

# Is Average Run Length to False Alarm Always an Informative Criterion?

Yajun Mei

Georgia Institute of Technology, Atlanta, Georgia, USA

**Abstract:** Apart from Bayesian approaches, the average run length (ARL) to false alarm has always been seen as the natural performance criterion for quantifying the propensity of a detection scheme to make false alarms, and no researchers seem to have questioned this on grounds that it does not always apply. In this article, we show that in the change-point problem with mixture pre-change models, detection schemes with finite detection delays can have infinite ARLs to false alarm. We also discuss the implication of our results on the change-point problem with either exchangeable pre-change models or hidden Markov models. Alternative minimax formulations with different false alarm criteria are proposed.

**Keywords:** Average run length; CUSUM; Expected false alarm rate; Quantile run length; Statistical process control; Surveillance.

**Subject Classifications:** 62L10; 62L15; 60G40.

## 1. INTRODUCTION

In sequential change-point detection problems, one seeks a detection scheme to raise an alarm as soon as unusual or undesired events happen at (unknown) time  $\nu$  so that appropriate action can be taken. Construction of the detection scheme is based on a sequence of (possibly vector) observations  $X_1, X_2, \dots$  which are observed sequentially, i.e., one at a time, and it is assumed that the distribution of the  $X$ 's will change if the undesired events occur. The

---

<sup>1</sup>Address correspondence to Yajun Mei, School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive NW, Atlanta, GA 30332-0205, USA; Fax: 404-894-2301; E-mail: ymei@isye.gatech.edu

decision whether to raise an alarm at time  $n$  will only depend on the first  $n$  observations. That is, our current decision only depends on our current and past observations, but not on future observations.

Such change-point problems are ubiquitous and, as a consequence, have many important applications, including statistical process control (SPC), industrial quality control, target or signal detection, and epidemiology, see, for example, Basseville and Nikiforov (1993), Lai (1995, 2001). There has been recently a surge of interest in application to information assurance, network security, health-care and public-health surveillance, particularly a relatively new area called syndromic surveillance. For recent reviews on new applications, see Fienberg and Shmueli (2005), Tartakovsky et al. (2006), Woodall (2006), and the references therein.

A tremendous variety of statistical methods and models have been developed for change-point detection. A partial list includes cumulative sum (CUSUM), Shewhart's control chart, exponentially-weighted moving average (EWMA) charts, Shiryayev-Roberts procedures, window-limited control charts, and scan statistics. See, for example, Shewhart (1931), Page (1954), Roberts (1966), Shiryayev (1963, 1978), Lorden (1971), Moustakides (1986), Pollak (1985, 1987), Ritov (1990), Lai (1995), Kulldorff (2001). New methods and new variants on existing methods are being developed all the time, but the discussion on formulating the right problem and assessing detection schemes under appropriate performance measures is rather limited.

There are two standard mathematical formulations for change-point problems. The first one is a Bayesian formulation, due to Shiryayev (1963), in which the change-point is assumed to have a known prior distribution. The second is a minimax formulation, proposed by Lorden (1971), in which the change-point is assumed to be unknown (possibly  $\infty$ ) but non-random. In the literature, both formulations have been extended to dependent observations by simply using new probability measures under which the observations may be dependent. See, for example, Bansal and Papantoni-Kazakos (1986), Basseville and Nikiforov (1993), Brodsky and Darkhovsky (1993, 2000), Yakir (1994), Beibel (1997), Lai (1995, 1998, 2001), Fuh (2003, 2004), Mei (2006b), Baron and Tartakovsky (2006). However, there seems to

be controversies on the optimality properties of widely used Page's CUSUM or Shiriyayev-Roberts procedures when observations are not independent, especially under the minimax formulation, as these procedures are (asymptotically) optimal in some situations but can be suboptimal in other situations.

In this article, instead of studying the optimality properties of CUSUM or other detection schemes, we take a further step to look at the appropriateness of the standard minimax formulation when observations are not independent. In the literature, the performance of a detection scheme is typically evaluated by two types of criteria, one being a measure of the detection delay after a change occurs, and the other being a measure of a frequency of false alarms. The importance of the appropriate definition of detection delay has gained a lot of attention in the literature, and there are several rigorous definitions of the detection delay under a minimax formulation that have been proposed, e.g., the "worst-case" detection delay in Lorden (1971), the "average" detection delay in Shiriyayev (1963) and Pollak (1985), the "exponential penalty" detection delay in Poor (1998). On the other hand, for the false alarm criterion, it is historically standard to use the "average run length" (ARL) to false alarm, which is the expected number of samples to be taken before a false alarm is signaled. Despite some concerns about the ARL to false alarm criterion and some alternative criteria proposed in the literature, see, for example, Barnard (1959), Brodsky and Darkhovsky (1993, 2000), Kenett and Zacks (1998), Lai (1998, 2001) and Tartakovsky et al. (2006), most researchers are still using the ARL to false alarm to evaluate the detection schemes, partly because the ARL, a simple function of the distribution of run length to false alarm, seems to be always well-defined. To the best of our knowledge, no researchers have questioned the appropriateness of the ARL to false alarm criterion on grounds that it does not always apply.

The primary goal of this article is to show that a detection scheme with finite detection delay can have infinite ARL to false alarm when the observations are not independent. Moreover, under the standard minimax formulation with the ARL to false alarm criterion, even if well-defined, we are in danger of finding a detection scheme that focuses on detecting larger changes instead of smaller changes when observations are not independent. We

illustrate this through a specific example with the “mixture” pre-change distribution, and also discuss the implication on the change-point problem either with the “exchangeable” pre-change distribution or in the hidden Markov models. While our example is only a theoretical example to show the invalidity of standard minimax formulation, particularly the ARL to false alarm criterion, we hope researchers pay more attention on the appropriateness of the performance criteria of detection schemes for dependent observations, especially for exchangeable pre-change models.

A closely related goal of this article is to propose two new minimax formulations for our example, since the standard minimax formulation is inappropriate. For that purpose, we introduce two new definitions: (1) “asymptotic efficiency” defined as the divergence rate of the logarithm of the ARL to false alarm; and (2) the “expected false alarm rate” (EFAR), which seems to be closely related to the quantiles of the distribution of the run length to false alarm. We acknowledge that our new formulations may not be applicable directly to the real-worlds problems. Neither will we attempt to develop a general theory in a general setting. Rather, by presenting new performance criteria for detection schemes in a specific example, we hope it will open new directions to study change-point problems and their applications with dependent observations, particularly with exchangeable pre-change models. In view of the history of sequential change-point detection problems for independent observations, which was studied by Shewhart (1931) and Page (1954) but not rigorously formulated and solved until between 1963 and 1990 (Shiryayev (1963), Lorden (1971), Moustakides (1986), Pollak (1985, 1987), Ritov (1990)), it should be anticipated that significant time and effort will be required for appropriate general theory of change-point problems with exchangeable pre-change distributions, or more generally, with any kinds of dependent observations.

The remainder of this article is organized as follows. Section 2 provides a motivation of our example where the pre-change distributions is a so-called “mixture” distribution, and presents the standard minimax formulation of the problem. Section 3 defines two families of detection schemes and studies their properties under the standard minimax formulation. Section 4 proposes alternative minimax formulations with different false alarm criteria for

our example, and proves the asymptotic optimality property of a class of detection schemes. Section 5 considers a further example with an exchangeable pre-change distribution and also discusses the implicit of our results on the change-point problems in Hidden-Markov Models (HMM). Section 6 includes some final remarks, and the Appendices include the proofs of Lemma 3.1 and Theorem 4.1.

## 2. MOTIVATION AND STANDARD FORMULATION

Before we state our example, let us first consider a classical change-point problem with independent observations. Suppose that  $X_1, X_2, \dots$  are independent normally distributed random variables with variance 1, and assume we want to detect a change from  $\theta$  to  $\lambda$  in the mean of the  $X$ 's, where  $\theta \leq 0 < \lambda$  and the post-change parameter  $\lambda$  is completely specified.

If the pre-change parameter  $\theta$  is also completely specified, say  $\theta = \theta_0$ , then a classical detection scheme is Page's CUSUM procedure, which would declare a change has occurred at the time

$$\begin{aligned} T_{CM}(\theta_0, b) &= \text{first } n \geq 1 \text{ such that } \max_{1 \leq k \leq n} \sum_{i=k}^n \log \frac{\phi_\lambda(X_i)}{\phi_{\theta_0}(X_i)} \geq b, \\ &= \text{first } n \geq 1 \text{ such that } \max_{1 \leq k \leq n} \sum_{i=k}^n \left[ X_i - \frac{\lambda + \theta_0}{2} \right] \geq \frac{b}{\lambda - \theta_0}, \end{aligned} \quad (2.1)$$

where the threshold  $b > 0$  is pre-specified, and  $\phi_\mu(x) = (1/\sqrt{2\pi}) \exp(-(x - \mu)^2/2)$  is the probability density function of a normal random variable with mean  $\mu$  and variance 1.

Now let us assume that we know the pre-change parameter  $\theta \leq 0$  but we do not know the exact value of  $\theta$ . This is one of key features arising from new applications such as syndromic surveillance, where the baseline model when there is no disease outbreak is not completely specified. Even in well-established applications such as quality control, although the assumption of known pre-change parameters seems to be reasonable as quality of products can be pre-defined, the issue of partially specified pre-change parameters has recently been recognized, see, for example, Jensen et al. (2006).

Several parametric approaches have been proposed to tackle this problem in the literature.

The first is to specify the nominal value  $\theta_0$  of the pre-change parameter. The choice of  $\theta_0$  can be made directly by considering a (pre-change) parameter which is close to the post-change parameters because it is always more difficult to detect a smaller change, e.g.,  $\theta_0 = 0$  or  $\lambda/2$  for this example. Alternatively,  $\theta_0$  can be estimated from a training sample. However, it is well-known that the performances of such procedures can be rather poor if the true pre-change parameter  $\theta$  is not  $\theta_0$ , see, for example, Stoumbos et al. (2000).

The second approach, proposed in Mei (2006a), is to specify a required detection delay at a given  $\lambda > 0$  while trying to maximize the ARLs to false alarm for all possible values of pre-change parameter  $\theta \leq 0$ . Mei (2006a) also developed a general theory for change-point problems when both the pre-change parameter  $\theta$  and the post-change parameter  $\lambda$  are only partially specified.

The third approach is to eliminate the nuisance, pre-change parameter  $\theta$ , see, for example, Pollak and Siegmund (1991), Yakir, Krieger and Pollak (1999), and Krieger, Pollak and Yakir (2003). All assume the availability of a training sample and eliminate the nuisance parameter via invariance. Besides invariance, another widely used method to eliminate the nuisance parameter is to integrate the nuisance parameter with respect to weight functions (or priors), see, for example, Wald (1947) and Kiefer and Sacks (1963) for the application in testing hypothesis problems, and Pollak (1987) and Lai (1998) in the change-point problems when the nuisance parameters are present in the *post-change* distribution. However, to the best of our knowledge, this method has not been applied to the change-point problems when the nuisance parameters are present in the *pre-change* distribution. This motivates us to consider the approach of eliminating the nuisance *pre-change* parameter via weight-functions, giving us a change-point problem with dependent observations.

Now let us state our example rigorously. Denote by  $\mathbf{P}_\theta$  and  $\mathbf{P}_\lambda$  the probability measures when  $X_1, X_2, \dots$  are independent and identically distributed (i.i.d.) normal random variables with means  $\theta$  and  $\lambda$  and variance 1, respectively. Assume  $\lambda > 0$  is completely specified and  $\theta \leq 0$  has a prior half-cauchy distribution with density

$$\pi(\theta) = \frac{2}{\pi(1 + \theta^2)} \quad \text{for } \theta \leq 0. \quad (2.2)$$

Different choices of  $\pi(\theta)$  and their implications will be explained in Section 5. Define a “mixture” probability measure  $\mathbf{P}_f = \int_{-\infty}^0 \mathbf{P}_\theta \pi(\theta) d\theta$ . That is, under  $\mathbf{P}_f$ ,  $X_1, \dots, X_n$  have a mixture joint density

$$f(x_1, \dots, x_n) = \int_{-\infty}^0 \left[ \prod_{i=1}^n \phi_\theta(x_i) \right] \pi(\theta) d\theta. \quad (2.3)$$

To emphasize the post-change distribution, we denote by  $\mathbf{P}_g$  the probability measure when  $X_1, X_2, \dots$  are i.i.d. normal with mean  $\lambda$  and variance 1, i.e.,  $g = \phi_\lambda$  and  $\mathbf{P}_g = \mathbf{P}_\lambda$ .

The problem we are interested in is to detect a change in distribution from the mixture distribution  $f$  in (2.3) to  $g = \phi_\lambda$ . Mathematically, for some unknown change-point  $\nu$  (possibly  $\infty$ ),  $X_1, \dots, X_{\nu-1}$  are distributed according to the joint mixture density  $f$  while  $X_\nu, X_{\nu+1}, \dots$  are independently distributed according to a common density  $g$ . Moreover, the post-change observations  $X_\nu, X_{\nu+1}, \dots$  are independent of the pre-change observations  $X_1, \dots, X_{\nu-1}$ . For  $1 \leq \nu < \infty$ , denote by  $\mathbf{P}^{(\nu)}$  and  $\mathbf{E}^{(\nu)}$  the probability measure and expectation when a change occurs at time  $\nu$ . We shall also use  $\mathbf{P}_f$  and  $\mathbf{E}_f$  to denote the probability measure and expectation when there is no change, i.e.,  $\nu = \infty$ , in which  $X_1, X_2, \dots$  are distributed with the mixture joint density  $f$ .

A detection scheme for detecting that a change has occurred is defined as a stopping time  $T$  with respect to  $\{X_n\}_{n \geq 1}$ . The interpretation of  $T$  is that, when  $T = n$ , we will raise an alarm at time  $n$  and declare that a change has occurred somewhere in the first  $n$  observations. We want to find a detection scheme which will raise an alarm as soon as possible after a change occurs, but will take observations as many as possible if no change occurs.

For a detection scheme  $T$ , the detection delay can be defined by the following “worst case” detection delay defined in Lorden (1971),

$$\bar{\mathbf{E}}_g(T) = \sup_{1 \leq \nu < \infty} \left( \text{ess sup } \mathbf{E}^{(\nu)}[(T - \nu + 1)^+ | X_1, \dots, X_{\nu-1}] \right).$$

It is worth pointing out that the definition of  $\bar{\mathbf{E}}_g(T)$  does not depend on the pre-change distribution  $f$  by virtue of the essential supremum, which takes the worst possible  $X$ 's before the change. In our results we can also use the “average” detection delay, proposed by

Shiryayev (1963) and Pollak (1985),  $\sup_{1 \leq \nu < \infty} \mathbf{E}^{(\nu)}(T - \nu | T \geq \nu)$ , which is asymptotically equivalent to  $\overline{\mathbf{E}}_g(T)$ .

It is important to mention the relationship between the detection delay  $\overline{\mathbf{E}}_g(T)$  with  $\mathbf{E}_g(T)$ . On the one hand, for many widely used detection schemes,  $\overline{\mathbf{E}}_g(T) = \mathbf{E}_g T$ , i.e., the worst case detection delay often occurs when the change-point  $\nu = 1$ , because it is often more difficult to detect when a change occurs at early stages than at latter stages. On the other hand,  $\overline{\mathbf{E}}_g(T)$  is a more rigorous measurement of detection delay in theory since  $\overline{\mathbf{E}}_g(T)$  takes into account probability measures  $\mathbf{P}^{(\nu)}$  which are included in the change-point problems.

The desire to have small detection delay  $\overline{\mathbf{E}}_g(T)$  must, of course, be balanced against the need to have a controlled false alarm rate. When there is no change,  $T$  should be as large as possible, hopefully infinite. However, Lorden (1971) showed that for independent observations, if  $\overline{\mathbf{E}}_g(T)$  is finite, then  $\mathbf{P}_f(T < \infty) = 1$ , i.e., the probability of ever raising a false alarm is 1. This means that we cannot use the probability of ever raising a false alarm as a false alarm criterion. Moreover, Lorden (1971) also showed that, for independent observations, an appropriate measurement of false alarms is  $\mathbf{E}_f(T)$ , the ARL to false alarm.

A good detection scheme  $T$  should have large values of the ARL to false alarm  $\mathbf{E}_f(T)$  while keeping the detection delay  $\overline{\mathbf{E}}_g(T)$  small. To balance the tradeoff between these two quantities, the standard minimax formulation of change-point problems for independent observations is then to seek a detection scheme  $T$  that minimizes the detection delay  $\overline{\mathbf{E}}_g(T)$  subject to  $\mathbf{E}_f(T) \geq \gamma$ , where  $\gamma$  is a given constant. In practice, due to the close relationship between  $\overline{\mathbf{E}}_g(T)$  and  $\mathbf{E}_g(T)$ , it often (but not always) suffices to study  $\mathbf{E}_f(T)$  and  $\mathbf{E}_g(T)$ .

Much research has been done in the literature to extend the standard minimax formulation to dependent observations by simply replacing the probability densities with the corresponding conditional densities, see, for example, Lai (1998). In particular, in our example where the pre-change distribution is the mixture distribution  $f$  in (2.3) and the post-change distribution is  $g = \phi_\lambda$ , the standard minimax formulation will evaluate the performance of a detection scheme  $T$  by the detection delay  $\overline{\mathbf{E}}_g(T)$  and the ARL to false alarm  $\mathbf{E}_f(T)$ .

### 3. INFINITE ARL TO FALSE ALARM

In this section, we illustrate that finite detection delay may be achieved even with infinite ARL to false alarm in the change-point problem with the mixture pre-change distribution  $f$  in (2.3). This, of course, is a severe criticism of the standard minimax formulation with the ARL to false alarm as an operating characteristic of a detection scheme, at least in the change-point problem with mixture pre-change distributions.

As mentioned in Section 2, the problem of detecting a change in distribution from the mixture distribution  $f$  in (2.3) to  $g = \phi_\lambda$  is motivated from that of detecting a change in the mean of independent normal observations from  $\theta \leq 0$  to  $\lambda$ . Hence detection schemes in the latter problem can be applied to the former problem, although its efficiency or optimality properties could be different. In the following we will consider two families of detection schemes, which correspond to the first two approaches of the latter problem mentioned in Section 2.

Let us first consider Page's CUSUM procedures  $T_{CM}(\theta_0, b)$  in (2.1) for a given  $\theta_0$ . That is, we will choose a nominal value  $\theta_0$  for the pre-change parameter, and then declare a change from the mixture distribution  $f$  in (2.3) to  $g = \phi_\lambda$  occurs if and only if  $T_{CM}(\theta_0, b)$  stops. Of course  $T_{CM}(\theta_0, b)$  is designed to detect a change in the mean from  $\theta_0$  to  $\lambda$ , and thus it *may* or *may not* be efficient to detect a change in distribution from  $f$  to  $g$ . Nevertheless, we can apply it to this problem and study its corresponding properties. For the sake of generality, we assume  $0 \leq \theta_0 < \lambda$ .

The following lemma, whose proof is in the Appendix, establishes the asymptotic performance of  $T_{CM}(\theta_0, b)$ .

**Lemma 3.1.** *For any  $b > 0$  and any  $\theta \leq (\theta_0 + \lambda)/2$ ,*

$$\mathbf{E}_\theta(T_{CM}(\theta_0, b)) \geq \exp\left(\frac{\lambda + \theta_0 - 2\theta}{\lambda - \theta_0}b\right), \quad (3.1)$$

where  $\mathbf{E}_\theta$  denotes the expectation when  $X_1, X_2, \dots$  are i.i.d. normal with mean  $\theta$  and variance 1, and as  $b \rightarrow \infty$

$$\overline{\mathbf{E}}_g(T_{CM}(\theta_0, b)) = \frac{b}{I(\lambda, \theta_0)} + O(1), \quad (3.2)$$

where

$$I(\lambda, \theta) = \mathbf{E}_\lambda \log(\phi_\lambda(X)/\phi_\theta(X)) = (\lambda - \theta)^2/2 \quad (3.3)$$

is the Kullback-Leibler information number.

The following theorem establishes the performance of  $T_{CM}(\theta_0, b)$  in the standard minimax formulation of the problem of detecting a change from the mixture distribution  $f$  in (2.3) to  $g = \phi_\lambda$ .

**Theorem 3.1.** *Assume  $0 \leq \theta_0 < \lambda$  and  $b > 0$ . Then the detection scheme  $T_{CM}(\theta_0, b)$  has a finite detection delay  $\bar{\mathbf{E}}_g(T_{CM}(\theta_0, b))$ , but has an infinite ARL to false alarm  $\mathbf{E}_f(T_{CM}(\theta_0, b))$ .*

*Proof.* By Lemma 3.1, the detection scheme  $T_{CM}(\theta_0, b)$  has a finite detection delay  $\bar{\mathbf{E}}_g(T_{CM}(\theta_0, b))$ , since the definition of  $T_{CM}(\theta_0, b)$  implies that its worst-case detection delay always occurs at  $\nu = 1$  regardless of pre-change distributions. To derive the ARL to false alarm, note that by the definition of the mixture pre-change  $f$  in (2.3), for any stopping time  $T$ ,

$$\mathbf{E}_f(T) = \int_{-\infty}^0 \mathbf{E}_\theta(T) \pi(\theta) d\theta, \quad (3.4)$$

where  $\pi(\theta)$  is defined in (2.2). Thus, by Lemma 1,

$$\begin{aligned} \mathbf{E}_f(T_{CM}(\theta_0, b)) &\geq \int_{-\infty}^0 \exp\left(\frac{\lambda + \theta_0 - 2\theta}{\lambda - \theta_0} b\right) \pi(\theta) d\theta \\ &= \int_{-\infty}^0 \exp\left(\frac{\lambda + \theta_0 - 2\theta}{\lambda - \theta_0} b\right) \left[\frac{2}{\pi(1 + \theta^2)}\right] d\theta \end{aligned}$$

which diverges for any  $b > 0$ . That is,  $\mathbf{E}_f(T_{CM}(\theta_0, b)) = \infty$  for any  $b > 0$ .  $\square$

To illustrate that  $\{T_{CM}(\theta_0, b)\}$  in (2.1) is not the only family of detection schemes with finite detection delay and infinite ARLs to false alarm, the second family of detection schemes we considered is those proposed in Mei (2006a), which is defined by

$$T^*(a) = \text{first } n \geq a \text{ such that } \max_{1 \leq k \leq n-a+1} \sum_{i=k}^n \left[X_i - \frac{\lambda}{2}\right] \geq \frac{\lambda}{2} a, \quad (3.5)$$

for  $a > 0$ . As shown in Mei (2006a),  $\{T^*(a)\}$  are asymptotically optimal solutions in the problem of maximizing the ARL to false alarm,  $\mathbf{E}_\theta T$ , for all possible values  $\theta \leq 0$  subject

to the constraint on the detection delay  $\bar{\mathbf{E}}_g T \leq \gamma$ . The following lemma establishes the asymptotic performance of  $T^*(a)$ . Since this lemma is a special case of Theorem 2.3 of Mei (2006a), we state it here without proof.

**Lemma 3.2.** *For any  $a > 0$  and any  $\theta \leq 0$ ,*

$$\mathbf{E}_\theta(T^*(a)) \geq \exp(I(\lambda, \theta)a), \quad (3.6)$$

where  $I(\lambda, \theta)$  is defined in (3.3) and as  $a \rightarrow \infty$

$$\bar{\mathbf{E}}_g(T^*(a)) \leq a + (C + o(1))\sqrt{a}, \quad (3.7)$$

where  $C = (1/\lambda)\sqrt{2/\pi}$ .

From Lemma 3.2 and relation (3.4), it is straightforward to show that for any  $a > 0$ , the detection scheme  $T^*(a)$  has a finite detection delay  $\bar{\mathbf{E}}_g(T^*(a))$ , but has an infinite ARL to false alarm  $\mathbf{E}_f(T^*(a))$ .

It is worth pointing out several implications of our example of detecting a change in distribution from the mixture distribution  $f$  in (2.3) to  $g = \phi_\lambda$ . First of all, besides  $T_{CM}(\theta_0, b)$  in (2.1) and  $T^*(a)$  in (3.5), it is easy to construct many other families of detection schemes with similar properties. With suitably chosen boundary values, these detection schemes can have the same detection delays, and have infinite ARL to false alarm. Hence the performance of these detection schemes are indistinguishable under the standard minimax formulation or the ARL to false alarm criterion.

Second, for dependent observations, the fact that a detection scheme  $T$  has infinite ARL to false alarm does not necessarily imply a small probability of ever raising a false alarm. In fact, in our example, any detection scheme  $T$  with finite detection delay  $\bar{\mathbf{E}}_g(T)$  will raise a false alarm with probability 1. To see this, fix a detection scheme  $T$  with finite detection delay  $\bar{\mathbf{E}}_g(T)$ . Applying the results of Lorden (1971) to the problem of detecting a change in the mean from  $\theta$  to  $\lambda$ , we have  $\mathbf{P}_\theta(T < \infty) = 1$ . By the definition of  $f$  in (2.3),  $\mathbf{P}_f(T < \infty) = \int_{-\infty}^0 \mathbf{P}_\theta(T < \infty)\pi(\theta)d\theta = 1$ , implying that the probability of  $T$  ever raising a

false alarm is 1. In particular, for the detection scheme  $T_{CM}(\theta_0, b)$  in (2.1) or  $T^*(a)$  in (3.5), although their ARLs to false alarm are infinite, they raise a false alarm with probability 1 even if there is no change from the mixture distribution  $f$  in (2.3).

Third, under the standard minimax formulation with the ARL to false alarm criterion, detecting larger changes will play a more important role in the change-point with mixture pre-change distributions, which may be undesirable. To see this, for a given detection scheme  $T$  in our example,  $\mathbf{E}_\theta(T)$  is generally an exponential decreasing function of  $|\theta|$  as  $\theta \leq 0$ , see, for example, relations (3.1) and (3.6) for the detection schemes  $T_{CM}(\theta_0, b)$  and  $T^*(a)$ , respectively. Hence, by (3.4), with suitably chosen  $\pi(\theta)$ , the main contribution to  $\mathbf{E}_f(T)$  will come from  $\mathbf{E}_\theta(T)$  with negatively large  $\theta$  values (far away from the post-change distribution  $g = \phi_\lambda$ ). This means that detecting larger changes will play a more important role under the standard minimax formulation. However, larger changes should be easily detected by any reasonable detection schemes, and one may be more interested in developing sophisticated schemes to detect a smaller change.

Finally, in the standard minimax formulation of change-point problems with dependent observations, an influential result is that of Lai (1998), which showed that

$$\bar{\mathbf{E}}_g(T) \geq (1 + o(1)) \frac{\log \mathbf{E}_f(T)}{I}$$

for any stopping time  $T$ , if the following sufficient conditions holds: there exists some constant  $I > 0$  such that for any  $\delta > 0$ ,

$$\lim_{n \rightarrow \infty} \sup_{\nu \geq 1} \text{ess sup} \mathbf{P}^{(\nu)} \left\{ \max_{t \leq n} \sum_{i=\nu}^{\nu+t} Z_i \geq I(1 + \delta)n | X_1, \dots, X_{\nu-1} \right\} = 0, \quad (3.8)$$

where  $Z_i = \log(g(X_i | X_1, \dots, X_{i-1}) / f(X_i | X_1, \dots, X_{i-1}))$  is the conditional log-likelihood ratio. That is, if Lai's sufficient condition in (3.8) holds, then any detection schemes with finite detection delay  $\bar{\mathbf{E}}_g T$  will have finite ARL to false alarm  $\mathbf{E}_f T$ . Combining this with Theorem 3.1 yields that in our example with the mixture distribution  $f$  defined in (2.3), there is no constant  $I > 0$  satisfying Lai's sufficient condition in (3.8). Therefore, although Lai's sufficient condition can be useful in some situations, e.g., it will be used in the Appendix to prove Theorem 4.1 of the present article, it may not be applicable in general.

## 4. ALTERNATIVE MINIMAX FORMULATIONS

The purpose of this section is to present two alternative minimax formulations for the change-point problem when  $f$  is the mixture distribution defined in (2.3) and  $g = \phi_\lambda$ , as the standard minimax formulation with the ARL to false alarm criterion is inappropriate here.

### 4.1. Relative Divergence Rate

From mathematical point of view, if some integrals or series sums go to  $\infty$ , it is natural to consider their divergence rates. It turns out that this strategy can be adopted to our example: suppose the problem of detecting a change in distribution from  $f$  to  $g$  can be treated as a limit of the problems of detecting changes from  $f_\xi$  to  $g$  with  $f_\xi$  converging to  $f$  as  $\xi \rightarrow -\infty$ , then for a given detection scheme  $T$ ,  $\mathbf{E}_f(T)$  is the limit of  $\mathbf{E}_{f_\xi}(T)$  as  $\xi \rightarrow -\infty$ . If  $\mathbf{E}_f(T) = \infty$ , it is possible that  $\mathbf{E}_{f_\xi}(T)$  is finite and well-defined for each  $\xi$ , but  $\mathbf{E}_{f_\xi}(T) \rightarrow \infty$  as  $\xi \rightarrow -\infty$ . In that case, it is natural to consider the divergence rate of  $\mathbf{E}_{f_\xi}(T)$ . However, we need to take into account the tradeoff between the detection delays and false alarms, which will likely provide a bound on the divergence rate of  $\mathbf{E}_{f_\xi}(T)$ . Thus, it will be more informative to treat that bound as a baseline rate and consider the relative divergence rate with respect to the baseline rate.

To present our ideas rigorously, first note that for independent observations, the asymptotic efficiency of a family of detection schemes  $\{T(a)\}$  in the problem of detecting a change in distribution from  $f$  to  $g$  can be defined as

$$e(f, g) = \liminf_{a \rightarrow \infty} \frac{\log \mathbf{E}_f(T(a))}{I(g, f) \bar{\mathbf{E}}_g(T(a))} \quad (4.1)$$

where  $I(g, f) = \mathbf{E}_g(\log(g(X)/f(X)))$  is the Kullback-Leiber information number, and  $\{\bar{\mathbf{E}}_g(T(a))\}$  is required to satisfy  $\bar{\mathbf{E}}_g(T(a)) \rightarrow \infty$  as  $a \rightarrow \infty$ . It is well-known (Lorden (1971)) that for independent observations,  $e(f, g) \leq 1$  for all families, and the equality can be achieved by Page's CUSUM procedures for detecting a change in distribution from  $f$  to  $g$ . This suggests to define a family of detection schemes  $\{T(a)\}$  is *asymptotically efficient* at  $(f, g)$  if  $e(f, g) = 1$ . It follows that Page's CUSUM procedure for detecting a change in distribution

from  $f$  to  $g$  is asymptotically efficient at  $(f, g)$  when observations are independent.

Now in our context when  $f$  is the mixture distribution defined in (2.3) and  $g = \phi_\lambda$ , the definition of the asymptotic efficiency  $e(f, g)$  in (4.1) cannot be applied directly for the following two reasons: (i)  $\mathbf{E}_f(T)$  can be  $\infty$  as shown in the previous section, and (ii) it is not clear how to define the information number  $I(g, f)$ .

Fortunately, this concept can be salvaged if we replace  $f$  by a sequence of distributions  $\{f_\xi\}$  with  $f_\xi$  converging to  $f$ . Define “new mixture” distributions

$$f_\xi(x_1, \dots, x_n) = \int_\xi^0 \left[ \prod_{i=1}^n \phi_\theta(x_i) \right] \pi_\xi(\theta) d\theta, \quad (4.2)$$

where

$$\pi_\xi(\theta) = \frac{\pi(\theta)}{\int_\xi^0 \pi(u) du} \quad \text{for } \theta \leq 0.$$

Note that as  $\xi \rightarrow -\infty$ ,  $\pi_\xi(\theta) \rightarrow \pi(\theta)$  for all  $\theta \leq 0$ , and  $f_\xi(x_1, \dots, x_n) \rightarrow f(x_1, \dots, x_n)$  for all  $n \geq 1$ . That is, the problem of detecting a change in distribution from  $f$  to  $g$  can be thought of as a limit of the problems of detecting changes from  $f_\xi$  to  $g$  as  $\xi \rightarrow -\infty$ .

Denote by  $\mathbf{P}_{f_\xi}$  and  $\mathbf{E}_{f_\xi}$  the probability measure and expectation, respectively, when  $X_1, X_2, \dots$  are distributed according to the mixture distribution  $f_\xi$ . The following theorem, whose proof is highly non-trivial and is presented in the Appendix, establishes the information bounds in the problem of detecting a change in distribution from  $f_\xi$  in (4.2) to  $g = \phi_\lambda$ .

**Theorem 4.1.** *For each pair  $(f_\xi, g)$ ,*

$$\overline{\mathbf{E}}_g(T) \geq (1 + o(1)) \frac{\log \mathbf{E}_{f_\xi}(T)}{I(\lambda, \xi)} \quad (4.3)$$

*for any stopping time  $T$  as  $\mathbf{E}_{f_\xi}(T)$  or  $\overline{\mathbf{E}}_g(T)$  goes to  $\infty$ , where  $I(\lambda, \xi) = (\lambda - \xi)^2/2$  is the Kullback-Leibler information number.*

We are now in a position to present the new minimax formulation in the problem of detecting a change in distribution from the mixture distribution  $f$  in (2.3) to  $g = \phi_\lambda$ . To accomplish this, define the asymptotic efficiency of a family of detection schemes  $\{T(a)\}$  at  $(f, g)$  as

$$e^*(f, g) = \liminf_{\xi \rightarrow -\infty} e^*(f_\xi, g), \quad (4.4)$$

where

$$e^*(f_\xi, g) = \liminf_{a \rightarrow \infty} \frac{\log \mathbf{E}_{f_\xi}(T(a))}{I(\lambda, \xi) \overline{\mathbf{E}}_g(T(a))} \quad (4.5)$$

is the asymptotic efficiency in the problem of detecting a change in distribution from  $f_\xi$  to  $g$ . Then Theorem 4.1 implies that  $e^*(f_\xi, g) \leq 1$  and  $e^*(f, g) \leq 1$  for all families of detection schemes, so we can define

**Definition 4.1.**  $\{T(a)\}$  is asymptotically efficient at  $(f, g)$  if  $e^*(f, g) = 1$ .

It is useful to note that our new definition of asymptotic efficiency,  $e^*(f, g)$  in (4.4), can be thought of as the relative divergence rate of  $\log \mathbf{E}_f(T(a))$  with respect to the baseline rate  $I(\lambda, \xi)$ , since we have  $\log \mathbf{E}_{f_\xi} T(a) \rightarrow \log \mathbf{E}_f T(a) = \infty$  and  $I(\lambda, \xi) = (\lambda - \xi)^2/2 \rightarrow \infty$  as  $\xi \rightarrow -\infty$ . Alternatively, the definition of  $e^*(f, g)$  can also be thought of as taking into account the difficulties of detecting different changes. This viewpoint can be very useful in theory and practice since it may allow one to detect both large and small changes efficiently. Of course the efficiencies will have different meanings for large and small changes.

It is not clear so far whether 1 is the *sharp* upper bound of  $e^*(f, g)$  over all families of detection schemes. In other words, can we find a family of detection schemes which is asymptotically efficient at  $(f, g)$  under this definition? It turns out that the answer is “yes,” and such asymptotically efficient detection schemes exist. Before offering such detection schemes, let us point out it is nontrivial to find detection schemes which are asymptotically efficient under our new definition.

For instance, the family of detection schemes  $T_{CM}(\theta_0, b)$  in (2.1) is not asymptotically efficient at  $(f, g)$ . In fact, the asymptotic efficiency of  $T_{CM}(\theta_0, b)$  is  $e^*(f, g) = 0$ . An outline of the proof is as follows. It is easy to show that  $\mathbf{E}_\theta(T_{CM}(\theta_0, b))$  is a decreasing function of  $\theta \in (-\infty, \theta_0]$ , since, intuitively, it will take longer to make a false alarm for a larger change. Then by relation (4.6) below, we have

$$\log \mathbf{E}_{f_\xi}(T_{CM}(\theta_0, b)) \leq \log \mathbf{E}_{\theta=\xi}(T_{CM}(\theta_0, b)) = (1 + o(1)) \left( \frac{\lambda + \theta_0 - 2\xi}{\lambda - \theta_0} b \right),$$

where the last equation follows from the classical results on Page's CUSUM procedures for independent observations. By (3.2) and the definition of  $e^*(f_\xi, g)$  in (4.5), we have

$$e^*(f_\xi, g) \leq \frac{(\lambda - \theta_0)(\lambda + \theta_0 - 2\xi)}{(\lambda - \xi)^2}.$$

Since the right-hand side of the above inequality converges to 0 as  $\xi \rightarrow -\infty$ , the asymptotic efficiency of the family of detection schemes  $T_{CM}(\theta_0, b)$  in (2.1) is  $e^*(f, g) = 0$  by the definition of  $e^*(f, g)$  in (4.4).

Now let us present asymptotic efficient detection under our new definition. It turns out that one family of such detection schemes is the detection scheme  $T^*(a)$  defined in (3.5). The following theorem establishes the asymptotic optimality properties of  $T^*(a)$  under our new definition in the problem of detecting a change in distribution from the mixture distribution  $f$  in (2.3) to  $g = \phi_\lambda$ .

**Theorem 4.2.** *The family of the detection schemes  $\{T^*(a)\}$  defined in (3.5) is asymptotically efficient at  $(f, g)$ .*

*Proof.* It suffices to show that for any arbitrary  $\xi \leq 0$ , the detection schemes  $\{T^*(a)\}$  satisfies  $e^*(f_\xi, g) \geq 1$ . Now fix  $\xi \leq 0$ . By the definition of  $f_\xi$  in (4.2), for any detection scheme  $T$ ,

$$\mathbf{E}_{f_\xi}(T) = \int_{\xi}^0 \mathbf{E}_{\theta}(T) \pi_{\xi}(\theta) d\theta, \quad (4.6)$$

Combining this with Lemma 3.2 yields

$$\mathbf{E}_{f_\xi}(T^*(a)) \geq \int_{\xi}^0 e^{I(\lambda, \theta)a} \pi_{\xi}(\theta) d\theta.$$

For any arbitrary  $\eta > 0$ , by the continuity of  $\pi(\theta)$  in (2.2) and  $I(\lambda, \theta)$  in (3.3) as well as the relation between  $\pi(\theta)$  and  $\pi_{\xi}(\theta)$ , there exists a positive number  $\delta$  such that  $\delta + \xi \leq 0$  and for all  $\theta \in (\xi, \xi + \delta)$ ,

$$I(\lambda, \theta) \geq (1 - \eta)I(\lambda, \xi) \quad \text{and} \quad \pi_{\xi}(\theta) \geq (1 - \eta)\pi_{\xi}(\xi).$$

The value of  $\delta$  may depend on  $\eta$  and  $\xi$ , but it does not depend on  $a$  or the stopping time  $T^*(a)$ . Hence,

$$\begin{aligned} \mathbf{E}_{f_\xi}(T^*(a)) &\geq \int_{\xi}^{\xi+\delta} e^{I(\lambda,\theta)a} \pi_\xi(\theta) d\theta \\ &\geq \int_{\xi}^{\xi+\delta} e^{I(\lambda,\xi)(1-\eta)a} (1-\eta) \pi_\xi(\xi) d\theta \\ &= e^{I(\lambda,\xi)(1-\eta)a} \delta (1-\eta) \pi_\xi(\xi). \end{aligned}$$

By Lemma 3.2, for sufficiently large value  $a$ ,

$$\frac{\log \mathbf{E}_{f_\xi}(T^*(a))}{I(\lambda, \xi) \overline{\mathbf{E}}_g(T^*(a))} \geq \frac{I(\lambda, \xi)(1-\eta)a + \log[\delta(1-\eta)\pi_\xi(\xi)]}{I(\lambda, \xi)(a + (C + o(1))\sqrt{a})},$$

which converges to  $1 - \eta$  as  $a \rightarrow \infty$  because  $\xi, \eta, \delta$  and  $C$  does not depend on  $a$ . Thus, for the detection schemes  $\{T^*(a)\}$ , we have

$$e^*(f_\xi, g) = \liminf_{a \rightarrow \infty} \frac{\log \mathbf{E}_{f_\xi}(T^*(a))}{I(\lambda, \xi) \overline{\mathbf{E}}_g(T^*(a))} \geq 1 - \eta.$$

Since  $\eta > 0$  is arbitrary, we have  $e^*(f_\xi, g) \geq 1$ . The theorem follows at once from the fact that  $\xi \leq 0$  is arbitrary.  $\square$

## 4.2. Expected False Alarm Rate

While the proposed asymptotic efficiency  $e^*(f, g)$  in Section 4.1 has theoretical meaning and seems to be good for evaluation of the overall quality of a detection scheme, it can be hardly considered as an appropriate characteristic for the false alarm rate only, and it perhaps will not satisfy practitioners who would like to see plots of the detection delay versus some reasonable false alarm rate measure in a particular problem. In addition, it is not clear how to run numerical simulations efficiently to calculate the asymptotic efficiency. For this reason, we propose another formulation with a new false alarm criterion: expected false alarm rate (EFAR).

To motivate our criterion, let us first look at the ARL to false alarm criterion when observations are i.i.d. under the pre-change distribution. As we mentioned earlier, even for independent observations, there are some concerns on the ARL to false alarm criteria, see

Barnard (1959), Kenett and Zacks (1998), Lai (2001), and Tartakovsky et al. (2006). But the general conclusion is that since the distribution of stopping times of widely used procedures such as Page's CUSUM and Shiriyayev-Roberts procedures are asymptotically exponential under the pre-change hypothesis, the ARL to false alarm criterion is a reasonable measure of the false alarm rate for independent observations.

In the following, we provide another viewpoint of understanding the ARL to false alarm criterion. This viewpoint allows us to show that for any stopping time  $T$  (no matter whether its distribution is asymptotically exponential or not), the reciprocal of the ARL to false alarm,  $1/\mathbf{E}_f(T)$ , can always be thought of as the false alarm rate when the observations are i.i.d. under the pre-change probability measure  $\mathbf{P}_f$ .

To see this, imagine *repeated* applications of a detection scheme (stopping time)  $T$  to the observations  $X$ 's under the probability measure  $\mathbf{P}_f$ . That is, define stopping times  $T_k$  inductively as follows:  $T_1 = T$ , and  $T_{k+1} =$  the stopping time obtained by applying  $T$  to the observations  $X_{T_1+\dots+T_{k+1}}, X_{T_1+\dots+T_{k+2}}, \dots$ . Since the  $X$ 's are i.i.d. under  $\mathbf{P}_f$ , the stopping times  $T_1, T_2, \dots$  are defined everywhere and are mutually independent with identical distribution which is the same as that of the detection scheme  $T$ . Now under  $\mathbf{P}_f$ , the new stopping times  $T_1, \dots, T_k$  are simply repeated application of the detection scheme  $T$ , and they raise  $k$  false alarms out of a total number  $T_1 + \dots + T_k$  of observations  $X$ 's. Thus under  $\mathbf{P}_f$ , the false alarm rate of  $T$  can be thought of as

$$\frac{k}{T_1 + \dots + T_k},$$

which converges to  $1/\mathbf{E}_f(T)$  as  $k$  goes to  $\infty$  by the strong law of large numbers. Note that this approach has also been used in Blackwell (1946) to provide an alternative proof of Wald's equation.

If the observations are not independent, then the above argument may still work if  $T_1, T_2, \dots, T_k$  are identically distributed and the strong law of large numbers for dependent observations is applicable to the  $T_k$ 's. In that case,  $1/\mathbf{E}_f(T)$  can still be thought of as the false alarm rate, even if the observations are not independent. The stationary autoregressive (AR) models in time series analysis seems to be one of such examples. Unfortunately, the

above arguments may not work in general when the observations are not independent, and it is not clear whether it is reasonable to treat  $1/\mathbf{E}_f(T)$  as the false alarm rate.

Nevertheless, this above approach shows that for the mixture pre-change distribution  $f$  defined in (2.3), while  $\mathbf{E}_f(T)$  may be misleading as a false alarm criterion under the mixture distribution  $f$ ,  $\mathbf{E}_\theta(T)$  remains informative as a function of  $\theta$ , as the observations are i.i.d. under  $\mathbf{P}_\theta$ . This motivates us to define the “expected false alarm rate” in the problem of detecting a change from the mixture pre-change distribution  $f$  in (2.3) to  $g = \phi_\lambda$ . Specifically, under  $\mathbf{P}_\theta$ , it is reasonable to define the the false alarm rate as  $\text{FAR}_\theta(T) = 1/\mathbf{E}_\theta(T)$ . Using the definition of the mixture distribution  $f$  in (2.3), it is nature to define the expected false alarm rate under  $\mathbf{P}_f$  as

$$\text{EFAR}_f(T) = \int_{-\infty}^0 \text{FAR}_\theta(T) \pi(\theta) d\theta = \int_{-\infty}^0 \frac{1}{\mathbf{E}_\theta(T)} \pi(\theta) d\theta,$$

and then the problem can be formulated as follows: Minimize the detection delay  $\bar{\mathbf{E}}_g(T)$  subject to the expected false alarm rate  $\text{EFAR}_f(T) \leq \alpha$  for some  $\alpha \in (0, 1)$ .

Note that by Jensen’s inequality, we have

$$\frac{1}{\mathbf{E}_f(T)} = \frac{1}{\int_{-\infty}^0 \mathbf{E}_\theta(T) \pi(\theta) d\theta} \leq \int_{-\infty}^0 \frac{1}{\mathbf{E}_\theta(T)} \pi(\theta) d\theta = \text{EFAR}_f(T).$$

That is, in our example, the reciprocal of the ARL to false alarm  $\mathbf{E}_f(T)$  is less than the proposed expected false alarm rate  $\text{EFAR}_f(T)$ . As we see in Section 2, the main contribution to  $\mathbf{E}_f(T)$  typically come from  $\mathbf{E}_\theta(T)$  with negatively large  $\theta$  values (far away from the post-change distribution  $g = \phi_\lambda$ ), which may be undesirable. On the other hand, the main contribution to the expected false alarm rate  $\text{EFAR}_f(T)$  tends to come from those  $\theta$  values which are close to the post-change distribution  $g = \phi_\lambda$ . Moreover, while there are schemes with infinite ARL to false alarm, there exist no schemes with zero expected false alarm rate  $\text{EFAR}_f(T)$ .

In statistics, besides mean, another widely used statistic to capture the distribution function of a random variable is median, or more generally, quantiles. It is interesting to point out that in our context,  $\text{EFAR}_f(T)$  is closely related to the “quantile run length” of

false alarm, which is defined as  $\xi_f^q(T)$  = the value  $u$  such that  $P_f(T \leq u) = 1 - P_f(T > u) = q$  for some  $0 < q < 1$ . Note the the idea of the quantile run length of false alarm was expressed in Barnard (1959). The close relationship between EFAR and the quantile run length can be seen from the following heuristic argument.

Recall that under  $\mathbf{P}_\theta$ , the observations are independent and thus  $T$  is generally asymptotically exponential distributed with mean  $\mathbf{E}_\theta(T)$ , implying that  $\mathbf{P}_\theta(T > u) \approx \exp(-u/\mathbf{E}_\theta(T))$  for  $u > 0$ . By the definition of the quantile  $\xi_f^q(T)$ , we have

$$1 - q = \mathbf{P}_f(T > \xi_f^q(T)) = \int_{-\infty}^0 \mathbf{P}_\theta(T > \xi_f^q(T))\pi(\theta)d\theta \approx \int_{-\infty}^0 \exp\left(-\frac{\xi_f^q(T)}{\mathbf{E}_\theta(T)}\right)\pi(\theta)d\theta,$$

if we assume the results on asymptotic exponential distributions hold uniformly for all  $\theta$ . Now let us assume  $\xi_f^q(T)/\mathbf{E}_\theta(T)$  is small, which seems to be reasonable for small  $q > 0$ . Then using the fact that  $\exp(-x) \approx 1 - x$  for small  $x$ , we have

$$1 - q \approx \int_{-\infty}^0 \left(1 - \frac{\xi_f^q(T)}{\mathbf{E}_\theta(T)}\right)\pi(\theta)d\theta.$$

Combining this with the fact that  $\int_{-\infty}^0 \pi(\theta)d\theta = 1$  yields

$$\xi_f^q(T) \approx q / \int_{-\infty}^0 \frac{1}{\mathbf{E}_\theta(T)}\pi(\theta)d\theta = \frac{q}{\text{EFAR}_f(T)}.$$

Therefore,  $\text{EFAR}_f(T)$  provides useful asymptotic information about the quantile run length to false alarm,  $\xi_f^q(T)$ , at least for small  $q > 0$ .

It is also worth pointing out that the above arguments suggest us to consider a general false alarm criterion of the form

$$H_f(T) = \int_{-\infty}^0 h(\mathbf{E}_\theta(T))\pi(\theta)d\theta, \tag{4.7}$$

where the non-negative function  $h(\cdot)$  preferably is decreasing or very slowly increasing so that the value of  $H_f(T)$  depends more on smaller values of  $\theta$ , e.g.,  $h(u) = 1/u$  or  $\exp(-u)$ . Note that the ARL to false alarm,  $\mathbf{E}_f(T)$ , can also be thought of as a special case with  $h(u) = u$ , but unfortunately, our results show that linear increase is too fast in our context.

It will be interesting to develop asymptotically optimal procedure under the expected false alarm rate criterion, or more generally under the criterion  $H_f(T)$  defined in (4.7). One

possible candidate is the detection schemes  $T^*(a)$  in (3.5), which asymptotically maximizes  $\mathbf{E}_\theta(T^*(a))$ , or equivalently, optimizes  $h(\mathbf{E}_\theta(T^*(a)))$ , for every pre-change  $\theta \leq 0$ , subject to a constraint on the detection delay  $\bar{\mathbf{E}}_g(T)$ . The detailed arguments are beyond the scope of this article and will be investigated elsewhere.

## 5. FURTHER EXAMPLES AND IMPLICATION

While the specific example we considered is in essence the problem of detecting a change in the mean of normal observations, we want to point out that the ideas and results can be generalized to many other problems. In the following, we discuss the implication of our example in change-point problems with either exchangeable pre-change models or hidden Markov models.

### 5.1. Exchangeable Pre-Change Models

It is straightforward to extend our ideas to the change-point problem in which the observations  $X_n$ 's are exchangeable under the pre-change distribution  $f$ . Here the  $X$ 's are exchangeable under a probability measure  $\mathbf{P}_f$  if for each  $n \geq 1$ , the  $n!$  permutations  $(X_{k_1}, \dots, X_{k_n})$  have the same  $n$ -dimensional joint probability distribution. To see this, by deFinetti's representation theorem for exchangeable distribution and its extension in Hewitt and Savage (1955), in most cases of interest, an exchangeable distribution for a data sequence is a mixture of i.i.d. distributions. That is, for exchange distribution  $f$ , in general, we have  $f(x_1, \dots, x_n) = \int_{\Theta} \prod_1^n p_\theta(X_i) d\Pi(\theta)$ , and thus our results for the mixture distribution can be easily extended to the problem with exchangeable pre-change distributions. Note that many widely used models or methods lead to exchangeable distributions, see, for example, linear random-effects models, or invariance reduction and weight-function approaches we mentioned in Section 2. These include the sequential  $\chi^2$ ,  $F$  and  $T^2$ -tests, as mentioned in Berk (1970).

In the following we discuss another concrete example to illustrate that the ARL to false alarm criterion may not be applicable for exchangeable pre-change models. Assume we are

interested in detecting a change in distribution in a sequence of observations  $X_1, X_2, \dots$ , in which, for some unknown time  $\nu$ ,

$$X_i = \begin{cases} Y_i - (a + b|Y|), & \text{if } i < \nu; \\ Y_i + (a + b|Y|), & \text{if } i \geq \nu. \end{cases}$$

where  $a, b > 0$  are two known constants,  $Y, Y_1, Y_2, \dots$  are unobservable, independent random variables,  $Y_i \sim N(0, 1)$ , and  $Y$  has a known density  $h(\cdot)$ .

In this problem, both the pre-change distribution  $f$  and the post-change distribution  $g$  are exchangeable. Let  $\mu_0 = a + b\mathbf{E}|Y|$  and  $\sigma_0^2 = b^2\text{Var}(|Y|)$ . Then it is straightforward to show that the  $X$ 's have mean  $-\mu_0$  and variance  $1 + \sigma_0^2$  under the pre-change distribution  $f$ , and have mean  $\mu_0$  and variance  $1 + \sigma_0^2$  under the post-change distribution  $g$ . Moreover, conditional on  $\nu$ ,  $\text{cov}(X_i, X_j) = -1 - \sigma_0^2$  if  $i < \nu < j$ , i.e., the pre-change and post-change observations are (negatively) correlated.

Thus, this problem can be thought of as detecting a change in the mean from  $-\mu_0$  to  $\mu_0$  (in which the observations are dependent). Motivated by Page's CUSUM procedure for independent observations, let us consider the detection scheme

$$T(\gamma) = \inf\{n : \max_{1 \leq k \leq n} \sum_{i=k}^n X_i \geq \gamma\}. \quad (5.1)$$

Then for  $\gamma > 0$ ,  $T(\gamma)$  will have finite detection delay  $\bar{\mathbf{E}}_g(T(\gamma))$  but  $\mathbf{E}_f(T(\gamma))$  can be infinite.

To see this, first note that the detection delay  $\bar{\mathbf{E}}_g(T(\gamma)) = \mathbf{E}_g(T(\gamma)) = \mathbf{E}(h(|Y|))$ , where  $h(z) = \mathbf{E}(N_z^{(1)})$ , and  $N_z^{(1)}$  is a new stopping time defined by

$$N_z^{(1)} = \inf\{n : \max_{1 \leq k \leq n} \sum_{i=k}^n (Y_i + a + bz) \geq \gamma\}.$$

Since  $Y_i$  are i.i.d.  $N(0, 1)$ , the property of  $N_z^{(1)}$  follows immediately from the classical results of Page's CUSUM procedure with independent observations. In particular, for  $z > 0$  we have  $h(z) \leq h(0) = \mathbf{E}(N_{z=0}^{(1)}) = \gamma/a + O(1)$ , implying that  $\bar{\mathbf{E}}_g(T(\gamma)) = \mathbf{E}_g(T(\gamma)) = \mathbf{E}(h(|Y|)) \leq h(0)$  is finite when  $a, b > 0$ .

Similarly, the ARL to false alarm  $\mathbf{E}_f(T(\gamma)) = \mathbf{E}(k(|Y|))$ , where  $k(z) = E(N_z^{(0)})$ , and  $N_z^{(0)}$  is a new stopping time defined by

$$N_z^{(0)} = \inf\{n : \max_{1 \leq k \leq n} \sum_{i=k}^n (Y_i - (a + bz)) \geq \gamma\}.$$

Since  $Y_i$  are i.i.d.  $N(0, 1)$ , and for  $z > 0$  we have  $k(z) \geq \exp(2\gamma(a + bz))$  by the classical result for Page's CUSUM procedure with independent observations, and

$$\mathbf{E}_f(T(\gamma)) = \mathbf{E}(k(|Y|)) \geq \mathbf{E}(\exp(2\gamma(a + b|Y|))).$$

Therefore, if the distribution of  $Y$  satisfies that  $\mathbf{E}(\exp(2\gamma(a + b|Y|))) = \infty$  for some  $a, b > 0$ , then  $\mathbf{E}_f(T(\gamma)) = \infty$  but  $\bar{\mathbf{E}}_g(T(\gamma))$  is finite.

Note that our proposed alternative false alarm criteria in Section 4 can be easily extended to this example. For instance, for the detection scheme  $T(\gamma)$  in (5.1), if  $\mathbf{E}_f(T(\gamma)) = \infty$ , then the asymptotic efficiency (or the relative divergence rate) can be defined by considering a sequence of problems in which  $Y$  is replaced by the truncated random variables  $Y_\xi = YI(|Y| \leq \xi)$  as  $\xi \rightarrow \infty$ . Here  $I(A)$  is the indicator function. Meanwhile, the expected false alarm rate can be defined by conditioning on  $Y$ . Specifically, using the function  $k(\cdot)$  defined in  $\mathbf{E}_f(T(\gamma))$ , we can define

$$\text{EFAR}_f(T(\gamma)) = \mathbf{E}\left(\frac{1}{k(|Y|)}\right) \leq \mathbf{E}(\exp(-2\gamma(a + b|Y|))).$$

It will be interesting to investigate whether  $T(\gamma)$  in (5.1) is asymptotically optimal under these two alternative formulations.

## 5.2. Hidden Markov Models (HMM)

Surprisingly, our example is also closely related to the problem of detecting changes in hidden Markov models (HMM) or state-space models, which have many important applications, including speech recognition and edge detection. Some general asymptotic theorems have been developed under the standard ARL to false alarm criterion, see, for example, Fuh (2003, 2004), and Lai (1995, 1998).

Let us first introduce hidden Markov models. Assume that  $U_1, U_2, \dots$  is an unobserved Markov chain with states  $\{1, 2, \dots, K\}$ , transition probability matrix  $M = [\alpha(i, j)]_{i, j=1, \dots, K}$ , and initial probability  $\pi = (\pi_1, \dots, \pi_K)$ . Given  $U_1, \dots, U_n$ , the observations  $X_i$ 's are conditionally independent, and given  $U_i$ ,  $X_i$  is independent of  $U_j$  for  $j \neq i$ . Moreover, the conditional distribution of  $X_n$  given  $U_n$  does not depend on  $n$ . We also assume the conditional distributions of  $Y_n$  given  $U_n = i$  are dominated by a  $\sigma$ -finite measure, and denote by  $h_i(\cdot)$  the corresponding conditional density.

For our purpose, here we will focus on a special scenario of change-point problems in the hidden Markov model. Assume that initially the observations  $X_1, X_2, \dots, X_{\nu-1}$  are from the hidden Markov model with initial probability  $\pi_0 = (\pi_1^0, \dots, \pi_{K-1}^0, 0)$  with  $\sum_{k=1}^{K-1} \pi_k^0 = 1$ . At some unknown time  $\nu$ , the observations  $X_\nu, X_{\nu+1}, \dots$  are from the hidden Markov model with another initial probability  $\pi_1 = (0, \dots, 0, 1)$ . Here it is useful to think the state  $U_n = K$  as an absorbing state. The transition probability matrices and conditional densities  $h_i(\cdot)$  are the same before and after the change. The problem is to detect the true change as soon as possible, while raising as few false alarms as possible.

Note that this setting is more general than it appears to be at first sight. For example, if  $K = 2$  and the transition probability matrix  $M =$  identity matrix, then the problem becomes the classical setting of detecting a change in distribution for independent observations, where the pre-change distribution is  $f(x) = h_1(x)$  and the post-change distribution is  $g(x) = h_2(x)$ . As another example, if  $K = 3$  and the transition probability matrix

$$M = \begin{pmatrix} p & 1-p & 0 \\ p & 1-p & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

for some  $0 < p < 1$ , then using the standard probability calculation method in Markov chain, the problem again becomes the classical setting of detecting a change in distribution for independent observations. The only difference is that the pre-change distribution  $f(x) = ph_1(x) + (1-p)h_2(x)$  and the post-change distribution  $g(x) = h_3(x)$ .

So far we have replicated the classical problem with independent observations, and thus

Page's CUSUM or Shiriyayev-Roberts procedures are (asymptotically) optimal under the standard minimax formulation for these two special cases. Now let us look at another special case where  $K = 3$  and the transition probability matrix  $M =$  identity matrix. Then the pre-change distribution is the mixture distribution  $f(X_1, \dots, X_n) = \pi_1^0 \prod_{i=1}^n h_1(X_i) + \pi_2^0 \prod_{i=1}^n h_2(X_i)$  with  $\pi_1^0 + \pi_2^0 = 1$ , whereas the observations are i.i.d. with density  $h_3(x)$  after a change occurs. In this case, Mei (2006a) showed that if both  $\pi_1^0$  and  $\pi_2^0$  are positive, Page's CUSUM or Shiriyayev-Roberts procedures are asymptotically suboptimal under the standard minimax formulation.

More generally, if the transition probability matrix  $M$  is identity matrix, and the number of state  $K$  is greater than 2 (possibly  $\infty$ ), then the pre-change distribution becomes a mixture (or exchangeable) distribution. This sheds light on the practical situation when the transition probability matrix  $M$  is close to the identity matrix, particularly when the total sample size of the observations is only moderately large. In these situations, under the ARL to false alarm criterion, detecting larger changes may play more important role in the problem of detecting changes in the hidden Markov models.

## 6. DISCUSSIONS

A different choice of a prior distribution  $\pi(\theta)$  in (2.2) will have an impact on whether or not a particular detection scheme  $T$  has infinite ARL to false alarm  $\mathbf{E}_f(T)$ . By (3.4), a detection scheme  $T$  has an infinite ARL to false alarm,  $\mathbf{E}_f(T)$ , if and only if  $\mathbf{E}_f(T) = \int_{-\infty}^0 (\mathbf{E}_\theta(T)) \pi(\theta) d\theta$  diverges. Thus, if  $\pi(\theta)$  converges to 0 very fast as  $\theta \rightarrow -\infty$ , then  $\mathbf{E}_f(T)$  can be finite.

For instance, in our example, if we redefine  $\pi(\theta)$  in (2.2) as  $\pi(\theta) = \sqrt{2/\pi} \exp(-\theta^2/2)$  for  $\theta \leq 0$ . Then for any  $b > 0$ , Page's CUSUM procedure  $T_{CM}(\theta_0, b)$  defined in (2.1) will have a finite ARL to false alarm. However,  $T^*(a)$  in (3.5) still has an infinite ARL to false alarm under this choice of  $\pi(\theta)$  if  $a \geq 1$ .

Meanwhile, if we choose  $\pi(\theta) = C_r \exp(-|\theta|^r)$  with  $r > 2$  and the constant  $C_r$  chosen so that  $\int_{-\infty}^0 \pi(\theta) d\theta = 1$ , then any detection scheme  $T$  with finite detection delay  $\bar{\mathbf{E}}_g(T)$  will

also have finite ARL to false alarm,  $\mathbf{E}_f(T)$ . To see this, for any stopping time  $T$  satisfying the detection delay  $\bar{\mathbf{E}}_g(T) = \lambda$ , the maximum value of  $\mathbf{E}_\theta(T)$  for each  $\theta \leq 0$  is of order  $\exp(I(\lambda, \theta)\gamma) = \exp((\theta - \lambda)^2\gamma/2)$ . Hence the choice of  $\pi(\theta) = C_r \exp(-|\theta|^r)$  with  $r > 2$  will guarantee that  $\mathbf{E}_f(T) = \int_{-\infty}^0 \mathbf{E}_\theta(T)\pi(\theta)d\theta$  converges, implying that  $\mathbf{E}_f(T)$  is finite.

Indeed, most of our results in this article are not invariant to the choice of the prior distribution  $\pi(\theta)$  of the unknown parameter of the distribution function of observations. One reviewer asked what should be done if we do not know the prior distribution  $\pi(\theta)$ . One possible approach is to use the general theory developed in Mei (2006a), which enables one to construct a family of detection schemes that asymptotically minimizes  $\mathbf{E}_\theta(T)$  for **every** (pre-change)  $\theta$  value without any knowledge of the prior  $\pi(\theta)$  in a general context (not necessarily for normal distributions or other exponential families). For instance,  $T^*(a)$  in (3.5) requires no knowledge of the prior  $\pi(\theta)$ , but it is asymptotically efficient under the divergence rate criterion and also seems to be asymptotically optimal in the sense that it asymptotically minimizes the expected false alarm rate (EFAR) subject to a constraint on detection delay.

In our contexts, even if the ARL to false alarm  $\mathbf{E}_f(T)$  is finite, it does not mean that it is an appropriate false alarm criterion, as illustrated in Section 4.2. Specifically, in our example with mixture pre-change distribution, the main contribution to  $\mathbf{E}_f(T)$  comes from  $\mathbf{E}_\theta(T)$  with negatively large  $\theta$  values (far away from the post-change distribution  $g = \phi_\lambda$ ), implying that the ARL to false alarm criterion pays more attention to detecting larger changes, which is undesirable. Similar conclusions also hold for other exchangeable pre-change distributions, or in the problem of detecting a change in hidden Markov models, especially if the transition matrix is close to identity matrix. Thus we should be cautious to use the ARL to false alarm criterion to assess detection schemes if we are more interested in detecting smaller changes.

Besides the minimax formulation, another widely used formulation in change-point problems is the Bayesian formulation, in which the change-point  $\nu$  is a random variable with known a prior distribution. It is interesting to note that in our example with the mixture pre-change distribution  $f$  in (2.3), although the standard minimax formulation is in-

appropriate, the Bayesian formulation can still be applied, and both Page's CUSUM and Shiryeyev-Roberts procedures are asymptotically optimal under the Bayesian formulation. This is because the false alarm criterion in the Bayesian formulation is  $\mathbf{P}_f(T \leq \nu)$ , which is asymptotically equivalent to  $\mathbf{P}_\theta(T \leq \nu)$  with  $\theta = 0$ . Thus, the Bayesian formulation of detecting a change from the mixture pre-change distribution  $f$  to  $g = \phi_\lambda$  is asymptotically equivalent to that of detecting a change from  $\theta = 0$  to  $\lambda$  in the mean of normally distributed random variables.

It is possible that the Bayesian formulation, the alternative formulations proposed here, or other formulations such as Quantile Run Length, may not be appropriate to other change-point problems with dependent observations. In fact, a message of this article is that well-known theorems and results in classical change-point problems could no longer hold for dependent observations, at least for exchangeable pre-change distributions. While this article is "incomplete" in the sense that it does not develop general theory in a general setting, we hope it is adventurous enough to give theoreticians and practitioners something serious to contemplate.

## APPENDIX A: PROOF OF LEMMA 3.1

Relation (3.1) with  $\theta = \theta_0$  and relation (3.2) are well-known for Page's CUSUM procedure  $T_{CM}(\theta_0, b)$  defined in (2.1), see, for example, Lorden (1971) or Siegmund (1985). It suffices to show relation (3.1) holds for general  $\theta$ .

To prove this, note that Page's CUSUM procedure  $T_{CM}(\theta_0, b)$  defined in (2.1) can be rewritten as

$$\begin{aligned} T_{CM}(\theta_0, b) &= \text{first } n \geq 1 \text{ such that } \max_{1 \leq k \leq n} \sum_{i=k}^n \left[ X_i - \frac{\lambda + \theta_0}{2} \right] \geq \frac{b}{\lambda - \theta_0} \\ &= \text{first } n \geq 1 \text{ such that } \max_{1 \leq k \leq n} \sum_{i=k}^n \log \frac{\phi_{\lambda_1}(X_i)}{\phi_\theta(X_i)} \geq \frac{\lambda + \theta_0 - 2\theta}{\lambda - \theta_0} b, \end{aligned}$$

where  $\lambda_1 = \lambda + \theta_0 - \theta$ . Hence,  $T_{CM}(\theta_0, b)$  can also be thought of as Page's CUSUM procedure in the problem of detecting a change in the mean from  $\theta$  to  $\lambda_1 = \lambda - \theta_0 - \theta$  with log-

likelihood boundary  $b_1 = b(\lambda + \theta_0 - 2\theta)/(\lambda - \theta_0)$ . Applying the classical results on Page's CUSUM procedures again,  $\mathbf{E}_\theta(T_{CM}(\theta_0, b)) \geq \exp(b_1)$ , completing the proof of (3.1).  $\square$

## APPENDIX B: PROOF OF THEOREM 4.1

Denote by  $\mathbf{P}_{f_\xi, g}^{(\nu)}$  the probability measure when the distribution of the observations  $X$ 's changes from  $f_\xi$  to  $g$  at time  $\nu$ . To simplify notations, we also denote by  $f_\xi(\cdot|X_1, \dots, X_{i-1})$  and  $g(\cdot|X_1, \dots, X_{i-1})$  the conditional density functions of  $X_i$  given  $X_1, \dots, X_{i-1}$  under the probability measure  $\mathbf{P}_{f_\xi}$  and  $\mathbf{P}_g$ , respectively. Note that  $g(\cdot|X_1, \dots, X_{i-1}) = g(\cdot) = \phi_\lambda(\cdot)$  because observations are independent under  $\mathbf{P}_g$ . Then the conditional log-likelihood ratio of  $X_i$  given  $X_1, \dots, X_{i-1}$  is  $Z_i = \log \left( g(X_i|X_1, \dots, X_{i-1})/f_\xi(X_i|X_1, \dots, X_{i-1}) \right)$ .

To prove (4.3), it suffices to show that the constant  $I = I(\lambda, \xi)$  satisfies Lai's sufficient condition in (3.8), with  $\mathbf{P}^{(\nu)}$  replacing by  $\mathbf{P}_{f_\xi, g}^{(\nu)}$ . To do so, under  $\mathbf{P}_{f_\xi, g}^{(\nu)}$ , define

$$\pi_\xi^\nu(\theta) = \frac{\phi_\theta(X_1) \cdots \phi_\theta(X_{\nu-1}) \pi_\xi(\theta)}{\int_\xi^0 [\phi_\theta(X_1) \cdots \phi_\theta(X_{\nu-1})] \pi_\xi(\theta) d\theta},$$

then  $\int_\xi^0 \pi_\xi^\nu(\theta) d\theta = 1$ , and thus

$$\begin{aligned} \sum_{i=\nu}^{\nu+t} Z_i &= \log \frac{\phi_\lambda(X_\nu) \cdots \phi_\lambda(X_{\nu+t})}{\int_\xi^0 [\phi_\theta(X_\nu) \cdots \phi_\theta(X_{\nu+t})] \pi_\xi^\nu(\theta) d\theta} \\ &\leq \log \frac{\phi_\lambda(X_\nu) \cdots \phi_\lambda(X_{\nu+t})}{\min_{\xi \leq \theta \leq 0} [\phi_\theta(X_\nu) \cdots \phi_\theta(X_{\nu+t})]} \\ &= \max_{\xi \leq \theta \leq 0} \sum_{i=\nu}^{\nu+t} \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)}. \end{aligned}$$

Since under  $\mathbf{P}_{f_\xi, g}^{(\nu)}$ ,  $X_\nu, X_{\nu+1}, \dots$  are i.i.d. normal with mean  $\lambda$  and variance 1, no matter what we observed  $X_1, \dots, X_{\nu-1}$ , Lai's sufficient condition in (3.8) holds if we can show that for all  $\delta > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbf{P}_g \left\{ \max_{t \leq n} \max_{\xi \leq \theta \leq 0} \sum_{i=1}^t \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} \geq I(\lambda, \xi)(1 + \delta)n \right\} = 0. \quad (\text{B.1})$$

Now it suffices to show that (B.1) holds.

To prove (B.1), let  $\bar{X}_t = \sum_{i=1}^t X_i/t$ , then it is easy to see that

$$\sum_{i=1}^t \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} = \frac{t}{2} [(\theta - \bar{X}_t)^2 - (\lambda - \bar{X}_t)^2], \quad (\text{B.2})$$

and the maximum of (B.2) over  $\xi \leq \theta \leq 0$  is attained at  $\theta = \xi$  if  $\bar{X}_t > 0$ . Moreover,  $\bar{X}_t$  will always be positive if  $t$  is sufficiently large. This suggests us splitting the probability in (B.1) into three parts, depending on whether  $t \leq \sqrt{n}$  or  $\min_{\sqrt{n} \leq t \leq n} \bar{X}_t > 0$ . Specifically, define

$$\begin{aligned} C1 &= \mathbf{P}_g \left\{ \max_{t \leq \sqrt{n}} \max_{\xi \leq \theta \leq 0} \sum_{i=1}^t \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} \geq I(\lambda, \xi)(1 + \delta)n \right\}, \\ C2 &= \mathbf{P}_g \left\{ \max_{\sqrt{n} \leq t \leq n} \max_{\xi \leq \theta \leq 0} \sum_{i=1}^t \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} \geq I(\lambda, \xi)(1 + \delta)n; \min_{\sqrt{n} \leq t \leq n} \bar{X}_t \geq 0 \right\}, \\ C3 &= \mathbf{P}_g \left\{ \max_{\sqrt{n} \leq t \leq n} \max_{\xi \leq \theta \leq 0} \sum_{i=1}^t \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} \geq I(\lambda, \xi)(1 + \delta)n; \min_{\sqrt{n} \leq t \leq n} \bar{X}_t < 0 \right\}. \end{aligned}$$

Then the probability in the left-hand side of (B.1) is less than or equal to  $C1 + C2 + C3$ . Therefore, it now suffices to show that each of  $C1, C2$  and  $C3$  goes to 0 as  $n \rightarrow \infty$ .

For the  $C1$  term, by (B.2), we have

$$\max_{\xi \leq \theta \leq 0} \sum_{i=1}^t \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} \leq \frac{t}{2} \max_{\xi \leq \theta \leq 0} (\theta - \bar{X}_t)^2 \leq \frac{t}{2} [(\xi - \bar{X}_t)^2 + (\bar{X}_t)^2]$$

since  $(\theta - \bar{X}_t)^2$  is a convex function of  $\theta$  and the maximum is attained at either  $\theta = \xi$  or  $\theta = 0$ . Thus

$$\begin{aligned} C1 &\leq \mathbf{P}_g \left\{ \max_{t \leq \sqrt{n}} \frac{t}{2} [(\xi - \bar{X}_t)^2 + (\bar{X}_t)^2] \geq I(\lambda, \xi)(1 + \delta)n \right\} \\ &\leq \mathbf{P}_g \left\{ \max_{t \leq \sqrt{n}} [(\xi - \bar{X}_t)^2 + (\bar{X}_t)^2] \geq 2I(\lambda, \xi)(1 + \delta)\sqrt{n} \right\} \end{aligned}$$

using the fact that  $t \leq \sqrt{n}$  in the  $C1$  term. As  $n \rightarrow \infty$ , the term  $(\xi - \bar{X}_t)^2 + (\bar{X}_t)^2$  converges to  $(\xi - \lambda)^2 + \lambda^2$  under  $\mathbf{P}_g$ , while  $2I(\lambda, \xi)(1 + \delta)\sqrt{n}$  goes to  $\infty$ . This implies that  $C1 \rightarrow 0$  as  $n \rightarrow \infty$ .

For the  $C2$  term, we have  $\bar{X}_t \geq 0$  and the maximum over  $\xi \leq \theta \leq 0$  is attained at  $\theta = \xi$ , so

$$C2 \leq \mathbf{P}_g \left\{ \max_{1 \leq t \leq n} \sum_{i=1}^t \log \frac{\phi_\lambda(X_i)}{\phi_\xi(X_i)} \geq I(\lambda, \xi)(1 + \delta)n \right\}.$$

The strong law implies that under  $\mathbf{P}_g$ ,

$$\frac{1}{n} \max_{1 \leq t \leq n} \sum_{i=1}^t \log \frac{\phi_\lambda(X_i)}{\phi_\xi(X_i)} \rightarrow I(\lambda, \xi)$$

with probability 1. Thus for any  $\delta > 0$ , the term  $C2 \rightarrow 0$  as  $n \rightarrow \infty$ .

For the  $C3$  term, we have  $C3 \leq \mathbf{P}_g \left\{ \min_{\sqrt{n} \leq t \leq n} \bar{X}_t < 0 \right\}$ . Using the strong law again, we have  $\min_{\sqrt{n} \leq t \leq n} \bar{X}_t \rightarrow \lambda > 0$  with probability 1 under  $\mathbf{P}_g$ , and thus the term  $C3 \rightarrow 0$  as  $n \rightarrow \infty$ . This completes the proof of Theorem 4.1.  $\square$

## ACKNOWLEDGMENTS

The author would like to thank his advisor, Dr. Gary Lorden, for his support, guidance, and encouragement, and Dr. Moshe Pollak for stimulating discussions. The author would also like to thank the reviewers for their thoughtful comments. This work was supported in part by Dr. Sarah Hotle's National Institutes of Health under Grant R01 AI055343.

## REFERENCES

- Bansal, R. K. and Papantoni-Kazakos, P. (1986). An Algorithm for Detecting a Change in a Stochastic Process, *IEEE Transactions on Information Theory* 32: 227-235.
- Barnard, G. A. (1959). Control Charts and Stochastic Processes, *Journal of Royal Statistical Society, Series B* 21: 239-271.
- Baron, M. and Tartakovsky, A. (2006). Asymptotic Bayesian Change-Point Detection Theory for General Continuous-Time Models, *Sequential Analysis* 25: 257-296.
- Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes: Theory and Applications*, Englewood Cliffs: Prentice-Hall.
- Beibel, M. (1997). Sequential Change-Point Detection in Continuous Time When the Post-Change Drift Is Unknown, *Bernoulli* 3: 457-478.

- Berk, R. H. (1970). Stopping Times of SPRTs Based on Exchangeable Models, *Annals of Mathematical Statistics* 41: 979-990.
- Blackwell, D (1946). On an Equation of Wald, *Annals of Mathematical Statistics* 17: 84-87.
- Brodsky, B. E. and Darkhovsky, B. S. (1993). *Non-Parametric Methods in Change-Point Problems*, Netherlands: Kluwer.
- Brodsky, B. E. and Darkhovsky, B. S. (2000). *Non-Parametric Statistical Diagnosis: Problems and Methods*, Netherlands: Kluwer.
- Fienberg, S. E. and Shmueli, G. (2005). Statistical Issues and Challenges Associated with Rapid Detection of Bio-terrorist Attacks, *Statistics in Medicine* 24: 513-529.
- Fuh, C.-D. (2003). SPRT and CUSUM in Hidden Markov Models, *Annals of Statistics* 31: 942-977.
- Fuh, C.-D. (2004). Asymptotic Operating Characteristics of an Optimal Change Point Detection in Hidden Markov Models, *Annals of Statistics* 32: 2305-2339.
- Hewitt, E. and Savage, L. J. (1955). Symmetric Measures on Cartesian Products, *Transactions of American Mathematical Society* 80: 470-501.
- Jensen, W., Jones-Farmer, L. A., Champ, C. W., and Woodall, W. H. (2006). Effects of Parameter Estimation on Control Chart Properties: A Literature Review, *Journal of Quality Control* 38: 349-364.
- Kenett, R. S. and Zacks, S. (1998). *Modern Industrial Statistics: Design and Control of Quality and Reliability*, Pacific Grove: Duxbury.
- Kiefer, J. and Sacks, J. (1963). Asymptotically Optimum Sequential Inference and Design, *Annals of Mathematical Statistics* 34: 705-750.
- Krieger, A. M., Pollak, M., and Yakir, B. (2003). Surveillance of a Simple Linear Regression, *Journal of American Statistical Association* 98: 456-469.

- Kulldorff, M. (2001). Prospective Time-Periodic Geographic Disease Surveillance Using a Scan Statistic, *Journal of Royal Statistical Society, Series A* 164: 61-72.
- Lai, T. L. (1995). Sequential Change-Point Detection in Quality Control and Dynamical Systems (with discussions), *Journal of Royal Statistical Society, Series B* 57: 613-658.
- Lai, T. L. (1998). Information Bounds and Quick Detection of Parameter Changes in Stochastic Systems, *IEEE Transactions on Information Theory* 44: 2917-2929.
- Lai, T. L. (2001). Sequential Analysis: Some Classical Problems and New Challenges, *Statistica Sinica* 11: 303-408.
- Lorden, G. (1971). Procedures for Reacting to a Change in Distribution, *Annals of Mathematical Statistics* 42: 1897-1908.
- Mei, Y. (2006a). Sequential Change-Point Detection When Unknown Parameters Are Present in the Pre-Change Distribution, *Annals of Statistics* 34: 92-122.
- Mei, Y. (2006b). Suboptimal Properties of Page's CUSUM and Shirayayev-Roberts Procedures in Change-Point Problems with Dependent Observations, *Statistica Sinica* 16, 883-897.
- Moustakides, G. V. (1986). Optimal Stopping Times for Detecting Changes in Distributions, *Annals of Statistics* 14: 1379-1387.
- Page, E. S. (1954). Continuous Inspection Schemes, *Biometrika* 41: 100-115.
- Pollak, M. (1985). Optimal Detection of a Change in Distribution, *Annals of Statistics* 13: 206-227.
- Pollak, M. (1987). Average Run Lengths of an Optimal Method of Detecting a Change in Distribution, *Annals of Statistics* 15: 749-779.
- Pollak, M. and Siegmund, D. (1991). Sequential Detection of a Change in a Normal Mean When the Initial Value Is Unknown, *Annals of Statistics* 19: 394-416.

- Poor, H. V. (1998). Quickest Detection with Exponential Penalty for Delay, *Annals of Statistics* 26: 2179-2205.
- Ritov, Y. (1990). Decision Theoretic Optimality of the CUSUM Procedure, *Annals of Statistics* 18: 1464-1469.
- Roberts, S. W. (1966). A Comparison of Some Control Chart Procedures, *Technometrics* 8: 411-430.
- Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*, New York: Van Nostrand.
- Shiryayev, A. N. (1963). On Optimum Methods in Quickest Detection Problems, *Theory of Probability and Its Applications* 8: 22-46.
- Shiryayev, A. N. (1978). *Optimal Stopping Rules*, New York: Springer.
- Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*, New York: Springer.
- Stoumbos, Z., Reynolds, M. R., Jr., Ryan, T. P., and Woodall, W. H. (2000). The State of Statistical Process Control as We Proceed into the 21st Century, *Journal of American Statistical Association* 95: 992-998.
- Tartakovsky, A., Rozovskii, B., Blazek, R., and Kim, H. (2006). Detection of Intrusion in Information Systems by Sequential Change-Point Methods (with discussions), *Statistical Methodology* 3: 252-340.
- Wald, A. (1947). *Sequential Analysis*, New York: Wiley.
- Woodall, W. H. (1983). The Distribution of the Run Length of One-Sided CUSUM Procedures for Continuous Random Variables, *Technometrics* 25: 295-301.
- Woodall, W. H. (2006). The Use of Control Charts in Health-Care and Public-Health Surveillance (with discussions), *Journal of Quality Technology* 38: 89-134.

- Woodroffe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*, Philadelphia: SIAM.
- Yakir, B. (1994). Optimal Detection of a Change in Distribution When the Observations Form a Markov Chain with a Finite State Space, in *Change-Point Problems*, E. Carlstein, H. Muller, and D. Siegmund, eds., pp. 346-358, Hayward: Institute of Mathematical Statistics.
- Yakir, B., Krieger, A. M., and Pollak, M. (1999). Detecting a Change in Regression: First Order Optimality, *Annals of Statistics* 27: 1896-1913.