This paper describes importance sampling techniques for Monte Carlo computation of the error probabilities and false alarm rates of sequential GLR (generalized likelihood ratio) tests and detection rules. It also discusses asymptotically optimal choice of the proposal distributions for importance sampling and gives numerical comparisons of the sequential GLR procedures with other sequential tests of composite hypotheses and detection rules in the literature.

Key Words: Sequential multiple hypothesis testing, sequential change-point detection, generalized likelihood ratio, importance sampling, multivariate exponential family.
1. INTRODUCTION

Let $X_1, X_2, \ldots$ be i.i.d. random vectors whose common density function belongs to a parametric family $\{f_\theta : \theta \in \Theta\}$ with respect to some dominating measure $\nu$ on $\mathbb{R}^d$. To test the simple null hypothesis $H_0 : \theta = \theta_0$ versus the simple alternative $H_1 : \theta = \theta_1$, Wald’s sequential probability ratio test (SPRT) stops sampling at stage

$$N = \inf\{n \geq 1 : L_n \notin (A, B)\},$$

rejecting $H_0$ if $L_N \geq B$, where $A < 1 < B$ and $L_n = \prod_{i=1}^n (f_{\theta_1}(X_i))/f_{\theta_0}(X_i)$ is the likelihood ratio statistic. A standard argument involving change of measures yields

$$P_{\theta_0}\{L_N \geq B\} \leq B^{-1}P_{\theta_1}\{L_N \geq B\}, \quad P_{\theta_1}\{L_N \leq A\} \leq AP_{\theta_0}\{L_N \leq A\},$$

in which $\leq$ can be replaced by $=$ if $L_N$ has to fall on either boundary exactly (i.e., there is no “overshoot”). Ignoring overshoots, Wald (1945) made use of (1.2) to obtain approximations to the error probabilities $P_{\theta_0}\{L_N \geq B\}$ and $P_{\theta_1}\{L_N \leq A\}$.

When $H_0$ is simple but $H_1$ is composite, Wald (1945) proposed to integrate the likelihood over the alternative hypothesis and to consider the integrated Type II error probability of the SPRT (1.1) with $L_n = \{\int \prod_{i=1}^n f_\theta(X_i)w(\theta)d\theta\}/\{|\prod_{i=1}^n f_0(X_i)\}$. Change of measures can again be used to yield (1.2) with $P_{\theta_0}$ replaced by $\int P_{\theta}w(\theta) \, d\theta$, which he used to approximate (by ignoring overshoots) the Type I and integrated Type II error probabilities of the test.

When $H_0 : \theta \in \Theta_0$ is composite, Pavlov (1990) proposed to replace the denominator $\prod_{i=1}^n f_0(X_i)$ of $L_n$ by the maximized likelihood $\sup_{\theta \in \Theta_0} \prod_{i=1}^n f_\theta(X_i)$, noting that for any $\theta_0 \in \Theta_0$,

$$\left\{ \prod_{i=1}^n f_{\theta_1}(X_i)/\sup_{\theta \in \Theta_0} \prod_{i=1}^n f_\theta(X_i) \geq B \right\} \subseteq \left\{ \prod_{i=1}^n f_{\theta_1}(X_i)/\prod_{i=1}^n f_0(X_i) \geq B \right\}. \quad (1.3)$$

If $H_1 : \theta \in \Theta_1$ is also composite, Wald’s weight function approach replaces $\prod_{i=1}^n f_{\theta_1}(X_i)$ by $\int_{\Theta_1} \{\prod_{i=1}^n f_\theta(X_i)\} w(\theta) \, d\theta$. Pavlov (1990), however, used an alternative approach which replaces $\theta$ at stage $i$ by the estimate $\hat{\theta}_{i-1} = \theta_{i-1}(X_1, \ldots, X_{i-1}) \in \Theta$ and stops sampling at stage

$$N^* = \inf\{n \geq 1 : \max_{j=0,1} L_{n,j} \geq B\}, \quad \text{where} \quad L_{n,j} = \left\{ \prod_{i=2}^n f_{\theta_{i-1}}(X_i) \right\}/\sup_{\theta \in \Theta_j} \prod_{i=2}^n f_\theta(X_i),$$

and which rejects $H_j$ if $L_{N^*,j} \geq B$. There are three reasons underlying Pavlov’s choice of the stopping rule. The first is that it can be easily extended to $k > 2$ hypotheses. The second is
that in the case \( \theta_i \equiv \theta \) and simple \( H_0 \) and \( H_1 \), (1.4) reduces to the 2-SPRT, which Lorden (1976) showed to be asymptotically optimal in minimizing \( E_\theta(T) \) among all tests with no larger Type I and Type II error probabilities. Replacing \( \theta \) by a consistent estimate \( \hat{\theta}_i \) can still preserve the asymptotic optimality property, as Pavlov showed under certain assumptions. The third reason is that \( \{ \prod_{i=1}^{n} (f_{\theta_{i-1}}(X_i)/f_{\theta_0}(X_i)), \ n \geq 1 \} \) is a nonnegative martingale with mean 1 under \( P_{\theta_0} \) when \( \theta_{i-1} \) depends only on \( X_1, \ldots, X_{i-1} \), and therefore

\[
P_{\theta} \left\{ \prod_{i=1}^{n} (f_{\theta_{i-1}}(X_i)/f_{\theta}(X_i)) \geq B \text{ for some } n \geq 1 \right\} \leq B^{-1}, \tag{1.5}
\]

as observed earlier by Robbins and Siegmund (1972, 1974). Combining (1.5) with (1.3) in which \( \theta_1 \) is replaced by \( \theta_{i-1} \) then yields the following simple bound on the error probabilities:

\[
\sup_{\theta \in \Theta} P_{\theta} \{ L_{N^*;j} \geq B \} \leq B^{-1}. \tag{1.6}
\]

Although these mixture likelihood ratios and adaptive likelihood ratios lead to simple extensions of Wald’s bounds for the error probabilities and to simple modifications of the SPRT when the hypotheses are composite, there are certain issues in applying them in practice. For mixture likelihood ratios, the choice of the function \( w \) implicitly weights the alternatives and how \( w \) should be chosen requires some prior knowledge of the actual alternative. For adaptive likelihood ratios, although at stage \( n \) one has an estimate \( \hat{\theta}_n \) that depends on all the available data \( X_1, \ldots, X_n \), one is only allowed to associate each \( X_i \) with \( \hat{\theta}_{i-1} \). Thus, some poor early estimates (for small \( i \)) can never be revised even though one has a large number \( n \) of observations. Moreover, since the denominator \( L_{n,j} \) in (1.4) already uses the maximum likelihood estimate constrained to lie in the set \( \Theta_j \), not using the maximum likelihood estimate (over the entire parameter space, as in the usual generalized likelihood ratio statistics) in the numerator seems to be somewhat contrived and inefficient even though (1.4) has been shown by Pavlov to be asymptotically efficient under certain conditions.

Another issue is that the error probability bound (1.2) and (1.6) may be too crude as approximations to the actual probabilities. Even for the simple case of Wald’s SPRT, ignoring the overshoot in Wald’s approximation can inflate the actual probability by a substantial factor, which led Siegmund (1975) to develop better approximations via asymptotic analysis of overshoots. Instead of analytic approximations, an alternative approach is to compute the error probabilities by Monte Carlo methods, and Siegmund (1976) made use of the likelihood ratio (or mixture likelihood ratio) structure of the sequential test statistics to find suitable proposal distributions for importance sampling in the Monte Carlo evaluation of boundary crossing probabilities of these statistics.
Instead of mixture or adaptive likelihood ratio statistics, we use herein generalized likelihood ratio (GLR) statistics
\[
\hat{L}_{n,j} = \left\{ \frac{\sup_{\theta \in \Theta} \prod_{i=1}^{n} f_\theta(X_i)}{\sup_{\theta \in \Theta_j} \prod_{i=1}^{n} f_\theta(X_i)} \right\}
\] (1.7)
to construct sequential tests of \(H_j: \theta \in \Theta_j\), and show in Section 2 that GLR sequential tests can have substantial savings in expected sample size over mixture and adaptive likelihood ratio counterparts. Although we no longer have the nice error bound (1.2) or (1.5) because the numerator of (1.7) is no longer a density, we can use Monte Carlo to compute the error probabilities of sequential GLR tests. Importance sampling techniques for efficient Monte Carlo computation of the error probabilities are described in Section 3. Section 4 extends the importance sampling methods of Section 3 to sequential change-point detection using window-limited GLR detection rules. With these importance sampling techniques for Monte Carlo computation of error probabilities and false alarm rates, critical thresholds for sequential GLR procedures can be determined as readily as for their fixed-sample-size counterparts in multiparameter and in time series testing and detection problems.

2. SEQUENTIAL GLR TESTS OF MULTIPLE HYPOTHESES

Lai (2003) has recently given a review of the theory of sequential GLR tests of the composite hypotheses \(H_0: \theta \in \Theta_0\) versus \(H_1: \theta \in \Theta_1\), where \(\Theta_0\) and \(\Theta_1\) are disjoint subsets of the natural parameter space of a multiparameter exponential family. When \(\Theta_0\) and \(\Theta_1\) are separated by a fixed indifference zone, Schwarz (1962), Fortus (1979) and Woodroofe (1980) have shown that sequential GLR tests with suitably chosen constant (i.e., time-invariant) stopping boundaries are asymptotically Bayes risk efficient as the sampling cost approaches 0. Lai (1988, 1997) and Lai and Zhang (1994) have shown how relatively simple time-varying boundaries can be constructed so that the corresponding sequential GLR tests still provide asymptotic solutions to the Bayes testing problem that incurs a sampling cost besides the loss due to incorrect terminal decision.

These sequential GLR tests and their underlying theory can be readily extended to \(k \geq 2\) hypotheses. For simplicity we shall consider the case of time-invariant stopping boundaries, assuming that any two such hypotheses are separated by an indifference zone. Let \(\Theta_1, \ldots, \Theta_k\) be disjoint subsets of \(\Theta\) and let \(H_j: \theta \in \Theta_j\) for \(1 \leq j \leq k\). Let \(B_j > 1\) be some pre-determined threshold level and let
\[
N_j = \inf\{n : \hat{L}_{n,j} > B_j\}
\] (2.1)
be the time at which the hypothesis $H_j$ is rejected by the sequential GLR test and eliminated from further consideration. As soon as there is only one hypothesis that has not been rejected, it is accepted and sampling terminates. More specifically, let $M = \max_{1 \leq j \leq k} N_j$. The $j$th hypothesis is accepted if $N_j = M$ and $\tilde{L}_{M,j} \leq B_j$. The terminal decision is $\delta = j$ if $H_j$ is accepted.

Instead of the GLR statistic, Pavlov (1990) and Dragalin and Novikov (1995) used the adaptive likelihood ratio (ALR) statistic $L_{n,j}$ defined in (1.4). Alternatively, one can follow Wald’s (1945) weight function approach and replace $\hat{L}_{n,j}$ in (2.1) by the mixture likelihood ratio statistic

$$\tilde{L}_{n,j} = \left\{ \int_{\Theta} \prod_{i=1}^{n} f_\theta(X_i) w(\theta) d\theta \right\} / \sup_{\theta \in \Theta_j} \prod_{i=1}^{n} f_\theta(X_i),$$

(2.2)

where $w$ is a weight function satisfying $\int_{\Theta} w(\theta) d\theta = 1$; see Robbins (1970) and Pollak (1978). From (1.5) and (1.3) (in which $f_\theta_1(X_i)$ is replaced by $f_{\theta_{i-1}}(X_i)$), it follows that

$$\sup_{\theta \in \Theta_j} P_\theta\{L_{n,j} \geq B_j \text{ for some } n \geq 1\} \leq B_j^{-1}.$$  

(2.3)

Similarly, (2.3) still holds if $L_{n,j}$ is replaced by the mixture likelihood ratio statistic $\tilde{L}_{n,j}$. Hence, the error probability $\sup_{\theta \in \Theta_j} P_\theta\{\delta \neq j\}$ does not exceed some prescribed bound $\alpha_j$ if we choose $B_j = 1/\alpha_j$ for the sequential ALR or mixture likelihood ratio test, abbreviated by “MIX” hereafter. However, the upper bound (2.3) is often substantially larger than $P_\theta\{\delta \neq j\}$ for $\theta \in \Theta_j$ and Monte Carlo provides a better way to evaluate the error probability. The following example gives more details about the Monte Carlo approach and its applications to compute the error probabilities and expected sample sizes of sequential GLR (or ALR or MIX) tests.

**Example 1.** Let $f_\theta$ be the trivariate normal density with mean $\theta = (\theta_1, \theta_2, \theta_3)$ and identity covariance matrix. Let $\Theta = \mathbb{R}^3$ and $\Theta_j = \{\theta : \theta_j \geq \max_{i \neq j} \theta_i + \Delta\}$ for $j = 1, 2, 3$, where $\Delta > 0$ so that $\Theta_j$ and $\Theta_i$ are separated by an indifference zone for $j \neq i$. To test $H_j$: $\theta \in \Theta_j$ sequentially by applying the stopping rule $T$ to GLR, or MIX, or ALR statistics subject to the constraint that the error probability of falsely rejecting $H_j$ be $\leq \alpha$ for $j = 1, 2, 3$, we need to determine the threshold $B$, noting that $\alpha_j \equiv \alpha$ implies that the $B_j$ in (2.1) have a common value $B$. Since

$$\sup_{\theta \in \Theta_j} P_\theta\{\delta \neq j\} = P_{(\Delta,0,0)}\{\delta \neq 1\} = P_{(0,\Delta,0)}\{\delta \neq 2\} = P_{(0,0,\Delta)}\{\delta \neq 3\},$$

we can determine $B$ via $P_{(\Delta,0,0)}\{\delta \neq 1\} = \alpha$. Letting $h(B)$ denote this error probability when the underlying threshold is $B$, we approximate $h(B)$ for $B < B'$ by the proportion $h_K(B)$
of incorrect terminal decisions in $K$ samples of $(X_1, X_2, \ldots)$ generated from $P_{(\Delta,0,0)}$ until $\max_j L_{n,j}$ (or its ALR and MIX counterparts) exceeds $B'$, and use Brent’s method to solve for $h_K(B) = \alpha$, in which equality is deemed to hold when $h_K(B) - \alpha$ is within 2 standard errors of the Monte Carlo estimate.

All three test statistics of $H_1$ have the same denominator which involves finding $\tilde{\theta} \in \Theta_1$ that is closest to $\bar{X}_n$. When $\bar{X}_n \in \Theta_1$, $\tilde{\theta} = \bar{X}_n$. When $\bar{X}_n \notin \Theta_1$, either $\tilde{\theta} = \frac{1}{2}(\bar{X}_{n1} + \bar{X}_{n2} + \Delta, \bar{X}_{n1} + \bar{X}_{n2} - \Delta, 2\bar{X}_{n3})$, which is the linear projection of $\bar{X}_n$ onto $\Theta_{12} := \{ \theta : \theta_1 = \theta_2 + \Delta \}$, or $\tilde{\theta} = \frac{1}{3}(\bar{X}_{n1} + \bar{X}_{n3} + \Delta, 2\bar{X}_{n2}, \bar{X}_{n1} + \bar{X}_{n3} - \Delta)$, which is the linear projection of $\bar{X}_n$ onto $\Theta_{13} := \{ \theta : \theta_1 = \theta_3 + \Delta \}$, or $\tilde{\theta} = \frac{1}{3}(\sum_j \bar{X}_{nj} + 2\Delta, \sum_j \bar{X}_{nj} - \Delta, \sum_j \bar{X}_{nj} - \Delta)$, which is the projection of $\bar{X}_n$ onto $\Theta_{12} \cap \Theta_{13}$. The GLR statistic has the simple form

$$\log \hat{L}_{n,1} = \frac{n}{2} \{ (\bar{X}_{n1} - \tilde{\theta}_1)^2 + (\bar{X}_{n2} - \tilde{\theta}_2)^2 + (\bar{X}_{n3} - \tilde{\theta}_3)^2 \}. \quad (2.4)$$

The numerator of the ALR statistic in (1.4) with $\tilde{\theta}_{i-1} = \bar{X}_{i-1}$ can be computed recursively. For the MIX statistic, if we choose the mixture density function $w$ to be trivariate normal with mean $\mathbf{0}$ and covariance matrix $\beta^{-1}I_3$, then the integral in the numerator of (2.2) has the closed-form expression

$$(2\pi)^{-3n/2} \left( \frac{\beta}{n + \beta} \right)^{1/2} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{3} \left( \sum_{i=1}^{n} (X_{ij} - \bar{X}_{nj})^2 + \frac{\beta_n}{n + \beta} \bar{X}_{nj}^2 \right) \right\}. \quad (2.5)$$

Choosing $\Delta = 0.2$, $\alpha = 0.05$, $K = 2000$ and $\beta = 1$, the last row of Table 1 gives the logarithm of the threshold $B$ for each of the three test statistics. If we had determined $B$ via the simple upper bound $\sup_{\theta \in \Theta_1} P_{\theta} \{ L_{n,1} \geq B \text{ for some } n \geq 1 \} \leq B^{-1}$ for the ALR statistics (see (1.5)), which also holds for MIX, then $B^{-1} = \alpha$ would have given log $B$ to be approximately 3, which is considerably larger than the corresponding values for MIX and ALR in Table 1. Not only does this simple upper bound ignore the overshoot, but it also considers only $L_{n,1}$ and discards the interplay between the three sequences $L_{n,1}, L_{n,2}$ and $L_{n,3}$ in the stopping and terminal decision rules. The large discrepancy between the thresholds of the sequential GLR and MIX tests in the last row of Table 1 can be explained by

$$\log \hat{L}_{n,j} - \log \tilde{L}_{n,j} = \frac{3}{2} \log \left( \frac{n + \beta}{\beta} \right) + \frac{1}{2} \left( \frac{\beta_n}{n + \beta} \right) (\bar{X}_{n1}^2 + \bar{X}_{n2}^2 + \bar{X}_{n3}^2),$$

which follows from (2.4) and (2.5).

Table 1 gives the error probabilities and expected sample sizes of the three sequential tests at various values of $(\theta_1,0,0)$. Each result is based on 2000 simulations. We observe that the sequential GLR test outperforms the other two tests, showing reductions in both the
### Table 1. Performance of sequential tests of multiple hypotheses at $\theta = (\theta_1, 0, 0)$, with standard errors in parentheses.

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>Error Probability</th>
<th>$E_\theta T$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLR</td>
<td>ALR</td>
</tr>
<tr>
<td>0.2</td>
<td>0.052</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>0.3</td>
<td>0.016</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>0.4</td>
<td>0.009</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.006</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.239</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>0</td>
<td>0.674</td>
<td>0.662</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>log $B$</td>
<td>4.45</td>
<td>2.7</td>
</tr>
</tbody>
</table>

error probability when $\theta_1 > 0.3$ and the expected sample size, and the reduction in expected sample size is considerable, particularly at $\theta_1 = 0$.

### 3. IMPORTANCE SAMPLING FOR SEQUENTIAL GLR TESTS

To simulate the Type I error probability $\alpha = P_{\theta_0}\{L_N \geq B\}$ of the SPRT with stopping rule (1.1), the direct Monte Carlo method generates $K$ realizations of $(X_1, \ldots, X_N)$ from $f_{\theta_0}$ and estimates $\alpha$ by the average $\hat{\alpha}$ of the $K$ realizations of the indicator variable $1_{\{L_N \geq B\}}$. This Monte Carlo method, however, is not efficient when $\alpha$ is small. In fact, s.e.$(\hat{\alpha})/\alpha \sim \alpha^{-1/2} \to \infty$ as $\alpha \to 0$, where s.e. refers to the standard error of an estimate.

A Monte Carlo method, or its associated estimator $\hat{\alpha}$, for providing an unbiased estimate of a probability $\alpha$ is said to be *asymptotically efficient* if

$$E\hat{\alpha}^2 = O(\alpha^2) \text{ as } \alpha \to 0;$$

see Sadowsky and Bucklew (1990) and Chan and Lai (2003b). Clearly (3.1) holds if there exists a nonrandom constant $C$ not depending on $\alpha$ such that

$$\hat{\alpha} \leq C\alpha \text{ a.s.}$$
In particular, to estimate the Type I error probability \( \alpha = P_{\theta_0} \{ L_N \geq B \} \) of the SPRT, Siegmund’s (1976) importance sampling method that estimates \( \alpha \) by the average \( \hat{\alpha}_{\theta_1} \) of \( L_{N}^{-1} \mathbf{1}_{\{ L_N \geq B \}} \) over \( K \) realizations of \( (X_1, \ldots, X_N) \) sampled from \( f_{\theta_1} \) satisfies (3.2) (and therefore also (3.1)) since \( \hat{\alpha}_{\theta_1} \leq B^{-1} \) and \( \alpha \) is asymptotically equivalent to constant times \( B^{-1} \), where the constant comes from excess over the boundary. Siegmund (1976, Theorem 1) has also shown that \( f_{\theta_1} \) is the only asymptotically efficient density in a one-parameter exponential family \( f_{\theta}(x) = \exp \left\{ \theta x - \psi(\theta) \right\} \) with respect to some dominating measure on the real line. Specifically, as \( B \to \infty \),

\[
E_{\theta_1}[\hat{\alpha}^2_{\theta_1}] / E_{\theta_2}[\hat{\alpha}^2_{\theta_2}] \to 0 \text{ for all } \theta \neq \theta_1. \tag{3.3}
\]

This theory of asymptotically efficient importance sampling schemes can be extended from the SPRT to sequential GLR tests. To fix the ideas, consider Monte Carlo evaluation of the error probability \( P_{\theta_0} \{ \hat{L}_{N,0} \geq B \} \) of a sequential GLR test of \( H_0: \theta \in \Theta_0 \), where \( \theta_0 \in \Theta_0 \) and

\[
N = \inf \{ n \geq n_0 : \hat{L}_{n,0} \geq B \} \land n_1. \tag{3.4}
\]

The test rejects \( H_0 \) if \( \hat{L}_{N,0} \geq B \). Whereas the stopping rule (1.1) of the SPRT involves \( f_{\theta_1} \) which turns out to be an asymptotically efficient importance density, the stopping rule (3.4) of the sequential GLR test does not specify an alternative density, but the asymptotically efficient importance density \( \prod_{i=1}^{n} f_{\theta_1}(x_i) \) for the SPRT can be extended to \( \int \prod_{i=1}^{n} f_{\theta}(x_i) \ dW(\theta) \) for some suitably chosen \( W \). How such \( W \) should be chosen has recently been addressed by Chan and Lai (2003b) in the context of a \( d \)-dimensional exponential family \( f_{\theta}(x) = \exp \{ (\theta - \theta_0)'x - \psi(\theta) \} \) of densities with respect to \( P_{\theta_0} \). For notational simplicity we shall assume that \( \theta_0 = 0 \) and reparameterize the model in terms of \( \mu = (\nabla \psi)(\theta) \).

Let \( c = \log B \),

\[
\theta_\mu = (\nabla \psi)^{-1}(\mu), \quad \phi(\mu) = \sup_{\theta \in \Theta} \{ \theta'\mu - \psi(\theta) \} = \theta_\mu' \mu - \psi(\theta_\mu),
\]

\[
g(\mu) = \phi(\mu) - \sup_{\theta \in \Theta_0} \{ \theta'\mu - \psi(\theta) \}. \tag{3.5}
\]

Then \( \log \hat{L}_{n,0} = ng(\bar{X}_n) \), so the sequential GLR test has a rejection region of the form

\[
\Gamma := \{ ng(\bar{X}_n) \geq c \text{ for some } n_0 \leq n \leq n_1 \} = \{ Ng(\bar{X}_N) \geq c \}. \tag{3.6}
\]

Chan and Lai (2003b) have developed asymptotically efficient importance sampling techniques to compute \( P_0(\Gamma) \) for smooth functions \( g \) (not necessarily of the GLR form in (3.5)). In particular, they have shown that under certain conditions on \( g \), an asymptotically
efficient importance density is of the form \( \int \prod_{i=1}^{n} f_{\theta_{i}}(x_i) w_c(\mu) \, d\mu \), where for some \( 0 < \epsilon < a_1^{-1} \),

\[
w_c(\mu) = \begin{cases} 
\beta_c \max\{a_0, 1/g(\mu)\}^{d/2} \exp\{ - \max[n_0, c/g(\mu)] \phi(\mu) \} & \text{if } g(\mu) > a_1^{-1} - \epsilon, \\
0 & \text{if } g(\mu) \leq a_1^{-1} - \epsilon,
\end{cases}
\]

in which \( \beta_c \) is a normalizing constant so that \( \int w_c(\mu) \, d\mu = 1 \). Let \( \tilde{\alpha} \) be the average of

\[
\left\{ \int \prod_{i=1}^{N} \left[ f_{\theta_i}(X_i)/f_0(X_i) \right] w_c(\mu) \, d\mu \right\}^{-1} 1_{\Gamma}
\]

over \( K \) realizations of \( (X_1, \ldots, X_N) \) sampled from the importance density \( \int \prod_{i=1}^{n} f_{\theta_{i}}(x_i) w_c(\mu) \, d\mu \). Chan and Lai (2003b) have shown that \( \tilde{\alpha} \) is asymptotically efficient for estimating \( P_0\{ng(\tilde{X}_n) \geq c \} \) for some \( n_0 \leq n \leq n_1 \) as \( c \to \infty \), \( n_0 \sim a_0 c \) and \( n_1 \sim a_1 c \).

To apply the preceding importance sampling method to the sequential GLR test with stopping rule (3.4), first consider the case where \( H_0 \) is simple so that \( \Theta_0 = \{0\} \), assuming without loss of generality that \( \theta_0 = 0 \). Then \( \phi = g \) and (3.7) can be rewritten as

\[
w_c(\mu) = \begin{cases} 
\beta_c e^{-c\gamma(\mu)} & \text{if } a_1^{-1} - \epsilon < \phi(\mu) < a_0^{-1}, \\
a_0^{-1/2} \beta_c \exp\{ -(a_0 + o(1))c\phi(\mu) \} & \text{if } \phi(\mu) \geq a_0^{-1},
\end{cases}
\]

where \( \gamma \) is integrable over the region \( \{ \mu : a_1^{-1} - \epsilon < \phi(\mu) < a_0^{-1} \} \). As noted by Chan and Lai (2003b), if \( \int \prod_{i=1}^{n} f_{\theta_{i}}(x_i) w_c(\mu) \, d\mu \) is an asymptotically efficient importance density, then so is \( \int \prod_{i=1}^{n} f_{\theta_{i}}(x_i) v(\mu) \, d\mu \) for any density \( v(\mu) \) satisfying \( v(\mu)/w_c(\mu) > \delta \) for some \( \delta > 0 \) and all \( \mu \). In the case \( g = \phi \), (3.9) implies that \( \beta_c e^{-c} \) is bounded by \( \left[ \int_{a_1^{-1} - \epsilon}^{a_0^{-1}} \gamma(\mu) \, d\mu \right]^{-1} \) and hence we can choose any density function \( v \) such that

\[
\inf_{a_1^{-1} - \epsilon < \phi(\mu) < a_0^{-1}} v(\mu) > 0, \quad \inf_{\phi(\mu) > a_0^{-1}} \left[ v(\mu) / \exp\{ -(a_0 + o(1))c\phi(\mu) + c \} \right] > 0
\]

for all large \( c \). This gives us considerable flexibility of choosing \( v \) such that \( \int \prod_{i=1}^{n} f_{\theta_{i}}(x_i) v(\mu) \, d\mu \) has a convenient form and is an asymptotically efficient importance density. In particular, when \( f_{\theta_{i}} \) is a normal density function with mean \( \mu \) and variance \( \sigma^2 \) (so \( \theta_{i} = \mu / \sigma^2 \)), we can choose \( v(\mu) \) to be any normal density, noting that \( \phi(\mu) = \mu^2 / 2\sigma^2 \). Thus a normal mixture of \( f_{\theta} \) is an asymptotically efficient importance density in this case.

**Example 2.** Let \( Y_i = (Y_{i1}, Y_{i2}) \), \( i = 1, 2, \ldots \) be i.i.d bivariate normal with \( EY_{ij} = \mu_j \) and \( \text{Var}(Y_{ij}) = \sigma_j^2 \) for \( j = 1, 2 \) and \( \text{Cov}(Y_{11}, Y_{12}) = 0 \). A sequential GLR test of \( H_0 : \mu_1 = \mu_2 \) with known \( \sigma_j^2 \) uses a rejection region \( \Gamma \) of the form (3.6) with \( X_i = (Y_{i1} - Y_{i2}) / \sqrt{\sigma_1^2 + \sigma_2^2} \).
and \( g(\mu) = \mu^2/2 \). Under \( H_0 \), \( X_i \) is univariate normal with mean 0 and variance 1. An asymptotically efficient importance density is \( (2\pi)^{-n/2} \int e^{-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2} v(\mu) \, d\mu \), where \( v \) is a normal density function with mean 0 and variance \( \beta^{-1} \). The importance weight, which is the multiplier of \( 1 \) in (3.8), has the simple form

\[
\left( \frac{\beta}{2\pi} \right)^{1/2} \left[ \int_{-\infty}^{\infty} \exp \left\{ N \left( \mu \bar{X}_N - \frac{\mu^2}{2} \right) - \frac{\mu^2 \beta}{2} \right\} \, d\mu \right]^{-1} = \left( \frac{\beta}{N + \beta} \right)^{1/2} \exp \left( \frac{N^2 \bar{X}_N^2}{2(N + \beta)} \right).
\]

### Table 2. Error probabilities (mean ± s.e.) of sequential test

<table>
<thead>
<tr>
<th>( c )</th>
<th>( n_0 )</th>
<th>( n_1 )</th>
<th>Direct MC</th>
<th>Importance sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>25</td>
<td>( (9.20 \pm 0.95) \times 10^{-3} )</td>
<td>( (9.45 \pm 0.13) \times 10^{-3} )</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>50</td>
<td>( (1.31 \pm 0.11) \times 10^{-2} )</td>
<td>( (1.388 \pm 0.016) \times 10^{-2} )</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>100</td>
<td>( (1.81 \pm 0.13) \times 10^{-2} )</td>
<td>( (1.811 \pm 0.021) \times 10^{-2} )</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>50</td>
<td>( (1.55 \pm 0.12) \times 10^{-2} )</td>
<td>( (1.539 \pm 0.017) \times 10^{-2} )</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>50</td>
<td>( (1.0 \pm 1.0) \times 10^{-4} )</td>
<td>( (0.838 \pm 0.012) \times 10^{-4} )</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>100</td>
<td>( (1.0 \pm 1.0) \times 10^{-4} )</td>
<td>( (1.246 \pm 0.015) \times 10^{-4} )</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>100</td>
<td>( (1.0 \pm 1.0) \times 10^{-5} )</td>
<td>( (0.860 \pm 0.012) \times 10^{-4} )</td>
</tr>
</tbody>
</table>

Importance sampling in this case means first drawing \( \mu \) from the \( N(0, \beta^{-1}) \) distribution and then \( X_1, \ldots, X_N \) from the \( N(\mu, 1) \) distribution. We performed 10,000 simulation runs each for both direct Monte Carlo and importance sampling for various combinations of \( c, n_0, n_1 \). The results summarized in Table 2 show a 50- to 100-fold variance reduction in using importance sampling to simulate probabilities of order 0.01. For probabilities of order \( 10^{-4} \), the direct Monte Carlo estimate is unreliable whereas importance sampling still gives accurate estimates of these small probabilities.

When \( \Theta_0 \) is a \( (d - q) \)-dimensional submanifold of \( \Theta \), the function \( g \) given by (3.5) for GLR statistics is equal to \( \phi \) on a \( q \)-dimensional submanifold of \( \Theta \); see Chan and Lai (2000, page 1644). In the preceding discussion, \( q = d \) as \( \Theta_0 = \{0\} \) corresponds to a simple null hypothesis. When \( q < d \) and there are difficulties working with the normalizing constant \( \beta_c \) in (3.9) or with the importance weights in (3.8), we can use a discrete approximation to the weighting measure \( W (dW(\mu) = w_c(\mu) \, d\mu) \) to circumvent these difficulties. Specifically, restrict \( \mu \) to a finite set \( S \) of points in \( \nabla \psi(\Theta) \) and sample \( \mu \) from \( S \) with probabilities proportional to (3.7). This is illustrated in the following example and is discussed further in Chan and Lai (2003b) who also propose another approximation that involves sequential importance sampling with resampling.
Example 3 (Sequential t-test). Let \( Y_{i1}, Y_{i2}, (i = 1, 2, \ldots) \) be independent normal variables with common variance \( \sigma^2 \) and \( E(Y_{i1}) = a_1, E(Y_{i2}) = a_2 \). To test \( H_0 : a_1 = a_2 \) with unknown \( \sigma \), define \( X_i = (Y_{i1}, Y_{i2}, Y_{i1}^2 + Y_{i2}^2) \). Since

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{\theta_1 y_1 + \theta_2 y_2 + \theta_3 (y_1^2 + y_2^2) - (y_1^2 + y_2^2)/2\sigma^2} dy_1 dy_2 < \infty \text{ iff } \theta_3 < (2\sigma^2)^{-1},
\]

the natural parameter space of \( X_i \) is \( \Theta = \{ \theta : \theta_3 < (2\sigma^2)^{-1} \} \) and \( H_0 \) corresponds to the submanifold \( \Theta_0 = \{ \theta \in \Theta : \theta_1 = \theta_2 \} \). The joint distribution of the sequential GLR statistics under \( H_0 \) does not depend on \((\sigma, a, a)\), which we can choose to be \((1, 0, 0)\) when simulating the Type I error probability of the sequential GLR test. Specializing (3.5) to the present setting yields

\[
g(\mu) = \log \left\{ 1 + \frac{(\mu_1 - \mu_2)^2}{2(\mu_3 - \mu_1^2 - \mu_2^2)} \right\}, \quad \phi(\mu) = \frac{\mu_3}{2} - 1 - \log \left\{ \frac{\mu_3 - \mu_1^2 - \mu_2^2}{2} \right\},
\]

which we substitute into (3.7) with \( d = 3 \). For a discrete approximation to the measure with density function (3.7), take \( \zeta > 0 \) and let \( \tilde{S} = \{ \mu \in (Z/\zeta)^3 : g(\mu) > c/n_0, \phi(\mu) < c/n_1 \} \), where \( Z \) denotes the set of integers. Define the normalized weights \( \tilde{w} \) at each \( \mu \in \tilde{S} \) to be proportional to (3.7). Let \( S = \{ \mu \in \tilde{S} : \tilde{w}(\mu) \geq 0.01 \} \), which corresponds to pruning \( \tilde{S} \) by removing points with weights less than 1%. The importance sampler first samples \( \mu \) from \( S \) with weights \( w \) renormalized from \( \tilde{w} \) and then generates \( X_1, \ldots, X_N \) independently from the density function \( f_{\theta_\mu} \). This is repeated \( K = 10,000 \) times and the average of (3.8) over the \( K \) simulation runs is taken as the estimate of the Type I error probability of the sequential GLR test. The results are summarized in Table 3, which shows that importance sampling has a much smaller standard error than direct Monte Carlo. Moreover, due to more frequent early stopping, the simulation time per run is also smaller in importance sampling, resulting in savings of computing time.

4. WINDOW-LIMITED SEQUENTIAL GLR RULES

Motivated by applications to statistical quality control and fault detection in the engineering literature, Lai (1995, 1998) developed a theory of sequential detection, with low false alarm rate, of parameter changes in stochastic systems. This theory leads to window-limited sequential GLR detection rules which are not too demanding in computational and memory requirements and yet are nearly optimal under several performance criteria. Lai and Shan (1999) introduced importance sampling techniques for determining the threshold of a sequential GLR detection rule to satisfy prescribed bounds on the false alarm rate. In
this section we establish the asymptotic efficiency of the importance density introduced by Lai and Shan (1999) by generalizing the corresponding theory of importance sampling for sequential GLR tests considered in the preceding section.

To begin with, consider the problem of detecting a change from a known baseline value $\theta_0$ (assumed to be 0 without loss of generality) of the parameter in the $d$-dimensional exponential family considered in Section 3. In the case $d = 1$, Lorden (1971) showed that Page’s (1954) CUSUM rule

$$ T = \inf \{ n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log(\frac{f_{\theta_1}(X_i)}{f_0(X_i)}) \geq c \} $$

is asymptotically optimal as $c \to \infty$ (in the sense of minimizing a worst-case detection delay subject to the false alarm constraint $E_0 T \geq \gamma$, with $c \sim \log \gamma$) for detecting a change of $\theta$ from 0 to some specified value $\theta_1$. In practice the post-change parameter value is typically unknown and he also showed that the asymptotic optimality property can be extended to the GLR detection rule

$$ T^* = \inf \{ n \geq 1 : \max_{1 \leq k \leq n} \sup_{|\theta| \geq \epsilon_c} \sum_{i=k}^{n} \log(\frac{f_{\theta}(X_i)}{f_0(X_i)}) \geq c \}, $$

where $\epsilon_c \sim 1/c$ is chosen to facilitate the analysis of $E_0 T^*$ in the asymptotic optimality proof. Instead of maximizing the log-likelihood ratio over $|\theta| \geq \epsilon_c$ in (4.2), Pollak and Siegmund (1975) proposed to integrate the likelihood ratio with respect to some probability distribution $G$, and showed that the mixture likelihood ratio (MIX) detection rule

$$ \tilde{T} = \inf \{ n \geq 1 : \max_{1 \leq k \leq n} \int \prod_{i=k}^{n} \frac{f_{\theta}(X_i)}{f_0(X_i)} \ dG(\theta) \geq \epsilon_c \} $$

<table>
<thead>
<tr>
<th>$c$</th>
<th>$n_0$</th>
<th>$n_1$</th>
<th>$\zeta$</th>
<th>Direct MC</th>
<th>Importance sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>25</td>
<td>3</td>
<td>$(1.91 \pm 0.14) \times 10^{-2}$</td>
<td>$(2.022 \pm 0.019) \times 10^{-2}$</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>50</td>
<td>3</td>
<td>$(2.37 \pm 0.15) \times 10^{-2}$</td>
<td>$(2.487 \pm 0.023) \times 10^{-2}$</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>100</td>
<td>3</td>
<td>$(3.20 \pm 0.18) \times 10^{-2}$</td>
<td>$(2.962 \pm 0.025) \times 10^{-2}$</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>50</td>
<td>3</td>
<td>$(3.91 \pm 0.19) \times 10^{-2}$</td>
<td>$(3.920 \pm 0.053) \times 10^{-2}$</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>50</td>
<td>5</td>
<td>$(1.0 \pm 1.0) \times 10^{-4}$</td>
<td>$(1.757 \pm 0.016) \times 10^{-4}$</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>100</td>
<td>5</td>
<td>$(2.0 \pm 1.4) \times 10^{-4}$</td>
<td>$(2.247 \pm 0.020) \times 10^{-4}$</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>100</td>
<td>5</td>
<td>$(2.0 \pm 1.4) \times 10^{-4}$</td>
<td>$(1.2157 \pm 0.0094) \times 10^{-4}$</td>
</tr>
</tbody>
</table>
is also asymptotically optimal when \( G(I) > 0 \) for every open interval \( I \). Another variant is Dragalin’s (1997) adaptive CUSUM (AdCUSUM) rule

\[
T' = \inf \{ n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log(f_{\theta_{k,i-1}}(X_i)/f_0(X_i)) \geq c \},
\]

in which \( \theta_{k,m} \) is some estimate of \( \theta \) based on \( X_k, \ldots, X_m \); see also Krieger, Pollak and Yakir (2003).

The CUSUM rule (4.1) and the adaptive CUSUM rule (4.4) can be written in the recursive form \( \inf \{ n : \ell_n \geq c \} \), where

\[
\ell_n = \begin{cases} \ell_{n-1} + \log(f_{\theta_1}(X_n)/f_0(X_n)) \end{cases},
\]

for CUSUM, with \( \theta_1 \) replaced by \( \theta_{\tau_n,n-1} \) for the adaptive CUSUM rule and \( \tau_n \) being the most recent time \( j \) prior to \( n \) such that \( \ell_j \neq 0 \). The GLR and MIX detection rules, however, do not have a simple recursive form, and the number of computations grow to \( \infty \) with \( n \) while maximization or integration of the likelihood function has to be performed at each possible change-time \( k \) between 1 and \( n \). To reduce the computational complexity, Lai (1995) has introduced a window-limited version of (4.2) in which \( \max_{1 \leq k \leq n} \) is replaced by \( \max_{n-k+1 \in \mathcal{N}} \), where

\[
\mathcal{N} = \{ \tilde{M}, \tilde{M} + 1, \ldots, M \} \cup \{ [b^jM] : 1 \leq j \leq J \},
\]

in which \( b > 1, J \leq \min\{i : b^iM \geq \gamma \} \), \( \gamma \) is a prescribed constraint on the expected duration to false alarm (i.e. \( E_0 \tilde{T} \geq \gamma \)) so that \( c \sim \log \gamma \), \( M \) is of order \( c \) and \( \tilde{M} \geq d \) represents a minimal sample size for reliable estimation of the \( d \) parameters of the exponential family. Besides considering the multivariate case, the theory of window-limited GLR rules in Lai (1995) also removes Lorden’s restriction \( |\theta| \geq \epsilon_c \) in (4.2) so that the GLR statistics have the convenient form given in Section 3. Thus the window-limited GLR detection rule can be expressed as

\[
\tilde{T} = \inf \{ n > \tilde{M} : \max_{k \leq n, n-k+1 \in \mathcal{N}} (n-k+1)g(\bar{X}_{k,n}) \geq c \},
\]

where \( g \) is given in (3.5) and \( \bar{X}_{k,n} = (\sum_{i=k}^{n} X_i)/(n-k+1) \) for \( k \leq n \). Since the baseline parameter is assumed to be known, we actually have \( g = \phi \) in this case.

Lai (1995) and Lai and Shan (1999) propose to use the false alarm rate \( P_0(\tilde{T} \leq m)/m \) in place of the expected duration \( E_0 \tilde{T} \) to false alarm. The two quantities are actually related via

\[
E_0 \tilde{T} \sim m/P_0(\tilde{T} \leq m) \quad \text{as} \quad c \to \infty \quad \text{and} \quad m/c \to \infty \quad \text{but} \quad (\log m)/c \to 0.
\]

Indeed, \( \tilde{T}/m \) behaves like a geometric random variable (whose mean is the reciprocal of the probability that the waiting time to success is 1). Lai (1995, p.631) points out certain
advantages of using the false alarm rate constraint $P_0(\hat{T} \leq m)/m \leq \alpha$ over the “average run length” constraint $E_0\hat{T} \geq \alpha^{-1}$, from both theoretical and practical considerations. In particular, for change-point detection in complex systems, one has to use Monte Carlo simulations to determine the threshold $c$ satisfying a false alarm constraint. Whereas simulating $E_0\hat{T} \geq \gamma$ would be computationally prohibitive if $\gamma$ is large (say $10^4$ or $10^6$, which will be considered in Example 4 below), it is relatively easy to simulate $P_0(\hat{T} \leq m)$ by importance sampling methods, as illustrated in Example 4.

While we have considered so far the case in which the baseline parameter $\theta_0$ is completely specified, the window-limited GLR scheme can be easily extended to the case in which the baseline involves certain unspecified nuisance parameters, as in the case of detecting shift in a normal mean from a target value $\mu_0$, with the variance being a nuisance parameter. In this more general setting, the window-limited GLR scheme still has the form (4.6), except that $g = \phi - \phi_0$, where $\phi_0(\mu) = \sup_{\theta \in \Theta_0}\{\theta'\mu - \psi(\theta)\}$; see (3.5). The following theorem gives an asymptotically efficient choice of the importance density for Monte Carlo evaluation of the false alarm probability $P_0(\hat{T} \leq m)$ of window-limited GLR detection rules. Roughly speaking, the importance density puts a uniform distribution on $\{1, \ldots, m\}$ for the change-time $k$ and uses the mixture density $\int \prod_{i=k}^m f_{\theta_0}(x_i)w_c(\mu)\,d\mu$ of Section 3 for the post-change distribution of $X_k, \ldots, X_n$.

**Theorem.** Let $\hat{M} \sim a_0c$ and $M \sim a_1c$ for $0 < a_0 < a_1$. Assume that $\mathcal{M}_\delta = \{\mu : a_1^{-1} - \delta < \phi(\mu) < a_0^{-1}, \phi_0(\mu) = 0\}$ is a smooth $g$-dimensional manifold with finite $g$-dimensional volume for every $0 \leq \delta < \delta_0$ and some $\delta_0 > 0$. Letting $c(\mu) = (e_1(\mu), \ldots, e_{d-q+1}(\mu))$, where $\{e_1(\mu), \ldots, e_{d-q+1}(\mu)\}$ is an orthonormal basis of the normal space of $\mathcal{M}_0$ at $\mu$, assume that $|c(\mu)\nabla^2\phi_0(\mu)e(\mu)| > 0$. Define $w_c(\mu)$ by (3.7). Then

$$m^{-1} \sum_{k=1}^n \int \prod_{i=1}^{k-1} f_0(x_i) \prod_{i=k}^m f_{\theta_0}(x_i)w_c(\mu)\,d\mu + (1 - n/m) \prod_{i=1}^n f_0(x_i)$$

(4.8)

is an asymptotically efficient importance density for simulating $P_0(\hat{T} \leq m)$ in the case $J = 0$, as $c \to \infty$, $m/c \to \infty$ but $(\log m)/c \to 0$.

**Proof.** The importance weight with respect to the importance density (4.8) is $L^{-1}_T$, where

$$L_n = m^{-1} \sum_{k=1}^n \int \prod_{i=1}^n f_{\theta_0}(x_i)/f_0(x_i)w_c(\mu)\,d\mu + (1 - n/m).$$

(4.9)

Since $\prod_{i=k}^n (f_{\theta}(x_i)/f_0(x_i)) = \exp\{(n - k + 1)[\theta'\hat{x}_n - \psi(\theta)]\}$, it follows from (4.9) that

$$L_T \geq m^{-1} \int \exp\{(\hat{T} - k + 1)[\theta'\hat{X}_{k,T} - \psi(\theta)]\}w_c(\mu)\,d\mu$$

(4.10)
for any $k \leq \hat{T}$. On \{\hat{T} \leq m\}, there exists $k \leq \hat{T}$ such that \((\hat{T} - k + 1)g(\hat{X}_{k,T}) \geq c\) and \(\hat{M} \leq \hat{T} - k + 1 \leq M\). Moreover, Chan and Lai (2003b) have shown that if \(ng(x) \geq c\) and \(c(a_0 + o(1)) \leq n \leq c(a_1 + o(1))\), then there exists \(A^* > 0\) such that

\[
\int \exp \{n[\theta^*_\mu x - \psi(\theta_\mu)]\} w_c(\mu) \, d\mu \geq A^* e^{-q/2} e^c. \tag{4.11}
\]

Putting $n = \hat{T} - k + 1$ and $x = \hat{X}_{k,T}$ in (4.11) and combining the result with (4.10) yield

\[
L_{\hat{T}} \geq m^{-1} A^* e^{-q/2} e^c \text{ on } \{\hat{T} \leq m\}. \tag{4.12}
\]

From Theorem 6 and the argument in the last paragraph of Section 4 of Chan and Lai (2003a), it follows that

\[
P_0(\hat{T} \leq m) \sim \zeta m e^{q/2} e^{-c} \tag{4.13}
\]

for some $\zeta > 0$. From (4.12) and (4.13), we can conclude that $L_{\hat{T}}^{-1}1_{\{\hat{T} \leq m\}} = O(m e^{q/2} e^{-c}) = O(P_0(\hat{T} \leq m))$, and therefore (4.8) is an asymptotically efficient importance density. \(\square\)

In the case $\Theta_0 = \{0\}$, $g = \phi$ and as noted in Section 3, we can replace $w_c(\mu)$ in the preceding by any density function $v$ satisfying (3.10). In particular, if $f_\theta$ is normal, then we can take $v$ to be any normal density. Moreover, in this case, the restriction that $J = 0$ in the preceding theorem can be removed as Chan and Lai (2002, Theorem 4) have shown that under $P_0$ the inclusion of the geometric window sizes $[b^j M]$ in (4.5) has negligible effect on the false alarm rate. The importance sampling procedure in this normal case can be described as follows:

1. Generate $\nu$ uniformly from $\{1, \ldots, m\}$.
2. Generate $\theta$ from the normal distribution $N(0, \beta^{-1})$.
3. Generate i.i.d. $X_1, \ldots, X_{\nu-1}$ with density $f_0$ and i.i.d. $X_{\nu}, \ldots, X_m$ with density $f_\theta$.

Repeating this procedure $K$ times, we obtain $K$ realizations of $L_{\hat{T}}^{-1}1_{\{\hat{T} \leq m\}}$, where

\[
L_n = m^{-1} \sum_{k=1}^{n} \left(\frac{\beta}{\beta + n - k + 1}\right)^{1/2} \exp \left(\frac{(n - k + 1)^2 \hat{X}_{k,n}^2}{2(n - k + 1 + \beta)}\right) + (1 - n/m) \tag{4.14}
\]

for all $n \leq m$. The average of these $K$ realizations of $L_{\hat{T}}^{-1}1_{\{\hat{T} \leq m\}}$ gives an estimate of $P_0(\hat{T} \leq m)$. Such importance sampling procedures were introduced by Lai and Shan (1999) and the preceding theorem establishes the asymptotic efficiency of these procedures.

Lai (1995) and Lai and Shan (1999) considered in particular sequential detection of changes in the parameters of an autoregressive model. Although this no longer conforms to
the case of independent observations from an exponential family treated in the preceding
theorem, the theory of asymptotically efficient importance distribution in Chan and Lai
(2003b) is developed in the more general Markov dependent setting and is therefore also
applicable to autoregressive models.

**Example 4.** Consider a stable AR(p) model

\[ Y_n = \theta_0'X_n + \epsilon_n \text{ for all } n, \quad (4.15) \]

where \( X_n = (Y_{n-1}, \ldots, Y_{n-p})' \), \( \theta_0 = (\theta_{01}, \ldots, \theta_{0p})' \) is the baseline parameter vector such that \( 1 - \theta_{01}z - \cdots - \theta_{0p}z^p \) has no roots inside the unit circle, and \( \epsilon_n \) are i.i.d. unobservable disturbances which we assume to be standard normal. The change-point model assumes the existence of (unknown) \( \nu \geq 1 \) and \( \theta \neq \theta_0 \) such that

\[ Y_n = \theta_0'X_n + \epsilon_n \text{ for } n < \nu, \quad Y_n = \theta'X_n + \epsilon_n \text{ for } n \geq \nu. \quad (4.16) \]

Let \( f_\theta(\cdot|\cdot) \) denote the conditional density of \( Y_n \) given \( X_n \) when the autoregressive parameter is \( \theta \). The window-limited GLR detection rule in this case can be expressed as

\[
\hat{T} = \inf \left\{ n : \sup_{k \leq n:n-k+1 \in \mathbb{N}} \sup_{\theta} \left[ \sum_{j=k}^{n} \log \left\{ \frac{f_\theta(Y_j|X_j)}{f_{\theta_0}(Y_j|X_j)} \right\} \right] > c \right\}
= \inf \left\{ n : \sup_{k \leq n:n-k+1 \in \mathbb{N}} \ell_n(k) > c \right\},
\]

(4.17)

where \( \ell_n(k) = (\hat{\theta}_{k,n} - \theta_0)'(\sum_{i=k}^{n} X_iX_i')^{-1}(\hat{\theta}_{k,n} - \theta_0) \) with \( \hat{\theta}_{k,n} = (\sum_{i=k}^{n} X_iX_i')^{-1}(\sum_{i=k}^{n} X_iY_i) \).

We can express \( \ell_n(k) = (n-k+1)g(\sum_{i=k}^{n} X_iX_i'/(n-k+1), \sum_{i=k}^{n} X_iY_i/(n-k+1)) \), with \( g(\mu) = (\mu_1^{-1}\mu_2 - \theta_0)'\mu_1(\mu_1^{-1}\mu_2 - \theta_0)/2 \). Note that \( X_i \) is a Markov chain having a stationary distribution under \( \theta_0 \).

Lai and Shan (1999) proposed the following importance sampling procedure to simulate \( P_0(\hat{T} \leq m) \):

1. Sample \( \nu \) from \( \{1, \ldots, m\} \) with probability \( 1/m \) for each possible value.
2. Generate \( \theta \) from the \( \text{N}(\theta_0, \beta^{-1}\mathbf{I}) \) distribution.
3. Generate recursively \( Y_n = \theta_0'X_n + \epsilon_n \text{ for } n < \nu, \text{ and } Y_n = \theta'X_n + \epsilon_n \text{ for } n \geq \nu. \)

Repeat this procedure \( K \) times and estimate \( P_0(\hat{T} \leq m) \) by the average of \( L_{\hat{T}}^{-1}\mathbf{1}_{\{\hat{T} \leq m\}} \) over the \( K \) realizations, where

\[
L_n = m^{-1}b^{p/2} \sum_{k=1}^{n} \{ \det \left( \sum_{i=k}^{n} X_iX_i' + \beta\mathbf{I} \right) \}^{-1/2} \exp[\ell_n(k)] + 1 - n/m \quad (4.18)
\]
the distribution of $X$ is a bivariate normal distribution with zero means, $\text{Var}(\theta)$ initialized at $I$. The adaptive CUSUM rule requires estimation of $\theta$, which we use recursive least squares initialized at $I/5$. Assuming that $\theta_0 = (-0.1, 0.1)'$, the distribution of $(X_0, X_{-1})$ is initialized at the steady state distribution under $\theta_0$, which is a bivariate normal distribution with zero means, $\text{Var}(X_i) = 1.0227$ and $\text{Cov}(X_0, X_{-1}) = -0.1136$.

To begin with, we need to determine the thresholds of all three sequential detection rules so that they have the same false alarm rate, which we measure by $P_0(\hat{T} \leq 100)/100$ instead of the average run length $E_0\hat{T}$ (and with $\hat{T}$ replaced by $\hat{T}$ and $T'$ for the other two rules). We consider the cases $P_0(\hat{T} \leq 100) = 10^{-2}, 10^{-3}$ and $10^{-4}$ and apply importance sampling to evaluate the probabilities. Brent’s method is then used to solve for the threshold. The importance density is asymptotically efficient.

The remainder of this example is concerned with a comparative study of the performance of the window-limited GLR rule with that of MIX and AdCUSUM (which we introduced in (4.3) and (4.4)) for sequential detection of parameter changes in an AR(2) model. For MIX, we use a window-limited version so that the computational complexity does not grow to $\infty$ with $n$. For a fair comparison with GLR, we use the same window sizes for both procedures, taking $N = \{3, \ldots, 5\} \cup \{50(1.5)^j : 1 \leq j \leq 15\}$.

The mixing distribution $G$ in MIX is chosen to be $N(\theta_0, \beta^{-1}I)$, for which

$$\hat{T} = \inf\{n : \sup_{k \leq n-1} \{\log \beta - \log\{\det(\sum_{i=k}^{n} X_iX_i' + \beta I)\}/2 + \ell_n(k)\} \geq c\}. \quad (4.19)$$

The adaptive CUSUM rule requires estimation of $\theta$, which we use recursive least squares initialized at $\theta_0$, with the covariance matrix initialized at $I/5$. Assuming that $\theta_0 = (-0.1, 0.1)'$, the distribution of $(X_0, X_{-1})$ is initialized at the steady state distribution under $\theta_0$, which is a bivariate normal distribution with zero means, $\text{Var}(X_i) = 1.0227$ and $\text{Cov}(X_0, X_{-1}) = -0.1136$.

To begin with, we need to determine the thresholds of all three sequential detection rules so that they have the same false alarm rate, which we measure by $P_0(\hat{T} \leq 100)/100$ instead of the average run length $E_0\hat{T}$ (and with $\hat{T}$ replaced by $\hat{T}$ and $T'$ for the other two rules). We consider the cases $P_0(\hat{T} \leq 100) = 10^{-2}, 10^{-3}$ and $10^{-4}$ and apply importance sampling to evaluate the probabilities. Brent’s method is then used to solve for the threshold. The

<table>
<thead>
<tr>
<th>Detection rule</th>
<th>Threshold $c$</th>
<th>Importance sampling</th>
<th>Direct Monte Carlo</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLR</td>
<td>9.6</td>
<td>$(0.998 \pm 0.035) \times 10^{-2}$</td>
<td>$(1.02 \pm 0.10) \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>12.0</td>
<td>$(1.00 \pm 0.07) \times 10^{-3}$</td>
<td>$(0.9 \pm 0.3) \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>14.3</td>
<td>$(1.02 \pm 0.06) \times 10^{-4}$</td>
<td>$(2 \pm 1) \times 10^{-4}$</td>
</tr>
<tr>
<td>Mix</td>
<td>5.8</td>
<td>$(1.06 \pm 0.02) \times 10^{-2}$</td>
<td>$(0.94 \pm 0.10) \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>$(1.03 \pm 0.02) \times 10^{-3}$</td>
<td>$(0.7 \pm 0.3) \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>10.2</td>
<td>$(1.02 \pm 0.02) \times 10^{-4}$</td>
<td>0</td>
</tr>
<tr>
<td>AdCUSUM</td>
<td>4.9</td>
<td>—</td>
<td>$(0.99 \pm 0.10) \times 10^{-2}$</td>
</tr>
</tbody>
</table>
Table 7: Average run length for various detection rules

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>( P_0{N \leq 100} )</th>
<th>GLR: ( c = 9.6 )</th>
<th>MIX: ( c = 5.8 )</th>
<th>AdCUSUM: ( c = 4.9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1362±19</td>
<td>1664±25</td>
<td>2754±59</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>178±2</td>
<td>183±3</td>
<td>197±3</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>62.7±0.8</td>
<td>58.5±0.8</td>
<td>57.3±0.9</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>28.3±0.4</td>
<td>27.0±0.4</td>
<td>24.6±0.4</td>
<td></td>
</tr>
<tr>
<td>-0.6</td>
<td>48.1±0.7</td>
<td>45.9±0.7</td>
<td>41.8±0.7</td>
<td></td>
</tr>
<tr>
<td>-0.4</td>
<td>153±2</td>
<td>155±2</td>
<td>155±3</td>
<td></td>
</tr>
<tr>
<td>-0.2</td>
<td>1314±19</td>
<td>1619±24</td>
<td>2284±48</td>
<td></td>
</tr>
</tbody>
</table>

The results are given in Table 6 for GLR and MIX. Also given for comparison are corresponding probabilities estimated by direct Monte Carlo (mean ± standard error). It is difficult to use importance sampling for adaptive CUSUM. Accordingly we only apply direct Monte Carlo for \( P(T' \leq 100) = 10^{-2} \) in that case. Each Monte Carlo computation of a probability is based on 10,000 simulations.

Table 7 compares the average run length \( E_{(\Delta,0.1)}N \) for \( N = \hat{T}, \tilde{T}, T' \). This corresponds to a detection delay when \( \theta_{01} \) changes from -0.1 to \( \Delta \geq 0 \) or \( \Delta \leq -0.2 \), with \( \theta_{02} \) unchanged and the change-point occurring at time 1. Each result is based on 2000 simulations. The table shows AdCUSUM to have the longest detection delay and that GLR to perform the best. Whereas AdCUSUM has been shown to perform well for detecting changes in a one-dimensional parameter by Dragalin (1997) and Krieger, Pollak and Yakir (2003), the price it pays for only using \( i - 1 \) observations to estimate a multidimensional \( \theta \) in \( f_\theta(X_i) \) even when \( n - i \) more observations are available can be quite substantial, as shown in Table 7.
ACKNOWLEDGEMENTS

Chan’s research was supported by the National University of Singapore and Lai’s research was supported by the National Science Foundation and the National Cancer Institute.

REFERENCES


