The Annals of Statistics
1994, Vol. 22, No. 2, 616–633

RECURSIVE TESTING OF MULTIPLE HYPOTHESES: CONSISTENCY AND EFFICIENCY OF THE BAYES RULE[1]

BY ANDREW L. RUKHIN

University of Maryland, Baltimore County

A version of the multiple hypotheses testing problem is studied in which the decision procedure is based only on the current observation and the previous decision. Conditions for inconsistency and consistency of the stepwise Bayes rule, which are related to the boundedness of the likelihood ratios, are given. The (typically slow) rate of convergence of the error probabilities of consistent procedures is determined, and a sharp lower bound for the Bayes risk in terms of bounds on the likelihood ratios is derived. A modification of the recursive Sakrison's procedure for a continuous estimation problem is obtained in this setting by embedding the discrete family of original probability distributions into an exponential family.

1. Introduction. Let $P_1, \ldots, P_m$ be a finite family of different probability distributions presenting alternative models for the distribution of each of observations $x_1, x_2, \ldots, x_n$. If $w_1, \ldots, w_m$ are prior probabilities, then the performance of a multiple hypotheses testing procedure $\delta_n = \delta_n(x_1, \ldots, x_n)$ under the zero–one loss is measured by its Bayes risk $\Sigma w_i P_i(\delta_n \neq i)$. It is known that the Bayes risk of the Bayes rule decreases exponentially fast as the sample size n increases. Also the optimal rate of exponential decay of this quantity is independent of positive prior probabilities and is determined by the information-type divergence between probability distributions. This fact leads to a notion of an asymptotically efficient multiple hypotheses testing procedure. Motivated by this concept of asymptotic optimality, one may ask if this optimality is preserved in the smaller class of following recursive procedures which are in general much easier to calculate. When the multiple decision making (classification) problem is being solved stepwise or recursively the allowable procedure δ_n depends only on the current observation x_n and the previous value δ_{n-1}.

The recursive estimation problem for a continuous parameter is well studied. See, for example, the monograph of Nevel'son and Hasminsky (1976), where the relationship between this problem and the stochastic approximation method is explored. One of the highlights of the corresponding theory is the construction of a sequence of recursive estimators which is asymptotically fully efficient. The first version of such a sequence was obtained by Sakrison (1965).

In the described setting with a discrete parameter our problem is related to multiple hypotheses testing with finite memory [cf. Cover (1969), Hellman

Received April 1992; revised April 1993.
[1]Research supported by NSF Grant DMS-88-03259.
AMS 1991 *subject classifications.* Primary 62F05; secondary 62F12, 62E20.
Key words and phrases. Asymptotic efficiency, consistency, error probabilities, exponential family, likelihood ratios, multiple hypothesis testing, recursive Bayes rule, Sakrison's estimator.

and Cover (1970), Cover, Freedman and Hellman (1976), Yakowitz (1974) and Bucklew and Ney (1991)]. In most of these papers decision δ_n is a stage-invariant function of the current observation x_n and the previous "state" δ_{n-1}, that is, $\delta_n = f(\delta_{n-1}, x_n)$, where f itself does not depend on n. Also the mentioned authors concentrate on the case $m = 2$. In our setting the number of possible states is the number (m) of alternative probability models, but the decision at time n depends on n. Thus we consider a finite time memory situation. When $m = 2$ this also corresponds to the classical setting of the finite state memory problem, in which case, as is shown by Cover, Freedman and Hellman (1976), the optimal recursive rule at stage n differs from the decision at stage $n-1$ only for extreme values of the likelihood ratio dP_2/dP_1. If this ratio is bounded, there is no consistent time-invariant recursive rule.

In this paper, in Sections 2 and 3 we prove similar results in our setup for the recursive Bayes procedure, which is based on the joint distribution of procedure derived at the previous stage and the current observation. This procedure turns out to be consistent if the support of the distribution of a pair of the likelihood ratios is unbounded, and in this case the (typically slow) rates of convergence to zero of the Bayes risk are determined. When the pairwise likelihood ratios are bounded, the procedure is inconsistent and we derive a sharp positive lower bound for the risk of Bayes procedures. In Section 4, by discretizing the Sakrison estimator constructed for the embedding of the original family of probability distributions in the exponential family, an asymptotically efficient procedure is obtained.

2. Consistency of recursive Bayes rules. Let p_i, $i = 1, \dots, m$, be the densities of distributions P_i with respect to measure μ over a measurable space $\mathcal{X}$. The recursive Bayes rule $\widetilde{\delta}_n$ is defined by the current observation x_n and the previous decision $\widetilde{\delta}_{n-1}$ as follows:

$$\begin{aligned}\{\widetilde{\delta}_n = i\} = \bigcup_{j=1}^m \Big\{ & \widetilde{\delta}_{n-1} = j,\ w_i p_i(x_n) P_i(\widetilde{\delta}_{n-1} = j) \\ & = \max_k w_k p_k(x_n) P_k(\widetilde{\delta}_{n-1} = j) \Big\}.\end{aligned} \tag{2.1}$$

Here the weights (prior probabilities) w_i are assumed to be positive. It is often convenient to let $\widetilde{\delta}_1$ be the Bayes procedure based on the first observation x_1,

$$\{\widetilde{\delta}_1(x_1) = i\} = \Big\{ w_i p_i(x_1) = \max_k w_k p_k(x_1) \Big\}.$$

Also for the sake of simplicity we assume that, for any $i \neq k$ and any real c,

$$\mu\Big(\{ p_i(x) = c p_k(x) \} \Big) = 0. \tag{2.2}$$

Otherwise a tie, $w_i p_i(x_n) P_i(\widetilde{\delta}_{n-1} = j) = w_k p_k(x_n) P_k(\widetilde{\delta}_{n-1} = j)$, may have a positive probability, and additional randomization is needed in (2.1) to define the value of $\widetilde{\delta}_n$.

It follows from (2.1) that

$$P_k(\tilde{\delta}_n = i)$$

$$= \sum_{j=1}^{m} P_k\Big\{w_i p_i(x_n) P_i(\tilde{\delta}_{n-1} = j) = \max_l w_l p_l(x_n) P_l(\tilde{\delta}_{n-1} = j)\Big\} \tag{2.3}$$

$$\times\ P_k(\tilde{\delta}_{n-1} = j).$$

Thus the distribution of $\tilde{\delta}_n$ is determined by the distribution of $\tilde{\delta}_{n-1}$.

The main questions about the recursive Bayes procedure are its consistency and, if this happens, the rate of convergence to zero of error probabilities.

To motivate the coming result, let us consider the simplest situation when $m = 2$, $w_1 = w_2 = 0.5$ and the recursive Bayes rule has equal error probabilities

$$P_1(\tilde{\delta}_n = 2) = P_2(\tilde{\delta}_n = 1) = y_n.$$

According to (2.3) this happens if

$$P_1\left(\frac{p_2(x)}{p_1(x)} \geq z\right) = P_2\left(\frac{p_1(x)}{p_2(x)} \geq z\right).$$

Let the distribution of likelihood ratio $\ell(x) = dP_2(x)/dP_1(x)$ under P_1 have the distribution function F and under P_2 the distribution function H. Then we demand that

$$1 - F(z) = P_1\big(\ell(X) \geq z\big) = P_2\big(\ell(X) < 1/z\big) = H\big(1/z\big).$$

Thus if F has density f, assume that, for all z,

$$\int_z^{\infty} dF(t) = \int_0^{z^{-1}} t\, dF(t). \tag{2.4}$$

Then in particular $\int_0^{\infty} t\, dF(t) = 1$ and $H(z) = \int_0^z t\, dF(t)$ is a distribution function. It is easy to check that if P is an absolutely continuous symmetric distribution over the real line and P_1 and P_2 are the shifts of P by b and $-b$, then the formulas above hold. Our assumptions imply that

$$y_1 = P_1(\tilde{\delta}_1 = 2) = P_1\big(\ell(x) \geq 1\big) = P_2\big(\ell(x) < 1\big) = P_2(\tilde{\delta}_1 = 1) < \tfrac{1}{2}$$

and because of (2.3), for any integer n,

$$y_n = y_{n-1} P_1\left(\ell(x) \geq \frac{y_{n-1}}{1 - y_{n-1}}\right) + (1 - y_{n-1}) P_1\left(\ell(x) > \frac{1 - y_{n-1}}{y_{n-1}}\right)$$

$$= 1 - (1 - y_{n-1}) F\left(\frac{1 - y_{n-1}}{y_{n-1}}\right) - y_{n-1} F\left(\frac{y_{n-1}}{1 - y_{n-1}}\right) = \psi(y_{n-1}).$$

If the support of distribution F is the whole positive half-line, then in the interval $(0, 0.5)$, $\psi(y) < y$ and the function $\psi(y)$ monotonically increases in this interval. It is easy to show that our assumptions imply differentiability of ψ and

$$\psi'(y) = F\left(\frac{1-y}{y}\right) - F\left(\frac{y}{1-y}\right) + f\left(\frac{1-y}{y}\right)\frac{1-y}{y^2} - f\left(\frac{y}{1-y}\right)\frac{y}{(1-y)^2}.$$

It follows from (2.4) that

$$f\left(\frac{1-y}{y}\right) = \frac{y^3}{(1-y)^3} f\left(\frac{y}{1-y}\right).$$

Therefore, for $0 \leq y \leq 0.5$,

$$\psi'(y) = F\left(\frac{1-y}{y}\right) - F\left(\frac{y}{1-y}\right) \geq 0.$$

Moreover,

$$\psi''(y) = -\frac{1}{(1-y)^3} f\left(\frac{y}{1-y}\right) \leq 0,$$

so that ψ is a concave function. Thus the sequence y_n, which can be viewed as an approximation to the solution of fixed-point problem $\psi(y) = y$, converges to the (unique) fixed point $y = 0$.

However, if the support of F is not the whole positive half-line (say, F has a positive jump at some $\bar{y} > 0$), then if $y_1 > \bar{y}$, sequence y_n converges to $\bar{y}$ and $\tilde{\delta}_n$ is not consistent.

Thus the consistency property of the recursive Bayes rule is clearly related to unboundedness of the support of the likelihood ratios. We formalize it as follows.

ASSUMPTION 1. There is a pair $r, q, 1 \leq r \neq q \leq m$, such that, for any positive t,

$$P_r\left(p_q(x)/p_r(x) > t\right) > 0.$$

THEOREM 2.1. *Under Assumption* 1 *the recursive Bayes rule* $\tilde{\delta}_n$ *is consistent.*

PROOF. Let $\Pi = \Pi_n$ denote the matrix formed by elements $\pi_{ki} = w_k P_k(\tilde{\delta}_n = i)$, $1 \leq i, k \leq m$. Define the matrix-valued mapping $\Psi = \Psi(\Pi)$ by the formula

$$\left[\Psi(\Pi)\right]_{ki} = \sum_j \pi_{kj} P_k\left(p_i(x)\pi_{ij} = \max_l p_l(x)\pi_{lj}\right).$$

Because of (2.3)

$$\Pi_n = \Psi(\Pi_{n-1}).$$

Any convergent subsequence Π_n must converge to a fixed point $\Pi_0 = (\pi^0_{ik})$ of Ψ, that is,

$$\Psi(\Pi_0) = \Pi_0. \tag{2.5}$$

Notice that $\Pi_0 = I$ solves (2.5), and Assumption 1 implies that this is the unique solution of (2.5). Indeed, assuming that $\pi^0_{ik} > 0$ for some $i \neq k$, one obtains the following from the proof of Lemma A.1 in the Appendix:

$$\sum w_k \pi^0_{kk} = \sum w_k \pi^0_{kj} P_k\Big(w_k \pi^0_{kj} p_k(x) = \max_l w_i \pi^0_{ij} p_i(x)\Big)$$
$$> \sum_{k,j} \max_i w_i \pi^0_{ij} P_i\Big(w_k \pi^0_{kj} p_k(x) = \max_l w_l \pi^0_{lj} p_l(x)\Big)$$
$$\geq \max_i \sum_k w_k \pi^0_{ik} \geq \sum_k w_k \pi^0_{kk}.$$

Therefore $\pi^0_{kj} = 0$ for $k \neq j$ and $\Pi_n \to I$, that is, $\widetilde{\delta}_n$ is consistent. □

Even if Assumption 1 holds and the recursive Bayes rule is consistent, the convergence rate of error probabilities is rather slow. Indeed, returning to the example before Theorem 2.1 we see that, under Assumption 1, $\psi'(0) = 1$ and, as $n \to \infty$,

$$\frac{\psi(y_{n-1})}{y_{n-1}} = \frac{y_n}{y_{n-1}} \to 1.$$

Thus convergence of sequence $y_n = \psi(y_{n-1})$ to the fixed point $y = 0$ is slow. In particular this convergence, which can be arbitrarily slow, cannot be of an exponential rate.

Let v_n be any concave sequence of positive reals which converges to zero and for which sequence $(v_{n+1} - v_n)/(v_n - v_{n-1})$ decreases and tends to 1. Then the function ψ linearly connecting points (v_n, v_{n+1}), $n = 1, 2, \ldots$, satisfies all conditions above so that the corresponding sequence $y_n = v_n$ indeed can tend to zero arbitrarily slowly.

The following rates for error probabilities can be derived via Lemma A.2 in the Appendix:

1. If $\psi(y) = y - ay^{p+1} + \varepsilon(y)$, where $p > 0$, $\varepsilon(y) y^{-(p+1)} \to 0$, then under mild additional assumptions about $\varepsilon(y)$ one has

$$y_n \sim [apn]^{-1/p}$$

[see De Bruijn (1958), Section 8.5, for $p = 1$].

2. If $\psi(y) = y - ay^{p+1} / |\log |y||^q + \varepsilon(y)$, then

$$y_n \sim \left[\frac{a}{n(\log n)^q p^{q+1}}\right]^{1/p}.$$

Function ψ of this kind occurs when P has a density of the form

$$p(u) = \frac{\beta \exp\{-\alpha \cosh(\beta u)\}}{2K_0(\alpha)},$$

where $2K_0(\alpha)$, the so-called MacDonald function, represents the normalizing factor [cf. Rukhin (1978)].

3. If $\psi(y) = y - a\exp\{-c|\log y|^p\}y^d|\log y|^{-q}$, $p > 1$, then

$$\log y_n = -\big[\log n/c\big]^{1/p} + \frac{(d-1)\big[\log n/c\big]^{(2-p)/p}}{cp} + O\left(\frac{\log\log n}{(\log n)^{(p-1)/p}}\right).$$

This function ψ with $p = 2, q = 2, c = \sigma^2/(8b^2)$ and $d = \frac{1}{2}$ corresponds to a normal distribution P with mean zero and variance σ^2 [cf. Rukhin and Shi (1992)] or, more generally, a density proportional to $\exp\{-c|u|^{p/(p-1)}\}$.

Indeed the heuristic meaning of Lemma A.2 is that, for the purpose of the asymptotics of sequence y_n, difference equation $y_n = \psi(y_{n-1}) = y_{n-1} - \varphi(y_{n-1})$ can be replaced by differential equation $y' = -\varphi(y)$, where y is treated as a function of large continuous argument x (replacing discrete n). If $1/\varphi$ is an integrable function, the needed solution of this differential equation has the form

$$x = \int_y^\infty \frac{dz}{\varphi(z)}.$$

If $\varphi(y) = y^{p+1}, y(x) = (px)^{-1/p}$ and this corresponds to item 1. When $\varphi(y) = y^{p+1}[\log|y|]^{-q}$, then $x \sim p^{-1}y^{-p}[\log|y|]^{-q}$ so that $y \sim p^{q-1}(\log x)^{-q/p}x^{-1/p}$, which leads to the formula in item 2.

At last, if $\varphi(y) = \exp\{-c|\log y|^p\}y^d|\log y|^{-q}$, then

$$\log y(x) \sim -\Big[\log\frac{x}{c}\Big]^{1/p} + \frac{(d-1)\big[\log x/c\big]^{(2-p)/p}}{cp},$$

which demonstrates the formula given in item 3.

3. Inconsistent recursive Bayes rules. In this section we consider the situation when the underlying distributions have pairwise bounded likelihood ratios so that the recursive Bayes rule is inconsistent.

ASSUMPTION 2. Let $p_1, \dots, p_m$ be probability densities over $(\mathcal{X}, \mu)$ whose ratios are bounded:

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

We assume that b_{ki} are the largest numbers for which these inequalities are true. Then, by comparing the two bounds for $p_i(x)/p_j(x) = (p_i(x)/p_k(x))/(p_k(x)/p_j(x))$, one obtains for $k \ne i, j$

$$b_{kj} > b_{ki}b_{ij}. \tag{3.1}$$

Further, B will denote the matrix (b_{ik}), $w = (w_1, \dots, w_m)^T$ will denote the vector of prior probabilities and $\langle z, w\rangle = \sum_k z_k w_k$ denotes the inner product of two vectors z and w. Also, the notation $z \ge 0$ means that all coordinates of the vector z are nonnegative, and $e = (1, \dots, 1)^T$.

Because of Lemma A.3 in the Appendix, the matrix B (as well as B^T) is nonsingular and there exists a vector $w, w > 0$, such that $B^{-1}w \geq 0$.

THEOREM 3.1. *Under Assumption 2, if*

$$(B^T)^{-1}w \geq 0, \tag{3.2}$$

then, for any n,

$$\rho_n = \sum_k w_k P_k(\tilde{\delta}_n = k) \leq \langle B^{-1}e, w\rangle. \tag{3.3}$$

PROOF. One has with $R_i = \{x\colon \tilde{\delta}_n(x) = i\}$, $i = 1, \ldots, m$,

$$\rho_n = \sum_i \int_{Ri} f_k \, d\mu,$$

where $f_i(x) = w_i p_i(x)$. Therefore conditions of Lemmas A.3 and A.4 in the Appendix are met with $a_{ik} = b_{ik}w_i/w_k$. Also, the determinants of 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 (A.7) is tantamount to (3.2).

Thus Lemma A.4 implies that

$$\rho_n \leq \langle A^{-1}w, e\rangle = \sum_{i,k} \frac{w_i B_{ki}}{|B|} = \left\langle (B^T)^{-1}w, e\right\rangle. \qquad \square$$

Theorem 3.1 is false without condition (3.2). Indeed, by letting w tend to a basis vector e_k one obtains $\rho_n \to 1$ since $\tilde{\delta}_n \to k$. However, $\langle (B^T)^{-1}e_k, e\rangle < 1$.

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

$$\rho_n \leq \frac{w_1(1 - b_{12}) + w_2(1 - b_{21})}{1 - b_{12}b_{21}}, \tag{3.4}$$

provided that

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

Also if $w_1 < b_{21}w_2$, $\rho_n \leq w_2$; and $\rho_n \leq w_1$ if $w_2 < b_{12}w_1$. For instance, if $w_1 = w_2 = \frac{1}{2}$,

$$\rho_n \leq \frac{1 - b_{12} - b_{21}}{2(1 - b_{12}b_{21})}.$$

As an example let P_i be two Bernoulli distributions with probabilities $p_i, i = 1, 2$. Then

$$b_{21} = \frac{p_2}{p_1}, \qquad b_{12} = \frac{1-p_1}{1-p_2}$$

and (3.4) reduces to an equality which shows that the bound (3.3) is sharp.

The next result gives an upper bound for ρ_n for a given value of ρ_{n-1}.

THEOREM 3.2. *Assume that, for a fixed n,*

$$\rho_{n-1} \geq 1 - \min_k w_k. \tag{3.5}$$

Then

$$\rho_n \leq \max_{k \neq j} \left[1 - \frac{w_j b_{jk}(1 - b_{kj})}{1 - b_{jk}b_{kj}} - \frac{(1 - \rho_{n-1})b_{kj}(1 - b_{jk})}{1 - b_{jk}b_{kj}} \right]. \tag{3.6}$$

Also, always

$$\rho_n \leq 1 - \max_{k \neq j} w_k b_{kj}. \tag{3.7}$$

PROOF. Let

$$y_{kj} = P_k\left(\tilde{\delta}_{n-1} = j\right),$$
$$z_{kj} = P_k\left(w_k y_{kj} p_k(x) = \max_l w_l y_{lj} p_l(x)\right),$$

so that $0 \leq y_{kj} \leq 1$, $0 \leq z_{kj} \leq 1$ and, for any j,

$$\sum_k y_{kj} = 1. \tag{3.8}$$

Then

$$\rho_{n-1} = \sum_k w_k y_{kk} \tag{3.9}$$

and

$$\rho_n = \sum_{j,k} w_k y_{kj} z_{kj} = \operatorname{tr}\left(WYZ^T\right),$$

where W is the diagonal matrix with elements (w_k), $Z = (z_{kj})$ and $Y = (y_{kj})$. Thus the problem of maximization of ρ_n for a given value of ρ_{n-1} is a problem of quadratic programming in matrix variables under the additional restraint $BZ \leq E$, where E denotes the matrix with all entries equal to 1. It is clear that

for a fixed Z the search for the extremum in this problem can be restricted to values of Y which are extreme points of the convex set defined by (3.8) and (3.9).

Denote $r = 1 - \rho_{n-1}$. Because of condition (3.5) it suffices to look at matrices Y such that, for some k,

$$w_k y_{kk} = \rho_{n-1} - \sum_{l:l \neq k} w_l = \rho_{n-1} - 1 + w_k = w_k - r$$

and $y_{ii} = 1$ for $i \neq k$. Thus

$$\begin{aligned}
\rho_n &\leq \max_k \left[\sum_i w_i y_{ii} z_{ii} + w_k \sum_{j \neq k} y_{kj} z_{kj} \right] \\
&\leq \max_k \sum_{i:i \neq k} w_i z_{ii} + (w_k - r) z_{kk} + w_k \sum_{j \neq k} y_{kj} z_{kj} \\
&= \max_k \max_{j:j \neq k} \left[\sum_{i \neq k} w_i z_{ii} + w_k - r + r z_{kj} \right] \\
&\leq \max_{k \neq j} \left[\sum_{i \neq k,j} w_i + w_k - r + \max \left[w_j z_{jj} + r z_{kj} \right] \right].
\end{aligned}$$

The maximization problem of a linear function $w_j z_{jj} + r z_{kj}$ under constraint $b_{ij} z_{jj} + b_{ik} z_{kj} \leq 1$, $i = 1, \ldots, m$, is a problem of linear programming. Because of (3.1) all restraints corresponding to the values of $i \neq j$, k are corollaries of the two inequalities for $i = j$ and $i = k$. Therefore (3.3) implies

$$\rho_n \leq \max_{k \neq j} \left[\sum_{i \neq j} w_i - r + \max \left[w_j, r, \frac{w_j(1 - b_{jk}) + r(1 - b_{kj})}{1 - b_{jk} b_{kj}} \right] \right],$$

which easily reduces to (3.6).

Inequality (3.7) follows from (3.6) if ρ_{n-1} in the right-hand side of (3.6) is replaced by ρ_n (which is larger because of Lemma A.1). Since $\max \operatorname{tr}(WYZ^T)$ under conditions (3.8), (3.9) and $BZ \leq E$ is a nondecreasing function of ρ_{n-1}, one can assume (3.5) to be valid when proving (3.7). □

It is easy to see that if, for example, $n = 2$ and $\rho_1 = \langle B^{-1} e, w \rangle$, then condition (3.5) of Theorem 3.2 is met if, for any j,

$$\sum_{k:k \neq j} b_{kj} w_k \geq w_j.$$

Because of (2.1), $\tilde{\delta}_{n-1} = \tilde{\delta}_n$ with probability 1 if, for all i,

$$w_i y_{ii} p_i(x) = \max_k w_k y_{ki} p_k(x)$$

and

$$w_j y_{ij} p_i(k) < \max_k w_k y_{kj} p_k(x),$$

which holds if and only if matrix $Y = (y_{kj} = P_k(\tilde{\delta}_n = j))$ belongs to the set

$$\mathcal{Y}\left\{Y: \min_k \frac{b_{ik} w_i y_{ii}}{w_k y_{ki}} \geq 1, \min_{j:j\neq i} \max_{k:k\neq i} \frac{b_{ki} w_k y_{kj}}{w_i y_{ij}} \geq 1,\ i = 1, \ldots, m\right\}. \tag{3.10}$$

It is easy to see that if Y is in $\mathcal{Y}$, then $\tilde{\delta}_p = \tilde{\delta}_n$ for any p larger than n. In other words no new data can make one change the previous decision.

On the other hand if Y does not belong to $\mathcal{Y}$, then the proof of Lemma A.1 shows that $\rho_{n-1} < \rho_n$. If in addition

$$\left(B^T\right)^{-1} \mathbf{w}_j \geq 0, \tag{3.11}$$

where $\mathbf{w}_j = (w_1 y_{1j}, \ldots, w_m y_{mj})^T$, $j = 1, \ldots, m$, then because of Theorem 3.1

$$\rho_n \leq \sum_j \left\langle B^{-1} e, \mathbf{w}_j \right\rangle = \left\langle B^{-1} e, w \right\rangle.$$

Thus ρ_n can exceed the bound of Theorem 3.1 only if (3.11) does not hold for some j, in which case (3.7) provides a sharp upper bound for ρ_n.

When $m = 2$ set (3.10), which can be identified as a subset of the unit square $\{0 \leq y_{11}, y_{22} \leq 1\}$, has the form

$$\mathcal{Y} = \left\{b_{12} w_1 y_{11} \geq w_2(1 - y_{22}),\ b_{21} w_2 y_{22} \geq w_1(1 - y_{11})\right\}.$$

We illustrate this by returning to the binomial example. Assume that $p_2 < p_1$. If $b_{21} \leq w_1/w_2 \leq 1/b_{12}$, then

$$\tilde{\delta}_1(x) = \begin{cases} 1, & x = 1, \\ 2, & x = 0, \end{cases}$$

so that $y_{11}(1) = p_1 = 1 - q_1$ and $y_{22}(2) = 1 - p_2 = q_2$. If $w_1 q_1 p_1 \leq w_2 q_2 p_2$, then $\tilde{\delta}_2(x_1, x_2) = 1$ iff $x_1 = x_2 = 1$ and $y_{11}(2) = p_1^2$, $y_{22}(2) = 1 - p_2^2$. Induction shows that if

$$w_1 q_1 p_1^n \leq w_2 q_2 p_2^n, \tag{3.12}$$

then $y_{11}(n) = p_1^n$ and $y_{22}(n) = 1 - p_2^n$. Matrix Y reaches set $\mathcal{Y}$ at the first moment (3.12) fails.

The typical behavior of the sequence $(y_{11}(n), y_{22}(n))$ for $w_1 = w_2 = 0.5$ is exhibited in Figures 1 and 2. Set $\mathcal{Y}$ corresponds to the right triangular region, and the evolution of sequence $(y_{11}(n), y_{22}(n))$ is such that it moves east increasing value $y_{11}(n) + y_{22}(n)$ until it reaches set $\mathcal{Y}$, where it stops. The degree of its penetration in this set cannot exceed bound (3.7) of Theorem 3.2.

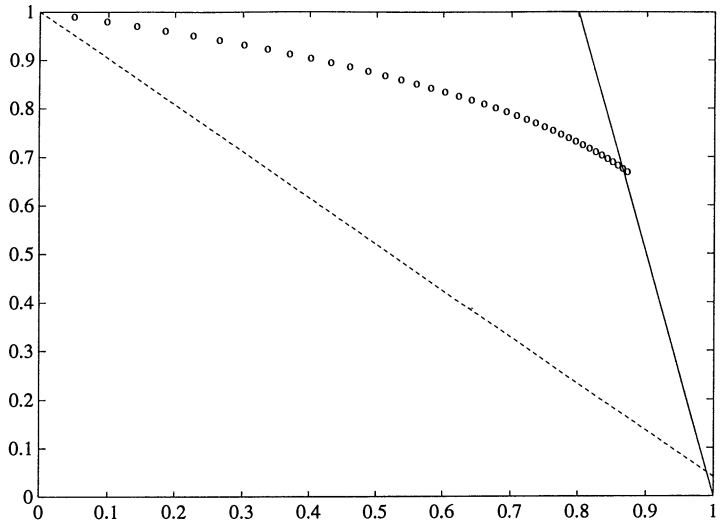

FIG. 1. *Graph of the probabilities of the correct decision for the recursive Bayes procedure for binomial distributions* ($p_1 = 0.1$, $p_2 = 0.01$).

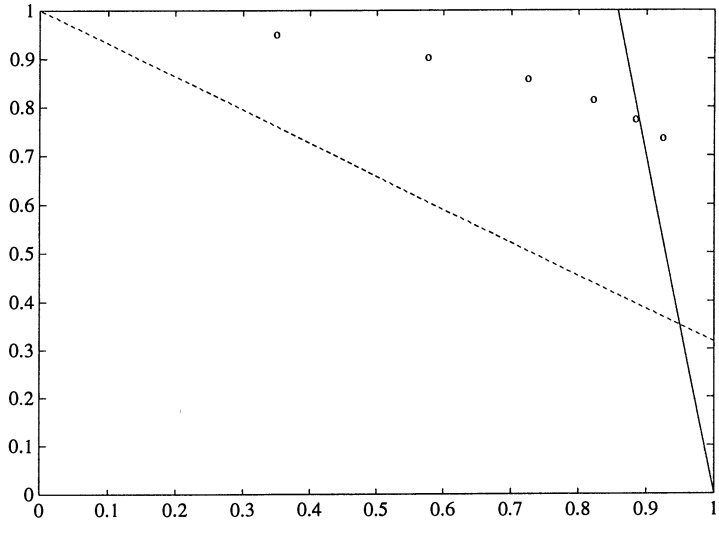

FIG. 2. *Graph of the probabilities of the correct decision for the recursive Bayes procedure for binomial distributions* ($p_1 = 0.35$, $p_2 = 0.05$).

A somewhat different situation occurs, for instance, when P_1 and P_2 are exponential distributions with parameters λ_1 and λ_2, that is, for $i = 1, 2$,

$$p_i(x) = \lambda_i \exp(-\lambda_i x), \qquad x > 0.$$

If $\lambda_1 < \lambda_2$, $b_{12} = \lambda_1/\lambda_2$, $b_{21} = 0$ and set $\mathcal{Y}$ coincides with segment $y_{11} = 1$; $1 - (\lambda_1 w_1)/(\lambda_2 w_2) \leq y_{22} \leq 1$. One can show that

$$y_{11}(n) \to 1, \qquad y_{22}(n) \to 1 - \frac{\lambda_1 w_1}{\lambda_2 w_2}.$$

4. Embedding in an exponential family. In view of results of the previous sections one might seek modifications of the recursive procedure which are asymptotically efficient. Here such a procedure is suggested. The main idea, which goes back to Wald, is to replace the discrete family of distributions $(P_1, \ldots, P_m)$ by a continuous one.

Namely, introduce the exponential family with density

$$p(x, \theta) = \exp\left\{\langle \theta, t(x)\rangle - \chi(\theta)\right\},$$

where

$$t(x) = \big(\log p_1(x), \ldots, \log p_m(x)\big),$$

and the parameter θ varies in the natural parameter space containing unit basis vectors e_i, $i = 1, \ldots, m$. We assume that this is a minimal exponential family. Clearly

$$p(x, e_i) = p_i(x).$$

Let

$$\xi = \nabla\chi(\theta) = E_\theta t(X), \tag{4.1}$$

which is known to be a one-to-one mapping from the natural space. Thus by setting $\theta = \theta(\xi)$ we employ the mean value parameterization of this exponential family with densities

$$f(x, \xi) = p\big(x, (x, \theta(\xi)\big).$$

To estimate a continuous vector parameter ξ, Sakrison (1965) suggested the following recursive procedure:

$$\delta_n = \delta_n(x_n, \delta_{n-1}) = \delta_{n-1} + I^{-1}(\delta_{n-1})\nabla \log f(x_n, \delta_{n-1})/n, \qquad n = 1, 2, \ldots,$$

with some fixed value δ_0. Here

$$I(\xi) = E_\xi \nabla \log f(X, \xi) \nabla^T \log f(X, \xi)$$

is the Fisher information matrix, which is supposed to be nonsingular.

Sakrison's procedure can be interpreted as a sequence of successive approximations of the Newton type to the solution of the likelihood equation. From this point of view it is just a modification of a stochastic approximation procedure which seeks the likelihood equation solution. Remarkably Sakrison's rule possesses all asymptotic optimality properties of the best nonrecursive rules: it is consistent and its asymptotic distribution is normal with the covariance matrix equal to $I^{-1}(\xi)$.

Under our parameterization

$$\nabla \log f(x,\xi) = D\theta\big[t(x) - \xi\big],$$

with $D\theta$ denoting the nonsingular matrix of partial derivatives of the function $\theta = \theta(\xi)$. Because of (4.1) this matrix coincides with the inverse of the Hessian $D_2\chi$ of vector function χ. Since

$$E_\xi\big[t(X) - \xi\big]\big[t(X) - \xi\big]^T = [D_2\chi]^{-1},$$

one obtains

$$I(\xi) = D\theta[D_2\chi]^{-1}[D\theta]^T = D\theta$$

so that

$$\delta_n = \delta_{n-1} + \big[t(x_n) - \delta_{n-1}\big]/n. \tag{4.2}$$

Thus in our situation if $\delta_0 = 0$, procedure δ_n coincides with maximum likelihood estimator $\sum t(x_j)/n$, which is known to be consistent and asymptotically efficient.

We show that the discretized version of δ_n leads to asymptotically efficient classification procedure δ_n^* in our setting:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\sum_i w_i P_i\big(\delta_n^* \neq i\big) &= \lim_{n\to\infty}\frac{1}{n}\log\max_i P_i\big(\delta_n^* \neq i\big)\\ &= \inf_{0<s<1}\max_{i\neq k}\log\int_\chi p_i^{1-s}(x)p_k^s(x)\,d\mu(x) = \alpha.\end{aligned} \tag{4.3}$$

Indeed the lower limit of the maximum of error probabilities logarithms is bounded from below by α according to a classical result of Rényi (1969) [see also Krafft and Puri (1974)]. Thus (4.3) can be taken as the definition of the asymptotic efficiency.

THEOREM 4.1. *Let*

$$\delta_n^* = \arg\min_i K(P_{\delta_n}, P_i),$$

where δ_n is defined by (4.2). *Then δ_n^* is an asymptotically optimal classification procedure in the sense that* (4.3) *holds.*

PROOF. One has, for a fixed i,

$$P_i(\delta_n^* \neq i) = P_i\big(K(P_{\delta_n}, P_i) \geq K(P_{\delta_n}, P_k) \text{ for some } k \neq i\big)$$
$$\leq \sum_{k: k \neq i} P_i\big(K(P_{\delta_n}, P_i) \geq K(P_{\delta_n}, P_k)\big).$$

Therefore

$$\lim_{n \to \infty} \frac{1}{n} \log P_i(\delta_n^* \neq i) \leq \max_{k: k \neq i} \lim_{n \to \infty} \frac{1}{n} \log P_i\big(K(P_{\delta_n}, P_i) \geq K(P_{\delta_n}, P_k)\big).$$

For fixed i and $j, i \neq j$ let $\Xi = \{Q: K(Q, P_i) \geq K(Q, P_j)\}$ be the closed convex set of all probability measures Q which are "closer" in terms of Kullback–Leibler divergence to P_j than to P_i. Obviously P_i does not belong to Ξ; denote by P_i^* its projection onto Ξ. This measure is defined uniquely by the property of minimizing the information number $K(Q, P_i)$ for all Q from Ξ [see Csiszar (1984)]. Also $K(P_i^*, P_i) \leq K(Q, P_i)$ for all Q from Ξ.

Therefore

$$P_i\big(K(P_{\delta_n}, P_i) \geq K(P_{\delta_n}, P_k)\big) \leq P_i\big(K(P_{\delta_n}, P_i) \geq K(P_i^*, P_i)\big).$$

Because of the theorem for the probabilities of large deviations due to Efron and Truax (1968), one has

$$\frac{1}{n} \log P_i\big(K(P_{\delta_n}, P_i) \geq K(P_i^*, P_i)\big)$$
$$\to -K(P_i^*, P_i) = -\inf_{Q: Q \in \Pi} K(Q, P_i)$$
$$= \inf_{s > 0} \log \int_{\chi} p_i^{1-s}(x) p_k^s(x)\, d\mu(x) = \alpha.$$

The last formula follows from Bahadur [(1971), Theorem 4.2]. Now it is immediate that

$$\lim_{n \to \infty} \frac{1}{n} \log P_i(\delta_n^* \neq i) \leq \alpha$$

and δ_n^* is asymptotically efficient as in (4.3). □

APPENDIX

In this section we establish four lemmas needed in Sections 2 and 3.

LEMMA A.1. *Let $\tilde{\delta}_n$ be the recursive Bayes procedure defined by* (2.1), *and set* $\rho_n = \sum_k w_k P_k(\delta_n = k)$. *Then, for any* $n \geq 2$,

$$\rho_n \geq \rho_{n-1}. \tag{A.1}$$

Under Assumption 1 *the inequality in* (A.1) *is strict.*

PROOF. For any positive $s_1, \dots, s_m$,

$$
\begin{aligned}
P_k\bigl(s_k p_k(x) &= \max_i s_i p_i(x)\bigr) \\
&= P_k(R_k) = \int_{R_k} p_k(x)\,d\mu(x) \\
&\geq \int_{R_k} \max_i s_i s_k^{-1} p_i(x)\,d\mu(x) \\
&\geq \max_i s_i s_k^{-1} P_i(R_k).
\end{aligned} \tag{A.2}
$$

Under Assumption 1 inequality (A.2) is strict for $k = q$. Indeed, because of this assumption and (2.2),

$$P_q(R_q) > \max_i s_i s_q^{-1} P_i(R_q) \geq s_r s_q^{-1} P_r(R_q) > 0.$$

Also, (2.1) and (A.2) imply, with $s_{ij} = w_i P_i(\tilde{\delta}_{n-1} = j)$,

$$
\begin{aligned}
\rho_n &= \sum_{j \neq k} P_k\bigl(s_{kj} p_k(x) = \max_i s_{ij} p_i(x)\bigr) s_{kj} \\
&\quad + \sum_j P_j\bigl(s_{jj} p_j(x) = \max_i s_{ij} p_i(x)\bigr) s_{jj} \\
&= \sum_j s_{jj} + \sum_{j \neq k} \Bigl[P_k\bigl(s_{kj} p_k(x) = \max_i s_{ij} p_i(x)\bigr) s_{kj} \\
&\qquad\qquad - P_j\bigl(s_{kj} p_k(x) = \max_i s_{ij} p_i(x)\bigr) s_{jj}\Bigr] \\
&> \sum_j s_{jj} = \rho_{n-1}.
\end{aligned}
$$

The same argument shows that (A.1) is implied by (A.2) without Assumption 1. □

The next result deals with the rate of convergence of error probabilities in a symmetric situation.

LEMMA A.2. *Let $\psi(y)$ be a function such that $\varphi(y) = y - \psi(y)$ is an increasing positive differentiable function on interval $(0, y_1)$. If $y_{n+1} = \psi(y_n), n = 1, \dots,$ and the sequence v_n is defined by the recurrent formula*

$$\int_{v_{n+1}}^{v_n} \frac{dt}{\varphi(t)} = 1, \qquad n = 1, 2, \dots; v_1 = y_1,$$

then, for all $n, y_n \leq v_n$.

If function ψ is concave and increasing and the sequence u_n is defined by the formula

$$u_{n+1} = u_n\psi'(v_n) + v_{n+1} - \psi(v_n), \qquad n = 1, 2, \ldots; u_1 = 0,$$

then $y_n \geq v_n - u_n$.

PROOF. According to the mean value theorem for some $\bar{v}, v_{n+1} < \bar{v} < v_n$,

$$1 = \frac{v_n - v_{n+1}}{\varphi(\bar{v})} \geq \frac{v_n - v_{n+1}}{\varphi(v_n)},$$

so that $\psi(v_n) = v_n - \varphi(v_n) \leq v_{n+1}$. In particular, $y_2 = \psi(y_1) = \psi(v_1) \leq v_2$.

Now if $y_n \leq v_n$ for some n, then

$$y_{n+1} = \psi(y_n) \leq \psi(v_n) \leq v_{n+1}.$$

To prove the second statement of Lemma A.2, assume that

$$y_n \geq v_n - u_n,$$

so that

$$y_{n+1} = \psi(y_n) \geq \psi(v_n - u_n).$$

Because of the concavity of ψ the inequality

$$\psi(v_n) - u_n\psi'(v_n) \geq v_{n+1} - u_{n+1} \tag{A.3}$$

will prove Lemma A.2 by induction; but (A.3) holds by the definition of u_n. □

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

$$\mu\{f_i \neq cf_k\} > 0. \tag{A.4}$$

Suppose that all the ratios f_i/f_k are bounded, that is,

$$a_{ki} \leq \frac{f_k(x)}{f_i(x)} \leq \frac{1}{a_{ik}} \quad \mu\text{-a.s.} \tag{A.5}$$

Moreover, assume that a_{ik} are the largest quantities satisfying (A.5), and let A be the $m \times m$ matrix formed by these numbers, $A = (a_{ik})$.

LEMMA A.3. *Under condition* (A.4) *matrix A is nonsingular.*

PROOF. It follows from (A.5) that, for all i, k, $a_{ik}a_{ki} \leq 1$ and (A.4) implies that, for $i \neq k, a_{ik}a_{ki} \leq 1 = a_{ii}a_{kk}$.

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

$$a_{ik}^{(n)} = a_{ik}^{(n-1)} - \frac{a_{in}^{(n-1)}}{a_{nn}^{(n-1)}}. \tag{A.6}$$

We assume here $a_{nn}^{(n-1)} \neq 0$. Positivity of these and all other coefficients is implied by the following inequality, which can be proven by induction: for $k \neq i, j$,

$$a_{ik}^{(n)} a_{kj}^{(n)} < a_{ij}^{(n)} a_{kk}^{(n)}.$$

The Gauss elimination algorithm shows that because of (A.6) matrix A is nonsingular. Actually it also shows that all principal minors of A are positive. □

One can prove that A^{-1} is an M-matrix [see Bellman (1970), Chapter 16, Example 13], that is, its off-diagonal elements are negative, or for a positive vector w, $(A^T)^{-1}w \geq 0$.

LEMMA A.4. *Under conditions of Lemma* A.3 *let* $R_k, k = 1, \ldots, m$, *be a partition of* $\mathcal{X}$. *Set* $w = (\int f_1 \, d\mu, \ldots, \int f_n \, d\mu)^T$ *and assume that*

$$(A^T)^{-1} e \geq 0. \tag{A.7}$$

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,

$$Az \leq w, \qquad z \geq 0 \tag{A.8}$$

and the determination of the maximum of linear function $\sum_k z_k = \langle z, e \rangle$ over the convex set determined by (A.8) presents a classical problem of linear programming.

Condition (A.7) guarantees that inequality

$$\langle z, e \rangle \leq \langle A^{-1} w, e \rangle$$

is a corollary of (A.8), and this proves Lemma A.4, which also follows from the duality theorem. □

REFERENCES

BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics. Regional Conferences Series in Applied Mathematics*. SIAM, Philadelphia.

BELLMAN, R. (1970). *Introduction to Matrix Analysis*, 2nd ed. McGraw-Hill, New York.

BUCKLEW, J. A. and NEY, P. E. (1991). Asymptotically optimal hypothesis testing with memory constraints. *Ann. Statist.* **19** 982–998.

COVER, T. M. (1969). Hypothesis testing with finite statistics. *Ann. Math. Statist.* **40** 828–835.

COVER, T. M., FREEDMAN, M. A. and HELLMAN, M. E. (1976). Optimal finite memory learning algorithms for the finite sample problem. *Inform. and Control* **30** 49–85.

CSISZAR, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.* **12** 768–793.

DE BRUIJN, N. G. (1958). *Asymptotic Methods in Analysis*. North-Holland, Amsterdam.

EFRON, B. and TRUAX, D. (1968). Large deviations theory in exponential families. *Ann. Math. Statist.* **39** 1402–1424.

HELLMAN, M. E. and COVER, T. M. (1970). Learning with finite memory. *Ann. Math. Statist.* **41** 765–782.

KRAFFT, O. and PURI, M. L. (1974). The asymptotic behavior of the minimax risk for multiple decision problems. *Sankhyā* **36** 1–12.

NEVEL'SON, M. B. and HASMINSKI, R. Z. (1976). *Stochastic Approximation and Recursive Estimation*. Amer. Math Soc., Providence, RI.

RÉNYI, A. (1969). On some problems of statistics from the point of view of information theory. In *Proceedings of the Colloquium on Information Theory* 343–357. Bolyai Math. Soc.

RUKHIN, A. L. (1978). Strongly symmetrical families and statistical analysis of their parameters. *J. Soviet Math.* **9** 59–88.

RUKHIN, A. L. and SHI, J. (1992). Recursive classification procedures: finite time memory and stepwise maximum likelihood procedure. Technical Report 92-03, Univ. Maryland, Baltimore County.

SAKRISON, D. J. (1965). Efficient recursive estimation; application to estimating the parameters of a covariance function. *Internat. J. Engrg. Sci.* **3** 461–483.

YAKOWITZ, S. (1974). Multiple hypothesis testing by finite memory algorithms. *Ann. Statist.* **2** 323–336.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF MARYLAND
BALTIMORE, MARYLAND 21228-5398