On the Possible Orderings in the Measurement Selection Problem

THOMAS M. COVER, FELLOW, IEEE, AND JAN M. VAN CAMPENHOUT

Abstract—**An aspect of the measurement selection problem—the existence of anomalous orderings on the probability of error obtained by selected subsets of measurements—is discussed. It is shown that for any ordering on the probability of error as a function of the subset of measurements (subject to an obvious set monotonicity condition), there exists a multivariate normal two-hypothesis problem $N(\mu,K)$ versus $N(-\mu,K)$ that exhibits this ordering. Thus no known nonexhaustive sequential k-measurement selection procedure is optimal, even for jointly normal measurements.**

I. Introduction

IN THE APPLICATION of classification or hypothesis testing, one is sometimes concerned with the problem of finding the best k-element subset of n measurements.

Suppose, for example, that measurements M_1,M_2,M_3 are available. The question arises as to whether such orderings on the probability of error are possible as

$$P_e(M_1) > P_e(M_2) > P_e(M_3) > P_e(M_2,M_3) > P_e(M_1,M_3) > P_e(M_1,M_2) > P_e(M_1,M_2,M_3).$$

It is well known that statistical dependence among the measurements can cause the best k-element subset not to be composed of the individually best measurements. Even conditionally independent measurements can exhibit such an anomalous behavior [1]–[3].

We will show that essentially all possible probability of error orderings can occur among subsets of n measurements, subject to a monotonicity constraint. This work characterizes all probability orderings and thus extends [1]–[3] from consideration of single-element subsets and their relations to the best k-element subset, to the mutual relationships of all subsets.

II. Preliminaries and Notation

Let the item upon which the measurements are performed belong to one of two classes, indexed by θ and labeled $\theta_1 = +1$ and $\theta_2 = -1$, which occur with the same prior probability 1/2. Let Ω denote the set $\{1, 2, \cdots, n\}$, and let S be a nonempty subset of Ω. Let $|S|$ denote the cardinality of S. Let $X_S = (X_{i_1}, X_{i_2}, \cdots, X_{i_{|S|}})$ be the ordered vector of random variables $X_i, i \in S, i_1 < i_2 < \cdots < i_{|S|}$. The minimal probability of error (Bayes risk) $P_e(S)$ in guessing the class θ using X_S is defined by

$$^{n}{}_e(S) = \Pr\{\theta^*(X_S) \neq \theta \,|\, X_S\} = \int \tfrac{1}{2} \min\{f_{\theta_1}(x_S), f_{\theta_2}(x_S)\}\, dx_S. \tag{1}$$

Here, $f_{\theta_i}(x_S)$, $i = 1,2$, is the class conditional density of X_S, and $\theta^*(x_S)$ is the Bayes decision based on the observed value of X_S:

$$\theta^*(x_S) = \begin{cases} \theta_1, & \text{if } f_{\theta_1}(x_S) \geq f_{\theta_2}(x_S) \\ \theta_2, & \text{if } f_{\theta_1}(x_S) < f_{\theta_2}(x_S). \end{cases} \tag{2}$$

If the densities $f_{\theta_i}(x_S)$ are normal with a common covariance matrix K_S and mean $\theta_i \mu_S$, i.e., if

$$f_{\theta_i}(x_S) = \phi(\theta_i \mu_S, K_S) = \frac{1}{(2\pi)^{|S|/2}\,|K_S|^{1/2}} \exp\{-\tfrac{1}{2}(x_S - \theta_i \mu_S)^t K_S^{-1}(x_S - \theta_i \mu_S)\}, \tag{3}$$

then $P_e(S)$ can be written

$$P_e(S) = \Pr\{f_{\theta_1}(X_S) \geq f_{\theta_2}(X_S),\ \theta = \theta_2\} + \Pr\{f_{\theta_1}(X_S) < f_{\theta_2}(X_S),\ \theta = \theta_1\}. \tag{4}$$

Since $\theta_1 = 1$, $\theta_2 = -1$, this reduces to

$$P_e(S) = \Pr\{\mu_S^t K_S^{-1} X_S < 0 \,|\, \theta = \theta_1\}. \tag{5}$$

Since $\mu_S^t K_S^{-1} X_S$ is a linear function of X_S, it is normally distributed with parameters

$$E(\mu_S^t K_S^{-1} X_S \,|\, \theta) = \mu_S^t K_S^{-1} \mu_S \theta$$

and

$$\operatorname{var}(\mu_S^t K_S^{-1} X_S \,|\, \theta) = \mu_S^t K_S^{-1} \mu_S. \tag{6}$$

Hence, $P_e(S) = \Phi(-(\mu_S^t K_S^{-1} \mu_S)^{1/2})$, where Φ is the standard cumulative normal distribution

$$\Phi(t) = (2\pi)^{-1/2} \int_{-\infty}^{t} \exp\left(-\frac{x^2}{2}\right) dx. \tag{7}$$

The quantity $(\mu_S^t K_S^{-1} \mu_S)^{1/2}$ is called the Mahalanobis distance between the class means, and will henceforth be denoted $d(S)$.

It is our purpose to investigate the existence of the possible orderings among the numbers $P_e(S)$, as a function of the subsets S of Ω. One can observe that, independently of the properties of $f_{\theta_i}(x_S)$, these orderings are subject to certain natural restrictions. Indeed, since additional information

Manuscript received November 8, 1976; revised April 4, 1977. This work was supported in part by Contract AF F44620-74-C-0068.

T. M. Cover is with the Department of Electrical Engineering and the Department of Statistics, Stanford University, Stanford, CA 94305.

J. M. Van Campenhout is with the Information Systems Laboratory, Stanford University, Stanford, CA 94305, on leave from the State University of Ghent, Ghent, Belgium.

Printed in U.S.A. Annals No. 709SM004

brought in by enlarging the set of measurements never increases the probability of error, it is clear that any ordering

$$\boldsymbol{P}\colon P_e(S_1) > P_e(S_2) > P_e(S_3) > \cdots > P_e(S_2n) \qquad (8)$$

on $P_e(S)$ is subject to the monotonicity condition

$$S' \subset S \Rightarrow P_e(S') \geq P_e(S). \qquad (9)$$

The primary result is the following.

Theorem 1: Every ordering on the Bayes risk $P_e(S)$ that satisfies the set inclusion monotonicity constraints in (9) is inducible. In particular, there exist n jointly normal random vectors

$$\boldsymbol{M}_i \sim N(\theta \boldsymbol{\mu}_i, \boldsymbol{K}), \qquad i = 1, 2, \cdots, n, \quad \theta \in \{-1,1\},$$

such that $P_e(S) = \Pr\{\theta^* \neq \theta \mid \boldsymbol{M}_i, i \in S\}$ has this ordering.

We shall prove this theorem by constructing a multivariate normal model inducing any given allowable ordering. Univariate measurements inducing the same ordering can be created by the artifice of using a standard invertible map, such as digit interleaving, from $\mathbb{R}^k$ to $\mathbb{R}$. In the Appendix, however, we will give a direct derivation of a univariate model.

III. An Analogous Idea

Suppose there are 2^n safety deposit boxes, one for each subset $S \subseteq \{1, 2, \cdots, n\} = \Omega$. Box S contains an amount of gold $g(S)$. Suppose also that there are n individuals $M_1, M_2, \cdots, M_n$. The ith individual has a ring of 2^{n-1} keys $K_{i,S}$, one for each box S such that $i \in S$. Box S has $|S|$ keyholes, one for each individual M_i such that $i \in S$. Box S will open if and only if all keys $K_{i,S}$, $i \in S$, are available.

We see that if the individuals M_i, where $i \in S$, gather together, they will be able to unlock an amount of gold

$$G(S) = \sum_{S' \subseteq S} g(S'). \qquad (10)$$

We ask whether all $(2^n)!$ orderings on $G(S)$, $S \subseteq \Omega$, are obtainable by a proper choice of $g(S)$, $S \subseteq \Omega$. The answer is yes if the values of $g(S)$ are allowed to be arbitrary real numbers. If $g(S) \geq 0$, for all S, then clearly only orderings on $G(S)$ satisfying $S' \subseteq S \Rightarrow G(S') \leq G(S)$ can be achieved, and the lemma proved below establishes that all such orderings can indeed be achieved. This model will now be translated into statistical terms.

The underlying idea in modeling the action of the keys $K_{i,S}$, $i \in S$, is described in the following example. Consider m jointly normal random variables T_i, $i = 1, 2, \cdots, m$, distributed according to

$$\boldsymbol{T} = (T_1, T_2, \cdots, T_m) \sim N(\theta \boldsymbol{\mu}, \boldsymbol{K})$$

where

$$\boldsymbol{\mu} = (1, 1, 1, \cdots, 1)^t$$

and

$$(\boldsymbol{K})_{ij} = \begin{cases} \sigma^2, & \text{if } i = j \\ \dfrac{-\sigma^2}{m-1}, & \text{if } i \neq j, \quad i, j = 1, 2, \cdots, m. \end{cases}$$

Let A_r denote the average of any r-element subset of the random variables T_i. Then we can write

$$E(A_r) = E\left(\frac{1}{r} \sum_{k=1}^{r} T_{i_k}\right) = \theta,$$

so the expected value of this average is θ, irrespective of r. The variance of A_r behaves much differently. Indeed, as one can easily verify, we have

$$\begin{aligned} \operatorname{var}(A_r) &= \sigma^2 \frac{(m-r)}{r(m-1)} \geq \left(\frac{\sigma}{m-1}\right)^2, \qquad 0 < r < m \\ \operatorname{var}(A_m) &= 0. \end{aligned} \qquad (11)$$

If σ is large, it is clear that any proper subset of the random variables T_i, $i = 1, 2, \cdots, m$ is virtually useless in estimating θ, while the empirical average of the entire population is precisely equal to θ.

The above distribution is degenerate, in the sense that $\det(\boldsymbol{K}) = 0$. A trivial modification can remove this degeneracy while retaining the essential idea.

Equation (7) establishes that the probability of error in a normal two-category case can be expressed as a monotone decreasing function of the Mahalanobis class distance $d(S)$. In the case above, $d(S)$ is arbitrarily small for all $S \subset \Omega$, but is indefinitely large for $S = \Omega$. One can also observe that in a multivariate normal two-category case, where the vector of random variables can be partitioned in mutually class-conditional independent subvectors, P_e is a monotone decreasing function of the sum of the squared Mahalanobis class distances computed over these subvectors.

With these ideas in mind, we can proceed with the construction of our normal model.

IV. The Normal Model

Let each measurement $\boldsymbol{M}_i$, $i \in \Omega$, be a vector of 2^{n-1} conditionally independent random variables $X_{i,S}$, where S ranges over all subsets of Ω containing i. Let the random variables $X_{i,S}$ that are indexed by the same fixed set S be grouped into a vector $\boldsymbol{X}_S$ where

$$\boldsymbol{X}_S = (X_{i_1,S}, X_{i_2,S}, \cdots, X_{i_{|S|},S}), \qquad i_1 < i_2 < \cdots < i_{|S|}, \quad i_j \in S, \qquad (12)$$

which is normally distributed with mean $\theta_i \boldsymbol{\mu}_S$ and covariance matrix $\boldsymbol{K}_S$,

$$f_{\theta_i}(\boldsymbol{X}_S) = \phi(\theta_i \boldsymbol{\mu}_S, \boldsymbol{K}_S), \qquad i = 1,2. \qquad (13)$$

The total population of all vectors $\boldsymbol{X}_S$, $\varnothing \neq S \subseteq \Omega$, is distributed according to

$$f_{\theta_i}(\boldsymbol{X}_{S_1}, \boldsymbol{X}_{S_2}, \cdots, \boldsymbol{X}_{S_{2^n-1}}) = \prod_{\varnothing \neq S \subseteq \Omega} \phi(\theta_i \boldsymbol{\mu}_S, \boldsymbol{K}_S). \qquad (14)$$

The analogy with the "gold in the boxes" model will now become apparent. The amount of gold $g(S)$ that a set of keys $K_S = \{K_{i,S} \colon i \in S\}$ can release corresponds to the Mahalanobis distance $d(S)$ computed over $\boldsymbol{X}_S$. As in the example, we can choose $\boldsymbol{K}_S$ such that the Mahalanobis distance computed over any proper subvector of $\boldsymbol{X}_S$ is arbitrarily small.

The probability of error incurred by using only measurements M_i for which $i \in S$ can then be written as a monotonically decreasing function of

$$D(S) = \sum_{\emptyset \neq S' \subseteq S} d^2(S') + \alpha(S), \qquad (15)$$

where $\alpha(S)$ represents the sum of the squared Mahalanobis distances arising from the partially known vectors $X_{S''}$ for which $S'' \cap S \neq \emptyset$ but $S'' \not\subseteq S$. The following lemma establishes the key fact needed to prove the theorem.

Lemma: For any fixed $\varepsilon > 0$, and for each linear ordering $\emptyset = S_0 \prec S_1 \prec S_2 \prec \cdots \prec S_{2^n-1} = \Omega$, $S_i \subseteq \Omega$, satisfying

$$S \subset S'' \Rightarrow S \prec S'', \qquad (16)$$

there exists a set of positive numbers $\{d^2(S) \colon S \subseteq \Omega\}$ such that

$$S \prec S'' \Rightarrow D(S) - \alpha(S) \leq D(S') - \alpha(S') - \varepsilon. \qquad (17)$$

Proof: For notational simplicity, denote $D(S) - \alpha(S)$ by $D^*(S)$. The set of equations (15) can be inverted by the Moebius inversion for the lattice to yield

$$d^2(S) = \sum_{S' \subseteq S} (-1)^{|S| - |S'|} D^*(S'), \qquad S \subseteq \Omega. \qquad (18)$$

Let us choose values $D^*(S)$ consistent with the set ordering in (17) in the following way:

$$D^*(S_{p+1}) = \max\left\{D^*(S_p), \sum_{S' \subset S_{p+1}} d^2(S')\right\} + \varepsilon, \qquad (19)$$

where $p = 0, 1, \cdots, 2^n - 2$ and $D^*(S_0) = D^*(\emptyset) = 0$. One observes the following.

i) $D^*(S_{p+1})$ is well defined, since by (18) and the monotonicity condition (16), the sum

$$\sum_{S' \subset S_{p+1}} d^2(S')$$

can be uniquely expressed in terms of $D^*(S_k)$, $0 \leq k \leq p$. Hence, (19) can be solved recursively.

ii) Equation (19) implies

$$D^*(S_{p+1}) \geq D^*(S_p) + \varepsilon, \qquad (20)$$

thus $S \prec S' \Rightarrow D(S) - \alpha(S) \leq D(S') - \alpha(S') - \varepsilon$ is satisfied.

iii) By (15) and (19) we can write

$$d^2(S) = D^*(S) - \sum_{S' \subset S} d^2(S') \geq \varepsilon > 0. \qquad (21)$$

Thus the values $d^2(S)$ are positive as required. Then i)–iii) imply the proof of the lemma. Now we are in a position to prove the main theorem.

V. Proof of the Ordering Property for the Normal Model

We will exhibit a choice for μ_S and K_S, $S \subseteq \Omega$, such that the numbers $d(S)$ as established by Lemma 1 can indeed be obtained.

It is here that we will use the properties of the simplex distribution demonstrated above, the idea being that each vector X_s only contributes negligibly towards classification unless it is completely known. On the other hand, when X_s is completely known, the reduction in the misclassification probability obtained by using X_s can be set to an arbitrary value for each $S \subseteq \Omega$. This can be achieved as follows.

Let

$$\mu_S = (1, 1, \cdots, 1)^t,$$

and

$$(K_S)_{ij} = -b_S + (|S| b_S + \sigma_S^2)\delta_{ij}, \qquad i, j \in \{1, 2, \cdots, |S|\}. \qquad (22)$$

Here, $b_S > 0$, and δ_{ij} is the Kronecker delta. It follows from the definition that

$$d(S) = (\mu_S^t K_S^{-1} \mu_S)^{1/2} = \left(\frac{|S|}{\sigma_S^2}\right)^{1/2}, \qquad (23)$$

while for any m-element subset of S, $0 \leq m < |S|$, the Mahalanobis class distance is given by

$$\hat{d}_m(S) = \left(\frac{m}{(|S| - m)b_S + \sigma_S^2}\right)^{1/2}. \qquad (24)$$

Thus $d(S) > 0$ can be chosen freely for each S, which makes the lemma applicable to the orderings on $D^*(S)$. An upper bound for $\alpha(S)$ is given by the sum of the maximum values of all terms that can appear in the expression for any $\alpha(S)$:

$$0 \leq \alpha(S) < \sum_{\emptyset \neq S' \subseteq \Omega} \frac{|S'| - 1}{b_{S'}}, \qquad \text{for all } S \subseteq \Omega. \qquad (25)$$

Hence, if we choose b_S such that

$$b_S > \frac{2^n(|S| - 1)}{\varepsilon}, \qquad \emptyset \neq S \subseteq \Omega, \qquad (26)$$

then $\alpha(S) < \varepsilon$, $S \subseteq \Omega$, and the ordering on $D^*(S)$ coincides with the ordering on $D(S)$. This completes the proof of the theorem.

VI. A Numerical Example with $n = 4$ for the Normal Model

In this example, we let $\varepsilon = 0.1$. Suppose we want to establish the ordering:

$$\begin{aligned} \boldsymbol{P}\colon P_e(1) &> P_e(2) > P_e(3) > P_e(2,3) > P_e(4) \\ &> P_e(3,4) > P_e(2,4) > P_e(1,4) > P_e(1,3) \\ &> P_e(1,2) > P_e(1,2,3) > P_e(1,2,4) > P_e(1,3,4) \\ &> P_e(1,2,3,4). \end{aligned}$$

Note that the best 2-element set is composed of the two individually worst measurements, and that neither of the best subsets of cardinality 3 or 2 includes the best subset of the next lower cardinality.

By applying Lemma 1, we obtain parameters and error probabilities as listed in Table I. These values are obtained by letting

$$b_S = \frac{2^n(|S| - 1)}{\varepsilon} + 10, \qquad \text{for all } S \subseteq \Omega. \qquad (27)$$

Fig. 1 graphically represents the monotonicity condition with the nodes ordered vertically according to $\boldsymbol{P}$.

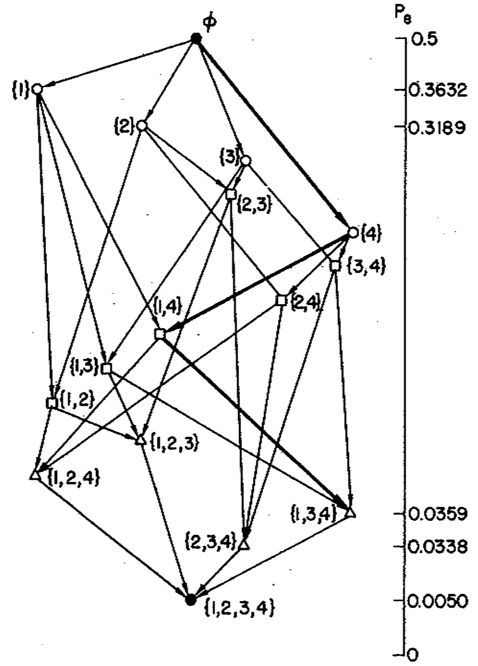

Fig. 1.

TABLE I
$P_e(S)$

S	σ_S^2	b_S	$\alpha_S + \sum_{S' \subseteq S} \frac{\lvert S' \rvert}{\sigma_{S'}^2}$	$P_e(S)$
{1}	10.00	---	.1224	.3632
{2}	5.00	---	.2218	.3189
{3}	3.33	---	.3215	.2854
{4}	1.43	---	.7204	.1980
{1,2}	1.67	170	1.5386	.1075
{1,3}	2.00	179	1.4385	.1151
{1,4}	4.00	170	1.3390	.1236
{2,3}	20.00	170	.6389	.2121
{2,4}	6.67	170	1.2386	.1329
{3,4}	20.00	170	1.1393	.1630
{1,2,3}	30.00	330	3.0399	.0437
{1,2,4}	30.00	330	3.1395	.0383
{1,3,4}	6.00	330	3.2395	.0359
{2,3,4}	1.88	330	3.3401	.0338
{1,2,3,4}	40.00	490	6.9000	.0050

VII. Conclusion

The existence of orderings as depicted in Fig. 1 (or even more anomalous orderings in cases with $n > 4$) indicates that published algorithms searching for the best k-element subset ($k < n$) by successively enlarging the best j-element set with the conditionally best measurement, for $j = 1, 2, \cdots, k - 1$, will in general not yield the correct answer. The heavy line in Fig. 1 represents the sequence of subsets so obtained. A closer observation reveals that not only does this procedure fail to be optimal, but so does the so-called 2-1 algorithm [5] which proceeds by enlarging the current set by the conditionally best pair and then discards the conditionally worst measurement. Also, the corresponding backwards procedure which starts out with Ω and eliminates measurements does not find the best 2-element set.

In the derivation of the monotonicity constraint, we tacitly assumed noninterfering measurements. However, measurements that involve the application of stimuli to the test item can interfere, because these stimuli might alter the response of the test item to simultaneous or subsequent stimuli. In those cases, the monotonicity condition need not be satisfied, and the existence of all $(2^n)!$ orderings can then be achieved.

Appendix
A Univariate Model

The "gold in the boxes" model is also very suitable for the construction of a univariate stochastic model exhibiting the anomalous ordering property. In this case, the dimensionality of the stochastic space on which the model is built is limited to the number of measurements n. Thus it is not possible to model the "deposit boxes" as disjoint subspaces spanned by independent sets of random variables. Instead we will use the following approach.

Consider n scalar-valued measurements $M_i, i \in \Omega = \{1, 2, \cdots, n\}$. Let $\mathscr{M}$ be the n-dimensional space spanned by these measurements and let $\boldsymbol{M} = (M_1, M_2, \cdots, M_n)$ denote measurement points in this space. Consider $2^n - 1$ disjoint regions R_S in $\mathscr{M}$, one for each nonempty subset S of Ω, such that its projections $R_{S|S'}$ on the subspaces spanned by M_i, $i \in S' \subset \Omega$, are disjoint.

Let these regions carry the total probability mass, i.e., let

$$p_S = \Pr\{\boldsymbol{M} \in R_S | \theta\}, \qquad \text{where } \sum_{S \subseteq \Omega} p_S = 1. \tag{A1}$$

In this case, the expression (1) for the probability of error can be written as

$$P_e(S) = \sum_{S' \subseteq \Omega} \int_{R_{S'|S}} \tfrac{1}{2} \min\left(dF_{\theta_1}(\boldsymbol{m}_S), dF_{\theta_2}(\boldsymbol{m}_S)\right) \tag{A2}$$

where

$$\boldsymbol{M}_S = (M_{i_1}, M_{i_2}, \cdots, M_{i_{|S|}}),$$
$$i_1 < i_2 < \cdots < i_{|S|}, \quad i_j \in S \subseteq \boldsymbol{\Omega},$$

and $F_{\theta_i}(\boldsymbol{m}_S)$ is the conditional probability distribution of $\boldsymbol{M}_S$.

The key idea then is to assign probabilities to the points in R_S, $S \subseteq \Omega$ such that the integrals in (A2) equal 0 on all regions $R_{S'|S}$ for which $S' \subseteq S$, and equal $\frac{1}{2}p_{S''}$ for all other regions $R_{S''|S}$. In that case, (A2) can be written as

$$P_e(S) = \frac{1}{2} \sum_{S' \not\subseteq S} p_{S'} = \frac{1}{2}\left(1 - \sum_{S' \subseteq S} p_{S'}\right), \tag{A3}$$

to which the lemma can be applied. Obviously, the numbers p_S^* generated by the lemma have to be scaled to satisfy

$\sum_{S\subseteq\Omega} p_S^* = 1$, but this in no way alters the ordering induced on $P_e(S)$.

We will now construct regions R_S and a probability assignment satisfying these requirements. The following two-dimensional example clarifies the underlying idea.

Consider a checkerboard aligned with the x and y axes and let a piece be on black if $\theta = \theta_1$ and on white otherwise. A two-dimensional view unambiguously determines the color, and hence the value of θ. However, in the projection on either axis, black and white squares overlay in the same number in all positions, and therefore no information on the value of θ can be obtained from a single coordinate of the piece. This idea easily generalizes to higher dimensions in the way described below.

Let the nonempty subsets of Ω be denoted $S_k, k = 1, 2, \cdots, 2^n - 1$, in lexicographic order. Let the regions R_S be the vertices of disjoint n-cubes, given by

$$R_{S_k} = \{2k, 2k+1\}^n, \qquad \text{for } k = 1, 2, \cdots, 2^n - 1. \tag{A4}$$

Let $I_k(\boldsymbol{m})$ denote the indicator function for R_{S_k}, i.e.,

$$I_k(\boldsymbol{m}) = \begin{cases} 1, & \text{if all components of } \boldsymbol{m} \text{ belong to} \\ & \quad \{2k, 2k+1\} \\ 0, & \text{otherwise.} \end{cases} \tag{A5}$$

Define the class conditional probability distribution of M by

$$P(\boldsymbol{m}\mid\theta) = 2^{-n} \sum_{k=1}^{2^n-1} I_k(\boldsymbol{m}) p_{S_k} \left(1 + \theta \prod_{i\in S_k} (-1)^{m_i}\right). \tag{A6}$$

If $\theta = 1$, this function assigns a probability $2^{-(n-1)} p_{S_k}$ to all vertices of R_{S_k} that have an even number of odd values among their coordinates m_i, $i \in S_k$, and zero to all other vertices of R_{S_k}. If $\theta = -1$, the latter vertices are assigned a probability $2^{-(n-1)} p_{S_k}$, and the former zero probability. The class conditional distributions of $\boldsymbol{M}_S$, $S \subset \Omega$, can then be shown to be

$$P(\boldsymbol{m}_S\mid\theta) = 2^{-|S|} \sum_{k=1}^{2^n-1} I_k(\boldsymbol{m}_S) p_{S_k} \cdot \left(1 + a_k(S)\theta \prod_{i\in S_k\cap S} (-1)^{m_i}\right) \tag{A7}$$

where

$$a_k(S) = \begin{cases} 0, & \text{if } S_k - S \neq \varnothing \\ 1, & \text{otherwise.} \end{cases}$$

Finally, the probability of error $P_e(S)$ is given by

$$P_e(S) = 2^{-|S|-1} \sum_{k=1}^{2^n-1} \sum_{\boldsymbol{m}_S \in R_{Sk|S}} p_{S_k} \cdot \min\left\{\left(1 + a_k(S) \prod_{i\in S_k\cap S} (-1)^{m_i}\right), \left(1 - a_k(S) \prod_{i\in S_k\cap S} (-1)^{m_i}\right)\right\}. \tag{A8}$$

This reduces to (A3) in a straightforward way.

References

[1] J. D. Elashoff, R. M. Elashoff, and G. E. Goldman, "On the choice of variables in classification problems with dichotomous variables," *Biometrika*, vol. 54, pp. 668–670, 1967.

[2] G. T. Toussaint, "Note on the optimal selection of independent binary-valued features for pattern recognition," *IEEE Trans. Information Theory*, vol. IT-17, p. 618, Sept. 1971.

[3] T. M. Cover, "The best two independent measurements are not the two best," *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-4, no. 1, pp. 116–117, Jan. 1974.

[4] A. N. Mucciardi and E. E. Gose, "A comparison of seven techniques for choosing subsets of pattern recognition properties," *IEEE Trans. Computers*, vol. C-20, pp. 1023–1031, Sept. 1971.

[5] S. D. Stearns, "On selecting features for pattern classifiers," in *Proc. Third Int. Joint Conf. Pattern Recognition*, Coronado, CA, Nov. 1976, IEEE Computer Society, pp. 71–75.