PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 119, Number 4, December 1993

LOWER BOUND ON THE ERROR PROBABILITY FOR FAMILIES WITH BOUNDED LIKELIHOOD RATIOS

ANDREW L. RUKHIN

(Communicated by Wei Y. Loh)

ABSTRACT. In the classification problem a new sharp lower bound for the error probability is derived. This bound depends only on the prior probabilities and on the support of pairwise likelihood ratios.

1. INTRODUCTION

Let $P_1, \ldots, P_m$ be a family of different probability distributions and let x be an observation on one of the corresponding random variables. In the classification problem a decision about the true distribution has to be made on the basis of x. If $\delta = \delta(x)$ is a classification procedure, i.e., δ takes values $1, \ldots, m$, and w_k, $k = 1, \ldots, m$, denote prior probabilities, then under zero-one loss the performance of δ is measured by Bayes risk (or error probability)

$$\rho(\delta) = \sum_i w_i P_i(\delta \neq i).$$

Lower bounds for error probability in terms of information measures have been obtained by a wealth of authors (see, e.g., [2, 4, 8]). Shannon, Gallager, and Berlekamp obtained a related inequality in the context of communication theory (see [6]). A good survey of these results and of their use in asymptotic statistical theory can be found in the monograph of Vajda [9].

In this paper the interest is in obtaining a lower bound on the error probability when all likelihood ratios dP_i/dP_k, $i \neq k$, are bounded. Intuitively it is clear that the closer the probability distributions are to one another, the more difficult the classification problem is and the larger the error probability is. This intuition is confirmed by the inequality derived in the paper. The motivation for the study of the bounded likelihood ratios family is in statistical inference with finite-memory systems [5]. For instance, Hellman and Cover [7] have shown that in the two-hypothesis testing problem a finite-memory test statistic can be consistent only if the distribution of the likelihood ratio is supported by

Received by the editors March 31, 1992.
1991 *Mathematics Subject Classification.* Primary 62C05; Secondary 62F11, 62F35.
Key words and phrases. Classification problem, error probability, likelihood ratios, linear programming, M-matrix.
Research supported by NSF grant #DMS-9000999.

the whole positive half-line. Thus in the bounded likelihood ratio situation the error probability is bounded from below by a positive constant and the mentioned inequality obtained in §3 shows how small this constant can be, i.e., it provides a lower bound on error probability for the optimal recursive procedure in the finite-memory setting at any observation stage. Also the choice of the prior distribution minimizing this bound is suggested. The matrix of bounding constants which solely determines this distribution is shown in §2 to have a very special nature; namely, its inverse is an M-matrix.

2. M-matrices and linear programming

In this section we establish two propositions needed in the proof of the main result but also of some independent interest.

Let $(\mathscr{X}, \mu)$ be a measurable space. In the following $f_1, \ldots, f_m$ are positive μ-integrable functions such that for any positive c and $i \neq k$

$$\mu\{x: f_i(x) \neq c f_k(x)\} > 0. \tag{2.1}$$

Suppose that all the ratios f_i/f_k are bounded, i.e.,

$$a_{ki} \leq \frac{f_k(x)}{f_i(x)} \leq \frac{1}{a_{ik}} \quad \mu\text{-a.s.} \tag{2.2}$$

Moreover, assume that a_{ik} are the largest quantities satisfying (2.2) and let A be the $m \times m$ matrix formed by these numbers, $A = (a_{ik})$.

Recall the definition of an M-matrix X ([1, Chapter 16, Exercise 13] or [3]). A matrix C such that $c_{ik} \leq 0$ for $i \neq k$ is called an M-matrix if one of the following equivalent conditions holds:

(1) C is nonsingular and all elements of C^{-1} are nonnegative;
(2) for some positive numbers $z_1, \ldots, z_m$, $\sum_{k=1}^{m} c_{ik} z_k > 0$, $i = 1, \ldots, m$;
(3) all principal minors of C are positive.

Proposition 2.1. *Under condition* (2.1) *the matrix* A *is nonsingular and* A^{-1} *is an M-matrix.*

Proof. It follows from (2.2) that, for all i, k, $a_{ik}a_{ki} \leq 1$ and if, for some $i \neq k$, $a_{ik}a_{ki} = 1$ then $f_k(x) = a_{ki}f_i(x)$ μ-a.s., which contradicts (2.1). $a_{ik}a_{ki} < 1 = a_{ii}a_{kk}$.

We start by establishing an analogue of this formula for a sequence of matrices arising in the Gauss elimination algorithm.

Let $a_{ik}^{(0)} = a_{ik}$ and define recursively

$$a_{ik}^{(n)} = a_{ik}^{(n-1)} - a_{in}^{(n-1)}/a_{nn}^{(n-1)}. \tag{2.3}$$

We assume here $a_{nn}^{(n-1)} \neq 0$. The positivity of these and all other coefficients is implied by the following inequality which is proven by induction:

$$a_{ik}^{(n)} a_{kj}^{(n)} < a_{ij}^{(n)} a_{kk}^{(n)} \tag{2.4}$$

where $k \neq i$ and $k \neq j$.

Assuming that (2.4) holds for $n-1$ it can be rewritten in the form

$$a_{nn}^{(n-1)}[a_{ij}^{(n-1)}a_{kk}^{(n-1)} - a_{ik}^{(n-1)}a_{kj}^{(n-1)}] - a_{in}^{(n-1)}[a_{nj}^{(n-1)}a_{kk}^{(n-1)} - a_{nk}^{(n-1)}a_{kj}^{(n-1)}] - a_{kn}^{(n-1)}[a_{nk}^{(n-1)}a_{ij}^{(n-1)} - a_{ik}^{(n-1)}a_{nj}^{(n-1)}] > 0. \tag{2.5}$$

If $a_{nk}^{(n-1)}a_{ij}^{(n-1)} > a_{ik}^{(n-1)}a_{nj}^{(n-1)}$, use the inequality

$$a_{in}^{(n-1)} < a_{ij}^{(n-1)}a_{nn}^{(n-1)}/a_{nj}^{(n-1)}$$

to estimate the left-hand side of (2.5) as

$$\begin{aligned} &a_{nn}^{(n-1)}[a_{ij}^{(n-1)}a_{kk}^{(n-1)} - a_{ik}^{(n-1)}a_{kj}^{(n-1)}] - a_{ij}^{(n-1)}a_{nn}^{(n-1)}a_{kk}^{(n-1)} \\ &\quad + a_{ok}^{(n-1)}a_{nn}^{(n-1)}a_{nk}^{(n-1)}a_{kj}^{(n-1)}/a_{nj}^{(n-1)} - a_{kn}^{(n-1)}[a_{nk}^{(n-1)}a_{ij}^{(n-1)} - a_{nj}^{(n-1)}a_{ik}^{(n-1)}] \\ &\quad = [a_{nk}^{(n-1)}a_{ij}^{(n-1)} - a_{ik}^{(n-1)}a_{nl}^{(n-1)}][a_{kj}^{(n-1)}a_{mm}^{(n-1)} - a_{kn}^{(n-1)}a_{nj}^{(n-1)}]/a_{nj}^{(n-1)} > 0. \end{aligned}$$

If $a_{nk}^{(n-1)}a_{ij}^{(n-1)} \le a_{ik}^{(n-1)}a_{nj}^{(n-1)}$, use the inequality

$$a_{ik}^{(n-1)} \ge a_{nk}^{(n-1)}a_{ij}^{(n-1)}/a_{nj}^{(n-1)}$$

and estimate the left-hand side of (2.5) from below as

$$\begin{aligned} &a_{nn}^{(n-1)}a_{ij}^{(n-1)}a_{kk}^{(n-1)} - a_{nn}^{(n-1)}a_{kj}^{(n-1)}a_{ij}^{(n-1)}/a_{nj}^{(n-1)} \\ &\quad - a_{in}^{(n-1)}a_{nj}^{(n-1)}a_{kk}^{(n-1)} + a_{in}^{(n-1)}a_{nk}^{(n-1)}a_{kj}^{(n-1)} \\ &\quad = [a_{kk}^{(n-1)}a_{nj}^{(n-1)} - a_{nk}^{(n-1)}a_{kj}^{(n-1)}][a_{nn}^{(n-1)}a_{ij}^{(n-1)} - a_{in}^{(n-1)}a_{nj}^{(n-1)}]/a_{nj}^{(n-1)} > 0. \end{aligned}$$

Thus (2.4) holds and it follows that $a_{nn}^{(n-1)} > 0$. In particular the matrix A is nonsingular and all principal minors of A are positive. To prove that A^{-1} is an M-matrix it suffices now to show that its off-diagonal elements are negative.

Let A_{k_j} be the k, j cofactor of matrix A. Since $A^{-1} = (A_{ki}/|A|)$, one must demonstrate that, for $i \neq k$, $A_{ki} < 0$.

Assume for concreteness sake that $i < k$. Transform elements of A repeatedly after (2.3), i.e., subtract the multiples of rows $1, \dots, k-1$ and $k+1, \dots, m$ so that the only nonzero element in the j th column is a_{jj}, $j = 1, \dots, i-1, k+1, \dots, m$, which is positive. Therefore, the sign of A_{ki} is that of the determinant of the matrix

$$\begin{pmatrix} a_{i\,i+1} & \cdots & a_{i\,k} \\ a_{i+1\,i+1} & \cdots & a_{i+1\,k} \\ \cdots & & \cdots \\ a_{k-1\,i+1} & \cdots & a_{i-1\,i} \end{pmatrix}$$

whose elements satisfy (2.4).

Applying the same procedure to columns of this matrix we see that if $i+k$ is even, the sign of its determinant coincides with the sign of a_{k-1k} (which is positive) and if $i+k$ is odd, this sign equals that of the determinant

$$\begin{vmatrix} a_{k-2\,k-1} & a_{k-1\,k-1} \\ a_{k-2\,k} & a_{k-1\,k} \end{vmatrix},$$

which is negative. □

Because of Proposition 2.1 one can always find a vector w with positive coordinates, $w > 0$, such that $(A^{\mathrm{T}})^{-1}w > 0$. Here A^{T} denotes the transposed matrix which is an M-matrix simultaneously with A.

In the sequel $\langle z, w\rangle = \sum_{i=1}^{m} z_i w_i$ will denote the inner product of two vectors $z = (z_1, \ldots, z_m)^{\mathrm{T}}$ and $w = (w_1, \ldots, w_m)^{\mathrm{T}}$. Also e_j, $j = 1, \ldots, m$, will denote the basis vectors and $e = \sum_j e_j = (1, \ldots, 1)^{\mathrm{T}}$.

Proposition 2.2. *Under the conditions of Proposition* 2.1 *let* R_k, $k = 1, \ldots, m$, *be a partition of* $\mathscr{X}$. *Put* $w = \left(\int f_1 \, d\mu, \ldots, \int f_n \, d\mu\right)^{\mathrm{T}}$ *and assume that*

$$(A^{\mathrm{T}})^{-1}e \geq 0. \tag{2.6}$$

Then

$$\sum_k \int_{R_k} f_k \, d\mu \leq \langle A^{-1}w, e\rangle.$$

Proof. One has

$$z_k = \int_{R_k} f_k \, d\mu \leq \frac{1}{a_{ik}} \int_{R_k} f_i \, d\mu,$$

so that

$$\sum_k a_{ik} z_k \leq \sum_k \int_{R_k} f_i \, d\mu = w_i.$$

In other terms

$$Az \leq w, \qquad z \geq 0, \tag{2.7}$$

and the determination of the maximum of the linear function $\sum_k z_k = \langle z, e\rangle$ over the convex set determined by (2.7) presents a classical problem of linear programming.

Condition (2.6) guarantees that the inequality $\langle z, e\rangle \leq \langle A^{-1}w, e\rangle$ is a corollary of (2.7), and this proves Proposition 2.2 which now also immediately follows from the duality theorem. □

If (2.6) does not hold then $\langle z, e\rangle$ takes values larger then $\langle A^{-1}w, e\rangle$ on the set (2.7). This fact follows from the Farkas Lemma, which together with the fundamental theorem of linear programming also can be used to show that

$$\max\langle z, e\rangle = \max\{\langle A_t^{-1}w^t, e^t\rangle : (A_t^{\mathrm{T}})^{-1}e^t \geq 0\}$$

where A_t denotes a submatrix of A obtained by deleting some rows and columns of A and e^t, w^t denote the corresponding subvectors of e and w. In particular for $m = 2$

$$\max\langle z, e\rangle = \begin{cases} \dfrac{(1 - a_{21})w_1 + (1 - a_{12})w_2}{1 - a_{12}a_{21}}, & a_{12}, a_{21} < 1, \\ w_1, & a_{12} > 1, \\ w_2, & a_{21} > 1. \end{cases}$$

It is worth noticing that the property that the set $\{x \geq 0, : C^{\mathrm{T}}x \leq y\}$ is bounded for any $y \geq 0$ and the nonsingularity of C characterizes M-matrices (see [3, p. 138]).

3. Main result

Here $p_1, \dots, p_m$ will be different probability densities over $(\mathscr{X}, \mu)$, whose ratios are bounded:

$$b_{ki} \le \frac{p_k(x)}{p_i(x)} \le \frac{1}{b_{ik}} \quad \mu\text{-a.s.}$$

and B will denote the matrix formed by the quantities b_{ik} which are assumed to be the largest possible. Also let $w = (w_1, \dots, w_m)^{\mathrm{T}}$ be the vector of prior probabilities.

If $\delta_o(x)$ is the Bayes estimator of the finite-valued parameter i, $i = 1, \dots, m$, then its Bayes risk $\rho(\delta_o)$ has the form

$$\rho(\delta_o) = \sum_i w_i P_i(\delta_o(x) \ne i) = 1 - \sum_i w_i P_i(\delta_o(x) = i).$$

Theorem 3.1. *Assume that*

$$(B^{\mathrm{T}})^{-1} w \ge 0. \tag{3.1}$$

Then

$$\rho(\delta_o) \ge 1 - \langle B^{-1} e, w \rangle. \tag{3.2}$$

Proof. One has with $R_i = \{x : \delta_o(x) = i\}$, $i = 1, \dots, m$,

$$\sum_i w_i P_i(\delta_o(x) = i) = \sum_i \int_{R_i} f_i \, d\mu$$

where $f_i(x) = w_i p_i(x)$. Therefore, conditions of Propositions 2.1 and 2.2 are met with $a_{ik} = b_{ik} w_i / w_k$. Also the determinants of the matrices A and B are equal, $|A| = |B|$, and if A_{ki} is the k, i cofactor of matrix A and B_{ki} is the same cofactor of B, then $A_{ki} = B_{ki} w_i / w_k$. Thus the inverse matrix A^{-1} has the form

$$A^{-1} = \left(\frac{A_{ki}}{|A|} \right) = \left(\frac{B_{ki}}{|B|} \frac{w_i}{w_k} \right),$$

and condition (2.6) is tantamount to (3.1).

Thus Proposition 2.2 implies that

$$\sum w_i P_i(\delta_o(x) = i) \le \langle A^{-1} w, e \rangle = \sum_{i,k} w_i B_{ki} / |B| = \langle (B^{\mathrm{T}})^{-1} w, e \rangle. \quad \square$$

Theorem 3.1 is false without condition (3.1). Indeed if w tends to a basis vector e_i, then $\rho(\delta_o) \to 0$ but $\langle B^{-1} e, e_i \rangle < 1$.

If p_i, $i = 1, \dots, m$, is the joint density for a sequence of n i.i.d. random variables with bounded likelihood ratios then $b_{ik} = \beta_{ik}^n$ and

$$\liminf_{n \to \infty} [\rho(\delta_o)]^{1/n} \ge \max_{i \ne k} \beta_{ik}.$$

Therefore, in the situation when a random sample is observed, the exponential rate of the error probability cannot exceed $-\log \max_{i \ne k} \beta_{ik}$ (see [8] for a similar inequality).

If $m = 2$, then according to Theorem 3.1

$$\rho(\delta_o) \geq 1 - \frac{w_1(1 - b_{12}) + w_2(1 - b_{21})}{1 - b_{12}b_{21}} = \frac{w_1 b_{12} + w_2 b_{21} - b_{12}b_{21}}{1 - b_{12}b_{21}}, \tag{3.3}$$

provided that

$$w_1 \geq b_{21}w_2 \geq b_{12}b_{21}w_1.$$

Also if $w_1 < b_{21}w_2$, $\rho(\delta_o) \geq 1 - w_2$, and if $w_2 < b_{12}w_1$, $\rho(\delta_o) \geq 1 - w_1$. For $w_1 = w_2 = \frac{1}{2}$

$$\rho(\delta_o) \geq \frac{b_{12} + b_{21} - 2b_{12}b_{21}}{2(1 - b_{12}b_{21})}. \tag{3.4}$$

When $b_{12} = b_{21} = b$, $0 < b < 1$, (3.4) reduces to the inequality

$$\rho(\delta_o) \geq \frac{b}{1 + b}$$

which shows that for b close to 1 the Bayes rule cannot be much more efficient than pure guessing (in which case the error probability is equal to $\frac{1}{2}$). The smaller the b, i.e., the wider the range of the likelihood ratio, the smaller the Bayes risk could be.

As an example let P_i, $i = 1, 2$, be two binomial distributions with probabilities p_i, $p_1 < p_2$. Then

$$b_{12} = \left(\frac{p_1}{p_2}\right)^n, \qquad b_{21} = \left(\frac{1 - p_2}{1 - p_1}\right)^n$$

and according to Theorem 3.1 one has $\{\delta_o(x) = 1\} = \{x : x \leq c\}$ with

$$c = n \log \frac{1 - p_1}{1 - p_2} \Big/ \log \left[\frac{p_2(1 - p_1)}{(1 - p_2)p_1}\right],$$

so that

$$P_1(\delta_o = 2) + P_2(\delta_o = 1) = \sum_{x>c} \binom{n}{x} p_1^x (1 - p_1)^{n-x} + \sum_{x \leq c} \binom{n}{x} p_2^x (1 - p_2)^{n-x}$$
$$\geq \frac{b_{12} + b_{21} - 2b_{12}b_{21}}{1 - b_{12}b_{21}}.$$

If $n = 1$ this reduces to an equality which shows that the bound (3.3) is sharp.

As another example consider the classification problem for two Cauchy distributions with the location parameter values θ and $-\theta$. An easy calculation shows that

$$b_{12} = b_{21} = 1 + 2\theta^2 - 2\theta\sqrt{1 + \theta^2}$$

and according to (3.4)

$$\rho(\delta_o) \geq \frac{1}{2} - \frac{\theta}{2\sqrt{1 + \theta^2}}.$$

This inequality shows that for small values of θ when the classification problem is "difficult", the Bayes rule cannot have the error probability much different from $\frac{1}{2}$, and it also gives the correct rate θ^{-2} of the error probability for large values of θ.

It is worth noticing that inequality (3.2) is useful even if some b_{ik} vanish. For instance when $p_i(x) = \lambda_i \exp(-\lambda_i x)$ for positive x with, say, $\lambda_1 < \lambda_2$, then $b_{21} = 0$, $b_{12} = \lambda_1/\lambda_2$ and, for any probability w_1, $\rho(\delta_o) \geq w_1\lambda_1/\lambda_2$.

Also observe that the bound (3.3) is larger than the lower bound for the error probability of the recursive time-invariant classification rule obtained by Hellman and Cover [7]. Indeed their bound has the form $(2\sqrt{b_{12}b_{21}w_1w_2} - b_{12}b_{21})/(1 - b_{12}b_{21})$ which cannot exceed (3.3).

Now we derive the form of the prior distribution which minimizes (3.2).

Theorem 3.2. *Under conditions of Theorem* 3.1

$$\langle B^{-1}e, w\rangle \leq \max_i \frac{1}{\sum_k b_{ik}} = \frac{1}{\sum_k b_{lk}}. \tag{3.5}$$

If l is defined by (3.5) *uniquely then the equality in* (3.12) *is attained if and only if* $w = B^{\mathrm{T}}e_l/(Be)_l$.

Proof. Let $y = (B^{\mathrm{T}})^{-1}w \geq 0$. Then

$$1 = \langle B^{\mathrm{T}}y, e\rangle = \langle y, Be\rangle. \tag{3.6}$$

The maximization of the linear function $\langle B^{-1}e, B^{\mathrm{T}}y\rangle = \langle e, y\rangle$ under condition (3.6) is also a linear programming problem whose solution is $y = e_l/(Be)_l$. □

If there are several values of l satisfying (3.5) then any vector w attaining equality in (3.5) must be a convex combination of the vectors $B^{\mathrm{T}}e_l/(Be)_l$.

Since B^{T} is a matrix with positive elements, it has the largest positive eigenvalue r with an eigenvector w with positive components. For such w

$$\langle B^{-1}e, w\rangle = \frac{1}{r}\langle e, w\rangle = \frac{1}{r}.$$

It follows from (3.5) that

$$\frac{1}{r} \leq \max_i \frac{1}{\sum_r b_{ik}},$$

i.e., $r \geq \min_i \sum_k b_{ik}$ which is of course a known bound on the largest eigenvalue [1, Chapter 16, §8].

If $m = 2$, the prior distribution from Theorem 3.2 has the form

$$w_1 = \frac{1}{1 + b_{12}}, \quad w_2 = \frac{b_{12}}{1 + b_{12}} \qquad \text{if } b_{12} < b_{21}$$

and

$$w_1 = \frac{b_{21}}{1 + b_{21}}, \quad w_2 = \frac{1}{1 + b_{21}} \qquad \text{if } b_{12} \geq b_{21}.$$

In the binomial example, if $n = 1$, $p_1 < p_2 < \frac{1}{2}$

$$w_1 = \frac{p_1}{p_1 + p_2}, \qquad w_2 = \frac{p_2}{p_1 + p_2}.$$

so that the "more difficult" value of p gets larger weight.

References

1. R. Bellman, *Introduction to matrix analysis*, 2nd ed., McGraw-Hill, New York, 1970.
2. M. Ben-Bassat and J. Raviv, *Renyi's entropy and the probability of error*, IEEE Trans. Inform. Theory **IT-24** (1978), 324–331.
3. A. Berman and R. J. Plemmons, *Nonngative matrices in the mathematical sciences*, Academic Press, New York, 1979.
4. J. T. Chu and J. C. Chueh, *Inequalities between information measures and error probability*, J. Franklin Inst. **282** (1966), 121–125.
5. T. M. Cover, M. A. Freedman, and M. E. Hellman, *Optimal finite memory learning algorithms for the finite sample problem*, Information and Control **30** (1976), 49–85.
6. R. G. Gallager, *Information theory and reliable communication*, Wiley, New York, 1968.
7. M. E. Hellman and T. M. Cover, *Learning with finite memory*, Ann. Math. Statist. **41** (1970), 765–782.
8. A. Renyi, *On some problems of statistics from the point of view of information theory*, Proceedings of the Colloquium on Information Theory (Debrecen, 1969), Bolyai Janos Matematikai Tarsulat, Bolyai Math. Soc.
9. I. Vajda, *Theory of statistical inference and information*, Kluwer, Dordrecht, 1989.

Department of Mathematics, University of Maryland Baltimore County, Baltimore, Maryland 21228
E-mail address: rukhin@math13.math.umbc.edu