The Annals of Statistics
1977, Vol. 5, No. 1, 172–181

ON THE EMPIRICAL BAYES APPROACH TO MULTIPLE DECISION PROBLEMS[1]

BY J. VAN RYZIN AND V. SUSARLA

University of Wisconsin, Madison

In the empirical Bayes approach to multiple decision problems, we obtain theorems and lemmas which can be used to obtain asymptotic optimality and rate results in *any* multiple decision empirical Bayes problem. Applications of these results to a classification problem, a monotone multiple decision, and a selection problem are given. In addition, a special lemma unique to the monotone multiple decision problem gives improved (exact) rate results in that case.

1. Introduction and summary. With $r(G)$ denoting the minimum Bayes risk in a decision problem, Robbins ([7] and [8]) proposed sequences of decision rules, based on data from n independent repetitions of the same decision problem, whose $(n+1)$st stage risk converges to $r(G)$ as $n \to \infty$. Such sequences of rules are called empirical Bayes rules. This paper presents a general empirical Bayes theory for multiple decision problems including rate results (not in [8]) and applies it to two important multiple decision problems: (i) a classification problem, and (ii) a monotone multiple decision problem. For a general discussion of multiple decision problems, we refer the reader to Chapter 6 of Ferguson [2].

In Section 2 we discuss the empirical Bayes multiple decision problem and give general rate results. Lemma 1 is a multiple decision generalization for *any* loss function of the useful Johns–Van Ryzin inequality (Lemma 1 of [5] and [6]) for the 2-decision problem. Section 3 applies these results to a classification problem. Lemma 3 in Section 4 provides a strengthening of the above generalization in the case of a monotone multiple decision problem which allows the *exact* generalizations of all the rate results of [5] and [6].

2. Some general results in the empirical Bayes multiple decision problem. Consider the following multiple decision problem. Let X be an observable random variable with values in a measurable space (χ, β) upon which is defined a σ-finite measure μ. On (χ, β) is defined a family $\mathscr{P} = \{P_\lambda \mid \lambda \in \Omega\}$ of probability measures dominated by μ and indexed by the parameter λ. Let $f_\lambda(x) = (dP_\lambda/d\mu)(x)$ be the μ-density of X when the parameter has value λ. Assume that the statistician is interested in an action space $A = \{a_0, \cdots, a_k\}$ consisting of a finite number of distinct actions. Associated with the problem is a specified loss function

Received August 1974; revised June 1976.

[1] Sponsored in part by the United States Army under Contract No. DAAG-29-75-C0024, by the National Science Foundation Grant NSF Contract No. GP-31931 and by a grant from the University of Wisconsin-Milwaukee.

AMS 1970 *subject classifications*. Primary 62C25; Secondary 62F99.

Key words and phrases. Empirical Bayes procedures, asymptotic optimality, rates of convergence to optimality, classification procedures, a monotone multiple decision problem.

$L(\lambda, a) \geqq 0$ on $\Omega \times A$. Finally, let Λ be an Ω-valued (unobservable) random variable which has a priori distribution G on Ω. The statistician chooses a decision rule $t(x) = (t(0 \mid x), \cdots, t(k \mid x))$, where $t(j \mid x) = \Pr\{\text{taking action } a_j \mid X = x\}$ and whose *Bayes risk* with respect to the a priori distribution G is

$$(1) \qquad r(G, t) = \sum_{j=0}^{k} \int t(j \mid x)[\int L(\lambda, a_j) f_\lambda(x)\, dG(\lambda)]\, d\mu(x) .$$

This risk is minimized by taking $t(j \mid x) = t_G(j \mid x)$, $j = 0, \cdots, k$, where $t_G(j \mid x)$ is the indicator function of the set

$$(2) \qquad S_j = \{x \mid j = \min \{l \mid \Delta_G(a_l, x) = \min_i \Delta_G(a_i, x)\}\}$$

with

$$(3) \qquad \Delta_G(a_j, x) = \int (L(\lambda, a_j) - L(\lambda, a_0)) f_\lambda(x)\, dG(\lambda) .$$

The rule $t_G(x) = (t_G(0 \mid x), \cdots, t_G(k \mid x))$ defined above is thus a *Bayes rule relative to G*, whose risk is

$$(4) \qquad r(G) = r(G, t_G) = \min_t r(G, t) .$$

Following Robbins ([7] and [8]), we seek empirical Bayes procedures not knowing G, which do almost as well as t_G in the $(n+1)$st problem as the number, n, of problems increases. Specifically, let $(X_1, \Lambda_1), (X_2, \Lambda_2), \cdots$, be a sequence of mutually independent pairs of random variables where each Λ_i is distributed as G on Ω and X_i has conditional density f_λ given $\Lambda_i = \lambda$. The empirical Bayes approach attempts to construct a decision procedure concerning Λ_{n+1} (unobservable) at stage $n+1$ based on $X_1, \cdots, X_{n+1}$, the data available at stage $n+1$. The $(\Lambda_1, \cdots, \Lambda_n)$ remain unobservable. Therefore, we consider decision rules of the form

$$(5) \qquad \begin{aligned} t_n(x) &= (t_n(0 \mid x), \cdots, t_n(k \mid x)) , \\ t_n(j \mid x) &= t_n(j \mid x_1, \cdots, x_n; x) , \end{aligned}$$

$j = 0, \cdots, k$ subject to $\sum_{j=0}^{k} t_n(j \mid x) = 1$ a.e. μ (for fixed $x_1, \cdots, x_n$), and take action a_j with probability $t_n(j \mid X_{n+1})$ at stage $n+1$. The risk at stage $n+1$ is given by

$$(6) \qquad r(G, t_n) = \sum_{j=0}^{k} E \int t_n(j \mid x)[\int L(\lambda, a_j) f_\lambda(x)\, dG(\lambda)]\, d\mu(x)$$

where E denotes expectation with respect to the n independent random variables $X_1, \cdots, X_n$ each with common μ-density

$$(7) \qquad f_G(x) = \int f_\lambda(x)\, dG(\lambda) .$$

Since the Bayes procedure $t_G(x)$ achieves the minimum Bayes risk $r(G)$ relative to G, we have $r(G, t_n) \geqq r(G)$, $n = 1, 2, \cdots$. Thus, the nonnegative difference $r(G, t_n) - r(G)$ is used as a measure of optimality of the sequence of procedures $\{t_n\}$ and we say:

DEFINITION 1 (Robbins [8]). The sequence of procedures $\{t_n\}$ is said to be *asymptotically optimal* (a.o.) relative to G if $r(G, t_n) - r(G) = o(1)$ as $n \to \infty$.

DEFINITION 2. The sequence of procedures $\{t_n\}$ is said to be *asymptotically optimal of order* α_n relative to G if $r(G, t_n) - r(G) = O(\alpha_n)$ as $n \to \infty$, where $\lim_{n\to\infty} \alpha_n = 0$.

In the remainder of the paper, we shall construct sequences of empirical Bayes rules for certain multiple decision problems. We shall do this by giving functions $\Delta_{j,n}(x) = \Lambda_{j,n}(x_1, \cdots, x_n; x)$ such that a.e. $(\mu)x$,

$$\Delta_{j,n}(x) \to_P \Delta_G(a_j; x) \qquad \text{as} \quad n \to \infty , \tag{8}$$

where $\to_P$ denotes convergence in probability with respect to the sequence of random variables $\{X_n\}$. The procedure $t_n(x) = (t_n(0 \mid x), \cdots, t_n(k \mid x))$ is then defined by taking $t_n(j \mid x)$ as the indicator function of the set

$$\hat{S}_j = \{x \mid j = \min \{l \mid \Delta_{l,n}(x) = \min_i \Delta_{i,n}(x)\}\} . \tag{9}$$

The following results are for general $\{t_n(x)\}$ of decision procedures.

We state the following lemma which generalizes Lemmas 1 of [5] and [6] to *any* general loss function. (Lemma 3 gives the *exact* generalization for the linear loss function similar to that in [5] and [6].)

LEMMA 1. *Let* $\{t_n(x)\} = \{(t_n(0 \mid x), \cdots, t_n(k \mid x))\}$ *where* $t_n(j \mid x)$ *is the indicator function of the set* $\hat{S}_j$ *in* (9). *Then,*

$$\begin{aligned}
(10)\qquad & 0 \leqq r(G, t_n) - r(G) \\
& \leqq \textstyle\sum_{l=0}^k \int_{S_l} \sum_{m=0}^k (\Delta_G(a_m, x) - \Delta_G(a_l, x)) \Pr\{\Delta_{m,n}(x) \leqq \Delta_{l,n}(x)\}\, d\mu(x) \\
(11)\qquad & \leqq \textstyle\sum_{l=0}^k \int_{S_l} \sum_{m=0}^k |\Delta_G(a_m, x) - \Delta_G(a_l, x)| \Pr\{|\Delta_{m,n}(x) - \Delta_{l,n}(x) \\
& \qquad - (\Delta_G(a_m, x) - \Delta_G(a_l, x))| \geqq |\Delta_G(a_m, x) - \Delta_G(a_l, x)|\}\, d\mu(x) ,
\end{aligned}$$

and for any $\delta > 0$,

$$\begin{aligned}
(12)\qquad r(G, t_n) - r(G) \leqq & \textstyle\sum_{l,m=0}^k \int_{S_m \cup S_l} |\Delta_G(a_m, x) - \Delta_G(a_l, x)|^{1-\delta} E[|\Delta_{m,n}(x) \\
& - \Delta_{l,n}(x) - (\Delta_G(a_m, x) - \Delta_G(a_l, x))|^\delta]\, d\mu(x) .
\end{aligned}$$

PROOF. By definition of t_G and (4), we have $r(G) = \sum_{l=0}^k \int_{S_l} \Delta_G(a_l, x)\, d\mu(x) + \int\int L(\lambda, a_0) f_\lambda(x)\, dG(\lambda)\, d\mu(x)$, Also from (3), (6) and the definition of $t_n(x)$, we have $r(G, t_n) = \sum_{l=0}^k \int \Delta_G(a_l, x) \Pr\{\hat{S}_{l|x}\}\, d\mu(x) + \int\int L(\lambda, a_0) f_\lambda(x)\, dG(\lambda)\, d\mu(x)$, where $\hat{S}_{j|x} = \{(x_1, \cdots, x_n) | j = \min_l \{l \mid \Delta_{l,n}(x_1, \cdots, x_n, x) = \min_i \Delta_{i,n}(x_1, \cdots, x_n, x)\}\}$. Hence, combining these two equalities we have

$$r(G, t_n) - r(G) = \int \textstyle\sum_{l=0}^k \Delta_G(a_l, x)[\Pr\{\hat{S}_{l|x}\} - I_{S_l}]\, d\mu(x) , \tag{13}$$

For x in S_l, the integrand of the rhs of (13) is $\sum_{m=0}^k \Delta_G(a_m, x) \Pr\{\hat{S}_{m|x}\} - \Delta_G(a_l, x) = \sum_{m=0}^k (\Delta_G(a_m, x) - \Delta_G(a_l, x)) \Pr\{\hat{S}_{m|x}\} \leqq \sum_{m=0}^k (\Delta_G(a_m, x) - \Delta_G(a_l, x)) \Pr\{\Delta_{m,n}(x) \leqq \Delta_{l,n}(x)\}$, where the equality follows from the fact that $\hat{S}_{0|x}, \cdots, \hat{S}_{k|x}$ is a partition of χ^n and the inequality is implied by the inequalities $\Delta_G(a_m, x) \geqq \Delta_G(a_l, x)$ for x in S_l and $\Delta_{m,n}(x) \leqq \Delta_{l,n}(x)$ on $\hat{S}_{m|x}$. This completes the proof of the result (10).

For x in S_l, $\Delta_G(a_m, x) \geqq \Delta_G(a_l, x)$. Therefore, for x in S_l and all m,

$$\begin{aligned}&\Pr\{\Delta_{m,n}(x) - \Delta_{l,n}(x) \leqq 0\}\\ &\quad \leqq \Pr\{|\Delta_{m,n}(x) - \Delta_{l,n}(x) - (\Delta_G(a_m, x) - \Delta_G(a_l, x))| \geqq |\Delta_G(a_m, x) - \Delta_G(a_l, x)|\}\,.\end{aligned}$$

Applying this inequality to each summand of the rhs of (10) gives the result (11).

The result (12) follows upon applying a Markov inequality to the rhs of (11) followed by grouping integrals of similar type. □

We now give two general theorems on asymptotic optimality which are applicable to an extensive variety of empirical Bayes multiple decision problems. These two theorems are direct consequences of Lemma 1. Before proving these theorems we mention the following result due to Robbins [8], an alternate proof of which follows immediately from Lemma 1 by using (12), (16), and the bounded convergence theorem.

THEOREM 1 (Robbins [8], Corollary 1). *Let G be such that*

$$\int L(\lambda, a_j)\, dG(\lambda) < \infty\,, \quad j = 0, \cdots, k \tag{14}$$

and let $\{t_n(x)\} = \{(t_n(0\,|\,x), \cdots, t_n(k\,|\,x)\}$ *be defined by* (9) *and satisfy*(8). *Then, the sequence* $\{t_n\}$ *of empirical Bayes rules is* a.o. *relative to* G.

THEOREM 2. *Let $h_j(x, y)$, $j = 1, \cdots, k$ be k real-valued measurable functions on $\chi \times \chi$ such that for $j = 1, \cdots, k$*

$$E[h_j(x, Y)] = \int h_j(x, y) f_G(y)\, d\mu(y) = \Delta_G(a_j, x) \qquad \text{a.e.} \quad \mu \tag{15}$$

and let

$$\Delta_{j,n}(x) = \frac{1}{n} \sum_{i=1}^n h_j(x, X_i)\,, \quad j = 1, \cdots, k\,; \qquad \Delta_{0,n}(x) \equiv 0 \quad \text{a.e.} \quad \mu\,. \tag{16}$$

Assume (14) *holds and that for $l, j = 0, \cdots, k$ and some δ in* $(0, 2)$,

$$\int |\Delta_G(a_j, x)|^{1-\delta} \sigma_l^{\delta}(x)\, d\mu(x) < \infty\,, \tag{17}$$

where $\sigma_j^2(x) = \mathrm{Var}\,\{h_j(x, Y)\}$. *Then the sequence of empirical Bayes rules defined by* (9) *with* $\Delta_{j,n}(x)$ *as in* (16) *is* a.o. *of order* $n^{-\delta/2}$ *relative to* G.

PROOF. This result is a consequence of (11) and the series of inequalities

$$\begin{aligned}&E[|\Delta_{m,n}(x) - \Delta_{l,n}(x) - (\Delta_G(a_m, x) - \Delta_G(a_l, x))|^{\delta}]\\ &\quad \leqq \max\{1, 2^{\delta-1}\} E\{[|\Delta_{m,n}(x) - \Delta_G(a_m, x)|^{\delta}] + E|\Delta_{l,n}(x) - \Delta_G(a_l, x)|^{\delta}]\}\\ &\quad \leqq \max\{1, 2^{\delta-1}\}\{\sigma_m^{\delta}(x) + \sigma_l^{\delta}(x)\}\,.\end{aligned}$$

□

If it is not possible to obtain unbiased estimates of $\Delta_G(a_j, x)$ as in (15) and (16), then the following theorem whose proof is similar to that of Theorem 1 will be found useful.

THEOREM 3. *Let* $\{h_{j,n}(x, y)\}$, $j = 1, \cdots, k$, *be k real-valued sequences of measur-*

able functions on $\chi \times \chi$ *and let*

$$(18)\qquad \Delta_{j,n}(x) = \frac{1}{n}\sum_{i=1}^{n} h_{j,n}(x, X_i)\,, \quad j = 1, \cdots, k\,; \qquad \Delta_{0,n}(x) \equiv 0 \quad \text{a.e.} \quad \mu\,.$$

Assume (14) *holds and for* $l, j = 0, \cdots, k$ *and some* δ *in* $(0, 2)$ *that*

$$(19)\qquad \int |\Delta_G(a_l, x)^{1-\delta}\sigma_{l,n}^{\delta}(x)\, d\mu(x) \leqq c_G a_n\,, \qquad \sigma_{l,n}^2(x) = \operatorname{Var}(h_{j,n}(x, Y))$$

and

$$(20)\qquad \int |\Delta_G(a_j, x)|^{1-\delta} b_{l,n}^{\delta}(x)\, d\mu(x) \leqq c_G' a_n'\,; \qquad b_{l,n}(x) = |Eh_{j,n}(x, Y) - \Delta_G(a_j, x)|\,.$$

Then the sequence of rules defined by (9) *with* $\Delta_{j,n}(x)$ *as in* (18) *is* a.o. *of order* $\alpha_n = \max\{n^{-\delta/2}a_n, a_n'\}$.

It is now clear from the above theorems that to construct empirical Bayes rules we need merely find the functions $h_j(x, y)$ or the sequences of functions $\{h_{j,n}(x, y)\}$ in Theorems 1 and 2. If we can then verify the conditions (8) and (14) we obtain asymptotic optimality via Lemma 1. If we can further verify (16) in Theorem 1 or (19) and (20) in Theorem 2, we then have a result on the rate of convergence to optimality. We shall do this in Sections 3 and 4 to illustrate the use of these general theorems in a classification problem and a monotone multiple decision problem. Of course, other applications are possible.

3. A classification problem. Consider now the following classification problem. Let $\{f_0(x), \cdots, f_k(x)\}$ be a set of $k+1$ known μ-densities, $\Omega = \{0, \cdots, k\}$ (the parameter space of class labels) and $A = \{a_0, \cdots, a_k\}$ the action space wherein action a_j represents classifying the observed random variable X as coming from the distribution with density f_j (that is, saying $\Lambda = j$). Furthermore, let $0 \leqq L(i, a_j) = 1_{ij} < \infty$, $i, j = 0, \cdots, k$ be the loss for misclassification of X as coming from f_j when in fact X came from f_i.

In the empirical Bayes setting we are confronted with a sequence of such classification problems and wish to decide about Λ_{n+1} (unobserved) based on previous observations $X_1, \cdots, X_n$ and the current observation X_{n+1} which is to be classified. To solve the empirical Bayes problem we must estimate (see (3)) for $j = 1, \cdots, k$

$$(21)\qquad \Delta_G(a_j, x) = \sum_{i=0}^{k} (1_{ij} - 1_{i0}) f_i(x) g_i\,,$$

where for $i = 0, \cdots, k$, $g_i = \Pr\{\Lambda = i\} \geqq 0$, $\sum_{i=0}^{k} g_i = 1$. Note that $G = (g_0, \cdots, g_k)$ is the unknown a priori distribution on $\Omega = \{0, \cdots, k\}$. To estimate (21) we construct the function $h_j(x, y)$ of Theorem 1 as follows. Assume there exist functions $\xi_j(y)$ (see discussion below), such that for $i, j = 0, \cdots, k$

$$(22)\qquad E_i\xi_j(Y) = \int \xi_j(y) f_i(y)\, d\mu(y) = 1 \quad \text{if } i = j$$
$$= 0 \quad \text{if } i \neq j\,.$$

Then define in Theorem 2 for $j = 1, \cdots, k$

$$h_j(x, y) = \sum_{i=0}^{k} (1_{ij} - 1_{i0}) f_i(x) \xi_i(y) \tag{23}$$

and

$$\Delta_{j,n}(x) = \frac{1}{n} \sum_{i=1}^{n} h_j(x, X_i) = \sum_{i=0}^{k} (1_{ij} - 1_{i0}) f_i(x) \bar{\xi}_i\,, \tag{24}$$

where $\bar{\xi}_i = n^{-1} \sum_{\nu=1}^{n} \xi_i(X_\nu)$. We can now state

THEOREM 4. *In the classification problem, the sequence of empirical Bayes rules* $\{t_n(x)\}$ *defined by* (9), (23) *and* (24) *is* a.o. *relative to any* $G = (g_0, \cdots, g_k)$ *if* (22) *holds, and is* a.o. *of order* $n^{-\frac{1}{2}}$ *relative to any* $G = (g_0, \cdots, g_k)$ *if for* $i, j = 0, \cdots, k$ (22) *holds and*

$$E_i \xi_j^2(Y) = \int \xi_j^2(y) f_i(y)\, d\mu(y) < \infty\,. \tag{25}$$

We note that asymptotic optimality for this problem was first shown by Robbins ([8], Section 7), but the rate result is new.

REMARKS. A set of appropriate functions $\xi_j(y)$ always exist and are easily constructable if $\{f_0, \cdots, f_k\}$ is a linearly independent set of functions in $L_2(\mu)$. In matrix notation, the functions are constructed by $\boldsymbol{\xi}^*(y) = B^{-1}\mathbf{f}(y)$ where $\boldsymbol{\xi}^*(y) = (\xi_0^*(y), \cdots, \xi_k^*(y))'$, $\mathbf{f}(y) = (f_0(y), \cdots, f_k(y))'$ ($'$ denotes transpose), B^{-1} is the inverse of a $(k + 1) \times (k + 1)$ matrix B whose (i, j)th element is $b_{ij} = \int f_i(x) f_j(x)\, d\mu(x)$, $i, j = 0, \cdots, k$, and (22) follows from the easily verified fact that $\boldsymbol{\xi}^*(y)$ is an unbiased estimator $\mathbf{g}$. Observe that the ξ_j^* functions so defined are the dual basis for the algebraic conjugate of the linear subspace of $L_2(\mu)$ spanned by $\{f_0, \cdots, f_k\}$, as discussed in Van Ryzin ([11], Section 3) or Robbins ([8], Section 7). However, the above matrix form is not discussed specifically in either reference. Finally, the invertibility of B follows from the linear independence of the set $\{f_0, \cdots, f_k\}$ in $L_2(\mu)$. (See, e.g., Taylor [9], Theorem 1.61-B.)

Hudimoto ([3], Section 6) has given another method for estimating $\mathbf{g} = (g_0, \cdots, g_k)'$: his method (in our notation) is to take $\boldsymbol{\xi}'(y) = (\xi_0'(y), \cdots, \xi_k'(y))' = A^{-1}\mathbf{F}(y)$, where $\mathbf{F}(y) = (F_0(y), \cdots, F_k(y))'$, F_i the distribution function associated with the density f_i, $i = 0, \cdots, k$ and A the $(k + 1) \times (k + 1)$ matrix whose (i, j)th element $a_{ij} = \int F_i(x)\, dF_j(x)$, $i, j = 0, \cdots, k$ which we assume to be invertible (see Lemma 2 below). Again, condition (22) follows by the unbiasedness of $\boldsymbol{\xi}'(y)$. For matrix conditions for invertibility of A for $k = 1, 2, 3$ ($r = 2, 3, 4$ in his paper) see Hudimoto ([3], Sections 2 and 6). However, he gave no general necessary and sufficient condition for invertibility as is given in the following lemma. Let P_i' be the unique (up to equivalence) Lebesgue-Stieltjes measure corresponding to F_i, $i = 0, \cdots, k$ and $\mu' = \sum_{i=0}^{k} P_i'$.

LEMMA 2. *The matrix* $A = (a_{ij}) = (\int F_i(x)\, dF_j(x))$ *is invertible if and only if* $(F_0, \cdots, F_k)$ *are linearly independent in* $L_2(\mu')$.

PROOF. The proof is a direct consequence of the following set of equivalent statements.

$$\sum_{i=0}^k \alpha_i F_i(x) = 0 \quad \text{a.e.} \quad \mu'$$
$$\Leftrightarrow \sum_{i=0}^k \alpha_i a_{ij} = 0 \quad \text{for} \quad j = 0, \cdots, k$$
$$\Leftrightarrow A'\boldsymbol{\alpha} = \mathbf{0}\,, \quad \boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_k)'\,, \quad \mathbf{0} = k+1 \text{ fold zero vector.} \ \square$$

Note that from Lemma 2 and the theorem of Yakowitz and Spragins [14], we see that the invertibility of A, the determinant condition of Teicher [10], identifiability of $\mathbf{g}$ and linear independence of $(F_0, \cdots, F_k)$ are all equivalent statements.

In passing, we also observe that by selecting the functions $\xi_j(Y)$ such that $\max_j |\xi_j(y)|$ is bounded a.e. μ, one can use the methods of Hudimoto ([3], Section 7) to show that for any $\varepsilon > 0$ and any a priori distribution $(g_0, \cdots, g_k)$ on $\Omega = \{0, \cdots, k\}$, there exist positive constants c_1 and c_2 such that

$$P\{r(G, t_n) - r(G) \geqq \varepsilon\} \leqq c_1 e^{-c_2 n}\,; \tag{26}$$

where $r(G, t_n) = \sum_{i,j=0}^k 1_{ij} g_i \int t_n(j\,|\,x) f_i(x)\, d\mu(x)$ (see (1)) is the conditional risk of misclassification in the $(n+1)$st problem given $X_1, \cdots, X_n$ using the empirical Bayes rule $t_n(X_{n+1})$ at stage $n + 1(t_n(x)$ defined by (9), (23) and (24)). Thus (26) says that the probability given the past n observations of the excess risk at stage $n+1$, using the empirical Bayes rule with *unknown* prior over the optimal risk with *known* prior, being arbitrarily small approaches zero at an exponential rate as n, the number of problems, increases. Finally, it is always possible to select the ξ_j such that $\max_j |\xi_j(Y)|$ is essentially bounded by taking the ξ_j as the ξ_j' functions above or as the ξ_j^* functions above with $\mu = \sum_{i=0}^k P_i$.

4. A monotone multiple decision problem. Consider the empirical Bayes multiple decision problem whose component problem is given as follows: Let $\Omega = (-\infty, \infty)$, $\chi = R$, $\mathscr{B} =$ Borel σ-field in R, and $-\infty = \lambda_{-1} < \lambda_0 < \cdots < \lambda_{k-1} < \lambda_k = \infty$ be known. Let action a_j correspond to deciding "the value of $\Lambda = \lambda$ is in the interval $[\lambda_{j-1}, \lambda_j]$, $j = 0, \cdots, k$." As a loss function, we take $L(\lambda, a_j)$ such that for $j = 0, \cdots, k-1$,

$$\begin{aligned} L(\lambda, a_{j+1}) - L(\lambda, a_j) &= c(\lambda_j - \lambda) \\ L(\lambda, a_0) &= 0 && \text{if} \quad \lambda \leqq \lambda_0 \\ &= c \sum_{i=1}^j (\lambda - \lambda_{i-1}) && \text{if} \quad \lambda_{j-1} < \lambda \leqq \lambda_j \end{aligned} \tag{27}$$

where $c\ (> 0)$ is a known constant. For $k = 1$, this reduces to the loss function considered by Johns [4] which is commonly used for empirical Bayes two-action problems. Without loss of generality we take $c = 1$ in what follows. Since $L(\lambda, a_{j+1}) - L(\lambda, a_j) \geqq$ or $\leqq 0$ according as $\lambda \leqq \lambda_j$ or $\lambda \geqq \lambda_j$, the decision problem is monotone (see, e.g., Ferguson [2], Definition 1, page 285). Assume that $E[|\Lambda|] < \infty$ so that the optimal Bayes risk $r(G)$ is finite.

Define, for $j = 0, \cdots, k$,

$$t_G(j\,|\,x) = I_{S_j} = I_{[\lambda_{j-1} f(x) < g(x) \leqq \lambda_j f(x)]}\,, \tag{28}$$

where

$$f(x) = \int f_\lambda(x)\, dG(\lambda) \qquad \text{and} \qquad g(x) = \int \lambda f_\lambda(x)\, dG(\lambda)\,. \tag{29}$$

To define $\{t_n(x)\}$ in a natural way, let $\hat{\alpha}_i$ be an estimate (based on $X_1, \cdots, X_n$) of

$$\alpha_i(x) = \lambda_i f(x) - g(x) = \Delta_G(a_{i+1}, x) - \Delta_G(a_i, x) \tag{30}$$

such that $\hat{\alpha}_i(x)$ is nondecreasing in i with $-\hat{\alpha}_{-1}(x) = \hat{\alpha}_k(x) = \infty$. Using $\hat{\alpha}_i$ $(i = -1, \cdots, k)$, let, for $j = 0, \cdots, k$,

$$t_n(j \mid x) = I_{\hat{S}_j} = I_{[\hat{\alpha}_{j-1}(x) < 0 \leqq \hat{\alpha}_j(x)]}\,. \tag{31}$$

LEMMA 3. *Let* $\{t_n(x)\} = \{(t_n(0 \mid x), \cdots, t_n(k \mid x))\}$ *be defined by* (31), *where* $\hat{\alpha}_i(x)$ *is increasing in* i. *Then*

$$\begin{aligned} 0 \leqq r(G, t_n) - r(G) = \textstyle\sum_{l=0}^{k} \int_{S_l} \{&\textstyle\sum_{m=0}^{l-1} |\alpha_m(x)| \Pr\{\hat{\alpha}_m(x) \geqq 0\} \\ &+ \textstyle\sum_{m=l}^{k-1} |\alpha_m(x)| \Pr\{\hat{\alpha}_m(x) < 0\}\}\, d\mu(x) \end{aligned} \tag{32}$$

and for $\delta > 0$,

$$0 \leqq r(G, t_n) - r(G) \leqq \textstyle\sum_{m=0}^{k-1} \int |\alpha_m(x)|^{1-\delta} E[|\alpha_m(x) - \hat{\alpha}_m(x)|^\delta]\, d\mu(x)\,. \tag{33}$$

PROOF. By (4), (6), (28) and (31), we have

$$r(G, t_n) - r(G) = \textstyle\sum_{l=0}^{k} \int \gamma_l(x)\{\Pr\{\hat{\alpha}_{l-1}(x) < 0 \leqq \hat{\alpha}_l(x)\} - I_{S_l}\}\, d\mu(x)\,, \tag{34}$$

where $\gamma_l(x) = \int L(\lambda, a_l) f_\lambda(x)\, dG(\lambda)$. For x in S_l, the integrand in (34) is

$$\begin{aligned} &\textstyle\sum_{m=0}^{k} \gamma_m(x) \Pr\{\hat{\alpha}_{m-1}(x) < 0 \leqq \hat{\alpha}_m(x)\} - \gamma_l(x) \\ &\quad = \textstyle\sum_{m=0}^{l-1} \gamma_m(x)(\Pr\{\hat{\alpha}_m(x) \geqq 0\} - \Pr\{\hat{\alpha}_{m-1}(x) \geqq 0\}) \\ &\qquad + \textstyle\sum_{m=l+1}^{k} \gamma_m(x)(\Pr\{\hat{\alpha}_{m-1}(x) < 0\} - \Pr\{\hat{\alpha}_m < 0\}) \\ &\qquad - \gamma_l(x)(\Pr\{\hat{\alpha}_l(x) < 0\} + \Pr\{\hat{\alpha}_{l-1}(x) \geqq 0\})\,. \end{aligned}$$

Rearranging this last expression according to terms involving $\Pr\{\hat{\alpha}_m(x) < 0\}$ and $\Pr\{\hat{\alpha}_m(x) \geqq 0\}$ and then recognizing that $\gamma_{m-1}(x) - \gamma_m(x) = -\alpha_{m-1}(x) > 0$ or $\leqq 0$ according as $m < l$ or $m \geqq l$ for x in S_l, the result follows.

The second result follows from the first result by using an argument similar to that given in Lemma 1. □

REMARKS. Lemma 3 is a strengthened version of Lemma 1 for the monotone multiple decision problem in the following sense: inequality (10) of Lemma 1 becomes an exact equality in (32). Moreover, (12) involves all the possible differences $|\Delta_G(a_m, x) - \Delta_G(a_l, x)|$ whereas (33) involves only terms of the type $|\Delta_G(a_m, x) - \Delta_G(a_{m-1}, x)|$. Also, note that the expression for the difference $r(G, t_n) - r(G)$ is an *exact* expression like (10) of [5] which is a main step of [5] for getting the exact rate results therein. Since Lemma 3 corresponds exactly to Lemma 1 and (10) of [5], it is obvious that all the rate results like $O(n^{-(1-\varepsilon)})$, $\varepsilon > 0$ of [5] and [6] (including even the exact rate result $O(n^{-1})$ in the geometric and Poisson cases) can be carried over to the case of $(k + 1)$ $(\geqq 3)$ actions with the obvious modifications in their statements and their proofs under certain moment conditions on the class of prior distributions.

5. Another brief example. The results of Section 2 can be applied (see [13], for example) when the component problem is a selection problem as described in Deely [1]. In a component selection problem, based on independent (observable) $X_1, \cdots, X_k$ random variables distributed as $f_{\lambda_1}, \cdots, f_{\lambda_k}$ respectively, the decision problem is to select that index j for which $\lambda_j = \max_i \lambda_i$ when the loss function is given by $L((\lambda_1, \cdots, \lambda_k), a_j) = \max_i \lambda_i - \lambda_j$ where $(\lambda_1, \cdots, \lambda_k) \sim G$ and a_j is the action deciding that $\lambda_j = \max_i \lambda_i$. Using the results of Section 2, it can be shown that one can obtain empirical Bayes procedures (when the component problem is a selection problem) which are a.o. $O(n^{-\frac{1}{2}})$ either when $f_\lambda = \lambda^u \beta(\lambda) h(u)$, $u = 0, 1, 2, \cdots$ wrt counting measure μ on $\{0, 1, 2, \cdots\}$ or when $f_\lambda = e^{-\lambda x}\beta(\lambda)h(u)I_{[u>a]}$ wrt Lebesgue measure μ on $(R, \mathscr{B})$ under some reasonable conditions on G and H. Deely [1] was the first to consider empirical Bayes selection problems using the loss L given above; and considered the two situations: (i) $G(\lambda_1, \cdots, \lambda_k) = \prod_{i=1}^k G_i(\lambda_i)$ where G_i are of a known parametric form, and (ii) a nonparametric case. The nonparametric case was studied and extended by Van Ryzin [12].

In conclusion, we point out that Theorems 1, 2 and 3 are most useful since they are applicable not only in the three examples pointed out here, but also in *any* multiple decision empirical Bayes problem.

REFERENCES

[1] DEELY, J. (1965). Multiple decision procedures from an empirical Bayes approach. Ph. D. thesis, Purdue Univ.

[2] FERGUSON, T. (1967). *Mathematical Statistics: A Decision Theory Approach.* Academic Press, New York.

[3] HUDIMOTO, H. (1968). On the empirical Bayes procedure (1). *Ann. Inst. Statist. Math.* **20** 169–185.

[4] JOHNS, M. V., JR. (1957). Nonparametric empirical Bayes procedures. *Ann. Math. Statist.* **28** 649–669.

[5] JOHNS, M. V., JR. and VAN RYZIN, J. (1971). Convergence rates for empirical Bayes two-action problems I. Discrete case. *Ann. Math. Statist.* **42** 1521–1539.

[6] JOHNS, M. V., JR. and VAN RYZIN, J. (1972). Convergence rates for empirical Bayes two-action problem II. Continuous case. *Ann. Math. Statist.* **43** 934–947.

[7] ROBBINS, H. (1955). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Prob.* 157–163, Univ. of California Press.

[8] ROBBINS, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20.

[9] TAYLOR, A. (1958). *Introduction to Functional Analysis.* Wiley, New York.

[10] TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34** 1265–1269.

[11] VAN RYZIN, J. R. (1966). The compound decision problem with $m \times n$ finite loss matrix. *Ann. Math. Statist.* **37** 412–424.

[12] VAN RYZIN, J. (1970). On some nonparametric empirical Bayes multiple decision problems. *Proc. First Internat. Symp. Nonparametric Techniques in Statist. Inference.* (M. L. Puri, ed.). 585–603, Cambridge Univ. Press.

[13] VAN RYZIN, J. (1970). Empirical Bayes procedures for multiple decision problems. Technical Report No. 249, Dept. of Statistics, Univ. of Wisconsin, Madison.

[14] YAKOWITZ, S. J. and SPRAGINS, J. D. (1968). On identifiability of finite mixtures. *Ann. Math. Statist.* **39** 209-214.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1210 WEST DAYTON STREET
MADISON, WISCONSIN 53706