

NEARLY-OPTIMAL SEQUENTIAL TESTS FOR FINITELY MANY PARAMETER VALUES¹

BY GARY LORDEN

California Institute of Technology

Combinations of one-sided sequential probability ratio tests (SPRT's) are shown to be "nearly optimal" for problems involving a finite number of possible underlying distributions. Subject to error probability constraints, expected sample sizes (or weighted averages of them) are minimized to within $o(1)$ asymptotically. For sequential decision problems, simple explicit procedures are proposed which "do exactly what a Bayes solution would do" with probability approaching one as the cost per observation, c , goes to zero. Exact computations for a binomial testing problem show that efficiencies of about 97% are obtained in some "small-sample" cases.

1. Introduction. The central idea of the present investigation is that, just as the SPRT is optimal and Bayes for testing 2 densities, combinations of one-sided SPRT's are "nearly optimal" and "nearly Bayes" for arbitrary testing problems involving $s \geq 2$ densities. The theorems supporting this contention are asymptotic, of the type originated by Chernoff (1959), wherein one assumes a cost per observation, c , which is made to go to zero. The present results, though limited to the finite case, are considerably sharper than their precursors in Schwarz (1962), Kiefer and Sacks (1963), Lorden (1967) and Wong (1968), all of which are asymptotically unaffected by insertion of arbitrary constant factors in the likelihood ratio inequalities defining the tests. In contrast with this, Theorems 1, 2, 3 and 4 below hinge on the asymptotic determination of the "right" critical value for each of the likelihood ratios. The latter are based on the determination in Lemma 1 of numbers $L(i, j)$ that yield optimal critical values for one-sided SPRT's in a Bayes context. As a consequence, for an arbitrary a priori distribution a family of multiple-SPRT's, $\{\delta(c)\}$, is specified and shown to attain the Bayes risk to within $o(c)$ (Theorem 1) and to "do exactly what a Bayes solution, $\delta^*(c)$, would do" with probability approaching one as $c \rightarrow 0$ (Theorem 3).

The most accurate asymptotic results previously obtained were those of Lorden (1967) which showed that the rule "stop when the a posteriori risk is less than a constant times c " is Bayes to within $O(c)$ in a general context and that combinations of SPRT's achieve the same results in the present finite case. Chernoff's and Kiefer and Sacks' investigations showed that the optimal Bayes risk itself is of order $c \log c^{-1}$, so that the present results improve the known efficiency of asymptotically Bayes tests from $1 - O((\log c^{-1})^{-1})$ to $1 - o((\log c^{-1})^{-1})$.

Received December 1974; revised June 1976.

¹ Research supported by the National Science Foundation under Grant GP-37819.

AMS 1970 subject classifications. Primary 62L10; Secondary 62F20.

Key words and phrases. Sequential probability ratio test, Bayes solution, asymptotic optimality.

Theorem 4 and its corollary are of a somewhat different (non-Bayes) character and for the problem of choosing which one of $s > 2$ densities is true provide an asymptotic generalization of the optimality property of the SPRT. A natural procedure using a combination of $\binom{s}{2}$ SPRT's is shown to minimize the s expected sample sizes to within $o(1)$ as the error probabilities go to zero. This result is in the same spirit as Wong (1968) and Lorden (1972), which were concerned with pointwise minimization of expected sample sizes for testing a continuous parameter θ subject to specified upper bounds on error probabilities (tending to zero). A result similar to those of the present paper appeared in Lorden (1976), concerned with the problem of minimizing a single expected sample size subject to two error probability bounds (which in some cases leads to minimization of the maximum expected sample size, the so-called Kiefer-Weiss problem).

The sharpness of the present asymptotic results tends to instill the hope that the procedures $\delta(c)$, or variants of them, will turn out to have high efficiency in practical, small-sample problems. A preliminary investigation was carried out in a simple case of testing two values of the binomial parameter, p , with a third value as "indifference point." The performance of the "pentagon-shaped" continuation regions of $\delta(c)$ is compared with that of Bayes solutions, $\delta^*(c)$, in Section 4. The results are encouraging.

Properties of the numbers $L(i, j)$, which play a fundamental role in the present investigation, are discussed in Section 3. Theorem 5 of that section and recent work by Siegmund (1975) and Lai and Siegmund (1976) suggest that these numbers are also fundamental in analyzing the effect of "excess over the boundaries" or "overshoot" in sequential analysis.

2. The general case. Independent and identically distributed random variables $X_1, X_2, \dots$, are observed sequentially, taking values in a measurable space on which given probability measures, $P_1, \dots, P_s$, are distinct and mutually absolutely continuous. Let $f_1, \dots, f_s$ denote the respective densities of $P_1, \dots, P_s$ with respect to a dominating sigma-finite measure, e.g., P_1 . Each observation costs c ($0 < c < 1$) and k terminal decisions ($k \geq 2$) are available. Let $W(i, d)$ denote the loss incurred when f_i is true and decision d is chosen, for $i = 1, \dots, s$, $d = 1, \dots, k$. Without loss of generality, assume that $\min_d W(i, d) = 0$ for all i . Decision d is said to be a correct decision for f_i if $W(i, d) = 0$. Thus, there is at least one correct decision for each f_i . To avoid trivialities, assume that $W(i, d) = W(i, d')$ for all i only if $d = d'$ and also that no decision is correct for all f_i .

Let E_i denote expectation under P_i , $i = 1, \dots, s$. Assume that

$$(1) \quad E_i[\log(f_i(X_1)/f_j(X_1))]^2 < \infty \quad i, j = 1, \dots, s.$$

Thus the information numbers

$$I(i, j) = E_i \log(f_i(X_1)/f_j(X_1)) \quad i, j = 1, \dots, s,$$

are finite.

Define the likelihoods

$$f_{in} = f_i(X_1) \cdots f_i(X_n) \quad i = 1, \dots, s, n = 1, 2, \dots,$$

setting $f_{i0} \equiv 1$ for all i . Given an *a priori* distribution λ_0 assigning weights $\lambda_0(1), \dots, \lambda_0(s)$ to $f_1, \dots, f_s$, define λ_n , the *a posteriori* distribution after n observations, by

$$(2) \quad \lambda_n(i) = \frac{\lambda_0(i)f_{in}}{\sum_{j=1}^s \lambda_0(j)f_{jn}} \quad i = 1, \dots, s; n = 1, \dots.$$

For any distribution, λ , on $\{f_1, \dots, f_s\}$, the risk of decision d is

$$(3) \quad r(\lambda, d) = \sum_{i=1}^s \lambda(i)W(i, d).$$

The *stopping risk* when λ is the *a priori* or a *a posteriori* distribution is

$$r(\lambda) = \min_{1 \leq d \leq k} r(\lambda, d).$$

A test δ consists of an extended stopping time, N , allowed to take the values $0, 1, \dots$, and ∞ , and a measurable terminal decision rule, D ; both may be randomized. The *integrated risk* of the test $\delta = (N, D)$ with respect to the *a priori* distribution λ_0 and cost c is

$$r_c(\lambda_0, \delta) = \sum_{i=1}^s \lambda_0(i) \{cE_i N + \sum_{d=1}^k W(i, d)P_i(D = d)\}.$$

For given λ_0 and c , the *Bayes risk*, $r_c(\lambda_0)$, is the minimum integrated risk over all tests. Fix λ_0 . Let $\{\delta^*(c)\}$ denote a family of *Bayes solutions*, i.e., tests attaining the minimum integrated risk, for $0 < c < 1$. Assume that $N^*(c)$ stops as soon as $r(\lambda_n)$ is equal to the Bayes risk for λ_n and $D^*(c)$ chooses a decision d attaining $\min_d r(\lambda_n, d)$.

Now consider the *a posteriori space*, the set of all $\lambda = (\lambda(1), \dots, \lambda(s))$ probabilities on $\{f_1, \dots, f_s\}$. Define the *Bayes stopping region*

$$S_c^* = \{\lambda \mid r(\lambda) = r_c(\lambda)\}.$$

Then $N^*(c)$ is the infimum of $n \geq 0$ such that $\lambda_n \in S_c^*$. Define the numbers

$$(4) \quad L(i, j) = \exp(-\sum_{n=1}^{\infty} n^{-1} [P_j(f_{jn} < f_{in}) + P_i(f_{in} \leq f_{jn})])$$

$$i \neq j = 1, \dots, s,$$

setting $L(i, i) \equiv 1$, $i = 1, \dots, s$, for convenience in the definition of $\delta(c)$, (12). Since $P_i(f_{in} = f_{jn}) = P_j(f_{in} = f_{jn})$, $L(i, j) = L(j, i)$. Like the information numbers, which lack this symmetry, the $L(i, j)$'s are positive for $i \neq j$. This fact is a consequence of Spitzer's identities (1956), as shown in the proof of Lemma 1. Note also that $L(i, j) \leq 1$ for $i \neq j$, and the value 1 is attained only if the support of f_j and the support of f_i are disjoint, so that the absolute continuity assumption is violated. The significance of the $L(i, j)$'s defined in (4) is established by the following lemma.

LEMMA 1. For $u, v > 0$, let

$$(5) \quad R(u, v) = \inf_{N \geq 0} [uE_1 N + vP_2(N < \infty)],$$

where the infimum is taken over all extended stopping times taking values $0, 1, \dots$, and ∞ .

The infimum is attained by

$$(6) \quad T\left(\frac{u}{v}\right) = \inf \{n \mid n \geq 0, v f_{2n} L(1, 2) \leq u f_{1n}\}.$$

PROOF. It is well known ([24], [15]) that the infimum in (5) is attained by stopping at the first $n \geq 0$ (or ∞ if there is no n) such that the “stopping risk” is less than or equal to the “continuation risk,” i.e., such that

$$(7) \quad v f_{2n} \leq R_1(u f_{1n}, v f_{2n}),$$

where

$$R_1(u, v) = \inf_{N \geq 1} [u E_1 N + v P_2(N < \infty)].$$

Similarly, $R_1(u, v)$ is attained by stopping at the first $n \geq 1$ (or ∞ if there is no n) such that (7) holds.

Since $R_1(u, v) = v R_1(u/v, 1)$, division of (7) by $v f_{2n}$ yields the equivalent inequality

$$(8) \quad 1 \leq R_1\left(\frac{u f_{1n}}{v f_{2n}}, 1\right).$$

Now, the function $R_1(t, 1)$ is the attained infimum of linear increasing functions of t and is therefore concave and increasing in t . Since $R_1(t, 1) \geq t$ by virtue of the restriction $N \geq 1$, evidently $R_1(t, 1) > 1$ for $t > 1$. For sufficiently small positive t , consideration of a stopping time such that $E_1 N < \infty$ and $P_2(N < \infty) < 1$ (e.g., that of a one-sided SPRT) shows that $R_1(t, 1) < 1$. By the continuity and monotonicity there is therefore a unique $t_1 \in (0, 1]$ such that

$$(9) \quad R_1(t_1, 1) = 1,$$

and (7) and (8) are equivalent to

$$(10) \quad t_1 \leq \frac{u f_{1n}}{v f_{2n}}.$$

It remains only to show that $t_1 = L(1, 2)$.

Now, $R_1(u, v)$ is attained by stopping at the first n such that (10) is satisfied. It is also optimal to stop when *strict* inequality holds in (10). Hence,

$$R_1(t_1, 1) = t_1 E_1 \tilde{N} + P_2(\tilde{N} < \infty),$$

where

$$\tilde{N} = \inf \{n \mid n \geq 1, f_{1n} > f_{2n}\}.$$

By (9), this last implies

$$(11) \quad t_1 = \frac{P_2(\tilde{N} = \infty)}{E_1 \tilde{N}}.$$

Note that $\tilde{N}$ is the time of first entry into $(0, \infty)$ of the random walk $S_n = \log f_{1n}/f_{2n}$. Hence, Spitzer's identities

$$P_2(S_n \leq 0 \text{ for } n \geq 1) = \exp(-\sum_{n=1}^{\infty} n^{-1} P_2(S_n > 0)),$$

(4.7) of [12], and

$$E_1 \tilde{N} = \sum_{n=0}^{\infty} P_1(\tilde{N} > n) = \exp(\sum_{n=1}^{\infty} n^{-1} P_1(S_n \leq 0)),$$

an immediate consequence of Corollary 2, page 330 of [12], yield $t_1 = L(1, 2)$, which proves the lemma.

REMARK. Since the lemma is of independent interest, it should be noted that the finite variance assumption, (1), is not needed, since only the mutual absolute continuity of the measures P_1 and P_2 is required. In fact, the lemma is easily shown to be true for *any* distinct probability measures P_1 and P_2 , the nonabsolutely continuous case being handled as follows. Let f_1 and f_2 be taken with respect to $P_1 + P_2$, let $f_{1n}/f_{2n} = +\infty$ when the denominator is zero, and $R_1(+\infty, v) = +\infty$. Evaluate $P_2(\tilde{N} = \infty)$ and $E_1 \tilde{N}$ using Corollary 2, page 330 of [12] (inequalities reversed), with t equal to $P_2(f_1 > 0)$ and $P_1(f_2 > 0)$, respectively. In both cases, the argument given above needs changing only if $t < 1$, in which case $\{S_k \leq 0\}$ is conditioned on $\{f_{1k} > 0\}$ (resp. $\{f_{2k} > 0\}$). The derivation of (4.7) is followed to complete the evaluation of $P_2(\tilde{N} = \infty)$.

The tests $\delta(c)$ are defined as follows. $N(c)$ is the infimum of $n \geq 0$ such that for some i and d

$$(12) \quad \lambda_n(j)W(j, d)L(i, j) \leq c\lambda_n(i) \quad \text{for all } j.$$

$D(c)$ chooses the smallest d attaining the stopping risk, $r(\lambda_{N(c)})$. To obtain a useful alternative characterization of $N(c)$, define the region $S_c(i, j, d)$ for $i, j = 1, \dots, s$, $d = 1, \dots, k$ as the set of λ 's satisfying

$$(13) \quad \lambda(j)W(j, d)L(i, j) \leq c\lambda(i).$$

Then by (12) $N(c)$ is the infimum of $n \geq 0$ such that $\lambda_n \in S_c$, where

$$(14) \quad S_c = \bigcup_{i,d} \bigcap_j S_c(i, j, d).$$

That the stopping regions S_c^* and S_c have a great deal in common is established by the following result.

LEMMA 2. Under assumption (1) there exist $Q_2 > Q_1 > 0$ such that for all $c > 0$

$$(15) \quad \{\lambda | r(\lambda) \leq Q_1 c\} \subset S_c \cap S_c^*$$

and

$$(16) \quad S_c \cup S_c^* \subset \{\lambda | r(\lambda) \leq Q_2 c\}.$$

PROOF. The choice $Q_1 = 1/s$ suffices for (15) by the following argument. If $r(\lambda) = r(\lambda, d)$ for a given d , then for all j the left-hand side of (12) is at most $r(\lambda)$ by virtue of (3) and the fact that the $L(i, j)$'s are ≤ 1 . Now, if $r(\lambda) \leq c/s$, then (12) holds for all j provided that $\lambda(i) \geq 1/s$, which must be true for some $i = 1, \dots, s$. Hence, $r(\lambda) \leq c/s$ implies $\lambda \in S_c$. A sufficient condition for $\lambda \in S_c^*$ is that $r(\lambda) \leq c$ since taking observations always yields risk $\geq c$. This proves (15) for $Q_1 = 1/s$.

To prove (16), let $\underline{L} > 0$ denote the minimum of $L(i, j)$ over $i, j = 1, \dots, s$. If $\lambda \in S_c$, then for some i and d (13) holds for all j , and replacing $L(i, j)$ by the lower bound, $\underline{L}$, and summing over j yields

$$r(\lambda, d) \cdot \underline{L} \leq sc\lambda(i) \leq sc,$$

whence $r(\lambda) \leq s\underline{L}^{-1}c$. By Lemma 2.2 of Lorden (1967) there is an $M^* > 0$ such that $\lambda \in S_c^*$ only if $r(\lambda) \leq M^*c$. Hence, $Q_2 = \max(M^*, s\underline{L}^{-1})$ suffices for (16) and the lemma is proved.

For $\rho > 1$, define

$$(17) \quad A_1(\rho) = \{\lambda \mid \lambda(i) \geq \rho\lambda(i+1), i = 1, \dots, s-1\}.$$

More generally, for a permutation π on $\{1, \dots, s\}$ define

$$A_\pi(\rho) = \{\lambda \mid \lambda(\pi(i)) \geq \rho\lambda(\pi(i+1)), i = 1, \dots, s-1\}$$

and let

$$A(\rho) = \bigcup_\pi A_\pi(\rho).$$

Let d_1 denote the decision that minimizes $(W(1, d), \dots, W(s, d))$ lexicographically, i.e., d_1 minimizes $x^s W(1, d) + x^{s-1} W(2, d) + \dots + x W(s, d)$ for large x . Let j_1 denote the smallest j such that $W(j, d_1) > 0$, noting that $1 < j_1 \leq s$. Let $\tilde{W}$ be the ratio of the largest $W(i, d)$ over all i and d to the smallest difference between any two distinct values of $W(i, d)$ over $i = 1, \dots, s, d = 1, \dots, k$. It follows from (3) and (17) that

$$(18) \quad r(\lambda) = r(\lambda, d_1) < \min_{d \neq d_1} r(\lambda, d) \quad \text{for } \lambda \in A_1(s\tilde{W})$$

and a fortiori for $\lambda \in A_1(\rho)$, $\rho > s\tilde{W}$.

Let $\bar{W}$ and $\underline{W}$ denote the largest and smallest, respectively, of the positive values of $W(j, d)L(i, j)$ over all i, j and d . A fact that will prove to be very useful in the sequel is

$$(19) \quad S_c \cap A_1(\bar{W}\underline{W}^{-1}) = S_c(1, j_1, d_1) \cap A_1(\bar{W}\underline{W}^{-1}) \quad \text{for all } c.$$

The left-hand side of (19) contains the right-hand side because whenever $\lambda \in A_1(\bar{W}\underline{W}^{-1})$ and (13) holds for $1, j_1$ and d_1 , it still holds with any other j in place of j_1 : for $j < j_1$ because $W(j, d_1) = 0$, and for $j > j_1$ because

$$\lambda(j)W(j, d_1)L(1, j) \leq \lambda(j)\bar{W} \leq \lambda(j_1)\underline{W} \leq \lambda(j_1)W(j_1, d_1)L(1, j_1).$$

The reverse inclusion in (19) follows from the fact that for all i and d

$$(20) \quad S_c(i, j(d), d) \cap A_1(\bar{W}\underline{W}^{-1}) \subset S_c(1, j_1, d_1) \cap A_1(\bar{W}\underline{W}^{-1}),$$

where $j(d)$ is the smallest j such that $W(j, d) > 0$. It suffices to consider $i = 1$ in (20), since changing from $\lambda(i)$, $i \neq 1$, to $\lambda(1)$ on the right-hand side of (13) introduces a factor of at least $\bar{W}\underline{W}^{-1}$, which amply compensates for the change from $L(i, j)$ to $L(1, j)$ on the left-hand side. Fix $i = 1$, then. By the definition of d_1 and j_1 , $j(d) \leq j_1$ and, if $j(d) = j_1$, then $W(j_1, d) \geq W(j_1, d_1)$, in which case (20) is clear from (13). In case $j(d) < j_1$, then (20) holds because the factor of

at least $\bar{W}W^{-1}$ picked up by switching from $\lambda(j(d))$ to $\lambda(j)$ in (13) amply compensates for switching from $W(j(d), d)L(1, j(d))$ to $W(j_1, d_1)L(1, j_1)$.

The following lemmas are the keys to the proof of Theorem 1.

LEMMA 3. Define $B(c) = A(\exp((\log c^{-1})^\dagger)) \cap \{\lambda | r(\lambda) \leq Q_2 c\}$. Fix λ_0 and let $M(c)$ denote the stopping time $\min(N(c), N^*(c))$. As $c \rightarrow 0$,

$$P_i(\lambda_{M(c)} \in B(c)) \rightarrow 1 \quad i = 1, \dots, s.$$

LEMMA 4. Define $B_1(c) = A_1(\exp((\log c^{-1})^\dagger)) \cap \{\lambda | r(\lambda) \leq Q_2 c\}$. Let

$$(21) \quad \check{R}(u, v) = \inf_{N \geq 0} [uE_1 N + vP_{j_1}(N < \infty)], \quad u, v > 0.$$

Set $W_1 = W(j_1, d_1)$. As $c \rightarrow 0$,

$$(22) \quad r_c(\lambda, \delta^*(c)) = \check{R}(c\lambda(1), \lambda(j_1)W_1) + o(c)$$

and

$$(23) \quad r_c(\lambda, \delta(c)) = \check{R}(c\lambda(1), \lambda(j_1)W_1) + o(c)$$

uniformly for $\lambda \in B_1(c)$.

REMARK. Analogs of (22) and (23) hold, of course, for $\lambda \in B(c)$. In Lemma 4, $\delta(c)$ and $\delta^*(c)$ denote the procedures based on the a priori distribution λ . The statement and proof of Theorem 1 are given next, with the proofs of Lemmas 3 and 4 deferred until afterward.

THEOREM 1. Suppose (1) holds and an a priori distribution, λ_0 , is given. Then

$$(24) \quad r_c(\lambda_0, \delta(c)) - r_c(\lambda_0, \delta^*(c)) = o(c) \quad \text{as } c \rightarrow 0.$$

PROOF. Assume without loss of generality that $\lambda(i) > 0$ for $i = 1, \dots, s$. Fix c . Both $\delta(c)$ and $\delta^*(c)$ are stationary with respect to the sequence of a posteriori distributions in the sense that they stop as soon as λ_n falls in their respective stopping regions, S_c and S_c^* . Thus, since $N(c) \geq M(c) = \min(N(c), N^*(c))$, conditioning on $M(c)$ and $\lambda_{M(c)}$ yields

$$r_c(\lambda_0, \delta(c)) = cE[M(c) | \lambda_0] + E[r_c(\lambda_{M(c)}, \delta(c)) | \lambda_0],$$

where $E[\cdot | \lambda_0]$ denotes $\sum_i \lambda_0(i)E_i[\cdot]$. On the right-hand side of this relation $\delta(c)$ is understood to be "starting all over again" with $\lambda_{M(c)}$ as a priori distribution, by virtue of the stationarity. A similar relation holds for $\delta^*(c)$ and subtraction of the two relations leads to

$$(25) \quad r_c(\lambda_0, \delta(c)) - r_c(\lambda_0, \delta^*(c)) = E[r_c(\lambda_{M(c)}, \delta(c)) - r_c(\lambda_{M(c)}, \delta^*(c)) | \lambda_0].$$

There is a constant $Q > 0$ such that

$$(26) \quad r_c(\lambda, \delta(c)) - r_c(\lambda, \delta^*(c)) \leq Qc \quad \text{for all } \lambda \text{ and } c.$$

To prove (26), consider the test $\tilde{\delta}(Q_1 c)$ with stopping region $\tilde{S}_c = \{\lambda | r(\lambda) \leq Q_1 c\}$ and terminal decision rule like $\delta(c)$'s. By Lemma 2, $\tilde{S}_c \subset S_c$, whence the sampling cost of $\tilde{\delta}(c)$ is no larger than that of $\delta(Q_1 c)$. The integrated risk of error

for $\delta(c)$ equals the $E[\cdot | \lambda_0]$ -expectation of the stopping risk, which is at most $Q_2 c$ by Lemma 2. Therefore,

$$r_c(\lambda, \delta(c)) - r_c(\lambda, \tilde{\delta}(Q_1 c)) \leq Q_2 c \quad \text{for all } \lambda \text{ and } c.$$

By Theorem 2.1 of Lorden (1967), there is a constant $M > 0$ such that

$$r_c(\lambda, \tilde{\delta}(Q_1 c)) - r_c(\lambda, \delta^*(c)) \leq Mc \quad \text{for all } \lambda \text{ and } c,$$

which, together with the preceding relation, proves (26) with $Q = Q_2 + M$.

Now, the difference of a posteriori risks

$$r_c(\lambda_{M(c)}, \delta(c)) - r_c(\lambda_{M(c)}, \delta^*(c))$$

appearing in the right-hand member of relation (25) is in any event at most Qc by (26) and is bounded above by $o(c)$ uniformly on the event $\{\lambda_{M(c)} \in B(c)\}$ by (22) and (23) and their analogs for $\lambda \in B(c) - B_1(c)$. Since this event has probability approaching one under every P_i by Lemma 3, the right-hand side of (25) is $o(c)$ as $c \rightarrow 0$ and the theorem is proved.

PROOF OF LEMMA 3. It is necessary first to establish that

$$(27) \quad \lim_{c \rightarrow 0} \frac{N(c)}{\log c^{-1}} = \lim_{c \rightarrow 0} \frac{N^*(c)}{\log c^{-1}} \\ = \min_d (\min_{j: W(j, d) > 0} I(i, j))^{-1} \quad \text{in } P_i\text{-probability.}$$

First consider $N(c)$ and note that

$$(28) \quad P_i(N(c) = \min_{d: W(i, d) = 0} N(c; i, d)) \rightarrow 1 \quad \text{as } c \rightarrow 0,$$

where

$$N(c; i, d) = \inf \{n \mid n \geq 0, \lambda_n \in \bigcap_{j=1}^n S_c(i, j, d)\}.$$

This is based first of all on the fact that the P_i -probability of $\{\lambda_n \in S_c(i', i, d') \text{ for some } n\}$ tends to zero if $i' \neq i$ and $W(i, d') > 0$, by Wald's error probability bound. Thus, only d 's such that $W(i, d) = 0$ need be considered. Now (28) follows since the P_i -probability of $\{\lambda_n \in \bigcap_j S_c(i', j, d) \text{ for some } i' \neq i \text{ before } \lambda_n \in \bigcap_j S_c(i, j, d)\}$ tends to zero because $N(c) \rightarrow \infty$ (clearly) and the last time, n , that $\lambda_n(i')/L(i', j) > \lambda_n(i)/L(i, j)$ for some j has a finite (nondefective) distribution under P_i .

To prove (27), fix d such that $W(i, d) = 0$. For j such that $W(j, d) > 0$ let $\underline{N}_j(c)$ denote the first time $\lambda_n \in S_c(i, j, d)$ and $\bar{N}_j(c)$ denote the last time $\lambda_n \notin S_c(i, j, d)$. Rewriting (12) in terms of $\log(f_{i_n}/f_{j_n})$ leads to

$$(29) \quad \lim_{c \rightarrow 0} \frac{\underline{N}_j(c)}{\log c^{-1}} = \frac{1}{I(i, j)} \quad \text{in } P_i\text{-probability}$$

by a well-known result for random walks (Chung (1969), page 127). Furthermore, $1 + \bar{N}_j(c) - \underline{N}_j(c)$ is nonnegative and bounded above by the smallest n such that

$$\inf_{m \geq n} \sum_{r=\underline{N}_j(c)+1}^{\bar{N}_j(c)+m} \log(f_i(X_r)/f_j(X_r)) \geq 0,$$

which has the same (nondefective) distribution for all c . Therefore, (29) holds for $\bar{N}_j(c)$ also and, since

$$\max_{j: W(j, d) > 0} \bar{N}_j(c) \leq N(c; i, d) \leq 1 + \max_{j: W(j, d) > 0} \bar{N}_j(c),$$

it follows that

$$\lim_{c \rightarrow 0} \frac{N(c; i, d)}{\log c^{-1}} = (\min_{j: W(j, d) > 0} I(i, j))^{-1} \quad \text{in } P_i\text{-probability.}$$

Using (28), (27) follows for $N(c)$. (Note that d 's such that $W(i, d) > 0$ play no role in (27) since $I(i, i) = 0$.) To establish (27) for $N^*(c)$, observe that $N(cQ_2/Q_1) \leq N^*(c) \leq N(cQ_1/Q_2)$ by Lemma 2.

Since $r(\lambda_{M(c)}) \leq Q_2 c$ by Lemma 2, it suffices for Lemma 3 to show that

$$(30) \quad P_i(|\log \lambda_{M(c)}(j) - \log \lambda_{M(c)}(j')| \leq (\log c^{-1})^{\frac{1}{2}}) \rightarrow 0$$

as $c \rightarrow 0$, for arbitrary $j \neq j'$, one of which may equal i . First consider the case where $I(i, j) - I(i, j') = \eta > 0$. Let T_1 denote the last time that $\log(\lambda_n(j)/\lambda_n(j')) + \frac{1}{2}\eta n$ is positive, which is finite a.s. (P_i) by the strong law of large numbers. Letting $\bar{I}(i) = 2 \max_j I(i, j)$, an easy calculation shows that if $\log c^{-1} > (2\bar{I}(i)/\eta)^{\frac{2}{3}}$, then the probability in (30) is at most $P_i(M(c) \leq \max(T_1, \bar{I}(i)^{-1} \log c^{-1}))$, which goes to zero by (27).

The case $\eta < 0$ is similar and in the case $\eta = 0$, i.e., $I(i, j) = I(i, j')$, (30) is proved as follows. By the central limit theorem for randomly stopped sums ([3], page 197), $\log(\lambda_{M(c)}(j)/\lambda_{M(c)}(j'))/(\log c^{-1})^{\frac{1}{2}}$ converges in distribution under P_i to a nondegenerate normal law with mean zero by virtue of (27) and the distinctness of f_j and $f_{j'}$. This completes the proof of Lemma 3.

PROOF OF LEMMA 4. Assume that $\exp((\log c^{-1})^{\frac{1}{2}}) > s\tilde{W}$, so that (18) applies and

$$(31) \quad r(\lambda) = r(\lambda, d_1) \leq Q_2 c.$$

It suffices to prove (22) with “ $\geq$ ” and (23) with “ $\leq$,” since $r_c(\lambda, \delta^*(c)) \leq r_c(\lambda, \delta(c))$ by the optimality of $\delta^*(c)$.

For the proof of “ $\geq$ ” in (22), define the (possibly infinite) stopping time

$$(32) \quad \bar{N} = \inf \left\{ n \mid n \geq 1, f_{j_n} \leq \frac{Q_1 \tilde{W}}{Q_2(1 + \tilde{W})} f_{1n} \text{ for } j = 2, \dots, s \right\}.$$

Since $\lambda \in A_1(s\tilde{W})$, calculation shows $\lambda(1) > \tilde{W}/(1 + \tilde{W})$ and, hence, if $\bar{N} < \infty$

$$\frac{\lambda_{\bar{N}}(j)}{\lambda(j)} = \frac{f_{\bar{N}j}}{f_{\bar{N}1}} \cdot \frac{\lambda_{\bar{N}}(1)}{\lambda(1)} \leq \frac{Q_2 \tilde{W}}{Q_2(1 + \tilde{W})} \cdot \frac{1}{\lambda(1)} < \frac{Q_1}{Q_2},$$

so that (31) implies $r(\lambda_{\bar{N}}, d_1) \leq Q_1 c$. It follows using Lemma 2 that $N^*(c) \leq \bar{N}$.

Define another extended stopping time

$$(33) \quad \begin{aligned} N'(c) &= N^*(c) & \text{if } D^*(c) = d_1, \\ &= \bar{N} & \text{if } D^*(c) \neq d_1, \end{aligned}$$

noting that $N'(c) \leq \bar{N}$, and also define the event

$$(34) \quad V_c = \left\{ \bar{N} < \infty \text{ and } \max_{i,j} \max_{n \leq \bar{N}} \log \frac{f_{in}}{f_{jn}} \leq \frac{1}{2}(\log c^{-1})^\dagger \right\}.$$

Note that, unlike $N^*(c)$ and $D^*(c)$, $\bar{N}$ and V_c do not depend on λ . Now, using primes to denote complements,

$$(35) \quad V_c' \cap \{\bar{N} < \infty\} \downarrow \emptyset \quad \text{as } c \downarrow 0,$$

since the likelihood ratios are finite for every n . Furthermore, on V_c , $\lambda_n \in A_1(\exp(\frac{1}{2}(\log c^{-1})^\dagger))$ for all $n \leq \bar{N}$ and for $N^*(c)$ in particular. Therefore, by (18)

$$(36) \quad V_c \subset \{D^*(c) = d_1\}$$

for $\lambda \in B_1(c)$, provided that $\exp(\frac{1}{2}(\log c^{-1})^\dagger) > s\tilde{W}$, which is assumed now and for the remainder of the proof. Using (33) and (36),

$$(37) \quad \begin{aligned} P_{j_1}(N'(c) < \infty) &\leq P_{j_1}(D^*(c) = d_1) + P_{j_1}(\{\bar{N} < \infty\} \cap V_c') \\ &= P_{j_1}(D^*(c) = d_1) + o(1) \end{aligned}$$

uniformly in λ as $c \rightarrow 0$ by (35) and the fact that $\bar{N}$ and V_c do not depend upon λ .

Also, by (33) and (36),

$$(38) \quad E_1 N'(c) \leq E_1 N^*(c) + E_1 \bar{N} 1\{V_c'\} = E_1 N^*(c) + o(1)$$

uniformly in λ as $c \rightarrow 0$ by (35) and monotone convergence, since $E_1 \bar{N} < \infty$. Putting together (37) and (38), and using the fact that $W_1 \lambda(j_1) \leq r(\lambda, d_1) \leq Q_2 c = O(c)$,

$$\begin{aligned} c\lambda(1)E_1 N^*(c) + W_1 \lambda(j_1)P_{j_1}(D^*(c) = d_1) \\ \geq c\lambda(1)E_1 N'(c) + W_1 \lambda(j_1)P_{j_1}(N'(c) < \infty) - o(c) \\ \geq \tilde{R}(c\lambda(1), W_1 \lambda(j_1)) - o(c), \end{aligned}$$

by the definition (21) of $\tilde{R}$, and the $o(c)$ term does not depend on λ . This shows that “ $\geq$ ” holds in (22) and it remains only to prove that “ $\leq$ ” holds in (23).

Observe that

$$(39) \quad N(c) \leq \bar{N}$$

by the same reasoning used for $N^*(c)$ and that, like (36),

$$(40) \quad V_c \subset \{D(c) = d_1\}$$

for $\lambda \in B_1(c)$.

The $E[\cdot | \lambda]$ -expectation of the stopping risk equals the part of $r_c(\lambda, \delta(c))$ due to error. Thus,

$$r_c(\lambda, \delta(c)) = cE[N(c) | \lambda] + E[r(\lambda_{N(c)}) | \lambda].$$

Using the fact that $r(\lambda_{N(c)}) \leq Q_2 c$,

$$(41) \quad \begin{aligned} r_c(\lambda, \delta(c)) &\leq cE[N(c) | \lambda] + \lambda(1)E_1 r(\lambda_{N(c)}) 1\{V_c\} + Q_2 c(1 - \lambda(1) + P_1(V_c')) \\ &= cE[N(c) | \lambda] + \lambda(1)E_1 r(\lambda_{N(c)}) 1\{V_c\} + o(c) \end{aligned}$$

uniformly in $\lambda \in B_1(c)$, since $1 - \lambda(1) \rightarrow 0$ uniformly and $P_1(V_e') \downarrow P_1(\bar{N} = \infty) = 0$ as $c \downarrow 0$ by (35). Define the extended stopping time $T_e = \inf \{n \mid n \geq 0 \text{ and } \lambda_n \in S_c(1, j_1, d_1)\}$, which in accord with Lemma 1 attains $\tilde{R}(c\lambda(1), \lambda(j_1)W_1)$, defined in (21). Assume that $\exp(\frac{1}{2}(\log c^{-1})^{\frac{1}{2}}) > \bar{W}\bar{W}^{-1}$. Then the fact that, on V_e , $\lambda_n \in A_1(\exp(\frac{1}{2}(\log c^{-1})^{\frac{1}{2}}))$ for $n \leq \bar{N}$ (hence, for $n \leq N(c)$) implies

$$(42) \quad N(c) = T_e \quad \text{on } V_e$$

by (19).

By (41) and the fact that T_e attains $\tilde{R}(c\lambda(1), \lambda(j_1)W_1)$, the “ $\leq$ ” in (23) is proved once it is established that

$$(43) \quad cE[N(c) \mid \lambda] \leq c\lambda(1)E_1T_e + o(c)$$

and

$$(44) \quad \lambda(1)E_1r(\lambda_{N(c)})1\{V_e\} \leq \lambda(j_1)W_1P_{j_1}(T_e < \infty) + o(c)$$

uniformly in $\lambda \in B_1(c)$.

By (39) and (42),

$$(45) \quad E_1N(c) \leq E_1T_e + E_1\bar{N}1\{V_e'\} = E_1T_e + o(1)$$

by monotone convergence, as in (38). To estimate $E_iN(c)$ for $i > 1$, note that $N(c)$ is at most the infimum of $n \geq 1$ such that $f_{j_n}\bar{W} \leq c\lambda(i)f_{i_n}$ for all $j \neq i$. The E_i -expectation of the latter stopping time is at most $M^*(1 + \log(\bar{W}/c\lambda(i)))$ for a suitable $M^* > 0$, as a consequence of Theorem 3.3 of Lorden (1967). Hence, letting $\phi(c) = \exp(-(\log c^{-1})^{\frac{1}{2}})$, which is greater than $\lambda(2), \dots, \lambda(s)$ for $\lambda \in B_1(c)$,

$$(46) \quad c\lambda(i)E_iN(c) \leq M^*c\phi(c)(1 + \log(\bar{W}/c\phi(c))) \quad i = 2, \dots, s,$$

using the fact that $x(1 + \log(\bar{W}/x))$ is increasing for $x < \bar{W}$. Since $\phi(c) \log c^{-1} \rightarrow 0$ as $c \rightarrow 0$, the right-hand side of (46) is seen to be $o(c)$ as well as not dependent on λ . Hence, (46) and (45) yield (43).

To prove (44), first observe that on V_e

$$(47) \quad r(\lambda_{N(c)}) = r(\lambda_{N(c)}, d_1) \leq W_1\lambda_{N(c)}(j_1) + Q_2cs\tilde{W}\exp(-\frac{1}{2}(\log c^{-1})^{\frac{1}{2}}) \\ = W_1\lambda_{N(c)}(j_1) + o(c)$$

by routine estimates making use of the facts that $\lambda_{N(c)} \in A_1(\exp(\frac{1}{2}(\log c^{-1})^{\frac{1}{2}}))$ on V_e and $r(\lambda_{N(c)}) \leq Q_2c$. Hence,

$$(48) \quad \lambda(1)E_1r(\lambda_{N(c)})1\{V_e\} \leq W_1\lambda(1)E_1\lambda_{N(c)}(j_1)1\{V_e\} + o(c)$$

uniformly for $\lambda \in B_1(c)$.

Now, by (42) and the fact that $T_e < \infty$ on V_e ,

$$E_1\lambda_{N(c)}(j_1)1\{V_e\} = E_1\lambda_{T_e}(j_1)1\{V_e\} \leq E_1\frac{\lambda_{T_e}(j_1)}{\lambda_{T_e}(1)}1\{T_e < \infty\} \\ = \frac{\lambda(j_1)}{\lambda(1)}P_{j_1}(T_e < \infty),$$

which combines with (48) to yield (44). Thus, the “ $\leq$ ” in (23) is established and the proof of Lemma 4 is complete.

The next theorem gives more information about the structure of the Bayes stopping region, S_c^* . As indicated in the remark that follows the proof, this result implies convergence of the stopping boundaries of $\delta(c)$ and $\delta^*(c)$ in the case of Koopman–Darmois families.

THEOREM 2. *Assume that (1) holds and that a function $\gamma(c)$ is given satisfying $\gamma(c)/\log c^{-1} \rightarrow \infty$ as $c \rightarrow 0$. Then, in the notation of Lemma 4,*

$$(49) \quad \lim_{c \rightarrow 0} \sup_{A_1(\gamma(c)) \cap S_c^*} \frac{\lambda(j_1)}{c\lambda(1)} = \lim_{c \rightarrow 0} \inf_{A_1(\gamma(c)) - S_c^*} \frac{\lambda(j_1)}{c\lambda(1)} = \frac{1}{W_1 L(1, j_1)}.$$

REMARK. Analogs of (49) hold, of course, for $A_\pi(\gamma(c)) \cap S_c^*$ and $A_\pi(\gamma(c)) - S_c^*$. If the f_i 's belong to a Koopman–Darmois family, then $\delta(c)$ and $\delta^*(c)$ are associated with continuation regions, $B(c)$ and $B^*(c)$, in the (n, S_n) plane. It follows easily from Theorem 2 that along a fixed ray from the origin the distance between the points of intersection with the boundaries of these two regions goes to zero with c , provided that the ray is not among the (finitely many) exceptional rays along which the a posteriori distributions are not in $B(c)$. Moreover, one can choose $\rho = \rho(c) \downarrow 1$ as $c \downarrow 0$ so that for $c \in (0, 1)$ $B(\rho c) \subset B^*(c) \subset B(\rho^{-1}c)$ in the region of the plane outside an exceptional set of points (n, S_n) whose distance from the closest exceptional ray is less than a constant times $\log n$.

PROOF. Assume that c is small enough so that

$$(50) \quad \gamma(c)^{\frac{1}{2}} > \max(s\tilde{W}, W\underline{W}^{-1}),$$

in order that (18) and (19) can be used where needed in the sequel. A straightforward calculation shows that for $\lambda \in A_1(\gamma(c))$ under the assumption (50), $W_1 \lambda(j_1)$ accounts for at least half of the stopping risk $r(\lambda) = r(\lambda, d_1)$. It follows easily that

$$(51) \quad r(\lambda) > c \max(Q_2, 2/L(1, j_1)) \Rightarrow \frac{\lambda(j_1)}{c\lambda(1)} > \frac{1}{W_1 L(1, j_1)} \quad \text{and} \quad \lambda \notin S_c^*.$$

Because of this, it turns out to be sufficient to prove (49) with $A_1(\gamma(c))$ replaced by $A_1(\gamma(c)) \cap \{\lambda \mid r(\lambda) \leq Q_3 c\}$, where $Q_3 = \max(Q_2, 2/L(1, j_1))$. “Putting back in” λ 's with $r(\lambda) > Q_3 c \geq Q_2 c$ doesn't add anything to S_c^* and the infimum in (49) is not reduced because of (51).

Let $\delta_1^*(c)$ denote the test which stops at the first $n \geq 1$ such that $\lambda_n \in S_c^*$ and chooses the smallest d minimizing the a posteriori risk. Then $\delta_1^*(c)$ attains the minimum integrated risk among all tests taking at least one observation, the so-called “continuation risk.” A slight modification of the argument used to prove Lemma 4 shows that

$$(52) \quad r_c(\lambda, \delta_1^*(c)) = \tilde{R}_1(c\lambda(1), \lambda(j_1)W_1) + o(c)$$

uniformly for $\lambda \in A_1(\gamma(c))$ satisfying $r(\lambda) \leq Q_3 c$, where $\tilde{R}_1$ is defined by modifying

(21) to require $N \geq 1$. The only changes necessary are using the $n \geq 1$ versions of $\delta^*(c)$, $\delta(c)$ and T_c , and the replacement of Q_2 by Q_3 and $\exp((\log c^{-1})^3)$ by $\gamma(c)$, with corresponding replacements of its square root and reciprocal. The fact that $\gamma(c) \rightarrow \infty$ as $c \rightarrow 0$ suffices for the argument until (46), where the assumption that $\gamma(c)/\log c^{-1} \rightarrow \infty$ is needed to make the argument about $\phi(c) = 1/\gamma(c)$ go through.

By arguing as in (47),

$$(53) \quad r(\lambda) = \lambda(j_1)W_1 + o(c),$$

uniformly for $\lambda \in A_1(\gamma(c))$ satisfying $r(\lambda) \leq Q_3 c$. By (52) and (53), the fact that $r(\lambda) \leq r_c(\lambda, \delta_1^*(c))$ for $\lambda \in S_c^*$ yields

$$\begin{aligned} \lambda(j_1)W_1 &\leq \tilde{R}_1(c\lambda(1), \lambda(j_1)W_1) + o(c) \\ &= \tilde{R}_1(c\lambda(1), \lambda(j_1)W_1) (1 + o(1)) \end{aligned}$$

uniformly for $\lambda \in A_1(\gamma(c)) \cap S_c^*$, using the fact that $\tilde{R}_1 \geq c\lambda(1)$, which is of order c since $\lambda(1) \rightarrow 1$ uniformly. Dividing by $\lambda(j_1)W_1$ yields

$$1 \leq \tilde{R}_1\left(\frac{c\lambda(1)}{\lambda(j_1)W_1}, 1\right) (1 + o(1)),$$

which implies that

$$(1 - o(1))L(1, j_1) \leq \frac{c\lambda(1)}{\lambda(j_1)W_1}$$

uniformly for $\lambda \in A_1(\gamma(c)) \cap S_c^*$, by the nondecreasing property of $\tilde{R}_1(t, 1)$ and the fact that this function takes the value 1 only at $t = L(1, j_1)$, by analogy with (9). This proves the “sup” part of (49), and a similar argument with reversed inequalities for $\lambda \notin S_c^*$ takes care of the “inf” part, proving the theorem.

The next result clarifies what it is that a test should do to “behave like $\delta^*(c)$ ” as $c \rightarrow 0$. Tests which meet the requirements of the theorem, such as $\delta(c)$, act “exactly as $\delta^*(c)$ would have acted” for the same observations $X_1, X_2, \dots$, with probability approaching one as $c \rightarrow 0$, for all P_i . Furthermore, such tests attain the Bayes risk within $o(c)$, as in Theorem 1. However, there is an additional assumption needed:

$$(54) \quad P_i(\log(f_i(X_1)/f_j(X_1))) = mb \text{ for some integer } m) < 1 \quad \text{for all } b > 0$$

for $i, j = 1, \dots, s$ such that $i \neq j$. This assumption rules out the so-called “lattice case” for the random walk $\{\log(f_{i_n}/f_{j_n})\}$ and excludes, for example, problems involving the binomial parameter p if two of its possible values, say p_i and p_j , are such that $\log(p_i/p_j)/\log((1-p_i)/(1-p_j))$ is a rational number.

For use in Theorem 3, it is helpful to consider the notion of a test defined by a stopping region, S , and an a priori distribution, in the sense that observations are stopped as soon as $\lambda_n \in S$ and the smallest d minimizing the a posteriori risk is chosen.

THEOREM 3. *Assume that (1) and (54) are satisfied and a fixed a priori distribution,*

λ_0 , is given. Let $\{\tilde{S}_c, 0 < c < 1\}$ denote a family of stopping regions satisfying (49) with $\gamma(c) = \exp((\log c^{-1})^{\frac{1}{3}})$ and also satisfying its analogs for $A_\pi(\gamma(c))$. Assume that there exist constants $\tilde{Q}_2 > \tilde{Q}_1 > 0$ such that

$$(55) \quad \{\lambda \mid r(\lambda) \leq \tilde{Q}_1 c\} \subset \tilde{S}_c \subset \{\lambda \mid r(\lambda) \leq \tilde{Q}_2 c\}.$$

Let $\tilde{N}(c) = \inf \{n \mid n \geq 0, \lambda_n \in \tilde{S}_c\}$.

Then, as $c \rightarrow 0$

$$(56) \quad P_i(\tilde{N}(c) = N^*(c) \text{ and } \tilde{D}(c) = D^*(c)) \rightarrow 1 \quad i = 1, \dots, s,$$

where $\tilde{D}(c)$ is the smallest d minimizing $r(\lambda_{\tilde{N}(c)}, d)$. Furthermore, letting $\tilde{\delta}(c) = (\tilde{N}(c), \tilde{D}(c))$,

$$r_c(\lambda_0, \tilde{\delta}(c)) - r_c(\lambda_0, \delta^*(c)) = o(c)$$

as $c \rightarrow 0$.

REMARKS. In the lattice case where (54) is not satisfied, the conclusions of the theorem hold provided that, as $c \rightarrow 0$, $\log c^{-1}$ remains bounded away from certain arithmetic sequences of exceptional points which can be determined from the proof. The fact that $\delta(c)$ satisfies the hypotheses of Theorem 3 is a straightforward consequence of (19) and its analogs. By similar arguments, many variants of $\delta(c)$ can be shown to satisfy the hypotheses. One such variant is obtained by stopping when (12) holds for the value of j maximizing $\lambda_n(j)$ subject to $W(j, d) > 0$. Another variant stops when, for the permutation π induced by λ_n , $r(\lambda_n) \leq c/L(\pi(1), \pi(j_1))$.

PROOF OF THEOREM 3. Let $\tilde{M}(c) = \min(\tilde{N}(c), N^*(c))$. Examination of the proof of Lemma 3 shows that as $c \rightarrow 0$

$$(57) \quad P_i(\lambda_{\tilde{M}(c)} \in \tilde{A}_i(\exp((\log c^{-1})^{\frac{1}{3}}))) \rightarrow 1 \quad i = 1, \dots, s,$$

where

$$\tilde{A}_i(\exp((\log c^{-1})^{\frac{1}{3}})) = \bigcup_{\pi: \pi(1)=i} A_\pi(\exp((\log c^{-1})^{\frac{1}{3}})).$$

Relation (57) also holds with $\tilde{M}(c)$ in place of $M(c)$ because (27) holds for $\tilde{N}(c)$ by virtue of (55). Since (49) holds with $\gamma(c) = \exp((\log c^{-1})^{\frac{1}{3}})$ both for S_e^* and $\tilde{S}_e$, it follows that $\lambda_{\tilde{M}(c)} \in \tilde{A}_1(\exp((\log c^{-1})^{\frac{1}{3}})) - (\tilde{S}_e \cap S_e^*)$ only if, for some $j > 1$ and d ,

$$|\log[(\lambda_{\tilde{M}(c)}(j)/c\lambda_{\tilde{M}(c)}(1))\dot{W}(j, d)L(1, j)]| \leq \varepsilon(c),$$

where $\varepsilon(c) \rightarrow 0$ as $c \rightarrow 0$. This last event requires that the random walk $\{\log(f_{1n}/f_{jn})\}$ hit one of a finite number of intervals of length $2\varepsilon(c)$, whose mid-points go to infinity with $\log c^{-1}$. Using the renewal theorem, this is seen to be an event whose probability goes to zero as $c \rightarrow 0$. Applying this result with the $\tilde{M}(c)$ -version of (57),

$$(58) \quad P_1(\lambda_{\tilde{M}(c)} \in \tilde{S}_e \cap S_e^*) \rightarrow 1 \quad \text{as } c \rightarrow 0.$$

The event in (58) is equivalent to $\tilde{N}(c) = N^*(c)$ and, from (18) and its analogs, it is clear that the probability that $r(\lambda_{\tilde{M}(c)}, d)$ is minimized by only one d approaches

one. Hence, (56) is proved for $i = 1$, and, by a similar argument, for all i . The conclusion about the integrated risk of $\tilde{\delta}(c)$ follows as in the proof of Theorem 1, with the assumption (55) sufficing for Lemma 2 and with Lemma 4 replaced by the fact that

$$r_e(\lambda_{\tilde{M}(c)}, \tilde{\delta}(c)) = r(\lambda_{\tilde{M}(c)}) = r_e(\lambda_{\tilde{M}(c)}, \tilde{\delta}^*(c))$$

with P_i -probability approaching one, as a consequence of (56). This completes the proof of Theorem 3.

The test $\delta(c)$ are always combinations of one-sided SPRT's, but they assume a particularly simple form for the s -decision problem of choosing which of given densities, $f_1, \dots, f_s$ is true. In this case, they become simple combinations of two-sided SPRT's and have an asymptotic property like the exact optimality of the SPRT (Wald and Wolfowitz, 1948).

To define such a test without reference to loss functions or a priori distributions, consider a *boundary matrix* $B = \|B_{ij}\|$ with positive entries except for the diagonal entries B_{ii} , which are immaterial.

For $0 < c < \min B_{ij}$ define the *matrix SPRT* $(T(c), D(c))$ as follows: stop at the first $n \geq 1$ such that for some i

$$(59) \quad B_{ji}f_{jn} \leq cf_{in} \quad \text{for all } j \neq i$$

and choose the unique i satisfying (59). The scale factor c can, of course, be eliminated by dividing it into the B matrix, but $c \rightarrow 0$ for fixed B leads to interesting asymptotic considerations.

For any test (N, D) consider weighted error probabilities of the form

$$(60) \quad \alpha_1 = \sum_{i=1}^s a_{i1} P_i(D = 1), \dots, \alpha_s = \sum_{i=1}^s a_{is} P_i(D = s),$$

where $A = \|a_{ij}\|$ is a given matrix of weights, positive except for the diagonal entries which are zero.

Choosing

$$(61) \quad B_{ij} = a_{ij} L(i, j) \quad i, j = 1, \dots, s,$$

leads to

THEOREM 4. Suppose that (1) holds, the matrix of weights A is given, and B is defined by (61). Let $\alpha_1(c), \dots, \alpha_s(c)$ denote the weighted error probabilities of $(T(c), D(c))$ defined as in (60). If a family of tests $\tilde{\delta}(c) = (\tilde{N}(c), \tilde{D}(c))$ has weighted error probabilities

$$(62) \quad \tilde{\alpha}_i(c) \leq \alpha_i(c) \quad i = 1, \dots, s,$$

then as $c \rightarrow 0$

$$(63) \quad E_i \tilde{N}(c) \geq E_i T(c) - o(1) \quad i = 1, \dots, s.$$

REMARK. Of course, for any B , the matrix A can be chosen to satisfy (61), yielding an asymptotic optimality property for $(T(c), D(c))$. Strengthening the hypothesis to eliminate A from the formulation, there is an immediate

COROLLARY. For any B , $(T(c), D(c))$ minimizes the expected sample sizes for $i = 1, \dots, s$ to within $o(1)$ as $c \rightarrow 0$ among all tests whose error probabilities $\{P_i(D = j), i, j = 1, \dots, s, i \neq j\}$ are less than or equal to those of $(T(c), D(c))$.

PROOF OF THEOREM 4. Since $L(i, j) = L(j, i)$ the order of these arguments will be interchanged whenever convenient. The conclusion (63) will be proved for $i = 1$, the other cases being similar.

Fix $0 < \varepsilon < 1$ and consider the a priori distribution

$$(64) \quad \lambda_0(1) = \frac{1}{1 + (s-1)\varepsilon}; \quad \lambda_0(i) = \frac{\varepsilon}{1 + (s-1)\varepsilon} \quad i = 2, \dots, s.$$

Define

$$(65) \quad W(i, j) = \frac{\lambda_0(j)a_{ij}}{\lambda_0(i)} \quad i, j = 1, \dots, s.$$

Assuming $c \leq s^{-1}\varepsilon(\min_{i \neq j} a_{ij})(\min L(i, j))$, it is easy to verify from (59) and the definitions that $(T(c), D(c))$ coincides with the procedure $\delta(c)$ of Theorem 1 for the problem specified by (64) and (65) with c as the cost per observation.

It follows from (59) and the choice of B that

$$a_{ji}P_j(D(c) = i) \leq a_{ji}cB_{ji}^{-1} = cL(i, j)^{-1},$$

so that

$$\alpha_i(c) \leq Mc \quad i = 1, \dots, s$$

where $M = s \max (L(i, j)^{-1})$. Also, for any test $\delta = (N, D)$ the integrated risk of error can be written using (60) and (65) as

$$e(\lambda_0, \delta) = \sum_i \lambda_0(i) \sum_j W(i, j) P_i(D = j) = \sum_{i,j} \lambda_0(j) a_{ij} P_i(D = j) = \sum_j \lambda_0(j) \alpha_j.$$

Hence, by (62) and the fact that $\alpha_j(c) \leq Mc$ for all j

$$(66) \quad e(\lambda_0, \tilde{\delta}(c)) \leq e(\lambda_0, \delta(c)) \leq Mc.$$

Before invoking Theorem 1, it is necessary to introduce a modification $\tilde{\tilde{\delta}}(c) = (\tilde{\tilde{N}}(c), \tilde{\tilde{D}}(c))$ of $\tilde{\delta}(c)$ by defining

$$\tilde{\tilde{N}}(c) = \min(\tilde{N}(c), T(\varepsilon c))$$

and

$$\begin{aligned} \tilde{\tilde{D}}(c) &= \tilde{D}(c) & \text{if } \tilde{N}(c) \leq T(\varepsilon c) \\ &= D(\varepsilon c) & \text{if } \tilde{N}(c) > T(\varepsilon c). \end{aligned}$$

Clearly,

$$e(\lambda_0, \tilde{\tilde{\delta}}(c)) \leq e(\lambda_0, \tilde{\delta}(c)) + e(\lambda_0, \delta(\varepsilon c)),$$

whence, by (66)

$$(67) \quad e(\lambda_0, \tilde{\tilde{\delta}}(c)) \leq e(\lambda_0, \delta(c)) + M\varepsilon c.$$

By Theorem 1

$$r_e(\lambda_0, \delta(c)) \leq r_e(\lambda_0, \tilde{\tilde{\delta}}(c)) + \varepsilon c$$

for sufficiently small c and, hence, using (67)

$$\sum_{i=1}^s \lambda_0(i) E_i T(c) \leq \sum_{i=1}^s \lambda_0(i) E_i \tilde{N}(c) + (M+1)\varepsilon.$$

Dividing by $\lambda_0(1)$ and using the definition of λ_0 leads to

$$E_1 T(c) \leq E_1 \tilde{N}(c) + \varepsilon \sum_{i=2}^s E_i (\tilde{N}(c) - T(c)) + (1 + (s-1)\varepsilon)(M+1)\varepsilon,$$

and the definition of $\tilde{N}(c)$ implies

$$E_1 T(c) \leq E_1 \tilde{N}(c) + \varepsilon \sum_{i=2}^s E_i (T(\varepsilon c) - T(c)) + s(M+1)\varepsilon.$$

Since ε is arbitrary, the conclusion (63) for $i = 1$ follows immediately once it is shown that there exist $M_1, M_2 > 0$ such that

$$(68) \quad E_i(T(\varepsilon c) - T(c)) \leq M_1 + M_2 \log \varepsilon^{-1}$$

for $i = 1, \dots, s$.

Since there are only a finite number of i values, it clearly suffices to prove (68) for one of them—say, $i = s$. To do this, note that $T(\varepsilon c) - T(c)$ is at most the number of observations required after time $T(c)$ for the random walks $\log(f_{sn}/f_{jn})$, $j = 1, \dots, s-1$, simultaneously to advance a distance

$$\begin{aligned} \zeta &= \log \varepsilon^{-1} && \text{if } D(c) = s \\ &= \log \left\{ \varepsilon c \min_{j < s} \left(\frac{f_{sT(c)}}{B_{js} f_{jT(c)}} \right)^{-1} \right\} && \text{if } D(c) \neq s. \end{aligned}$$

By Theorem 3.3 of Lorden (1967) the conditional (on ζ) E_s -expectation of the number of observations required is at most $V_1 + V_2 \zeta$, for some $V_1, V_2 > 0$.

Hence,

$$\begin{aligned} E_s(T(\varepsilon c) - T(c)) &\leq V_1 + V_2 E_s \zeta \\ &\leq V_1 + V_2 \log \varepsilon^{-1} + V_2 \log (\max_j B_{js}) \\ &\quad + V_2 (\log c^{-1}) P_s(D(c) \neq s) + V_2 \sum_{j=1}^{s-1} E_s \log \left(\sup_{n \geq 1} \frac{f_{jn}}{f_{sn}} \right), \end{aligned}$$

which suffices for (68) since

$$(\log c^{-1}) P_s(D(c) \neq s) \leq c (\log c^{-1}) \sum_{j=1}^{s-1} B_{sj}^{-1} \leq \frac{1}{e} \sum_{j=1}^{s-1} B_{sj}^{-1}$$

and

$$E_s \log \left(\sup_{n \geq 1} \frac{f_{jn}}{f_{sn}} \right) = \int_0^\infty P_s \left(\sup_{n \geq 1} \frac{f_{jn}}{f_{sn}} > e^t \right) dt \leq \int_0^\infty e^{-t} dt = 1.$$

REMARK. The full strength of (62) is not needed to prove that (63) holds for $i = 1$. Relation (67), with a larger M , can be derived straightforwardly under the assumptions that (62) is true for $i = 1$ and $\tilde{\alpha}_2(c), \dots, \tilde{\alpha}_s(c)$ are $O(c)$. The conclusion for $i = 1$ follows as before.

3. The numbers $L(i, j)$. If instead of Lemma 1 a simple calculation is made “neglecting excess over the boundary” for one-sided SPRT’s, then $R_1(u, v)$ is attained by using $u/v I(i, j)$ in place of $u/v L(i, j)$. This suggests that $L(i, j)/I(i, j)$

is close to 1 when the "excess" is small, which is borne out by the following result, a one-sided variant of the error probability approximations for (two-sided) SPRT's in Siegmund (1975).

THEOREM 5. *Suppose $I(i, j)$ is positive and (1) holds. Then if the random walk $\{\log(f_{in}/f_{jn})\}$ is not concentrated on a lattice,*

$$\frac{L(i, j)}{I(i, j)} = E_i e^{-\phi},$$

where $\phi > 0$ has the limit distribution as $\gamma \rightarrow \infty$ of the excess of $\{\log(f_{in}/f_{jn})\}$ over the boundary γ .

REMARK. If the random walk is concentrated on a lattice of width h and $\gamma \rightarrow \infty$ through multiples of h , then $E_i e^{-\phi} = h(e^h - 1)^{-1} L(i, j)/I(i, j)$ provided strict crossing of the boundary is required. In either case, $L(i, j) \leq I(i, j)$, and since $L(i, j) = L(j, i)$, $L(i, j) \leq \min(I(i, j), I(j, i))$.

PROOF. In the nonlattice case, it is well known ([4], pages 355–357) that ϕ has probability density function $P_i(\xi > x)/E_i \xi$, $0 < x < \infty$, where ξ denotes the first positive ladder variable of $\{\log(f_{in}/f_{jn})\}$. Thus, by Fubini's theorem,

$$E_i e^{-\phi} = \frac{\int_0^\infty e^{-x} P_i(\xi > x) dx}{E_i \xi} = \frac{E_i(\int_0^\infty e^{-x} dx)}{E_i \xi} = \frac{1 - E_i e^{-\xi}}{E_i \xi}.$$

Letting $\tilde{N}$ denote the time of first entry of $S_n = \log(f_{in}/f_{jn})$ into $(0, \infty)$, i.e., $\tilde{N} = \inf\{n | n \geq 1, f_{in} > f_{jn}\}$, and using Wald's equation, it follows that

$$I(i, j) E_i e^{-\phi} = \frac{1 - E_i e^{-\xi}}{E_i \tilde{N}} = \frac{1 - E_i(f_{j\tilde{N}}/f_{i\tilde{N}})}{E_i \tilde{N}} = \frac{P_j(\tilde{N} = \infty)}{E_i \tilde{N}}.$$

This last is seen to equal $L(i, j)$ exactly as in the proof of Lemma 1.

In the case of testing the mean of a normal distribution with variance one, if f_i and f_j have means θ_i and θ_j , then $I(i, j) = \frac{1}{2}|\theta_i - \theta_j|^2$ and

$$L(i, j) = \exp(-2 \sum_{n=1}^\infty n^{-1} \Phi(-\frac{1}{2} n^{1/2} |\theta_i - \theta_j|)).$$

TABLE 1
L-numbers in Normal case

$ \theta_i - \theta_j $	L	L/I	$\exp(-.58 \theta_i - \theta_j)$
.1	.004717	.9434	.9437
.2	.01780	.8901	.8905
.3	.03779	.8398	.8403
.4	.06339	.7924	.7929
.5	.09346	.7477	.7483
.6	.1270	.7056	.7061
.7	.1631	.6659	.6663
.8	.2012	.6286	.6288
.9	.2403	.5934	.5933
1.0	.2802	.5603	.5599

Table 1 illustrates some values of L and L/I together with a useful approximation to L/I , given in the last column. Table 1 of [11] is more extensive, listing $A = -\frac{1}{2} \log L(i, j)$ as a function of $w = \frac{1}{2}|\theta_i - \theta_j|$.

For the Wiener process, a direct argument based on the operating characteristics of one-sided SPRT's with critical likelihood ratios near 1 shows that $R(u, v)$ is attained by using $L(i, j) = I(i, j)$, which is to be expected since there is no excess over the boundary.

4. A binomial example. Computations of the operating characteristics of $\delta(c)$ and $\delta^*(c)$ were carried out in several cases of the following problem. The observations take only the values 0 and 1, with probabilities $1 - p$ and p , respectively. Values $0 < p_1 < p_0 < p_2 < 1$ are given. The terminal decisions are " p_1 is true" and " p_2 is true," with 0-1 loss, p_0 being an indifference point. The cost per observation is $c > 0$ and a priori probabilities μ_i , $i = 0, 1, 2$, are assigned to the corresponding p_i 's. All cases considered are symmetric in the sense that $p_2 = 1 - p_1$ and $p_0 = \frac{1}{2}$.

The method used for computing the Bayes solutions, $\delta^*(c)$, and the operating characteristics of $\delta^*(c)$ and $\delta(c)$ was the standard "backward induction" algorithm, which is thoroughly described for binomial cases in Weiss (1962). With this algorithm, each computer run for fixed c yields the operating characteristics for a number of a priori distributions. The binomial case is particularly simple computationally because the results for the point (t, y) are quickly derived from those for $(t + 1, y)$ and $(t + 1, y + 1)$, these being the only points immediately accessible from (t, y) in the course of sampling.

TABLE 2
Operating characteristics of $\delta(c)$ and $\delta^*(c)$; $p_1 = .35$, $\mu_i \equiv \frac{1}{3}$

	c	Error probabilities (%)	EN at p_1, p_2	EN at p_0	Integrated risk	Efficiency (%)
$\delta^*(c)$	.007	8.81	10.54	13.53	.1395	97.21*
$\delta(c)$		7.38	12.41	15.59	.1435	
$\delta^*(c)$	.005	6.71	12.60	16.86	.1148	96.67
$\delta(c)$		5.28	15.19	19.78	.1188	
$\delta^*(c)$	.004	5.27	14.21	19.91	.0996	97.61
$\delta(c)$		4.35	16.34	22.18	.1021	
$\delta^*(c)$	.003	4.16	15.92	23.08	.0826	97.10
$\delta(c)$		3.23	18.73	26.08	.0851	
$\delta^*(c)$	.002	2.85	18.26	28.48	.0623	98.17
$\delta(c)$		2.31	20.70	30.76	.0635	
$\delta^*(c)$	.001	1.37	23.20	38.86	.0375	98.95
$\delta(c)$		1.25	24.59	39.62	.0379	
$\delta^*(c)$	.0005	.68	27.64	49.32	.0220	97.59
$\delta(c)$		.57	30.56	51.14	.0225	

* The efficiencies were computed as the ratio of integrated risks accurate to six significant figures.

Runs were made with $p_1 = .35$ for $c = .007, .005, .004, .003, .002, .001, .0005$ and with $p_1 = .4$ for $c = .004$ and $.002$. The above table shows the results for $p_1 = .35$ with respect to the a priori distribution assigning equal weights of $\frac{1}{3}$ to the three points.

The results for $p_1 = .4$ and $c = .004$ and $.002$ are similar, giving efficiencies of 97.37 % and 97.17 % with $\delta(c)$ error probabilities of 9.81 % and 4.99 %, respectively.

Although there is an increasing trend as c decreases, the sequence of efficiencies in Table 2 is not monotonic. This is in contrast with similar results in Lorden (1975) and seems to be due to the fact that the set of lattice points falling in the continuation region of $\delta(c)$ (or $\delta^*(c)$) changes in a discontinuous and somewhat irregular fashion as c varies. As c is reduced the continuation regions of $\delta(c)$ and $\delta^*(c)$ grow and a given lattice point may first be included in one of the regions and, later, for somewhat smaller c , is finally included in both regions. Thus, for some values of c the symmetric difference of the two continuation regions may contain a relatively large number of lattice points and, for other values of c , a relatively small number. This affects the closeness of $\delta(c)$'s performance to that of $\delta^*(c)$.

Among all the results for a priori distributions satisfying the symmetry condition $\mu_1 = \mu_2$ the lowest efficiencies obtained with $p_1 = .35$ for $c = .001, .003$ and $.007$, for example, were 98.43 %, 96.37 % and 96.88 %, respectively, omitting cases where error probabilities exceeded .15. For $p_1 = .4$, the lowest symmetric-case efficiencies were 96.61 % for $c = .004$ and 97.01 % for $c = .002$, with cases involving high error probabilities (i.e., very small sample sizes) once again omitted.

Acknowledgments. The author is grateful to David Siegmund and Michael Klass for stimulating and helpful conversations, to the Associate Editor for a careful reading of the manuscript and useful suggestions, and especially to Jack Kiefer for his enthusiastic and inspiring guidance of the author's thesis research, of which the present work is an outgrowth.

REFERENCES

- [1] CHERNOFF, H. (1959). Sequential design of experiments. *Ann. Math. Statist.* **30** 755-770.
- [2] CHOW, Y. S., ROBBINS, H. and SIEGMUND, D. (1971). *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, Boston.
- [3] CHUNG, K. L. (1968). *A Course in Probability Theory*. Harcourt, Brace and World, New York.
- [4] FELLER, W. (1966). *An Introduction to Probability Theory and its Applications* 2. Wiley, New York.
- [5] KIEFER, J. and SACKS, J. (1963). Asymptotically optimum sequential inference and design. *Ann. Math. Statist.* **34** 705-750.
- [6] LAI, T. and SIEGMUND, D. (1976). A nonlinear renewal theory with applications to sequential analysis I. To appear.
- [7] LORDEN, G. (1967). Integrated risk of asymptotically Bayes sequential tests. *Ann. Math. Statist.* **38** 1399-1422.

- [8] LORDEN, G. (1972). Likelihood ratio tests for sequential k -decision problems. *Ann. Math. Statist.* **43** 1412-1427.
- [9] LORDEN, G. (1976). 2-SPRT's and the modified Kiefer-Weiss problem of minimizing an expected sample size. *Ann. Statist.* **4** 281-291.
- [10] SCHWARZ, G. (1962). Asymptotic shapes of Bayes sequential testing regions. *Ann. Math. Statist.* **33** 224-236.
- [11] SIEGMUND, D. (1975). Error probabilities and average sample number of the sequential probability ratio test. *J. Royal Statist. Soc. Ser. B* **37** 394-401.
- [12] SPITZER, F. (1956). A combinatorial lemma and its application to probability theory. *Trans. Amer. Math. Soc.* **82** 332-339.
- [13] WALD, A. (1947). *Sequential Analysis*. Wiley, New York.
- [14] WALD, A. and WOLFOWITZ, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* **19** 326-339.
- [15] WALD, A. and WOLFOWITZ, J. (1950). Bayes solutions of sequential problems. *Ann. Math. Statist.* **21** 82-99.
- [16] WEISS, L. (1962). On sequential tests which minimize the maximum expected sample size. *J. Amer. Statist. Assoc.* **57** 551-566.
- [17] WONG, S. P. (1968). Asymptotically optimum properties of certain sequential tests. *Ann. Math. Statist.* **39** 1244-1263.

DEPARTMENT OF MATHEMATICS
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125