SIDE INFORMATION FOR THE DISCRIMINATION OF SIMPLE HYPOTHESES (*)

by MARIA ASSUNTA CISEK and ANDREA SGARRO (in Trieste) (**)

SOMMARIO. — *Si considera un problema di discriminazione statistica subordinata. Si effettuano delle osservazioni congiunte* (x_i, y_i); *le osservazioni* x_i *sono usate ai fini della discriminazione solo come informazione collaterale. Si paragonano i risultati con quelli della teoria classica di Neyman e Pearson. Il lavoro vuol essere un contributo alla costruzione di una teoria delle decisioni a più terminali.*

SUMMARY. — *A problem of conditional hypothesis discrimination is considered. A sample of joint observations* (x_i, y_i) *is taken. The* x_i*-observations are used only as side information for discrimination. Comparisons are made with standard Neyman-Pearson theory. This paper is meant as a contribution towards a theory of multi-terminal decision making.*

1. Introduction.

In this paper we consider a problem of statistical hypothesis discrimination over product sets. Assume that n joint observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) are taken independently from a finite product set $\mathcal{X} \times \mathcal{Y}$; for example, x_i and y_i might be correlated attributes of the i-th item of a certain lot. The decision-maker knows that either (P, V) or (Q, W) is the true probability distribution ruling the random behaviour of the sample (above the pro-

(*) Pervenuto in Redazione il 26 febbraio 1981.
Lavoro eseguito nell'ambito del Gruppo Nazionale per l'Informatica Matematica del C.N.R.

(**) Indirizzo degli Autori: Istituto di Matematica dell'Università - Piazzale Europa 1 - 34100 Trieste.

bability vectors P and Q correspond to marginal distribution over $\mathcal{X}$ while the stochastic matrices V and W correspond to conditional distributions over $\mathcal{Y}$). The n observations are used to discriminate the simple hypotheses (P, V) and (Q, W).

A classical solution is the following: after fixing an allowed first error probability α, $0 < \alpha < 1$, a dichotomy $(\mathcal{A}_n, \mathcal{B}_n)$ of $\mathcal{X}^n \times \mathcal{Y}^n$ is constructed according to Neyman-Pearson criterion with $Prob_{(P,V)}(\mathcal{B}_n) \leqslant \alpha$; (P, V) is chosen if (x^n, y^n) belongs to $\mathcal{A}_n$; else (Q, W) is chosen $(x^n = x_1 x_2 \ldots x_n,\ y^n = y_1 y_2 \ldots y_n)$. It is known that this decision rule is optimal in the sense that the second error probability $Prob_{(Q,W)}(\mathcal{A}_n)$ is minimal.

So far no use of the product structure of the observations has been made; instead, in our model such a structure is explicitely used. First, we assume that $P = Q$, so that the (unconditional) behaviour of the $\mathcal{X}$-observations is perfectly known; by themselves they give no information for discrimination. Second, we assume that some sort of processing is needed to obtain the $\mathcal{Y}$-observations so that, at first, only the non-informative marginal sample $x^n = x_1 x_2 \ldots x_n$ is known to the decision-maker. He might now argue that the probability behaviour of the $\mathcal{Y}$-observations is ruled either by the conditional distribution $V^n(\cdot \mid x^n)$ or by the conditional distribution $W^n(\cdot \mid x^n)$; therefore he might prefer to use Neyman-Pearson criterion for discriminating the conditional distributions over $\mathcal{Y}^n$ rather than the joint distributions over $\mathcal{X}^n \times \mathcal{Y}^n$. In this case he will construct a (conditionally optimal) dichotomy $(\mathcal{A}_n, \mathcal{B}_n)$ of $\mathcal{Y}^n$, $V^n(\mathcal{B}_n \mid x^n) \leqslant \alpha$, with the decision rule: choose (P, V) if y^n belongs to $\mathcal{A}_n$, else choose (P, W). (We remark that the use of an «objective) criterion such as the one of Neyman and Pearson is consistent, at least within the bounds of classical statistics, because the marginal observation x^n is non-informative by itself.)

Of course the second error probability is a function of the $\mathcal{X}$-observations; in order to assess the overall effectiveness of our conditional decision rule it is more appropriate to consider this error probability averaged with respect to the probability of the $\mathcal{X}$-observations. In this paper we give an asymptotic evaluation of the average second error probability.

Note that $x^n = x_1 x_2 \ldots x_n$ is used somehow as *side information* for discriminating (P, V) and (P, W). The notion of side information has inspired much research work in probabilistic information theory (cf. [1] where this notion was first introduced in a source coding problem, or [2]). Since the problems of source coding and hypothesis discrimination are deeply related (cf., e.g., [3] or [2]), it is to be hoped that side information may have to play an

important role also in statistics. More generally there might arise a theory of *multi-terminal decision-making*, as there is now a theory of multi-terminal information theory. This paper is meant as a (very small) contribution in this direction.

In section 2 we give some technical preliminaries and notations. In section 3 the problem is formally described. In section 4 we prove our main result. Its consequences are discussed in section 5; comparisons are made with a standard use of Neyman-Pearson theory for discriminating (P, V) and (P, W) or for discriminating the corresponding marginal probability distributions over $\mathcal{Y}$ when no side information is available.

2. Preliminaries and notations.

Let $\mathcal{X} = \{a_1, a_2, ..., a_K\}$ and $\mathcal{Y} = \{b_1, b_2, ..., b_H\}$ be two alphabets, that is two non-empty finite sets. Probability distributions (p.d.'s) over $\mathcal{X}$ and over $\mathcal{Y}$ will be identified with the corresponding probability vectors. We shall consider also stochastic matrices with rows indexed in $\mathcal{X}$ and columns indexed in $\mathcal{Y}$. Such matrices will be denoted by symbols like $V : \mathcal{X} \to \mathcal{Y}$; the entry corresponding to row a_i and column b_j will be denoted by $V(b_j \mid a_i)$. A p.d. P over $\mathcal{X}$ and a stochastic matrix $V : \mathcal{X} \to \mathcal{Y}$ identify a joint p.d. (P, V) over the Cartesian product $\mathcal{X} \times \mathcal{Y}$; the corresponding marginal p.d. over $\mathcal{Y}$ will be denoted by PV.

We need also stationary memoryless extensions of p.d.'s and of stochastic matrices. For example, if $P = (p_1, p_2, ..., p_K)$ is a p.d. over $\mathcal{X}$, P^n is the p.d. over the Cartesian power $\mathcal{X}^n$ defined by

$$P^n(a_{j_1} a_{j_2} \dots a_{j_n}) = \prod_{i=1}^{n} P(a_{j_i}) = \prod_{i=1}^{n} p_{j_i}, \quad a_{j_1} a_{j_2} \dots a_{i_n} \in \mathcal{X}^n.$$

If $\mathcal{C}_n$ is a subset of $\mathcal{X}^n$, $P^n(\mathcal{C}_n)$ means $\sum_{x^n \in C^n} P^n(x^n)$.

In the language of information theory, $\mathcal{X}^n$ and $\mathcal{Y}^n$ would be called the input alphabet and the output alphabet, P^n and $(PV)^n$ the input p.d. and the output p.d. and V^n the channel connecting input and output.

A *dichotomy* $(\mathcal{A}, \mathcal{B})$ of a set $\mathcal{X}$ is a partition of $\mathcal{X}$: $\mathcal{X} = \mathcal{A} \cup \mathcal{B}$, $\mathcal{A} \cap \mathcal{B} = \varnothing$; $|\mathcal{C}|$ denotes the cardinality of a set $\mathcal{C}$.

The *divergence* of two p.d.'s $P = (p_1, p_2, ..., p_K)$ and $Q = (q_1, q_2, ..., q_K)$ is defined as

$$D(P \,||\, Q) = \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i}$$

if P is absolutely continuous with respect to Q; else set $D(P \,||\, Q)$

equal to $+\infty$ (logs and exps are taken, e.g., to the base 2). For the properties of the divergence we refer, e.g., to [2]; here we recall only that $D(P \,||\, Q) \geqslant 0$, with equality iff $P = Q$.

If P is a p.d. over $\mathfrak{X}$ and V and W are stochastic matrices $\mathfrak{X} \to \mathcal{Y}$, we shall use also the *average divergence*

$$D(V \,||\, W \,|\, P) = \sum_{i=1}^{K} p_i D(V_i \,||\, W_i)$$

V_i and W_i being the i-th row of V and of W, respectively.

Consider the following equivalence over $\mathfrak{X}^n$: two sequences are equivalent iff they are a permutation of one another. The corresponding equivalence classes will be called *types* of order n. A type is identified by the numbers $N(a \,|\, x^n)$ where x^n is any sequence in the type and $N(a \,|\, x^n)$ is the number of occurrences of letter a in sequence x^n. By dividing these numbers by the sequence length n, one obtains a p.d. over $\mathfrak{X}$, $T_n = \{n^{-1} N(a_1 \,|\, x^n), n^{-1} N(a_2 \,|\, x^n), \ldots, n^{-1} N(a_K \,|\, x^n)\}$. Also such p.d.'s will be referred to as types and the same symbols will be used as for sets; this will cause no misunderstanding.

If T_n is a type in $\mathfrak{X}^n$ and P is a p.d. over $\mathfrak{X}$, one has:

$$(1) \quad (n+1)^{-|\mathfrak{X}|} \exp\{-n D(T_n \,||\, P)\} \leqslant P^n(T_n) \leqslant \exp\{-n D(T_n \,||\, P)\}$$

Let γ be a positive real number. A sequence x^n in $\mathfrak{X}^n$ is called a P-typical sequence of constant γ if

$$|N(a_i \,|\, x^n) - n p_i| \leqslant n^{3/4} \gamma, \quad 1 \leqslant i \leqslant K.$$

Clearly the set of P-typical sequences of constant γ is a union of types. (For details about typicality cf. [2].)

3. The problem stated.

Let $P = (p_1, p_2, \ldots, p_K)$ be a p.d. over $\mathfrak{X}$ and V and W two stochastic matrices $\mathfrak{X} \to \mathcal{Y}$ (equivalently, let (P, V) and (P, W) be two p.d.'s over $\mathfrak{X} \times \mathcal{Y}$).

We want to discriminate the two hypotheses $H_V =$ «(P, V) is the true p.d.» and $H_W =$ «(P, W) is the true p.d.»; this amounts to discriminating the channels V and W. We shall not follow the standard Neyman-Pearson approach. Rather, as explained in the introduction, we assume that the $\mathfrak{X}^n$-observation is available before the $\mathcal{Y}^n$-observation. Suppose $x^n = x_1 x_2 \ldots x_n$ is observed. At this step the random behaviour of the $\mathcal{Y}^n$-observation is ruled either by the conditional p.d. $V^n(\cdot \,|\, x^n)$ or by the conditional p.d.

$W^n(\cdot \mid x^n)$, according whether H_V or H_W is true. We now choose a dichotomy $(\mathcal{A}_n, \mathrm{B}_n)$ of $\mathcal{Y}^n$, $\mathcal{A}_n = \mathcal{A}_n(x^n)$, $\mathcal{B}_n = \mathcal{B}_n(x^n)$; if the $\mathcal{Y}^n$-observation belongs to $\mathcal{A}_n$ we decide that H_V is true, otherwise we decide that H_W is true. The dichotomy of $\mathcal{Y}^n$ will be chosen using the criterion of Neyman and Pearson for discriminating $V^n(\cdot \mid x^n)$ and $W^n(\cdot \mid x^n)$. Namely, let α be a real number, $0 < \alpha < 1$. $(\mathcal{A}_n, \mathcal{B}_n)$ is any dichotomy verifying the following requirements:

i) $V^n(\mathcal{A}_n \mid x^n) \geqslant 1 - \alpha$;

ii) $W^n(\mathcal{A}_n \mid x^n) = \min W^n(\mathcal{C}_n \mid x^n)$, the minimum being taken with respect to all subsets $\mathcal{C}_n$ of $\mathcal{Y}^n$ which verify requirement *i.*

For reasons explained in the introduction, the following average error probability is of concern to us:

$$P_e(n) = P_e(n, P, V, W) = \sum_{x^n \in \mathcal{X}^n} P^n(x^n)\, W^n(\mathcal{A}_n \mid x^n).$$

In section 4 we shall investigate the asymptotic behaviour of $P_e(n)$.

4. Result.

In order to prove the theorem below we use as a lemma the following known fact (for a proof cf., e.g., chapter 1 of [2]).

LEMMA. *Let* $\{E_i\}_{i \geqslant 1}$ *and* $\{F_i\}_{i \geqslant 1}$ *be two sequences of p.d.'s over the alphabet* $\mathcal{Y}$. *Let* $E^{(n)}$ *and* $F^{(n)}$ *be the product p.d.'s over* $\mathcal{Y}^n$ *defined by* $E^{(n)} = \prod_{i=1}^{n} E_i$ *and* $F^{(n)} = \prod_{i=1}^{n} F_i$. *Set* $s(n, \alpha) = \min F^{(n)}(\mathcal{C}_n)$, *the minimum being taken with respect to all subsets* $\mathcal{C}_n$ *of* $\mathcal{Y}^n$ *such that* $E^{(n)}(\mathcal{C}_n) \geqslant 1 - \alpha$. *Assume that* $|\log F_i(b_j)|$ *is bounded by a constant* γ *independent of* i *and* b_j. *Then, for all* δ, $0 < \delta < 1$, *there exists an* $n_0 = n_0(|\mathcal{Y}|, \gamma, \alpha, \delta)$ *such that, for* $n \geqslant n_0$

$$|\log s(n, \alpha) + D(E^{(n)} \| F^{(n)})| \leqslant n\delta.$$

(Note that $\{E_i\}_{i \geqslant 1}$ and $\{F_i\}_{i \geqslant 1}$ identify two memoryless non stationary stochastic processes over $\mathcal{Y}$.)

We stress that n_0 above depends on $\{E_i\}_{i \geqslant 1}$ and $\{F_i\}_{i \geqslant 1}$ only through $|\mathcal{Y}|$ and γ.

THEOREM. *If the entries of* W *are all strictly positive*

$$\lim_n n^{-1} \log P_e(n) = -\min_R D(R, V \| P, W)$$

the minimum being taken with respect to all p.d.'s R over $\mathfrak{X}$.
Proof.

Fix δ, $0 < \delta < 1$. Using the lemma, for any $x^n, x^n \in \mathfrak{X}^n$, one has:

$$(2) \qquad |n^{-1} \log W^n(\mathfrak{A}_n \mid x^n) + D(V \| W \mid T_n)| \leqslant \frac{\delta}{2}$$

if $n \geqslant n_0 = n_0(|\mathfrak{Y}| \mid \gamma, \alpha, \delta)$; above $\gamma = \max_{(a,b)\in\mathfrak{X}\times\mathfrak{Y}} \log|W(b \mid a)| < +\infty$ and T_n is the type of x^n. We have used the fact that

$$D(V^n(\cdot \mid x^n) \| W^n(\cdot \mid x^n)) = \sum_{i=1}^{n} D(V(\cdot \mid x_i) \| W(\cdot \mid x_i)) =$$
$$= \sum_{a\in\mathfrak{X}} N(a \mid x^n) D(V(\cdot \mid a) \| W(\cdot \mid a)) = n D(V \| W \mid T_n).$$

If $\mathfrak{T}_n$ denotes the set of types of order n, one has:

$$P_e(n) = \sum_{x^n\in\mathfrak{X}^n} P^n(x^n) W^n(\mathfrak{A}_n \mid x^n) =$$
$$= \sum_{T_n\in\mathfrak{T}_n} \sum_{x^n\in T_n} P^n(x^n) W^n(\mathfrak{A}_n \mid x^n) \leqslant$$
$$\overset{(a)}{\leqslant} \sum_{T_n\in\mathfrak{T}_n} P^n(T_n) \exp\{-n[D(V \| W \mid T_n) - \frac{\delta}{2}]\} \leqslant$$
$$\overset{(b)}{\leqslant} \sum_{T_n\in\mathfrak{T}_n} \exp\{-n[D(T_n \| P) + D(V \| W \mid T_n) - \frac{\delta}{2}]\}$$

for $n \geqslant n_0$, n_0 not depending on T_n. Above (a) follows from (2) and (b) follows from (1). Since $|\mathfrak{T}_n|$ is trivially bounded by $(n+1)^{|\mathfrak{X}|}$, one has:

$$(3) \qquad P_e(n) \leqslant (n+1)^{|\mathfrak{X}|} \max_{T_n} \exp\{-n[D(T_n\|P) + D(V\|W|T_n) - \frac{\delta}{2}]\} =$$
$$= (n+1)^{|\mathfrak{X}|} \exp\{-n[\min_{T_n} D(T_n, V \| P, W) - \frac{\delta}{2}]\}.$$

(An easy computation shows that $D(R \| P) + D(V \| W \mid R) = D(R, V \| P, W)$ for any *p.d.* R over $\mathfrak{X}$.)

A lower bound for $P_e(n)$ is obtained much in the same way. One has, for $n \geqslant n_0$:

$$(4) \quad P_e(n) \overset{(c)}{\geqslant} \sum_{T_n\in\mathfrak{T}_n} P^n(T_n) \exp\{-n[D(V \| W \mid T_n) + \frac{\delta}{2}]\} \geqslant$$
$$\overset{(d)}{\geqslant} \sum_{T_n\in\mathfrak{T}_n} (n+1)^{-|\mathfrak{X}|} \exp\{-n[D(T_n, V \| P, W) + \frac{\delta}{2}]\} \geqslant$$
$$\geqslant (n+1)^{-|\mathfrak{X}|} \exp\{-n[\min_{T_n} D(T_n, V \| P, W) + \frac{\delta}{2}]\}.$$

Above (c) follows from (2) and (d) follows from (1). We now put together (4) and (5). Using standard approximation arguments (we recall that types are dense in the set of all p.d.'s) one obtains, for n large enough:

$$(n+1)^{-|\mathcal{X}|} \exp\{-n[\min_R D(R, V \| P, W) + \delta]\} \leqslant P_e(n) \leqslant$$

$$\leqslant (n+1)^{|\mathcal{X}|} \exp\{-n[\min_R D(R, V \| P, W) - \delta]\}.$$

This gives soon the theorem, because δ is arbitrary and

$$\lim n^{-1} \log (n+1)^{|\mathcal{X}|} = 0.$$

QED

5. **Final remarks.**

When H_V and H_W are discriminated using Neyman-Pearson criterion for the product set $\mathcal{X}^n \times \mathcal{Y}^n$ one obtains the optimal exponent $D(P, V \| P, W)$ (cf., e.g., [2]), which obviously can be strictly greater than $\min_R D(R, V \| P, W)$. This is no surprise since our partitions of $\mathcal{Y}^n$ induce partitions over $\mathcal{X}^n \times \mathcal{Y}^n$ which are constrained to have the product structure

$$\left(\bigcup_{x^n \in \mathcal{X}^n} \{x^n\} \times \mathcal{A}_n(x^n), \quad \bigcup_{x^n \in \mathcal{X}^n} \{x^n\} \times \mathcal{B}_n(x^n) \right)$$

From a different standpoint it can be of some interest to com pare our exponent with the case obtained using Neyman-Pearson criterion to discriminate the marginal distributions over $\mathcal{Y}^n$, that is with $D(P\,V \| P\,W)$. Below we give an example where $D(P\,V \| P\,W)$ is strictly smaller than $\min_R D(R, V \| P, W)$: in a way, in this case the side information x^n is actually useful for discriminating $P\,V$ and $P\,W$.

Example. Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, $P = \left(\frac{1}{2}, \frac{1}{2}\right)$, $V_\varepsilon(\cdot \mid 0) = W_\varepsilon(\cdot \mid 1) = (1-\varepsilon, \varepsilon)$, $V_\varepsilon(\cdot \mid 1) = W_\varepsilon(\cdot \mid 0) = (\varepsilon, 1-\varepsilon)$. Then $D(P\,V \| P\,W) = 0$ while $\lim_{\varepsilon \to 0} D(R, V_\varepsilon \| P, W_\varepsilon) = +\infty$ uniformly in R.

It is interesting to observe that by a «slight» modification of our decision procedure our exponent can be improved as to reach the asymptotical theoretical maximum, $D(P, V \| P, W)$. Consider the set $\mathcal{T}_n$ of $\mathcal{X}^n$-sequences which are P-typical in composition for a chosen constant. It is well-known that the probability of this set goes to 1 when n goes to infinity (cf. [2]). If the side information x^n belongs to $\mathcal{T}_n$ we do not modify our decision rule. If however

x^n does not belong to $\mathcal{T}_n$ we choose the dichotomy $\mathcal{A}_n = \varnothing$, $\mathcal{B}_n = \mathcal{Y}^n$, that is we decide that H_W is true whatever the $\mathcal{Y}^n$-observation. By using the standard properties of P-typical sequences it can be shown that the modified error exponent $-n^{-1} \log P_e^*(n)$ goes to $D(P, V \| P, W)$ as n goes to infinity (for details see [4]).

Some questions were left open. The theorem was proved on the assumption that all the entries of W be strictly positive: this is a rather ad hoc condition, depending only on our proving technique. Moreover it would be interesting to know whether the inequality $\min_R D(R, V \| P, W) \geqslant D(P V \| P W)$ is generally true (cf. the example above). This would much help in assessing the effectiveness of our decision procedure.

REFERENCES

[1] D. SLEPIAN and J. K. WOLF, *Noiseless coding of correlated information sources.* IEEE Transactions on Information Theory (IT 19) 4 (1973) 471-480.

[2] I. CSISZÁR and J. KÖRNER, *Information Theory,* Academic Press (1981).

[3] G. LONGO and A. SGARRO, *The error exponent for the testing of simple statistical hypotheses: a combinatorial approach.* Journal of Combinatorics, Information & System Sciences (5) 1 (1980) 58-67.

[4] M. A. CISEK, *L'informazione collaterale nella discriminazione di ipotesi statistiche.* Thesis, Istituto di Matematica, Università degli Studi di Trieste (1980).