}$$

Our assumption of the mutual independence of the components (the probability low being either P or Q) implies that

$$(4.4) \qquad h_a(P, Q) = h_a(1) + h_a(2) + \ldots + h_a(m).$$

Our second assumption that the components are equally distributed implies moreover that $h_a(1) = h_a(2) = \ldots = h_a(m)$, so that from (4.4) it follows that

$$(4.5) \qquad h_a(P, Q) = mh_a(1),$$
$$h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r}) = (m - r)\, h_a(1).$$

On the base of (4.2) and (4.5) we obtain (cf. (3.6))

$$(4.6) \qquad R(j_1, \ldots, j_r) = \frac{-(m - r)\, h_a(1)}{k\,.\,(m - r)} = \frac{-h_a(1)}{k},$$

whatever be the index set $(j_1, \ldots, j_r)$ of the rejected components.

However, from the admissibility point of view (delay bounded from above by N, level of the logarithm of the error probability bounded from above by $nh_a(P, Q)$), the number of rejected components is bounded from above by the inequality

$$(4.7) \qquad N\,.\,(m - r)\,.\,h_a(1) \leqq nh_a(P, Q) = nmh_a(1),$$

(note that, by assumption, $h_a(P, Q)$ and, thus, also $h_a(1)$ are strictly negative), i.e.

$$(4.8) \qquad r \leqq m\,.\left(1 - \frac{n}{N}\right).$$

Case 2. As in Case 1, the components $u_{i,1}, u_{i,2}, \ldots, u_{i,m}$ are supposed to be mutually independent (both with respect to P and to Q) but not necessarily equally distributed. The cost function $K(r)$ is again assumed to be of the type (4.2).

It is not a restriction to assume that

$$(4.9) \qquad h_a(1) \leqq h_a(2) \leqq \ldots \leqq h_a(m).$$

Since, by the independence hypothesis, (4.4) remains valid and, moreover, for $r = 0, 1, \ldots, m - 1$, the equality

$$(4.10) \qquad h_a(P^{j_1 \ldots j_r}, Q^{j_1 \ldots j_r}) = h_a(j_{r+1}) + \ldots + h_a(j_m)$$

holds, it follows that

$$(4.11) \qquad R(j_1, \ldots, j_r) = -\frac{h_a(j_{r+1}) + \ldots + h_a(j_m)}{k\,.\,(m - r)}.$$

 On the base of (4.9) and (4.11) one obtains that

$$\max_{(j_1,\ldots,j_r)} R(j_1, \ldots, j_r) = R(m - r + 1, m - r + 2, \ldots, m) \tag{4.12}$$

$$= - \frac{h_a(1) + \ldots + h_a(m - r)}{k \,.\, (m - r)}\,.$$

Obviously, $R(m - r + 1), m - r + 2, \ldots, m)$ is an increasing (non-decreasing) function of r. Thus, its absolute maximum is obtained for r maximum, i.e. for $r = m - 1$, and equals $-h_a(1)/k$. It is obtained by rejecting all the components except the first one.

However, for the same admissibility reasons as in Case 1, the number of rejected components is bounded from above by the inequality

$$N \,.\, (h_a(1) + \ldots + h_a(m - r)) \leqq nh_a(P, Q)\,. \tag{4.13}$$

If r_0 is the maximum r satisfying (4.13), the optimal admissible reduction is obtained by rejecting all the components corresponding to the index set $(m - r_0 + 1, m - r_0 + 2, \ldots, m)$, the indexing being that satisfying (4.9). This results from the fact that $R(m - r + 1, \ldots, m)$ is an increasing (non-decreasing) function of r and that by using a delay (sequence length) smaller than N (greater than N is not admissible) the maximum r satisfying the corresponding inequality (4.13) will be smaller than or equal to r_0. The corresponding inequality (4.13) where N is replaced by the smaller delay must be satisfied as before in order to ensure an admissible level of error probability. This proves the optimality of the alphabet reduction above. It must be combined with the observation of a sequence of length N.

Obviously, in the Case 1 the inequality (4.13) reduces to (4.7) or, what is the same, to (4.8). Let $r(N)$ be the maximum r satisfying (4.8) and let $r(n')$ be the maximum r satisfying the analogue of (4.8) when we replace N by n'. It is possible to see that (asymptotically at least) $n'(m - r(n'))$ equals to nm for any $n' \geqq n$. As a consequence, the total cost $n'(m - r(n')) \,.\, k$ of discriminating the statistical hypotheses H_P and H_Q on the base of a sequence of n' random variables resulting from the initial ones by rejecting $r(n')$ components (maximal admissible alphabet reduction) does not depend on n' and equals to nmk, i.e. the total cost corresponding to the discrimination on the base of a sequence of n unreduced random variables. Thus, from the cost point of view, all the versions of data reduction of the above type $(n', r(n'))$ are in the Case 1 equivalent. They are admissible for $n' \leqq N$ and optimal.

Remark 1. If the cost function $C(j_1, \ldots, j_r)$ is additive but more general than of the type $K(r) = k \,.\, (m - r)$ as before, namely, if

$$C(j_1, j_2, \ldots, j_r) = k(j_{r+1}) + k(j_{r+2}) + \ldots + k(j_m)\,, \tag{4.14}$$

where $k(j)$ is the cost corresponding to the j-th component, $j = 1, 2, \ldots, m$, and if the indexing is such that

$$k(1) \leqq k(2) \leqq \ldots \leqq k(m)\,, \tag{4.15}$$

then, in the Case 1, the optimal version of data reduction is obtained by rejecting the greatest possible number of components in the order $m, m - 1, m - 2, \ldots$, of decreasing (non-increasing) cost (cf. (4.15)), i.e. this version will be of the type $(N, r(N))$.

Remark 2. In the Case 2 but with cost function of the type (4.14), the optimal version of compression remains, obviously, the same as for a cost function of the type $K(r) = k\,.\,(m - r)$ provided that, for the same indexing, (4.9) and (4.15) hold simultaneously.

Remark 3. In the general case where $R(j_1, \ldots, j_r)$ is given by (3.6), if the number m of components is relatively large, the maximization of $R(j_1, \ldots, j_r)$ by considering all the admissible versions of $(j_1, \ldots, j_r)$ becomes practically impossible as compared with the computing capacities at our disposal. Indeed, the total number of these versions in passing from m components to $m - r$ components is given by

$$W_{m,m-r} = \binom{m}{m - r}\,. \tag{4.16}$$

Moreover, since the optimal admissible number, r_0, of rejected components is a priori unknown, it will be, in general, necessary to test the situation for $r = 1, 2, \ldots$ $\ldots, r'$, where r' may attain the value $m - 1$. Thus, in the exhaustive case the total number of alternatives to be considered in the process of optimization will be of the order

$$Z_{m,m-r'} = W_{m,m-1} + W_{m,m-2} + \ldots + W_{m,m-r'}\,. \tag{4.17}$$

As said in Section 3, this situation leads us to proceed approximately by applying non-exhaustive methods as that described there. The total number of alternatives to be considered in this case is of the order

$$Q_{m,m-r'} = \left(m - \frac{r' - 1}{2}\right) r'\,, \tag{4.18}$$

where r' has the same meaning as before.

Paper [6] studies in more detail this question of comparison of the numbers of alternatives to be considered in the exhaustive and non-exhaustive case, and the analogue question arising in reducing finite alphabets (cf. end of section 3).

(Received December 27, 1973.)

[1] H. Chernoff: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Statistics *23* (1952), 493—507.

[2] L. H. Koopmans: Asymptotic rate of discrimination for Markov processes. Ann. Math. Statistics *31*, (1960), 982—994.

[3] A. Perez: Asymptotic discernibility of two stationary Markov chains. Paper presented at the Second International Symposium on Information Theory in USSR, Tsahkadsor, 1971.

[4] A. Perez: Generalization of the Chernoff's result on the asymptotic discernibility of two random processes. Paper presented at the 1972 European Meeting of Statisticians, Budapest.

[5] A. Perez: Extensions of Shannon-McMillan's limit theorem to more general stochastic processes. In: Trans. Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes 1962. Praha 1964, 545—574.

[6] L. Baladová: Number of alternatives in reducing finite spaces and vector spaces. To appear in Kybernetika.

Dr. Albert Perez, DrSc.; Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 180 76 Praha 8. Czechoslovakia.